Development of AI-based Tools for Functional Annotation of lncRNA and its Applications

by

Shubham Choudhury (PhD20210)

Under supervision of

Prof. Gajendra P.S. Raghava

DEPARTMENT OF COMPUTATIONAL BIOLOGY

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY

NEW DELHI- 110020

Development of AI-based Tools for Functional Annotation of lncRNA and its Applications

by

Shubham Choudhury (PhD20210)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

to the

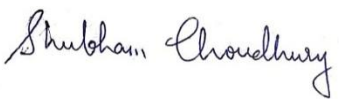
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY

July, 2026

Declaration

I acknowledge that I am fully responsible for the entire content of my thesis, including any sections assisted by online tools, including Artificial Intelligence-based tools. I accept full accountability for any violations of ethical standards in publications arising from the use of such tools.

July, 2026



Shubham Choudhury

PhD20210

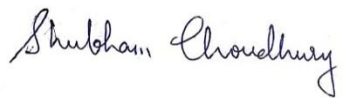
Advisor – Prof. Gajendra P.S. Raghava

Indraprastha Institute of Information Technology, Delhi

New Delhi-110020

Declaration regarding use of AI

I acknowledge that I am fully responsible for the entire content of my thesis, including any sections assisted by online tools, including Artificial Intelligence-based tools. I accept full accountability for any violations of ethical standards in publications arising from the use of such tools.

Signature of PhD Student: 

Name of PhD Student: **Shubham Choudhury**

Roll Number: PhD20210

Certificate

This is to certify that the thesis titled “Development of AI-Based Tools for Functional Annotation of lncRNA and its applications” being submitted by Mr. Shubham Choudhury to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of Doctor of Philosophy, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

July, 2026

Gajendra P.S. Raghava

Professor

Department of Computational Biology

IIIT-Delhi, 110020

Acknowledgement

As I finish this journey, I have a lot to be thankful for.

I would start with my PhD supervisor - Professor Gajendra P. S. Raghava. I am really grateful for his guidance as well as for the confidence and autonomy he bestowed upon me during my doctoral studies. He gave me the freedom to experiment with concepts on my own while providing assistance when it was really necessary. I also want to express my gratitude to Dr. Jaspreet Kaur Dhanjal and Dr. Arjun Ray, members of my PhD Monitoring Committee, whose insightful criticism improved my work at every turn.

I also want to express my gratitude to those who worked behind the scenes to make sure that research goes well. I am appreciative of Mrs. Shipra Jain, as well as the Academic and IRD Staff - Mr. Kapil Dev Garg, Mrs. Sarika, Mr. Imran Khan, Mr. Mohit, Mr. Raju Biswas, and Mrs. Anshu Dureja, for their continuous assistance with financial and administrative issues. I also like to thank Mr. Bhawani, Mr. Adarsh, and Mr. Vinod of the IT team for their timely help with any technical IT-related issues. Additionally, I want to thank the FMS department for their hard work in maintaining the institute's infrastructure and housekeeping. Their efforts made sure that the space was tidy, practical, and cozy so that academic and research activity could continue unhindered.

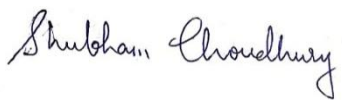
I am indebted to IIIT-Delhi for providing the research atmosphere and infrastructure that enabled this study, as well as the Department of Science and Technology's DST-INSPIRE Fellowship for their financial support.

Raghava Lab has been like a second home for the past five years. I am especially grateful to all the seniors - Dr. Sumeet Patiyal, Dr. Anjali Dhall, Dr. Neelam Sharma, and Dr. Leimarembi Devi Naorem, for their constant guidance and mentoring. I would also take this opportunity thank my lab colleagues, Dr. Ritu, Dr. Akanksha, Nishant, Nisha, Anand, Shivani, Naman, Pratik, Saloni, Pankaj, and Sahil, for keeping the lab a happy and sane place for research and personal growth.

In addition to professional side, I've been extremely blessed to have friends who have given this period of life meaning. For their constant support and company throughout difficult times, I am indebted to Nishant, Dr. Sumeet, Dr. Anjali, Dr. Neetesh, Dr. Mansi, Shruti, Nisha, Pankaj, Pratik, Naman, Dr. Ritu, Dr. Madhu, Rudrakshi, Shiva, Sneha, and Rajeshwari. Over the years, my college friends from IISER - Biswa, Rajat, Saurav, Abhishek, Balakrishna, Baishali, Yoti, Akhil, and

Sumeet have consistently been a source of emotional and mental support. Additionally, I want to thank Mr. Rajesh, whose tea became a vital component of my everyday schedule. During long workdays, his constantly stimulating tea and the quick chats that accompanied it provided much-needed breaks. Also, special mention to the tennis coaching group for that inspiring morning coaching session - Mr. Afzal (Coach), Naman, Vikrant, Spoorthy, and Pankaj.

Above everything, I would want to thank my parents, Mamata Choudhury and Samarendra Choudhury. Their unwavering faith in me, free from criticism or uncertainty, has been my most solid foundation. I also like to take this opportunity to thank my maternal grandpa, Late Er. Krushna Chandra Sahoo, who did not survive to see this accomplishment. I handle some of my hardest situations based on his patient and calm take on any adverse situation in life. I hope he would have been proud of this work and the person I have become.



Shubham Choudhury

Abstract

Long non-coding RNAs (lncRNAs) are a significant component of the human transcriptome and are well known for their ability to regulate gene expression. They have been implicated in a variety of disease-associated pathways. However, functional annotation of these lncRNAs remains a major challenge in the era of genomics and metagenomics, where large-scale genome sequencing has become routine. In this thesis, we have attempted to develop computational resources for annotating lncRNAs and to explore their applications in cancer diagnostics. In the first part of this thesis, we compiled all lncRNAs resources that includes experimental techniques, prediction methods and data repositories. These resources were manually curated from the literature and made available to the scientific community through a web platform called lncInfo. Next, we developed CytoLNCpred, a cell line–specific computational method for identifying cytoplasm-associated lncRNAs across 15 human cell lines. We implemented multiple computational approaches, including machine learning, deep learning, and large language models; in order to achieve improved prediction accuracy. The second part of the thesis describes a method, lncrnaPI, developed to predict lncRNA–protein interacting pairs using transformer-based sequence embeddings as features. In the final section of the thesis, we explored the cancer diagnostic potential of lncRNAs based on their transcriptomic profiles. Specifically, we analyzed lncRNA transcriptomic profiles in patients with pancreatic ductal adenocarcinoma (PDAC). We explored two independent approaches for identifying lncRNA-based biomarkers capable of discriminating PDAC patients from healthy individuals. First, we developed machine learning–based models using traditional approaches for biomarker identification. Second, we investigated a new concept in which large language models were used for mining transcriptomic data. One key objective of this research work is to facilitate the scientific community working in the field of lncRNA biology. Therefore, we developed web-based servers and standalone tools that are publicly available fostering open science. The work presented in this thesis provides novel computational strategies for enhancing the functional annotation of lncRNAs, decoding their molecular interactions, and exploiting their expression patterns for clinically meaningful applications.

Table of contents

List of Abbreviations

Abbreviation	Full Form
AI	Artificial Intelligence
APAAC	Amphiphilic Pseudo-Amino Acid Composition
AUC	Area Under the ROC Curve
BAC	Bacterial Artificial Chromosome
BLAST	Basic Local Alignment Search Tool
CAGE	Cap Analysis of Gene Expression
ceRNA	Competing Endogenous RNA
CNRCI	Cytoplasm-to-Nucleus Relative Concentration Index
CNN	Convolutional Neural Network
CP	Chronic Pancreatitis
DACC	Dinucleotide-based Auto-Cross Covariance
DL	Deep Learning
DNA	Deoxyribonucleic Acid
EMT	Epithelial-Mesenchymal Transition
EV	Extracellular Vesicle
FISH	Fluorescence In-Situ Hybridization
GEO	Gene Expression Omnibus
HC	Healthy Controls
HGP	Human Genome Project
LGBM	Light Gradient Boosting Machine
LLM	Large Language Model
lncRNA	Long non-coding RNA
LPI	lncRNA-Protein Interaction
MCC	Matthews Correlation Coefficient
ML	Machine Learning
mRMR	Minimum Redundancy Maximum Relevance
mRNA	Messenger RNA
PDAC	Pancreatic Ductal Adenocarcinoma

PSSM	Position-Specific Scoring Matrix
RCI	Relative Concentration Index
RNA	Ribonucleic Acid
SPC	Shannon Entropy of Physicochemical Property
SVM	Support Vector Machine
TPM	Transcripts Per Million

List of Figures

List of Tables

List of Publications

Publication from Thesis

1. **Choudhury, S.**, Bajiya, N., & Raghava, G. P. (2025). Prediction of lncRNA-protein interacting pairs using LLM embeddings based on evolutionary information. *bioRxiv*, 2025–11.
2. Choudhury, S., Mehta, N. K., & Raghava, G. P. (2025). A large language model for predicting pancreatic ductal adenocarcinoma patients from blood-derived exosomal transcriptomics data. *bioRxiv*, 2025–03.
3. **Choudhury, S.**, Mehta, N. K., & Raghava, G. P. (2025). CytoLNCpred-a computational method for predicting cytoplasm associated long non-coding RNAs in 15 cell-lines. *Frontiers in Bioinformatics*, *5*, 1585794.
4. Choudhury, S., Rathore, A. S., & Raghava, G. P. (2024). Compilation of resources on subcellular localization of lncRNA. *Frontiers in RNA Research*, *2*, 1419979.

Other Publications

1. Aggarwal, S., Dhall, A., Patiyal, S., **Choudhury, S.**, Arora, A., & Raghava, G. P. (2023). An ensemble method for prediction of phage-based therapy against bacterial infections. *Frontiers in Microbiology*, *14*, 1148579.
2. Bajiya, N., **Choudhury, S.**, Dhall, A., & Raghava, G. P. (2024). AntiBP3: A method for predicting antibacterial peptides against Gram-positive/negative/variable bacteria. *Antibiotics*, *13*(2), 168.
3. Bajiya, N., Najrin, S., Kumar, P., **Choudhury, S.**, Tomer, R., & Raghava, G. P. (2025). CPPsite3: An updated large repository of experimentally validated cell-penetrating peptides. *Drug Discovery Today*, 104421.
4. **Choudhury, S.**, Bajiya, N., Patiyal, S., & Raghava, G. P. (2024). MRSLpred—A hybrid approach for predicting multi-label subcellular localization of mRNA at the genome scale. *Frontiers in Bioinformatics*, *4*, 1341479.
5. Dhall, A., Patiyal, S., **Choudhury, S.**, Jain, S., Narang, K., & Raghava, G. P. (2023). TNFepitope: A webserver for the prediction of TNF-α inducing epitopes. *Computers in Biology and Medicine*, *160*, 106929.

6. Gahlot, P. S., **Choudhury, S.**, Bajiya, N., Kumar, N., & Raghava, G. P. (2025). Prediction of Plant Resistance Proteins Using Alignment-Based and Alignment-Free Approaches. *Proteomics*, *25*(5–6), e202400261.
7. Jaganath, D., Sieberts, S. K., Raberahona, M., Huddart, S., Omberg, L., Rakotoarivelo, R., Lyimo, I., Lweno, O., Christopher, D. J., & Nhung, N. V. (2025). *Accelerating cough-based algorithms for pulmonary tuberculosis screening: Results from the CODA TB DREAM Challenge*. *12*(10), ofaf572.
8. Kumar, N., Choudhury, S., Bajiya, N., Patiyal, S., & Raghava, G. P. (2025). Prediction of Anti-Freezing Proteins From Their Evolutionary Profile. *Proteomics*, *25*(3), e202400157.
9. Kumar, N., Patiyal, S., **Choudhury, S.**, Tomer, R., Dhall, A., & Raghava, G. P. (2023). DMPPred: A tool for identification of antigenic regions responsible for inducing type 1 diabetes mellitus. *Briefings in Bioinformatics*, *24*(1), bbac525.
10. Rathore, A. S., **Choudhury, S.**, Arora, A., Tijare, P., & Raghava, G. P. (2024). ToxinPred 3.0: An improved method for predicting the toxicity of peptides. *Computers in Biology and Medicine*, *179*, 108926.
11. Rathore, A. S., Jain, S., Choudhury, S., & Raghava, G. P. (2025). A large language model for predicting neurotoxic peptides and neurotoxins. *Protein Science*, *34*(8), e70200.
12. Rathore, A. S., Kumar, N., **Choudhury, S.**, Mehta, N. K., & Raghava, G. P. (2025). Prediction of hemolytic peptides and their hemolytic concentration. *Communications Biology*, *8*(1), 176.
13. Tomer, R., Patiyal, S., Kaur, D., **Choudhury, S.**, & Raghava, G. P. (2024). Genome-based solutions for managing mucormycosis. *Advances in Protein Chemistry and Structural Biology*, *139*, 383–403.

1

Introduction

1.1 Historical Background and Discovery of lncRNAs

Investigations into the "C-value paradox," which states that biological complexity is not correlated with genome size, led to the discovery of long non-coding RNAs (lncRNAs) [1–3]. The total DNA content (C-value) of organisms with a similar number of protein-coding genes might vary considerably [2]. Despite having over 20,000 protein-coding genes, humans and simpler organisms like the marbled lungfish have very different morphological complexity. The percentage of non-coding genes in their genomes, as shown in Table 1.1, is the one of the causes of this discrepancy. These non-coding genomes were initially labelled as “junk DNA” but were later found to be transcriptionally active. These genes were transcribed to produce a new class of RNAs with no protein-coding potential. While these non-coding transcripts were initially dismissed as functional "noise," the discovery of *H19* and its essential role in murine embryonic development shifted the research paradigm in 1980 [4]. The characterization of *Xist* as an lncRNA that facilitates dosage compensation through the recruitment of Polycomb Repressive Complexes 1 and 2 (PRC1 and PRC2) [5–7]. These pioneering studies established lncRNAs as crucial components of organism development and modulators of genome architecture.

Table 1.1 - *Genome size and proportion of non-coding DNA across representative species.*

Species	Genome Size (Approximate)	Non-Coding Genome (%)	Protein-Coding Genes (Approximate)
Bacteria (*E. coli*)	4.6 Mb	12	4,300
Maize (*Z. mays*)	2,300 Mb (2.3 Gb)	98.5	32,000
Arabidopsis (*A. thaliana*)	135 Mb	75	27,000
Plasmodium (*P. falciparum*)	23 Mb	50	5,000
Ganoderma (*G. lucidum*)	43 Mb	60	10,000
Yeast (*S. cerevisiae*)	12 Mb	30	6,000
Mosquito (*A. gambiae*)	250 Mb	90	13,000

Fruit Fly (*D. melanogaster*)	140 Mb	80	14,000
Nematode (*C. elegans*)	100 Mb	75	20,000
Zebrafish (*D. rerio*)	1,400 Mb (1.4 Gb)	98	25,000
Marble lungfish (*P. aethiopicus*)	~130,000 Mb (130 Gb)	99.5	20,000
Frog (*X. tropicalis*)	1,700 Mb (1.7 Gb)	98	20,000
Chicken (*G. gallus*)	1,000 Mb (1.0 Gb)	97	18,000
Opossum (*M. domestica*)	3,600 Mb (3.6 Gb)	98	20,000
Cow (*B. taurus*)	2,700 Mb (2.7 Gb)	98.5	20,000
Rat (*R. norvegicus*)	2,800 Mb (2.8 Gb)	98.5	22,000
Mouse (*M. musculus*)	2,700 Mb (2.7 Gb)	98.5	22,000
Macaque (*M. mulatta*)	2,900 Mb (2.9 Gb)	98.5	20,000
Chimpanzee (*P. troglodytes*)	3,300 Mb (3.3 Gb)	98.5	20,000
Human (*H. sapiens*)	3,200 Mb (3.2 Gb)	98.5	20,000

1.2 Evolutionary Dynamics of lncRNAs

During the course of evolution, the genomic landscape has transitioned from compact prokaryotic genomes to heavily regulatory genomes of complex eukaryotes. The number of protein coding genes has remained nearly constant (around 20,000) in higher multicellular organisms like nematodes, chicken, mouse, chimpanzee and humans. However, the proportion of non-coding genome has increased exponentially [8]. This drastic increase in non-coding genome, suggests that the organismal complexity is not linked to the protein coding genes. The evolutionary progress is rather tied to the expansion of this non-coding regulatory network, which precisely controls how, when, and where those proteins are expressed. Among the non-coding genes, lncRNAs have emerged as one of the key players in the regulatory network.

The most important characteristic of lncRNA evolution is how their primary sequence rapidly diverges, even as their tissue-specific expression and regulatory functions are maintained. It has

been observed that lncRNA sequences are poorly conserved across species but their promoters are strongly conserved. Even among closely related mammals, the sequence identity frequently falls below 70%. [9]. Despite this sequence-level divergence, their genomic position relative to neighboring genes are conserved. This preserves the thermodynamic stability required for protein-RNA and DNA-RNA interactions [9,10]. Furthermore, many lncRNAs are "born" through the exaptation of transposable elements (TEs), which provide ready-made regulatory motifs for transcription initiation and splicing, driving lineage-specific such as primate-specific cognitive expansion [8]. Recent multidimensional analyses confirm that while sequence conservation is sparse, "positionally conserved" lncRNAs are often linked to critical developmental transcription factor loci, maintaining their regulatory roles across millions of years of vertebrate evolution [11,12].

1.3 Early Characterization

The initial lncRNAs *H19* and *Xist*, were identified using labor intensive, locus-specific methods such as subtractive hybridization and Northern blotting [6,13]. The early discoveries were limited by the lack of any genomic reference. Sanger sequencing required individual clones of target genes, which was a slow and costly process [14]. Most lncRNAs lack an open reading frame and had very low transcript abundance, leading them to be classified as transcriptional noise or experimental artifacts. In the late 1990s, the National Institute of Health, USA, initiated the Human Genome Project (HGP) which aimed to sequence the entire human genome and generate a detailed genomic reference [15]. They utilized Bacterial Artificial Chromosome (BAC) physical maps and capillary Sanger sequencing to create the reference scaffold against which all RNA could be mapped [16]. One of the major findings of the HGP was the significantly lower protein coding genes identified at the end of project in 2004 [15]. Based on the repartition of the CpG islands the number of protein coding genes was predicted to be around 70,000 – 80,000, but the final number of protein coding genes identified came out to be 22,857. The timeline of these discoveries is shown visually in Figure 1.1.

Functional Annotation of the Mammalian Genome (FANTOM) project improved the annotation of lncRNAs in mammals [17]. They used Cap Analysis of Gene Expression (CAGE) and full-length cDNA library cloning, where the 5′ ends of capped transcripts are sequenced and this

unveiled nearly 23,000 novel ncRNAs. Further research conducted by the ENCODE consortium, identified that only 1.2% of the genome are protein coding whereas about 93% of the genome was being transcribed [18]. In the 2012 report of the ENCODE Project, they annotated 9,640 manually curated lncRNA loci in the human genome. These pilot projects highlighted that pervasive transcription products are not junk, they play an important role in cell physiology.

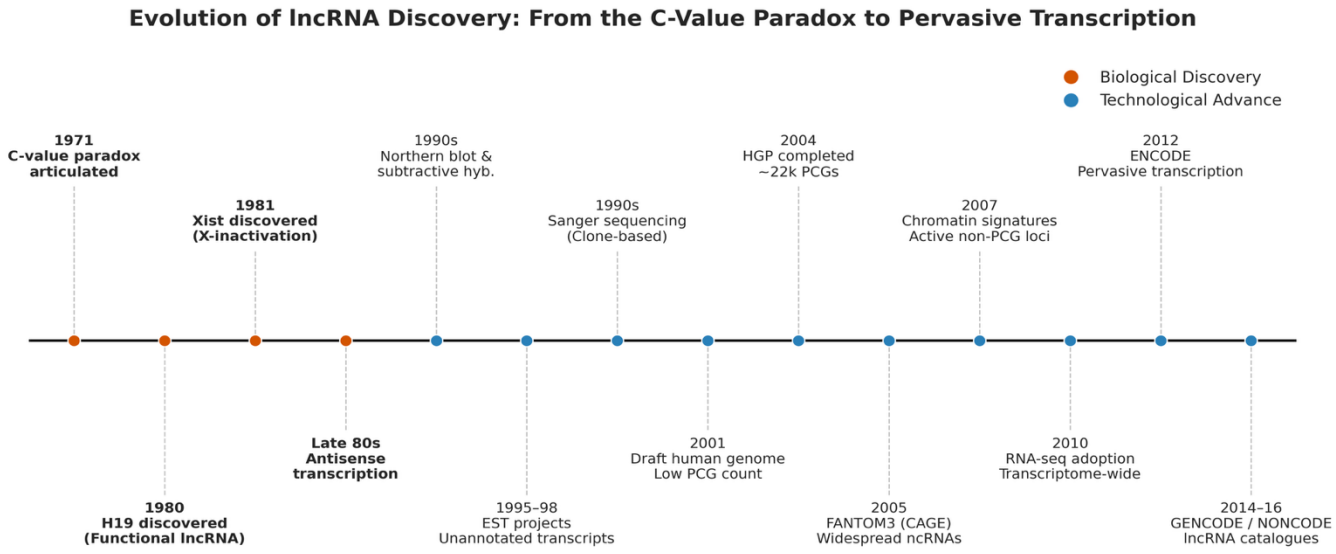


Figure 1.1 - *Chronology of key biological milestones and technological advances in lncRNA research*

1.4 Biogenesis of lncRNAs

The transcription of long non-coding RNAs (lncRNAs) is primarily mediated by RNA polymerase II (Pol II), originating from diverse genomic loci including intergenic regions, enhancers, and the antisense strands of protein-coding genes [19,20]. In their nascent state, most lncRNAs undergo canonical mRNA-like processing, characterized by the addition of a 5'-methylguanosine cap and 3'-polyadenylation [21]. However, lncRNAs biogenesis frequently deviates from the standard mRNA pathway as many lncRNA transcripts exhibit inefficient splicing kinetics. This incomplete

processing often exposes cis-acting retention signals within retained introns, thereby facilitating the nuclear sequestration of the transcript rather than its export to the cytoplasm.

A number of specialized biogenesis pathways have been identified in addition to canonical processing. Some lncRNAs rely on different stabilizing mechanisms because they do not have a poly-A tail. For example, instead of polyadenylation, the abundant nuclear transcript *MALAT1* undergoes 3'-end processing via RNase P cleavage, creating a stable triple-helical structure that shields the transcript from exonucleolytic destruction [22]. Furthermore, distinct classes of lncRNAs arise from intron-derived fragments, such as sno-lncRNAs, which are capped by small nucleolar RNAs (snoRNAs) at both termini [23]. Additionally, circular RNAs (circRNAs) are generated through non-canonical "back-splicing" events, wherein a downstream splice-donor site is covalently ligated to an upstream splice-acceptor site, forming a covalently closed loop that confers exceptional resistance to degradation [24,25].

1.5 Major function of lncRNAs

lncRNAs are known to be involved in a wide array of cellular process ranging from chromatin reorganization to post-transcription regulation. Since by definition lncRNA are non-coding, they generally perform their intended functions by interacting with proteins or DNA. Figure 1.2 provides an overview of the major function performed by lncRNAs.

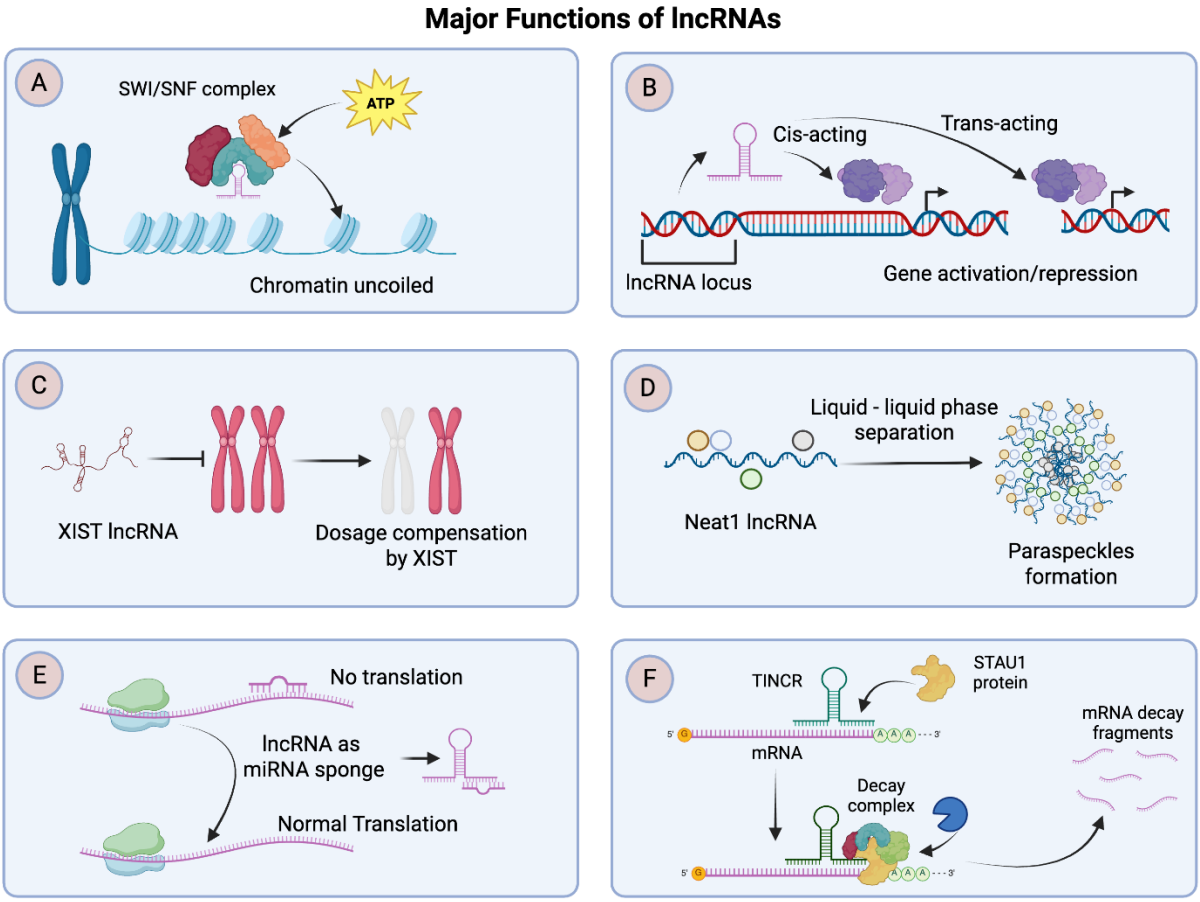


Figure 1.2 - *Regulatory mechanisms and functional diversity of long non-coding RNAs (lncRNAs). (A) Modulation of chromatin architecture through the recruitment of remodeling complexes such as SWI/SNF. (B) Transcriptional control of gene expression via cis- or trans-acting mechanisms. (C) Dosage compensation mediated by XIST-induced X-chromosome inactivation. (D) Structural scaffolding of nuclear bodies (e.g., paraspeckles) through Neat1-driven liquid-liquid phase separation. (E) Sequestration of microRNAs (miRNA sponging) to facilitate target mRNA translation. (F) Regulation of mRNA stability and degradation via interaction with the STAU1-mediated decay complex.*

1.5.1 Modification of Chromatin Architecture

Nuclear lncRNAs are known to alter the chromatin environment by interacting with DNA (both in cis- and trans-) [26]. These lncRNAs primarily localize on the chromatin, where they form complexes with proteins and can enhance or inhibit their function at targeted DNA regions [27].

The expression levels of a given lncRNA in relation to the factors it interacts with can define the extent of the effects that lncRNAs exert on the targeted chromatin [27]. By interacting with chromatin modifiers and directing them to target-gene promoters, lncRNAs have the ability to either activate or repress gene transcription in cis or trans at a distance. [26,28]. lncRNAs can act as decoys of specific chromatin modifiers and prevent them from binding to promoter regions of target genes [27,29]. lncRNAs can also transcriptionally form RNA–DNA hybrid (like R-loops) by interacting with DNA. These hybrid structures work in conjunction with chromatin modifiers or transcription factors to control expression of the target genes. [29,30]. Few studies have reported that RNA–DNA–DNA triplexes can play a crucial role in gene silencing or activation [32]. In this case, the propensity for triplex formation is mostly determined by the RNA sequence.

1.5.2 Transcription regulation

The most common and studied mechanism of gene repression by lncRNAs are linked to dosage compensation [27]. *Xist* coats the entire length of one X chromosome in female mammals to recruit silencing complexes, ensuring females have the same dosage of X-linked gene products as males [33]. lncRNAs can also regulate gene expression by meddling with the transcription machinery. They perform this function either by affecting the recruitment of transcription factors or RNA polymerase II or by altering histone modification to change chromatin accessibility [34]. lncRNAs can promote the expression of protein-coding genes (PCGs) through preformed chromatin loops that recruit activating complexes. A key example is *HOTTIP* which coordinates the formation of a loop to bring *WDR5/MLL* complexes into close proximity with *HOXA* genes to drive their transcription [35].

1.5.3 Roles in scaffolding and condensates

LncRNAs act as structural skeletons, or scaffolds, that physically hold different proteins and molecules together to help them interact and perform downstream functions [36]. By forming biomolecular condensates, which are discrete droplets inside the cell generated without a membrane, they are able to organize the inside of the cell thanks to their scaffolding ability [37]. To speed up reactions or store ingredients for later use, lncRNAs can draw particular elements into

these droplets. To deal with cellular stress, *NEAT1* recruits particular proteins to form nuclear structures known as paraspeckles [38]. Similar to this, *MALAT1* arranges nuclear speckles to concentrate splicing tools in a single location, guaranteeing effective gene processing [39].

1.5.4 Post-transcriptional regulation

LncRNAs regulate the preservation, translation, and destruction of messenger RNAs (mRNAs) in the cytoplasm and are extremely active long after transcription is complete. Serving as a "sponge" or competitive endogenous RNA (ceRNA) is one of their most frequent functions. lncRNAs like *PTENP1* act as decoys that soak up microRNAs [40]. These bound miRNAs are unable to bind to their intended mRNA targets and destroy them. lncRNAs can also physically interact with mRNAs to determine their stability. *BACE1-AS* lncRNA wraps around the *BACE1* mRNA to form a protective double-stranded shield, blocking degradation enzymes and causing a buildup of proteins associated with Alzheimer's disease [41]. On the other hand, some lncRNAs trigger destruction of mRNAs. The *TINCR* lncRNA binds to specific mRNAs and recruits the Staufen1 protein to break them down, ensuring that certain differentiation genes are turned off rapidly [42]. lncRNAs can also control the level of gene expression by attaching directly to ribosomes or translation factors [43]. In this case, they act as a dimmer switch to increase or decrease the actual production of proteins from the mRNA strand.

1.6 lncRNAs implicated in diseases

Recent studies have reported the role of lncRNAs in various diseases. Figure 1.3 summarizes commons diseases with their associated lncRNAs. Their dysregulation has been recognized as a primary driver in oncogenesis, where they may function as oncogenes or tumor suppressors. The lncRNA *HOTAIR* recruits PRC2 and silences the metastasis-suppressor genes in breast and colorectal cancers [44]. Similarly, *MALAT1* is significantly upregulated in non-small cell lung cancers and it is known to promote cell migration and invasion [45]. As a result, these cancer-causing transcripts are being investigated not only as therapeutic targets but also as non-invasive biomarkers for early discovery. The expression of *PCA3* in urine samples has been approved by the FDA as a non-invasive test for early detection of prostate cancer [46].

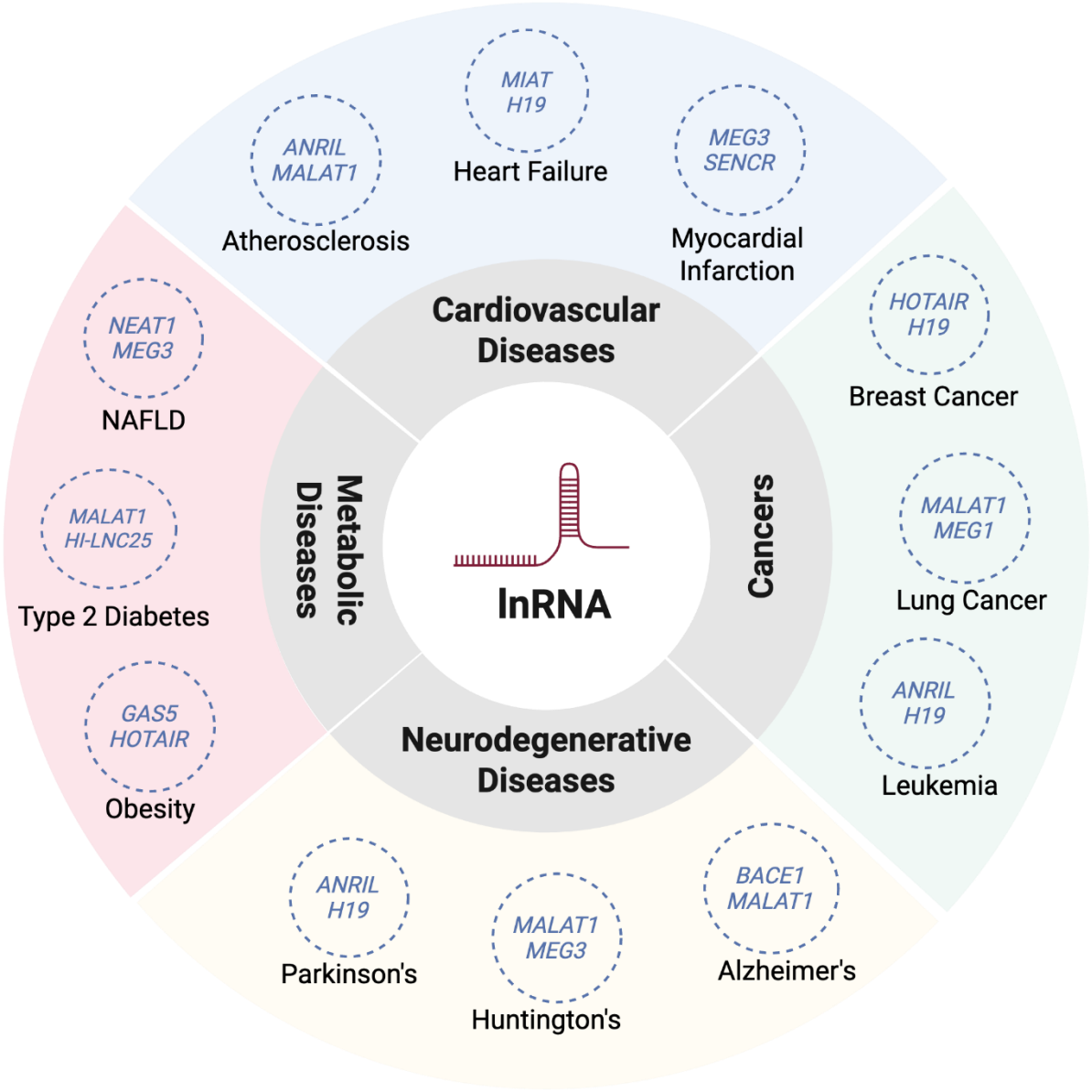


Figure 1.3 - *An overview of lncRNAs implicated in various diseases*

In the central nervous system, lncRNAs are critical for synaptic plasticity and neuronal development, and their malfunction is intimately linked to neurodegenerative disorders. In Alzheimer's disease, the lncRNA *BACE1-AS* is upregulated and forms a duplex with *BACE1* mRNA [41]. This stabilizes the *BACE1* mRNA leading to accumulation of beta-amyloid plaques, which is a hallmark of Alzheimer's disease. The lncRNA *HTT-AS* acts as a repressor of the *HTT*

gene and it is downregulated in the case of Huntington's disease [47]. This results in increased expression of the mutant huntingtin protein which accelerates neuronal toxicity. These mechanisms highlight the regulatory role of lncRNAs in maintaining neuronal homeostasis and their potential utility in RNA-based therapeutics.

lncRNAs also play pivotal roles in cardiovascular and metabolic pathophysiology. According to genome-wide association studies (GWAS), the *ANRIL* locus is a major genetic risk factor for coronary artery disease and atherosclerosis [48]. By recruiting epigenetic modifiers to the *p15/CDKN2B* tumor suppressor locus, *ANRIL* affects vascular wall cell proliferation and senescence. Microvascular dysfunction and diabetic retinopathy have been linked to the long noncoding RNA *MIAT* [49,50]. It controls the migration and proliferation of endothelial cells and functions as a competitive endogenous RNA (ceRNA) for *miR-150*. These instances demonstrate the crucial role lncRNAs play in the vascular issues linked to metabolic disorders.

1.7 Lack of Functional Annotation of lncRNAs

Majority of lncRNAs remain functionally unannotated despite their growing importance. Unlike their protein-coding counterparts, lncRNAs lack conserved domains, exhibit rapid evolutionary divergence, and display highly context-dependent expression patterns. It is challenging to deduce function solely from sequence because of this tendency. Because there is so little information now available on lncRNAs, it is challenging to apply traditional annotation techniques like homology, domain prediction, or structural similarity. Experimental characterization by subcellular imaging, CLIP-based mapping, overexpression systems, or knockdowns can be time-consuming, expensive, and frequently complicated by overlapping genetic features or low expression levels. Additionally, lncRNAs behave differently depending on the cell type, therefore functional roles that are established in one cell line cannot be assumed in another. Only a small percentage of the thousands of lncRNAs listed in public databases have mechanistic evidence to support them. The incorporation of lncRNAs into a diagnostic or therapeutic paradigm has been hampered by this annotation bottleneck, which also renders transcriptome analysis results challenging to understand. Consequently, lncRNA's functional discovery will be greatly accelerated exponentially by the advancement of computational techniques.

1.8 Proposal Origin

This thesis is based on the increasing understanding that long non-coding RNAs (lncRNAs) are essential regulators of cellular signaling, RNA processing, chromatin architecture, and gene expression. Some lncRNAs can be so essential for biological functions that they can be compared to genes that code for proteins. Their biological and clinical significance is highlighted by their tissue-specific expression, dynamic subcellular localization, and extensive participation in development and illness. Most lncRNAs are still functionally unannotated, despite their essential role in cellular physiology. Advancements in biomarker integration, interaction profiling, and localization mapping are lagging behind, in spite of the fact that data generation has increased at a faster rate. The current lot of computational methods suffer from multiple drawbacks. Among these are out-of-date datasets, lncRNA databases containing mRNA data, and webservers that aren't working. This gap between functional understanding of lncRNAs and the abundance of available data calls for the development of computational methods that are backed by biology.

LncRNAs are functionally involved in a variety of biological processes, have dynamic subcellular localization, and exhibit tissue-specific expression patterns. lncRNAs are known to be involved in developmental processes, disease pathways and immunological dysregulation. Despite of their biological and therapeutic relevance, our knowledge on the functional aspect of lncRNAs is still inadequate. The major issue with the current lot of computational tools is their reliance on incomplete or out-of-date reference datasets. Majority of lncRNA-protein interaction tools suffer from this issue, and it hinders their ability to make accurate predictions. Moreover, the expression of lncRNA is not yet fully explored in the context of development of non-invasive biomarkers for cancers.

In order to address these issues, there is a need for computational methods that are systematic, scalable, and based on cellular biology. This need drives the creation of reliable methods for modeling lncRNA–protein interactions, predicting lncRNA subcellular localization, and using expression profiles for non-invasive cancer screening. drives the creation of reliable techniques for modeling lncRNA–protein interactions, predicting lncRNA subcellular localization, and using expression profiles for non-invasive cancer screening.

1.9 Objective of the Thesis

The goal of this thesis is to fill significant gaps in lncRNA functional annotation by developing computational techniques that focus on three crucial aspects of lncRNA biology: subcellular localization, protein interaction prediction, and cancer biomarker detection. Our first goal was to methodically compile a knowledgebase of relevant sources. Using state-of-the-art machine learning techniques, larger and cleaner interaction datasets, and contemporary sequence embeddings, our second goal is to create an approach to predict lncRNA–protein interactions. Third, we investigate the translational potential of lncRNAs by employing novel representation learning techniques and lncRNA expression signatures to create reliable, non-invasive cancer detection methods, specifically for pancreatic ductal adenocarcinoma. All of these goals work together to enhance the use of lncRNAs in fundamental research and precision medicine by offering integrative methods and insights that facilitate a deeper functional knowledge of these molecules.

1.10 Organization of the Chapters

The entire work has been structured into eight chapters to provide a clear and logical progression of the research carried out in this thesis.

Chapter 1***:*** This chapter provides a comprehensive overview of long noncoding RNAs. We've discussed their discovery, biological relevance, and emerging role in regulation and disease. It explains why understanding lncRNA localization, interactions, and clinical potential is critical. This chapter discusses the motivation for conducting this research.

Chapter 2: This chapter provides a thorough assessment of the existing literature on lncRNA biology. It includes previous research on their subcellular distribution, interactions with proteins, and use as cancer biomarkers. It outlines existing computational and experimental work in each domain, identifies methodological shortcomings, and provides a scientific rationale for the current investigation.

Chapter 3: In this chapter, all of the resources currently available on lncRNA subcellular localization are systematically gathered, curated, and integrated. It creates a single online resource by combining databases, tools, and experimental data. The foundation for the following work is established in this chapter.

Chapter 4: Computational methods for predicting lncRNA localization in various cell types are the main topic of this chapter. It outlines the feature engineering, machine learning, and dataset preparation techniques used, along with performance reviews and comparisons to other approaches.

Chapter 5: This chapter gives a summary of the computational methods used to model interactions between lncRNA and proteins. It includes machine learning pipelines, embedding techniques, dataset construction, and benchmarking against existing LPI tools.

Chapter 6: Using lncRNA expression data, a diagnostic model for pancreatic ductal adenocarcinoma is developed. The potential of lncRNAs for non-invasive cancer detection is highlighted in this chapter.

Chapter 7: This chapter investigates the application of large language models for the classification of cancer based on transcriptomes. The process of transforming expression data into sequence-like representations is described in detail, and the effectiveness of sophisticated LLMs for reliable cancer prediction is assessed.

Chapter 8: In this chapter, the main conclusions of the thesis are compiled, using knowledge from all methodological sections. It lists the work's primary contributions and talks about its main drawbacks. Additionally, this chapter suggests future research avenues in translational applications and functional annotation of lncRNA.

2

Review of Literature

Long non-coding RNAs (lncRNAs) are defined as transcripts exceeding 200 nucleotides in length that lack functional protein-coding [9]. According to GENCODE annotations the number of lncRNA genes in the human genome is reported over 16,000, while some other databases report figures exceeding 100,000 [51,52]. These molecules regulate gene expression through interactions with DNA, RNA, and proteins, and depending on their subcellular localization, can modulate chromatin architecture, regulate nuclear body assembly, alter mRNA stability and translation, and interfere with signaling pathways [9]. The tissue-restricted expression of lncRNAs—roughly 78% of them are categorized as tissue-specific, compared to just 19% of mRNAs—is a particularly noteworthy feature that makes them highly desirable as potential biomarkers [53]. As a result of their close involvement in the spatial regulation of gene expression during development, many lncRNAs are transcribed from enhancers, associate with chromatin-modifying complexes, and induce phase separation of nuclear condensates [54]. Certain members of this class exert regulatory effects on a scale comparable to protein-coding genes; a single lncRNA can regulate the expression of hundreds of protein-coding genes, meaning its dysregulation could produce cascading effects across entire gene expression programmes.

The pathological importance of lncRNAs cannot be ignored anymore. In cancers, it has been observed that lncRNAs regulate tumor cell behavior including invasion, proliferation, and epithelial-mesenchymal transition [55]. They also function as key regulators of immune cells such as T cells, B cells, macrophages, and dendritic cells, which are closely tied to anti-tumor immunity and immune evasion. The most studied lncRNAs –*HOTAIR* and *NEAT1* have been linked to poor prognosis and treatment resistance across multiple malignancy types [56,57]. Apart from cancer, the lncRNA *NEAT1* has also shown great diagnostic and prognostic potential in inflammatory conditions and neurodegenerative diseases [58]. Dysregulated lncRNA expression has been linked to various autoimmune and immune-related conditions like systemic lupus erythematosus, rheumatoid arthritis, and multiple sclerosis [58–60]. Neurological associations have been reported in glioblastoma, Parkinson's disease, and Alzheimer's disease, where lncRNA signatures were found to correlate with infiltration of immune cells patterns and stratify survival for patients [61–63].

Despite this breadth of involvement, the functional characterization of lncRNAs has not kept pace with their rate of discovery [64,65]. The majority of transcripts that have been cataloged lack a

biological purpose, and there are still significant limitations in the computational infrastructure that can be used to fill this gap [66]. Subcellular localization prediction tools often use reference datasets that are either out-of-date or no longer represent the state of annotation [67,68]. Most lncRNA-protein interaction tools are usually trained on small experimental datasets that are biased towards well-characterized protein-coding [69]. In cancer biomarker research, lncRNA expression values are often reduced to numerical inputs for classification models, stripped of the regulatory context that gives those values biological meaning [70]. This thesis takes that persistent gap as its starting point, and is structured around three interconnected objectives: predicting lncRNA subcellular localization, modelling lncRNA interactions with proteins, and evaluating lncRNA expression profiles as a basis for non-invasive cancer diagnostics.

2.1 Understanding the Role of Subcellular Localization

Where a lncRNA resides within the cell is not incidental to its function - it is, in many respects, the primary determinant of it [71]. The subcellular localization of a lncRNA governs which molecular partners it can access and which biological processes it is positioned to influence, with nuclear and cytoplasmic populations performing fundamentally distinct regulatory roles. This principle was established early in the history of lncRNA research. The first functionally characterized lncRNAs were predominantly nuclear, among them *XIST*, which orchestrates X-chromosome inactivation through chromatin silencing, and *MALAT1* and *NEAT1*, identified in an early screen for nuclear-enriched polyadenylated transcripts [72,73]. *MALAT1* accumulates in nuclear speckles, while *NEAT1* forms the structural scaffold of paraspeckles, with both regulating splicing and transcription [73,74]. These early characterizations, however, created an impression that lncRNAs are predominantly nuclear, an assumption that subsequent transcriptome-wide mapping has substantially revised. Subcellular RNA localization is not a stochastic process but a dynamic and tightly regulated one, responsive to homeostatic conditions, external stimuli, and cellular stress. Cytoplasmic lncRNAs are now recognized as a substantial and functionally diverse population, performing roles in mRNA stability, translational regulation, and post-transcriptional gene control. In the cytoplasm, lncRNAs act as sponges for microRNAs and interact with RNA-binding proteins to modulate mRNA stability and translation, while also serving as scaffolds that nucleate protein complexes involved in signal transduction [9,71]. Depending on its compartment,

a single transcript can exhibit different behaviors. For example, *PYCARD-AS1* uses chromatin-modifying enzymes to decrease translation of its target mRNA in the cytoplasm and silence gene transcription in the nucleus. Location is not limited to the nucleus-cytoplasm binary [75]. Through mechanisms that are still unclear, ribosome-associated lncRNAs have been linked to translational control, mitochondrial lncRNAs take role in metabolic regulation, and those linked to stress granules react to proteotoxic challenge.

This localization, however, is not fixed. The biogenesis of lncRNAs is distinct from that of mRNAs and is linked to their specific subcellular localization and functions, with cell type, developmental stage, and pathological context all capable of shifting the distribution of a given transcript between compartments [71]. A lncRNA that is predominantly nuclear in one cell type may be enriched in the cytoplasm in another, driven by differences in splicing efficiency, RNA-binding protein availability, post-transcriptional modifications, and sequence-embedded localization signals [76]. This context-dependence is biologically important but also presents a significant challenge for experimental and computational characterization alike, because only a relatively small proportion of lncRNAs have annotations regarding their subcellular localization despite the vast number that have been identified.

The databases that exist to address this gap each carry limitations that constrain their utility. LncATLAS, one of the most widely used resources, represents 6,768 GENCODE-annotated lncRNAs across compartments of 15 cell lines, captures only a fraction of the annotated lncRNA transcriptome and is restricted to a small number of cancer-derived and immortalized cell lines that may not reflect the localization patterns of lncRNAs in differentiated tissues [77]. LncATLAS localizations derived from cell lines may not be predictive of localization in differentiated tissue or living organisms. RNALocate v2.0, a broader repository integrating manually curated literature alongside RNA-sequencing datasets, contains over 213,000 RNA subcellular localization entries across 171 locations and 104 species, but encompasses multiple RNA classes alongside lncRNAs and applies discrete qualitative labels that are not directly comparable to the quantitative metrics used by cell-fractionation approaches [78]. The number of lncRNA sequences in RNALocate v2 has decreased significantly relative to the first release, leaving researchers with reduced localization information rather than expanded coverage. The majority of computational models created for localization prediction were trained using datasets from either of these sources, thus

they share the same cell-line biases, coverage gaps, and class imbalances. According to recent benchmarking studies, models tested on unfiltered lncRNA data only attain about 60% accuracy, suggesting that predicting lncRNA subcellular localization from sequence features is a much more difficult task than commonly reported performance figures imply. No single resource currently integrates experimental localization data with comprehensive metadata across cell types, tissue contexts, and compartments in a manner that would adequately support the development of generalizable predictive models.

2.1.1 Existing Tools for Subcellular Localization

The evolution of computational techniques for predicting the subcellular localization of lncRNA has matched the field's overall growth. The first predictor created for lncRNA subcellular localization was lncLocator, which was released in 2018 [68]. This was followed by a number of tools, such as iLoc-lncRNA, DeepLncRNA, Locate-R, LncLocation, and a number of deep learning variations that continued until 2024 [79–82]. Using nucleotide composition data like k-mer frequencies and pseudo-nucleotide compositions, early methods framed the challenge as a multi-class classification task and paired them with traditional machine learning classifiers like random forests and support vector machines. By directly applying deep learning to lncRNA transcript sequences, DeepLncRNA marked an initial attempt to go beyond hand-crafted features and achieved an area under the ROC curve of 0.787 spanning nuclear and cytosolic classes [80]. By adding structural descriptors, word embedding representations, and graph-based sequence encodings, other tools aimed to build upon this basis. LncLocFormer, a transformer-based model that uses a localization-specific attention mechanism, and SGCL-LncLoc, an interpretable deep learning framework that uses supervised graph contrastive learning, are more recent additions to this domain [83,84]. They were both published in 2023 and 2024, respectively.

Even with the development of these methods, there are still some issues that are yet to be addressed. The primary concern here is the training dataset that is used in these methods. Tools such as DeepLncLoc, GM-lncLoc, and GraphLncLoc, were trained on datasets obtained from RNALocate v1 [85–88]. This means that the same dataset has been repurposed across multiple tools. The dataset created from RNALocate v1 suffers from redundant data, inconsistent annotation standards, and mRNA contamination. These technical limitations undermine the quality of the

datasets and raise serious questions about the validity of the methods built on these datasets. Datasets created from lncAtlas, where they have quantitatively defined subcellular localization, also has its own limitations. Since, these datasets have numerical values, they deal with issues like incorrect normalization, non-standard thresholding, or label assignment procedures. These issues can introduce bias in the models and decrease the biological validity of downstream models. Also, some studies have chosen to drop lncRNAs that do not have extreme localization values. Oversampling strategies like SMOTE, have been used in a variety of tools to resolve class imbalance [89]. However, these strategies tend to introduce the risk of overfitting and can mislead models by distorting the true distribution of localization patterns.

Most of the tools also fail to account for the cell-line specific localization of lncRNAs. Analyses across 15 cell lines have identified a category of lncRNAs that switch compartments between cell types, complicating any attempt to assign a fixed localization label to a given transcript and undermining the assumption of a stable, sequence-determined localization that most current tools implicitly rely upon. Existing tools that rely on traditional encoding techniques such as k-mer profiles and one-hot representations overlook the richer biological information embedded in RNA sequences, and there remains a significant gap in the application of pre-trained RNA language models to the localization prediction problem. Taken together, these limitations point to a need for tools trained on cleaner, more representative datasets, designed with cell-type context as a first-class consideration, and evaluated under conditions that reflect the true difficulty of the prediction task rather than artificially curated subsets.

Table 2.1 *- Computational Tools for Predicting lncRNA Subcellular Localization*

Tool	Approach	Features	Compartments	Limitations	Citation
lncLocator (2018)	RF + SVM	k-mer + autoencoder	5	No cell-type specificity; outdated data	[68]
iLoc-lncRNA (2018)	SVM	PseKNC	4	Removes multi-localized RNAs	[79]
DeepLncRNA (2018)	Deep NN	k-mer + motifs	2	Only nuclear/cytosolic	[80]

lncLocation (2020)	SVM	sequence + structure	4	Limited generalizability	[82]
Locate-R (2020)	LD-SVM	k-mer + n-gap	4	Oversampling artifacts	[81]
lncLocPred (2020)	Logistic regression	PseDNC + k-mer	4	Noisy training sets	[89]
lncLocator 2.0 (2021)	CNN-BLSTM	word embeddings	2	Discards borderline-localized lncRNAs	[90]
TACOS (2022)	AdaBoost	physicochemical + compositional	2	Includes mRNAs; not cell-specific	[91]

2.2 lncRNA–Protein Interactions

The vast majority of functional lncRNAs exert their biological effects through interactions with RNA-binding proteins, assembling into ribonucleoprotein complexes that are then recruited to specific nucleic acid targets or directed to defined subcellular compartments [92,93]. Early studies have classified lncRNAs into five classes based on their function - guides, scaffolds, decoys, enhancers, and signals [92]. Guide lncRNAs recruit chromatin-modifying enzymes or transcription factors to their target genomic loci. Scaffold lncRNAs assemble multi-protein complexes that enable enzymatic cascades or coordinate chromatin remodeling. Decoy lncRNAs bind to targets of transcription factors or RNA-binding proteins, preventing them from binding with their usual targets [30].

The range of protein partners involved is correspondingly broad, encompassing chromatin modifiers such as PRC2 and MLL, sequence-specific transcription factors including p53 and NF-κB, and large families of RNA-binding proteins such as hnRNP and PTBP1 [30,54]. A given lncRNA can contain multiple structural or sequence domains that interact with multiple proteins at the same time or selectively depending on conditions, meaning these functional categories are not mutually exclusive. *HOTAIR*, for instance, not only associates with PRC2 to promote histone methylation at target loci but also interacts with hnRNP B1, which remodels *HOTAIR* structure to

relieve RNA-mediated inhibition of PRC2 catalytic activity, a mechanism that depends on RNA-RNA base-pairing between HOTAIR and nascent transcripts at target genes [94].

Mapping these interactions experimentally has required the development of dedicated methodologies. Protein-centric techniques such as RIP-seq and CLIP-seq, alongside RNA-centric approaches including CHART and ChIRP, have collectively identified thousands of lncRNA-protein associations across diverse cell types and conditions [95–98]. A number of databases, including StarBase, RNAInter, POSTAR, NPInter, and RAIN, have been created to record these interactions [99–103]. These databases combine data from many experimental sources with computational predictions and literature mining.

Although computational methods for predicting lncRNA-protein interactions (LPIs) have advanced significantly, there are still face some challenges. Prediction bias is introduced by most available approaches as have only been tested on one LPI dataset (mostly NPInter 2.0). Also, some methods are unable to make an accurate prediction on novel lncRNAs or proteins, that were not present in their training data. Without taking into consideration higher-order structure, cellular context, or the modular domain architecture that controls many lncRNA-protein contacts, early machine learning techniques framed LPI prediction as a sequence-matching problem, depending on k-mer composition and physicochemical features taken from lncRNA and protein sequences. By directly learning hierarchical sequence representations from data, deep learning techniques have significantly increased predictive performance; yet, they remain hindered by the training sets they have been trained on. Only a few hundred of the estimated 2,000 human RNA-binding proteins have sufficiently extensive interaction datasets, emphasizing the basic lack of data that restricts the applicability of existing models. One of the most important limitations in the field is this disparity between biologically scope of the issue and the empirical resources available to solve it; any computational framework for LPI prediction must directly address this issue.

2.2.1 Existing Computational Work on LPI Prediction

In the past ten years, computational methods for predicting lncRNA–protein interactions (LPIs) have gone through several stages of development, each based on the machine learning frameworks that were popular at the time. Initial research regarded LPI identification predominantly as a binary

classification challenge utilizing sequence-derived data. In these methods, lncRNA and protein sequences were converted into numerical attributes, including k-mer composition, physicochemical descriptors, and anticipated secondary structural features. Tools like lncPro, RPI-Pred, and RPISeq used these representations along with traditional machine learning models, especially support vector machines and random forests, to tell the difference between pairs that interact and pairs that don't [104–106]. These studies showed that it was possible to computationally predict LPIs, but the models relied a lot on manually designed features. These kinds of representations often couldn't account for the long-range dependencies or the structural flexibility that lncRNAs have, since their length and shape variability make it hard to encode features in a simple way.

Acknowledgment of these constraints resulted in the integration of biological network information into predictive models. Interaction networks show functional relationships that can't always be captured out just by looking at sequence properties. So, methods like LPIHN used random walk strategies across integrated co-expression and protein–protein interaction networks to find possible lncRNA–protein connections based on topological proximity and network similarity [107]. This network-based approach provided additional biological context, but its effectiveness was dependent on the completeness of the interaction information. These networks are still incomplete since many lncRNAs are poorly understood. Later studies used ensemble frameworks, which merged many classifiers and varied feature sets to improve prediction stability. LPI-BLS, RPI-SE, LPI-HYADBS, and SAWRPI all utilized this method and reported that their performance improved on benchmark datasets [108–111]. However, the comparatively shallow representations produced from manually constructed features continued to limit these models' ability to make predictions.

Recent advancements in deep learning have presented an alternative approach for modeling LPI prediction. Neural network architectures can learn hierarchical representations directly from raw sequence data instead of using predefined descriptors. Convolutional neural networks are often used to find local sequence motifs, while recurrent models like bidirectional LSTMs or GRUs look at the sequence as a whole to find larger contextual relationships. By combining these architectures, it is possible to present both short- and long-range sequence patterns in a single model. Attention mechanisms have enhanced performance in multiple studies by emphasizing informative sequence regions during training. Building on these advancements, graph-based deep

learning techniques have been proposed to represent the intricate connections found in biological interaction networks. This concept is extended by graph convolutional networks, graph autoencoders, and more recently graph transformers, which use attention-based reasoning over interaction graphs to model multi-step associations between lncRNAs and proteins. [112].

Despite methodological advances, most published lncRNA–protein interaction (LPI) predictors exhibit key limitations. Many tools are evaluated on a single dataset, leading to dataset-specific bias and limited evidence of generalizability. Two interconnected problems limit the discovery potential of existing models of lncRNA-protein interactions. First, they perform poorly when they come across novel proteins or lncRNAs, making it difficult for them to generalize beyond known interactions. Second, rather than being challenged with truly novel data, the field has used a limited set of benchmark resources, primarily NPInter v2.0, which has led to a training scenario where numerous models are iteratively built on the same annotated interaction dataset. [113]. In addition to limiting dataset diversity, this approach raises the possibility of overestimating performance. The problem of data leakage is another major worry. A number of research divide training and test sets without taking protein clustering or sequence similarity into consideration. Highly identical or homologous proteins may therefore show up in both sets, increasing evaluation metrics and masking the model's actual ability to generalize to new sequences.

The absence of uniform benchmarking methods also makes it more difficult to compare published techniques objectively. Differences in dataset compilation, preprocessing, negative sampling tactics, and evaluation protocols impede repeatability and fairness in performance assessment. Most of these methods are also not available as functional web servers, which limits their applicability to researchers without the computing capacity to install them locally. All of these limitations suggest that LPI prediction frameworks must be trained on carefully selected, non-redundant datasets, tested in scenarios intended to assess true generalization, and made available via user-friendly, maintained interfaces.

Table 2.2 - *Summary of Existing Computational Tools for Predicting lncRNA–Protein Interactions*

Category	Tool	Year	Features	Model Framework	Performance	Citation

Sequence-based (early phase)	lncPro	2013	Physicochemical properties + secondary structure	Discriminative scoring model	Acc 90.30%	[114]
	RPI-Pred	2015	Reduced alphabet sequence + structural blocks	Support Vector Machine	AUC 0.89–0.95	[105]
Network-based	LPIHN	2015	Co-expression + PPI network similarity	Random Walk with Restart	AUC 0.96	[107]
Ensemble learning	LPI-BLS	2019	k-mer + pseudo composition	Broad Learning stacked ensemble	Acc 0.902–0.927	[108]
	RPI-SE	2020	PWM + k-mer features	Stacking ensemble (XGB/SVM/ExtraTree)	AUC 0.904–0.994	[109]
	LPI-HyADBS	2021	Multi-source sequence descriptors	Hybrid DNN/XGB/SVM ensemble	AUC 0.559–0.854	[110]
	SAWRPI	2022	3-mer + GloVe embeddings	Stacked ensemble	AUC 0.746–0.992	[111]
Deep learning	LGFC-CNN	2021	Multi-modal (raw sequence + structure features)	Multi-module CNN	AUC 0.978–0.998	[115]
Graph-based (GAE/GNN)	FMSRT-LPI	2024	Multi-source similarity + topology fusion	Graph Autoencoder	AUC 0.990–0.992	[116]
Graph transformer	GHCT	2025	Topology-driven identity features	Attention-based Graph Transformer	AUC 0.975–0.981	[112]

2.3 Role of lncRNAs in Cancers

The involvement of lncRNAs in cancer biology has proven to be neither peripheral nor incidental. These molecules function as oncogenes or tumor suppressors whose activity can influence all

hallmarks of cancer, and their regulatory reach extends across nearly every level at which tumor biology operates, from chromatin remodeling and transcriptional control to post-transcriptional regulation, splicing, and translational efficiency [117]. What makes this particularly significant is that lncRNAs display highly tissue-, cell type-, and developmental-specific expression, regulating biological processes that control differentiation and development, and are highly dysregulated in cancers where they play crucial roles in cancer development and progression through oncogenic and tumor suppressor pathways, cellular metabolism, and the immune response and tumor microenvironment [118]. The relationship between a given lncRNA and cancer outcome is rarely straightforward. Analysis of 50 cancer-related lncRNAs across 24 cancer types in The Cancer Genome Atlas revealed that certain lncRNAs are consistently associated with worse or better survival across multiple cancers, while others adopt opposing prognostic roles depending on the disease context [119]. A cell-line model of PDAC has demonstrated that *HOTAIR* promotes tumorigenesis, demonstrating that a lncRNA's functional identity is inseparable from its cellular context [120]. *HOTAIR*, for example, is overexpressed across a variety of malignancies and linked to apoptosis inhibition, cellular proliferation, and genomic instability. Another avenue of lncRNA involvement in cancer that has only lately drawn significant interest is the immunological component. It has been shown that immune evasion in triple-negative breast cancer is facilitated by lncRNA-dependent downregulation of antigenicity and intrinsic tumor suppression, which is mediated by the degradation of the antigen peptide-loading complex and tumor suppressors Rb and p53 [121]. Resistance to PD-1 blockade is correlated with increased expression of the responsible lncRNA.

Table 2.3 *- Summary of lncRNAs acting as biomarker in cancers*

lncRNA Name	Cancer Type	Biomarker Potential: (Role)	Citation
MACC1-AS1	Pancreatic Cancer	Prognostic: High expression correlates with worse survival outcomes.	[122]
AFAP-AS1	Breast Cancer	Predictive: Upregulation is linked to radioresistance and tumor growth.	[123]
CBR3-AS1	Breast Cancer	Diagnostic: High expression distinguishes malignant tissues from normal controls (Sensitivity: 0.9).	[124]

MALAT1	Lung Cancer (NSCLC)	Diagnostic: Elevated levels in plasma/serum; acts as a minimally invasive liquid biopsy marker.	[125]
ANRIL	Colorectal Cancer	Predictive: Binding to miR-181a-5p suppresses radiosensitivity.	[123]
HOTAIR	Breast Cancer	Prognostic: High expression is a strong predictor of metastasis and poor clinical outcome.	[126]
PCA3	Prostate Cancer	Diagnostic: Detected in urine; currently one of the most clinically validated lncRNA biomarkers.	[127]
LINC00976	Pancreatic Cancer	Oncogenic: High expression drives cancer progression; potential therapeutic target.	[122]
BLACAT1	Head & Neck SCC	Predictive: Knockdown of this lncRNA enhances sensitivity to radiation therapy.	[123]
H19	Colorectal Cancer	Prognostic: Aberrant expression profiles serve as indicators for tumor progression.	

Pancreatic ductal adenocarcinoma (PDAC) sits at the intersection of many of these challenges and represents one of the most compelling contexts in which to study the diagnostic and prognostic potential of lncRNAs [128]. With a median survival of less than six months from diagnosis and a five-year survival rate of roughly 7.7%, PDAC continues to rank among the top causes of cancer-related mortality globally [129,130]. The disease's aggressive biology, the dense surrounding stromal milieu that hinders drug delivery, and a nearly universal resistance to immunotherapy and chemotherapy are all contributing factors to its lethality [131,132]. Importantly, these biological challenges are made worse by the absence of trustworthy early detection methods. By the time of diagnosis, around 80% of patients with pancreatic cancer have lost the chance for surgical resection; by then, the illness has usually progressed to a stage where there is little chance of recovery [133]. The most often used blood biomarker for PDAC, CA19-9, is not sufficiently specific for early-stage identification and is not exclusively elevated in malignancy. In light of this, lncRNAs have become very interesting contenders [134]. Transcripts such as *HOTAIR, H19, PVT1, LINC01133,* and *MALAT1* consistently exhibit differential expression in PDAC tissue compared to non-malignant pancreatic tissue, and their abundance has been linked in several independent cohorts to tumor grade, nodal involvement, and overall survival [135]. Multi-transcript signatures independently linked to PDAC patient survival have been found through integrative analysis of lncRNA and mRNA expression data from large-scale cancer genome

atlases. This suggests that lncRNA expression patterns carry prognostic information not captured by protein-coding gene profiles alone. Because of their active mechanistic involvement in tumor biology and statistical correlation with disease outcomes, lncRNAs are elevated above conventional expression biomarkers. This means that their dysregulation represents actual changes in regulatory circuitry rather than passive correlates of disease state.

2.3.1 Existing methods on blood-based biomarkers for PDAC

The biological basis for using blood-based RNA biomarkers to detect PDAC is well established. Extracellular vesicles (EVs) carry a variety of molecular cargo, such as proteins, lipids, DNA, and RNA, so analyzing EV-associated transcripts in peripheral blood can reveal information about tumor biology without requiring tissue biopsy [136]. Tumor-derived EVs display the signaling states and gene expression patterns of the parent cells. This is relevant to the diagnosis of PDAC, which is hindered by the risk of invasive sampling and the pancreas' physical inaccessibility. EV-based RNA biomarkers demonstrated a pooled sensitivity of 84% and specificity of 89%, according to a meta-analysis that examined 39 studies from all stages of the malignancy (gastric cancer) [137].

This area of research has advanced thanks to Yu et al.'s groundbreaking work on PDAC. They discovered an eight-transcript panel comprising *FGA, KRT19, HIST1H2BK, ITIH2, MARCH2, CLDN1, MAL2,* and *TIMP1* genes [138]. Using plasma-derived extracellular vesicles from the GSE133684 cohort, this biomarker panel obtained an AUC of 0.936. This finding showed that PDAC could be successfully distinguished from both healthy individuals and patients with chronic pancreatitis using a compact, EV-derived RNA signature. In the past, it has been challenging to distinguish this difference using even well-known protein markers. In order to address the problem from a different angle, Reese et al. combined serum exosomal miR-200b and miR-200c with CA19-9 [139]. They reported an AUC of 0.97, supporting the claim that combining RNA markers with already used clinical tests can significantly increase discriminative power. A four-lncRNA panel was recently constructed by He et al. from plasma exosomal transcripts, *LINC01268, LINC02802, AC124854.1,* and *AL132657.1* [140]. The panel achieved an AUC of 0.848 and directly demonstrated that circulating vesicles containing lncRNAs carry a detectable and diagnostically relevant disease signal. The lncRNA *HEVEPA* was found to be significantly

expressed in serum EVs from PDAC patients in a related line of work. When combined with HULC and CA19-9, it increased detection sensitivity beyond what was possible with either the lncRNA or the protein marker alone [141].

Variability in preliminary handling and analytical platforms has impeded reproducibility across research, despite the identification of several EV-derived proteins, microRNAs, long non-coding RNAs, and DNA changes associated with early carcinogenesis and treatment resistance. Protein or RNA biomarkers alone have repeatedly failed to meet the sensitivity and specificity standards needed for population-level screening. The capacity to accurately differentiate chronic pancreatitis from cancer is still the most clinical hurdle in the development of PDAC biomarkers, and it has not been sufficiently addressed by CA19-9 alone or the majority of single-transcript approaches. The trajectory of the field points towards multi-transcript panels, ideally integrating lncRNA and mRNA features, evaluated through robust computational classification frameworks, as the most promising route towards diagnostically actionable blood-based tests.

Table 2.4 *- RNAs as potential non-invasive pancreatic cancer biomarkers*

Study	Sample source	Biomarkers	Performance (AUC)	Citation
Yu et. al., 2020	EVs from plasma	RNA: *FGA, KRT19, HIST1H2BK, ITIH2, MARCH2, CLDN1, MAL2, TIMP1*	0.936	[138]
Resse et. al., 2020	Exosomes from blood serum	miRNA: *miR-200b, miR-200c + CA19-9*	0.970	[139]
He et. al., 2024	Exosomes from plasma	lncRNA: *LINC01268, LINC02802, AC124854.1, AL132657.1*	0.848	[140]

2.4 Conclusion

A growing divide between transcript identification and functional characterization defines lncRNA research. Only a small percentage of identified lncRNAs have been biologically investigated, despite the fact that high-throughput sequencing has significantly increased the catalogue. This difference reflects long-standing problems such as insufficient annotation, diversity in transcript

classifications between platforms, low and cell-type-specific expression, and context-dependent regulation activity. Many computational tools inherit dataset-specific biases instead of correcting them, existing resources are still fragmented, and annotation standards are inconsistent, all of which hinder the generalizability and robustness of downstream analysis.

Important functional dimensions include subcellular localization and lncRNA–protein interactions, however both fields are hindered by diverse and inconsistent datasets. Although most prediction models are trained on a limited number of cell lines without explicitly accounting for cell-type-specific patterns, localization evidence varies among cell types and experimental setups. Similarly, prediction methods for lncRNA–protein interactions lack consistent evaluation processes and rely on out-of-date datasets. Although computational techniques are essential, their ability to generalize outside the training dataset is limited by improper training algorithm and lack of variability in the training dataset.

The development of lncRNA-based blood biomarkers for PDAC has great translational potential but yet they have unresolved limitations. Due to delayed diagnosis and ineffective early detection techniques, PDAC mortality is still quite high. Extracellular vesicle-derived RNA biomarker panels have shown a lot of promise, but variations in sequencing platforms and sample handling can make them difficult to replicate. Additionally, it can be difficult to distinguish between PDAC and chronic pancreatitis due to their similar symptoms. All of these gaps highlight the need for computationally rigorous, biologically informed frameworks that are based on carefully selected data and tested in real-world scenarios.

3

Compilation of Resources on lncRNA Subcellular Localization

3.1 Introduction

The function of lncRNA relies heavily on its subcellular location. Unlike mRNAs, which require cytoplasmic translation, lncRNAs localize to specific cellular compartments according to their regulatory activities [142]. While cytoplasmic lncRNAs interact with translation complexes, signaling pathways, and stress granules, nuclear lncRNAs interact with chromatin and transcriptional machinery to regulate gene expression and genome structure [143]. Other subsets have extremely limited localization to compartments such as dendrites, axons, the nuclear periphery, and the nucleolus, indicating tight spatial regulation. As a result, describing subcellular localization provides a mechanical foundation for functional inference rather than being only descriptive.

Despite its importance, it is currently challenging to map LncRNA localization experimentally on a large scale, both logistically and technically. Nuclear-cytoplasmic fractionation paired with RNA sequencing offers scalable compartment-level resolution, although it is limited by low fractionation efficiency and spatial precision. While high-resolution methods, such as fluorescent in situ hybridization, are more precise, their throughput is limited and their technical complexity is higher. As a result, existing localization data is scarce, heterogeneous, and spread among multiple studies with different cell types and reporting standards. Uneven coverage and inaccurate compartment definitions are two instances of limitations seen in public sources such as RNALocate. These constraints have a direct impact on computational predictors, which frequently neglect multi-compartment transcripts, treat localization as a mutually exclusive single-label feature, and rely on out-of-date datasets. To address these difficulties, we need to use rigorous dataset curation, standardized annotation frameworks, and model development methodologies that account for biological complexity.

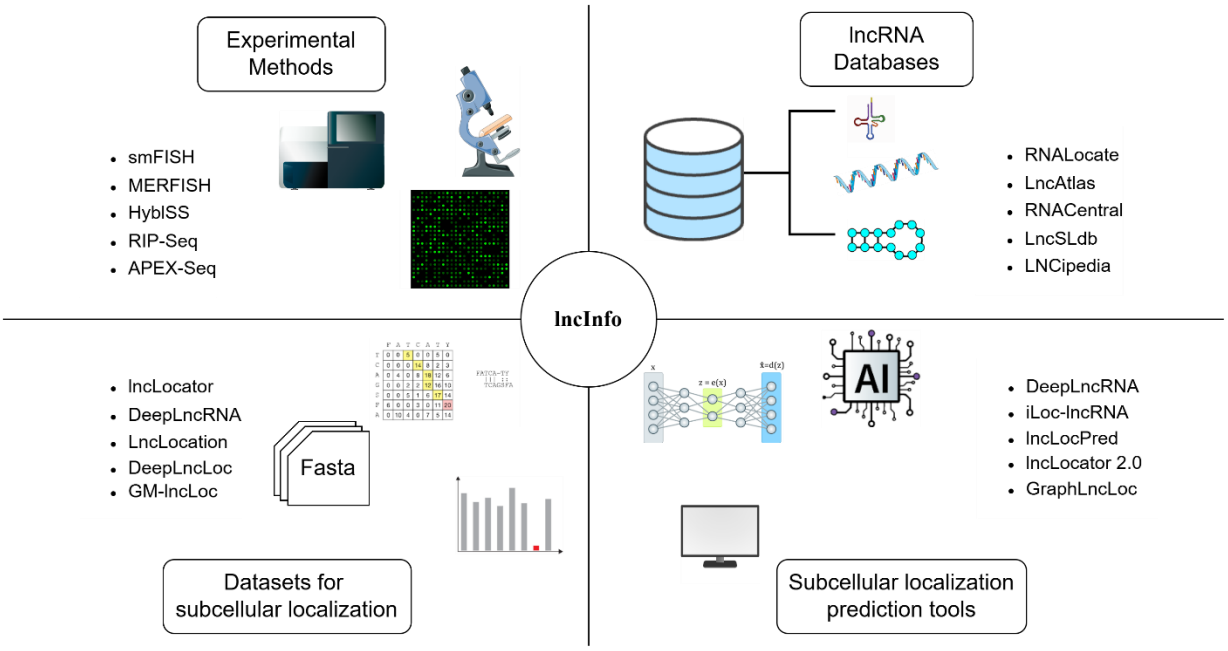


Figure 3.1 *– An overview of the modules covered in lncInfo*

3.2 Experimental Techniques to Investigate Subcellular Localization

A range of experimental methods have been employed by researchers to study lncRNA localization. This method includes fluorescence in-situ hybridization (FISH) and its variants, such as smFISH, smiFISH, SNV-FISH, inoFISH, MERFISH, SABER-FISH, and seqFISH [144–150]. These methods enable one to directly observe RNA molecules inside fixed cells, which has been essential in making significant discoveries. Despite their advantages, FISH-based methods need precise experimental setup, precise probe design, and sophisticated microscopy. Signal crowding, high operational costs, and reduced detection efficacy during multiplexing remain significant disadvantages.

Molecular labeling and high-resolution imaging are used in image-based biochemical techniques like STARmap, FISSEQ, HybISS, and CRISPR-based RNA imaging to investigate the distribution of RNA in intact cells and tissues [151–154]. These methods shed light on RNA quantity, spatial organization, and transcriptional activity in many cellular compartments. Nevertheless, these techniques frequently need complex experimental configurations and have trouble with low-abundance lncRNAs.

RNA-seq–based techniques also contribute valuable information. Fractionation-based RNA-seq separates cytoplasmic and nuclear compartments before sequencing, enabling transcriptome-wide assessment of localization. Other sequencing-based techniques such as GRO-seq, CAGE, PARE, TIF-seq, SHAPE, PARS, and RIP-seq investigate transcription dynamics, RNA structure, or RNA-protein interactions [155–161]. While these methods offer large-scale data, they depend on accurate fractionation and deep sequencing coverage, which can be limiting for lncRNAs expressed at low levels. A summary of the existing experimental methods is provided in Table 3.1.

Table 3.1 - *A summary of the existing experimental techniques to investigate lncRNA*

Methods	Brief description of the method	Citation
FISH based methods		
smFISH	Experimental technique smFISH (single-molecule FISH) is used for visualization and quantification of RNA molecules in cells.	[144,162,163]
smiFISH	Single molecule inexpensive FISH is easy to uses and flexible techniques used for quantification of RNA molecules in cells.	[164]
SNV FISH	Single nucleotide variant FISH can detect single-base changes in DNA sequences within cells or tissues.	[146,165]
inoFISH	inosine FISH allow one to visualize and quantify adenosine-to-inosine edited transcripts in situ.	[147]
SABER-FISH	Signal amplification by exchange reaction FISH uses a hybridization-based signal amplification to detect low-abundance RNA targets in cells and tissues.	[149]
ClampFISH	Click-amplifying FISH is an improved version of FISH, which have high specificity and signal amplification.	[166]
seqFISH+	Spatially resolved transcriptomics by FISH enables simultaneous detection of thousands of RNA molecules in situ.	[150]
MERFISH	Multiplexed Error-Robust FISH can detect and localize thousands of RNA molecules simultaneously within cells.	[148,167]
General imaging methods		
STARmap	STARmap is hybrid technique that combines hydrogel-tissue chemistry and DNA sequencing to detect expression of genes	[151]

HyblSS	Hybridization-based in situ sequencing combines in situ hybridization with DNA sequencing to identify RNA molecules at the single-cell level.	[153]
FISSEQ	Fluorescent In Situ Sequencing combines FISH with next generation sequencing for detecting RNA molecules.	[152]
RNA stem-loop system	The RNA stem-loop system developed for controlled expression of RNA molecules in cells by using a stem-loop structure.	[168]
Cas system	CRISPR-mediated RNA imaging technique involves the use of a modified CRISPR-Cas system to target RNA molecules.	[154]
lncRNA profiling		
RNA-seq	RNA sequencing is a next generation sequencing technology that can be used to sequence lncRNA in a given cell location.	[169]
Microarray	Microarray RNA profiling is a commonly used technique for measuring expression RNA transcripts	[170]
Tiling arrays	Tiling arrays are used to determine genome binding in ChIP assays or to identify transcribed regions	[171]
SAGE CAGE	Cap analysis gene expression, is an extension of SAGE, used for quantifying transcripts including lncRNA.	[156,172]
PARE	Parallel Analysis of RNA Ends allows one to quantify RNA molecules in a sample	[157]
GRO-seq	GRO-seq (Global Run-On sequencing) is a method used to study transcriptional activity of the genome at a global scale.	[155]
RIP-Chip	RIP-Chip developed for measuring interaction between RNA and proteins.	[158]
TIF-seq	Transcription initiation footprinting and sequencing is used to study the initiation of transcription in a genome.	[173]
SHAPE	Selective 2'-hydroxyl acylation by primer extension is a high-resolution technique to measure RNA structure.	[159]
PARS	Parallel Analysis of RNA Structure allow to study structure and function of wide range of RNAs at genome scale.	[160]
APEX-RIP	Ascorbate Peroxidase Proximity-Dependent Biotinylation RIP allow to identify RNA-protein interactions at genome scale.	[174]
Biochemical Fractionation + RNA-Seq	It is a powerful experimental method for subcellular localization of RNA molecules.	[175]

Note: Reprinted from [175] *under the Creative Commons Attribution License (CC BY).*

3.3 Existing Databases on lncRNA

A wide range of databases has been established to support lncRNA research, spanning general annotation, expression, sequence, subcellular localization, and disease association. Table 3.2 provides a brief summary of these resources.

General databases such as FANTOM, GENCODE, ENCODE, GTEx, LncExpDB, and LncBook provide foundational information including transcript structure, expression patterns, regulatory annotations, and curated gene models [18,50,176–179]. As of right now, GENCODE lists 62,703 genes, 19,393 of which code for proteins and 19,928 of which are long non-coding RNAs. Numerous collections of non-coding RNA sequences from various species are available in sequence annotation databases such as LNCipedia, CANTATAdb, NONCODE, RNAcentral, and CHESS [51,180–183]. These databases minimize redundancy, combine sequences from several sources, and often offer secondary structural information. In contrast to RNAcentral, which compiles more than 18 million ncRNA sequences from 44 sources, NONCODEV6 alone has 644,510 lncRNAs, including 173,112 human lncRNA transcripts.

Subcellular localization is the focus of several databases. Using a Relative Concentration Index (RCI), LncATLAS reports localization information compiled from RNA-seq data based on ENCODE fractionation [77]. LncSLdb manually selects localization evidence from high-throughput experiments and literature, encompassing over 11,000 lncRNA transcripts from fruit flies, mice, and humans [184]. With around 200,000 entries covering numerous RNA types and dozens of subcellular compartments, RNALocate is the most extensive database of RNA localization. Notwithstanding its size, RNALocate has changed over time, occasionally lowering the quantity of available sequences and causing discrepancies across tools that depend on various releases [88].

Databases linking lncRNAs to diseases, such as LncRNADisease2 and Lnc2Cancer, catalog experimentally validated associations with cancer, cardiovascular disease, and neurological disorders [185,186]. LncRNADisease2 documents 205,959 lncRNA-disease associations and 2,297 lncRNA causative associations. These resources provide insights into how mislocalization might contribute to pathology and are useful for identifying disease-specific regulatory mechanisms.

Table 3.2 - *Brief summary of existing databases on lncRNA*

Database	Brief description	Website link	Citation	Year
General Databases				
FANTOM Cat	A database for functional annotation of the mammalian genome	https://fantom.gsc.riken.jp/cat/	[176]	2017
GENCODE	A database that provides comprehensive gene annotation for the human and mouse genome	https://www.gencodegenes.org/	[187]	2022
Expression databases				
ENCODE	A database of functional annotation based on sequencing data	https://www.encodeproject.org/	[188]	2018
lncExpDB	A comprehensive database of experimentally validated lncRNA and circRNA.	https://ngdc.cncb.ac.cn/lncexpdb/	[178]	2021
LNCbook 2.0	A database long non-coding RNAs and their functions.	https://ngdc.cncb.ac.cn/lncbook/	[179]	2023
GTEx	Genotype-Tissue Expression database	https://gtexportal.org/home/	[177]	2020
TANRIC	A database of non-coding RNAs in cancer.	https://www.tanric.org/	[189]	2022
GEO	A public repository for gene expression.	https://www.ncbi.nlm.nih.gov/geo/	[190]	2013
Sequence Annotation				
LNCipedia	A database for lncRNA sequences and their annotation.	https://lncipedia.org/	[180]	2018
CANTATAdb	A database of long non-coding RNAs in plants.	http://cantata.amu.edu.pl/	[181]	2024
NONCODEV6	A database developed to maintain non-coding RNAs and their annotation.	http://www.noncode.org/	[52]	2021
RNACentral	A comprehensive ncRNA sequence collection representing all ncRNA types from a broad range of organisms	https://rnacentral.org/	[182]	2021
CHESS	A repository that contains human genes and transcripts.	http://ccb.jhu.edu/chess	[183]	2023
Subcellular Localization				
LncAtlas	A quantitative resource for lncRNA subcellular localization	https://lncatlas.crg.eu/	[77]	2017

LncSLdb	A manually curated database of RNA subcellular localization.	http://bioinformatics.xidian.edu.cn/lncSLdb	[184]	2018
RNALocate v3	Curation of RNA subcellular localization from public resources like literature, RNA-seq datasets	http://www.rna-society.org/rnalocate/	[191]	2025
Disease Association				
LncRNADisease v3.0	A repository that contains lncRNA and associated diseases	http://www.rnanut.net/lncrnadisease/	[185]	2024
Lnc2Cancer v3.0	A resource for experimentally supported lncRNA/circRNA cancer associations	http://bio-bigdata.hrbmu.edu.cn/lnc2cancer/	[186]	2021

Note: Reprinted from [175] *under the Creative Commons Attribution License (CC BY).*

3.4 Creation of Datasets by Different Tools

The training datasets for the majority of computational predictors of lncRNA subcellular localization are created using RNALocate or fractionation-based RNA-seq studies [88]. lncLocator, iLoc-lncRNA, LncLocation, Locate-R, and lncLocPred generally acquire sequences from RNALocate, use CD-HIT to reduce redundancy, and reject transcripts associated with multiple compartments [68,79,81,82,89]. Although these preprocessing steps increase dataset consistency, they also significantly lower sample size and remove physiologically relevant multi-localized lncRNAs. Furthermore, early versions of RNALocate contained mixed RNA biotypes and ambiguously annotated entries, which could introduce noise into downstream models.

The original lncLocator dataset, taken from RNALocate v1, contains 612 lncRNA sequences spread across five cellular compartments after filtering multi-localized entries and applying an 80% similarity criterion to CD-HIT. For iLoc-lncRNA, a similar pipeline yielded 655 non-redundant sequences; this dataset was then utilized in LncLocation, Locate-R, and lncLocPred. KD-KLNMF used a similar dataset of 644 sequences and employed SMOTE oversampling after feature extraction to correct class imbalance. While oversampling improves numerical balance, it alters the natural distribution of localization classes and may bias model learning.

RNA-seq–derived datasets, used in tools such as DeepLncRNA and lncLocator 2.0, rely on quantitative cytoplasmic-to-nuclear expression ratios. DeepLncRNA analyzed 93 ENCODE RNA-seq samples across 14 cell lines and assigned nuclear or cytosolic labels using log2 fold-change

thresholds, resulting in 4,380 cytosolic and 4,298 nuclear lncRNAs [80,90]. LncLocator 2.0 obtained localization annotations from lncATLAS and excluded transcripts with Cytoplasm/Nucleus RCI values between −1 and 1, retaining only strongly compartment-biased lncRNAs; this dataset was later adopted by TACOS with restriction to ten cell lines [77,91]. Although RNA-seq–based datasets capture cell-line–specific localization patterns, they are typically limited to binary classification, constraining their utility for multi-compartment prediction. Table 3.3 summarizes the datasets underlying existing localization prediction tools.

Table 3.3 - *Summary of datasets used in existing subcellular localization tools*

Tools		Cytoplasm	Cytosol	Exosome	Nucleus	Ribosome	Total
lncLocator	Benchmark	301	91	25	152	43	612
iLoc-lncRNA	Benchmark	426		30	156	43	655
DeepLncRNA	Benchmark		4380		4298		8678
LncLocation	Benchmark	426		240	344	314	1324
Locate-R	Benchmark	426		240	314	344	1324
lncLocPred	Benchmark	426		30	156	43	655
	Independent	199		16	82	99	396
KD-KLNMF	Benchmark	417		417	417	417	1668
	Independent	14		35	45	84	178
DeepLncLoc	Benchmark	328	88	28	325	88	857
	Test	20	10	7	20	10	67
GM-lncLoc	Dataset1	292	292	292	292	292	1460
	Dataset2	417		417	417	417	1668
	Independent set	198		16	82	99	395
GraphLncLoc	Benchmark	328		28	325	88	769
	Independent	20		7	20	10	57

Note: Reprinted from [175] *under the Creative Commons Attribution License (CC BY).*

3.5 Existing Prediction Methods and Their Performance

Over time, computational techniques for lncRNA localization prediction have undergone substantial development. Nucleotide composition, k-mer frequencies, pseudo-nucleotide

compositions, physicochemical descriptors, and secondary structural features were among the handcrafted features used by earlier tools, which mainly relied on traditional machine learning models—support vector machines, logistic regression, and random forests. While these approaches produced interpretable models, their performance was limited by the quality and variability of training datasets.

In addition to conventional classifiers, lncLocator employed a stacked autoencoder on k-mer features to achieve an overall accuracy of 59.1% [68]. iLoc-lncRNA achieved an overall accuracy of 86.72% by combining Pseudok-tuple Nucleotide Composition with an SVM ensemble and iterative feature selection [79]. lncLocation, Locate-R, and lncLocPred used comparable methodologies and achieved accuracies of 87.78%, 90.69%, and 92.37% on their training datasets [81,82,89]. KD-KLNMF used Kullback-Leibler divergence-based non-negative matrix factorization (KLNMF) to choose features with an RBF-SVM model and achieved an outstanding 97.24% accuracy on a jackknife test [192].

Deep learning techniques like autoencoders, convolutional neural networks, and sequence embedding models have increased prediction accuracy by learning high-level features from sequences. DeepLncLoc used subsequence embedding and a textCNN architecture, whereas lncLocator 2.0 predicted continuous CNRCI values using GloVe embeddings and a CNN-BiLSTM-MLP pipeline [85]. Recently, graph-based models such as GM-lncLoc and GraphLncLoc depict RNA sequences as graphs in order to capture structural dependencies. GM-lncLoc achieved 94.2% accuracy on its benchmark dataset by combining a graph convolution network with model-agnostic meta-learning [86,87].

Despite great performance on training data, most available techniques perform poorly on independent datasets. On their own training dataset, lncLocPred, Locate-R, and iLoc-lncRNA achieved accuracies of 92.37%, 90.69%, and 86.72%, but declined to 44.44%, 38.64%, and 35.86% on their independent dataset, respectively. All of the tools assessed by GM-lncLoc showed a similar pattern. This persistent performance disparity points to dataset bias, model overfitting, and inadequate training data diversity. A thorough overview of current tools and their stated performance indicators can be found in Tables 3.4 and 3.5.

Table 3.4 - *Summary of the existing subcellular localization prediction tools*

Method	Year	Feature	Algorithm	Number of subcellular compartments	Citation
lncLocator[#1]	2018	k-mer	Random forest, SVM, Autoencoder	5 (Nucleus, Ribosome, Cytoplasm, Exosome, Cytosol)	[68]
iLoc-lncRNA	2018	PseKNC	SVM	4 (Nucleus, Ribosome, Cytoplasm, Exosome)	[79]
DeepLncRNA	2018	k-mer, Genome loci, RNA binding motifs	Neural Networks	2 (Nucleus/Cytosol)	[80]
LncLocation	2020	k-mer, physico-chemical properties	SVM	4 (Nucleus, Ribosome, Cytoplasm, Exosome)	[82]
Locate-R[#2]	2020	k-mer, n-gapped k-mer	Locally deep SVM	4 (Nucleus, Ribosome, Cytoplasm, Exosome)	[81]
lncLocPred	2020	k-mer, PseDNC	Logistic regression	4 (Nucleus, Ribosome, Cytoplasm, Exosome)	[89]
KD-KLNMF	2020	k-mer, Di-nucleotide-based spatial autocorrelation	SVM	4 (Nucleus, Ribosome, Cytoplasm, Exosome)	[192]
lncLocator 2.0	2021	Word embedded sequences	GloVe + CNN + BiLSTM + MLP	2 (Nucleus/Cytoplasm)	[90]
DeepLncLoc	2022	k-mer, Subsequence embedding	TextCNN	5 (Nucleus, Ribosome, Cytoplasm, Exosome, Cytosol)	[85]
TACOS	2022	Composition-based, Dinucleotide	AdaBoost	2 (Nucleus/Cytoplasm)	[91]

		physicochemical properties			
GM-lncLoc[#2]	2023	k-mer	GCN based on MAML	4/5 (Nucleus, Ribosome, Cytoplasm, Exosome)/Cytosol	[86]
GraphLncLoc	2023	de Brujin Graphs	Graph Convolution Networks	4 (Nucleus, Ribosome, Cytoplasm, Exosome)	[87]

#1: SOS oversampling; #2: SMOTE oversampling

Note: Reprinted from [175] *under the Creative Commons Attribution License (CC BY).*

Table 3.5 - *Performance of existing subcellular localization prediction tools*

Metrics		Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)	MCC	F1-score	AUC
lncLocator	Overall	36.3			59.1		0.367	
iLoc-lncRNA	Nucleus	77.56	97.59			0.796		
	Cytoplasm	99.06	67.68			0.742		
	Ribosome	46.51	99.83			0.652		
	Exosome	16.67	1			0.4		
	Overall				86.72			
DeepLncRNA		83	62.4		72.4	0.451	0.744	0.787
LncLocation	Nucleus	74.19		95.83				
	Cytoplasm	100		85				
	Ribosome	55.56		100				
	Exosome	33.33		100				
	Overall	87.78						
Locate-R	Nucleus	65.92	95.15			0.658		
	Cytoplasm	84.74	89.1			0.725		
	Ribosome	100	98.37			0.97		
	Exosome	100	99.17			0.978		
	Overall				90.69			
lncLocPred	Nucleus	96.79	96.79			0.915		
	Cytoplasm	99.06	85.59			0.876		
	Ribosome	60.47	99.84			0.751		
	Exosome	20	100			0.439		

	Overall				92.37			
KD-KLNMF	Nucleus	90.65	99.52			0.928		
	Cytoplasm	98.56	96.8			0.93		
	Ribosome	99.76	100			0.998		
	Exosome	100	100			1		
	Overall				97.24			0.9981
lncLocator 2.0	15 cell lines (min-max)							0.6088 - 0.8499
DeepLncLoc	Nucleus							0.67
	Cytoplasm							0.76
	Ribosome							0.657
	Exosome							0.804
	Cytosol							0.806
	Overall				54.8		0.421	0.82
TACOS	Overall (10 cell-lines)	77.72	72.81		75.26	0.5064		0.8339
GM-lncLoc	Overall (Dataset1)	93.3			93.4		0.933	
	Nucleus (Dataset2)	88.85	98.21			0.889		
	Cytoplasm (Dataset2)	93.21	96.06			0.879		
	Ribosome (Dataset2)	96.8	98.99			0.959		
	Exosome (Dataset2)	99.07	99.38			0.982		
	Overall (Dataset2)				94.2			
GraphLncLoc	Overall	47.5		69.1	61.2		0.506	

Note: Reprinted from [175] *under the Creative Commons Attribution License (CC BY).*

3.6 Challenges and Future Perspectives

Major challenges continue to exist on every stage despite improvements in databases, computational models, and experimental methods. Because of their high costs and decreasing detection performance in multiplexed environments, FISH-based technologies are difficult to scale. Signal enhancement based on amplification has the potential to increase false positives and introduce artifacts. In multiplexed RNA-FISH techniques, decoding barcodes necessitates sophisticated software and significant processing power. Despite their strength, sequencing-based techniques are constrained by the low transcript abundance of many lncRNAs and the challenge of achieving clean, repeatable fractionation.

Numerous databases that are currently in operation need to be updated frequently but have not been kept up to date. For example, LNCipedia and LncRNADisease2 have not been updated since 2018. Inconsistencies in naming nomenclature among annotation databases make data integration even more difficult. The usefulness of lncATLAS is limited because it only covers the cytoplasm and nucleus on the localization side (it was last updated in 2017). In order to produce models with limited generalizability, computational techniques frequently rely on these noisy or out-of-date datasets, lack cell-line specificity, or eliminate multi-localized transcripts. Further complexity is introduced by the fact that localization patterns seen in fixed cells may not always correspond to those in live cells, and the cell-line-specific localization information in ENCODE is still underutilized.

It will take coordinated efforts to increase and diversify training datasets in order to overcome these constraints. A particularly interesting route forward is provided by recent developments in spatial transcriptomics, which allow for the subcellular characterization of transcripts. More precise quantification of subcellular localization would be made possible by the availability of MERFISH datasets. When taken as a whole, these advancements suggest that better algorithms, cleaner training datasets, and thorough validation on separate data are required.

3.7 Conclusion

Understanding lncRNA function requires an understanding of subcellular localization, and a rich ecosystem of computational and experimental tools has been developed to assist in this effort. Significant obstacles still exist in dataset development, curation standards, and predictive modeling

techniques despite methodological advancements. Class imbalance handling, compartment definition, annotation quality, and evaluation protocol limitations all continue to impact the biological interpretability and dependability of localization models. Addressing these challenges demands a systematic approach rather than gradual model modification.

This chapter objectively integrates current computational tools for lncRNA subcellular localization, benchmark datasets, public repositories, and significant experimental techniques to create a systematic framework for future study. These elements are combined into a single analytical framework which serves as the foundation for increasing biological relevance and predicting accuracy. This overview immediately supports the thesis's subsequent objectives and contributes to a larger endeavor to understand the regulatory activities of lncRNAs.

4

Cell-line specific Subcellular Localization of lncRNA

4.1 Introduction

The subcellular localization of lncRNAs is closely linked to their function within the cell [193]. Nuclear lncRNAs are known to regulate gene expression and chromatin organization, while cytoplasmic lncRNAs take part in signal transduction, translational control, and post-transcriptional regulation [194–197]. Some lncRNAs show multiple localizations and functional diversification depending on the cell type, reflecting their ability to adapt to different cellular environments. Understanding where a lncRNA is located therefore provides direct information about its function.

Experimental studies suggest that subcellular localization depends on a variety of biological factors. One key cis-element that drives subcellular localization of lncRNA is the presence of zipcode elements [198]. Zipcode elements are cis-acting sequence signals that range from a few bases to a kilobase in their length and are mostly found in the 3'-UTR regions. Motifs like the Alu derived C-rich motifs present in several lncRNAs are also known to favour nuclear localization [199]. It should be noted here that a majority of the experimental studies are investigating what drives nuclear retention of lncRNAs but other localization patterns are severely neglected.

Although the primary nucleotide sequence of a given lncRNA remains invariant across different cell-lines, yet there are numerous other factors that regulate their localization. The spatial fate of a lncRNA relies on the availability of RNA-binding proteins (RBPs) and variations in the concentration of nuclear retention factors, such as hnRNP K, U1 snRNP, etc. [199,200]. Secondly, cell-line-specific alternative splicing generates distinct transcript isoforms that may lack secondary structures required for anchoring in the nucleus [201]. Ineffective splicing of lncRNAs generally leads to a poor export efficiency and more nuclear retention [202,203]. Thirdly, the epitranscriptomic changes like the N6-methyladenosine (m6A) modifications can act as a spatial barcode that recruits transport proteins at varying efficiencies depending on the cellular state [204]. Finally, localization is also driven by target availability, meaning that if a lncRNA functions by anchoring to a specific chromatin locus, epigenetic silencing or structural inaccessibility of that target in a particular cell-line would prevent nuclear retention.

Previous methods developed for cell-line specific subcellular localization has multiple drawbacks. Some methods are trained on datasets that included mRNA sequences, which introduced noise into models built for lncRNAs. Others excluded lncRNAs which had equal chance of localizing to two

compartments, which removed biologically meaningful transcripts. Neither lncLocator 2.0 nor TACOS, adequately addressed these issues [90,91]. Also, it was observed that the same lncRNA can be nuclear in one cell line and cytoplasmic in another, as shown clearly in Figure 4.1. Cell-line specificity is therefore not optional but essential to understand subcellular localization.

To address these gaps, we developed CytoLNCpred, a method for predicting cytoplasm-associated lncRNAs across 15 human cell lines. The goal was to build cleaner datasets, test a broader range of modeling strategies, and produce a tool that meets industry standards for validation. We compared classical machine learning using composition and correlation features, embedding-based approaches using DNABERT-2, and a fine-tuned DNABERT-2 model. Among all strategies, correlation-based features combined with machine learning algorithms gave the best performance on held-out validation data. CytoLNCpred is publicly available at https://webs.iiitd.edu.in/raghava/cytolncpred/.

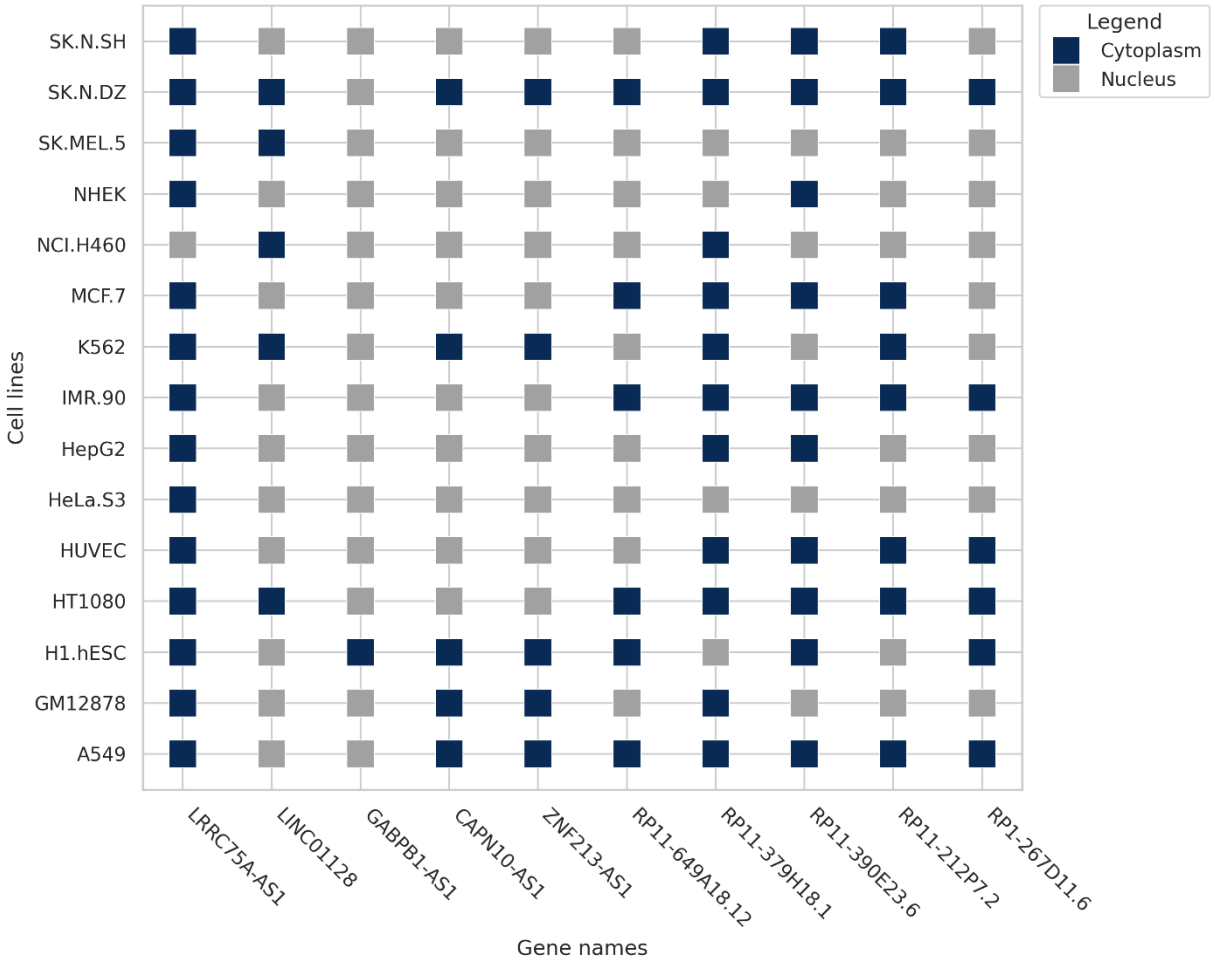


Figure 4.1 *– Bubble plot showing the variation in localization of a single lncRNA transcript across multiple cell-lines*

4.2 Materials and Methods

The overall architecture of the CytoLNCpred workflow is illustrated in Figure 4.2. The key stages are dataset preparation, feature generation, model training, and evaluation.

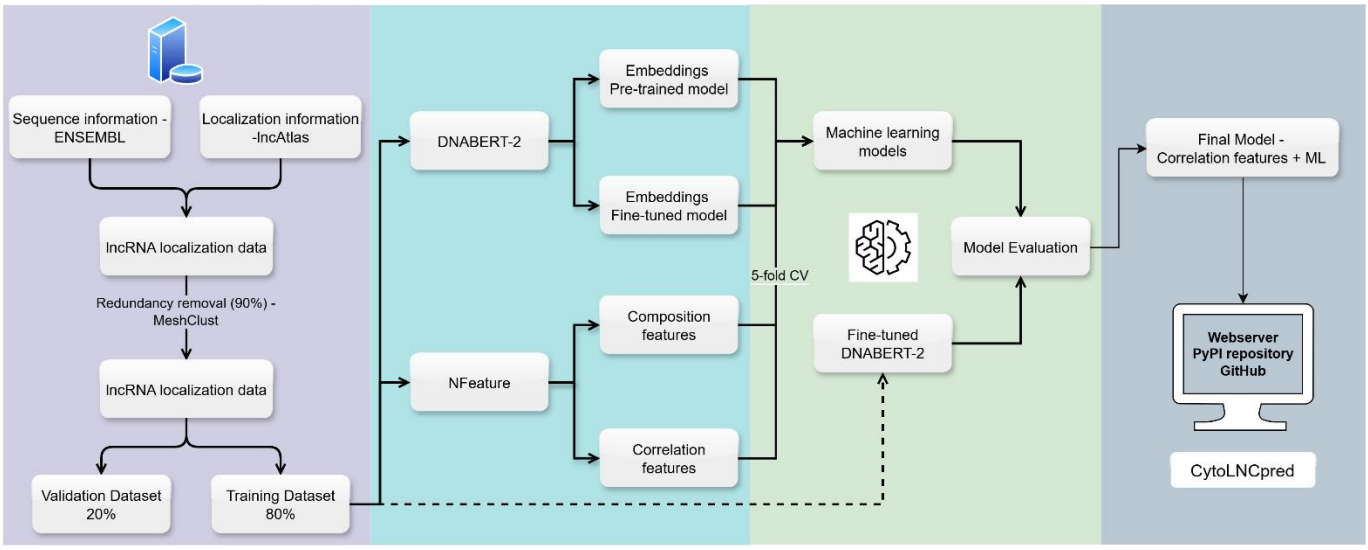


Figure 4.2 *– A graphical overview of the methodology adopted in CytoLNCpred*

4.2.1 Dataset creation

Localization data was sourced from lncAtlas, which provides the Cytoplasm to Nucleus Relative Concentration Index (CNRCI) derived from ENCODE RNA-seq experiments across human cell lines [205]. CNRCI is defined as:

$$CNRCI = log_2(Cytoplasmic\ expression\ (FPKM)\ /\ Nuclear\ expression\ (FPKM))$$

Sequences were assigned a cytoplasmic label if CNRCI > 0 and a nuclear label if CNRCI < 0. Sequence information was retrieved from ENSEMBL (version 112), and lncRNAs without available sequences were excluded [206]. Redundancy was removed using MeshClust at 90% sequence identity [207]. Sequences longer than 10,000 nucleotides were also removed, as very long sequences introduced noise in machine learning models and were computationally prohibitive for large language models. Each cell-line dataset was split 80:20 into training and validation sets. The full dataset summary across all 15 cell lines is provided in Table 4.1, and the dataset creation process is illustrated in Figure 4.3.

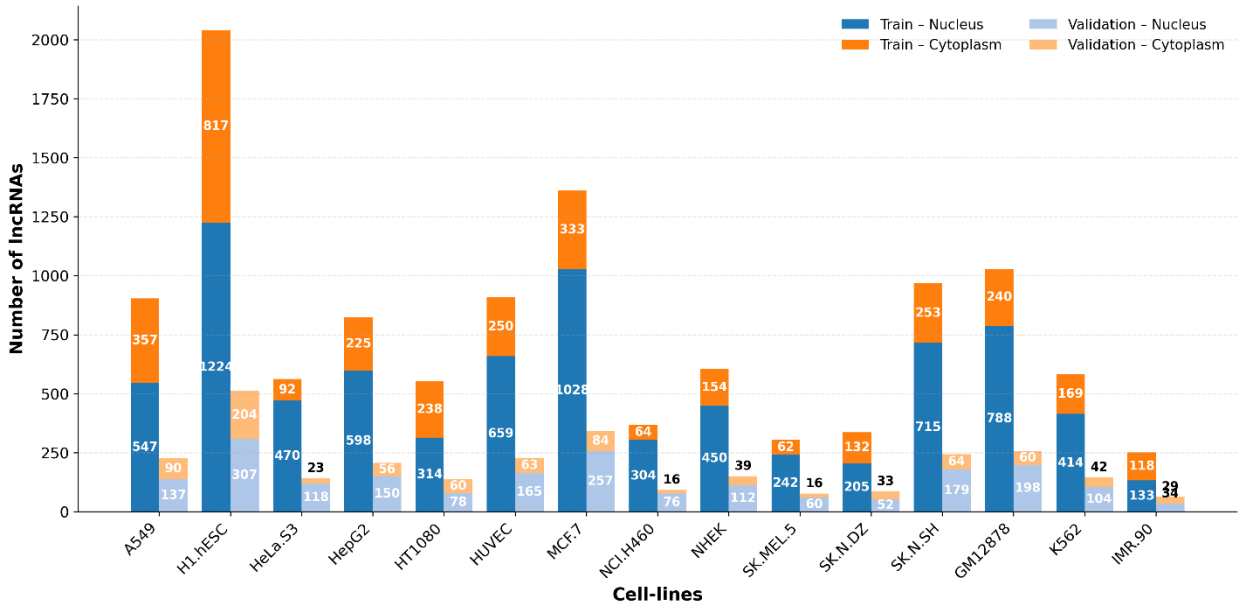


Figure 4.3 - *Distribution of nuclear and cytoplasmic lncRNAs across training and validation datasets for each cell line, shown as stacked bars with counts indicated for each segment.*

4.2.2 Feature generation – Composition and correlation-based

All composition and correlation features were generated using the in-house tool Nfeature [208]. Composition-based features represent quantitative properties of the nucleotide sequence like - nucleic acid composition, distance distribution of nucleotides (DDN), nucleotide repeat index (NRI), pseudo composition, and entropy metrics. Combined together, these composition-based yielded a total of 123 descriptors.

Correlation-based features capture the interdependencies between nucleotide properties along the sequence. Autocorrelation measures the relationship of a feature with itself at different lags, while cross-correlation captures relationships between two distinct properties. These descriptors normalize variable-length sequences into fixed-length vectors, making them compatible with standard machine learning algorithms. Thirteen types of correlation descriptors were computed, including trinucleotide cross-correlation, auto-dinucleotide cross-correlation, Moran and Geary autocorrelation, and various pseudo-correlation features, producing 1,100 descriptors in total. Combined, the two feature types produced 1,223 descriptors.

4.2.3 Embeddings extracted using DNABERT-2

DNABERT-2 is an adaptation of the BERT architecture designed specifically for DNA sequences [209]. It uses Byte Pair Encoding for tokenization, which performs better than k-mer tokenization for capturing sequence context. DNABERT-2 generates high-dimensional embeddings that encode not just individual bases but their biological context in terms of structure, function, and inter-nucleotide relationships.

In this study, embeddings were extracted from both the pre-trained DNABERT-2 model and a version fine-tuned on our training data. These embeddings were derived from the final hidden layer of the DNABERT-2 model using max pooling and a 768-dimensional vector per sequence was obtained. These embeddings were then used as features for downstream machine learning classifiers. We also fine-tuned the complete DNABERT-2 model using our training data and then evaluated it directly on the validation set.

4.2.4 Five-fold cross-validation

We applied five-fold cross-validation for evaluating model performance fairly and to avoid overfitting. The training dataset (80%) was further divided into five stratified subsets/folds, such that each fold maintained the same class balance. For each iteration, four folds were used for model training and the remaining fold for validation, and this was repeated until every fold had been used once as a validation set. This process generates a set of 5 performance metrics and their average value is reported as the training metric. Cross-validation ensures that our comparisons across different feature sets and models were consistent and reliable. Figure 4.4 explains the complete model evaluation process graphically.

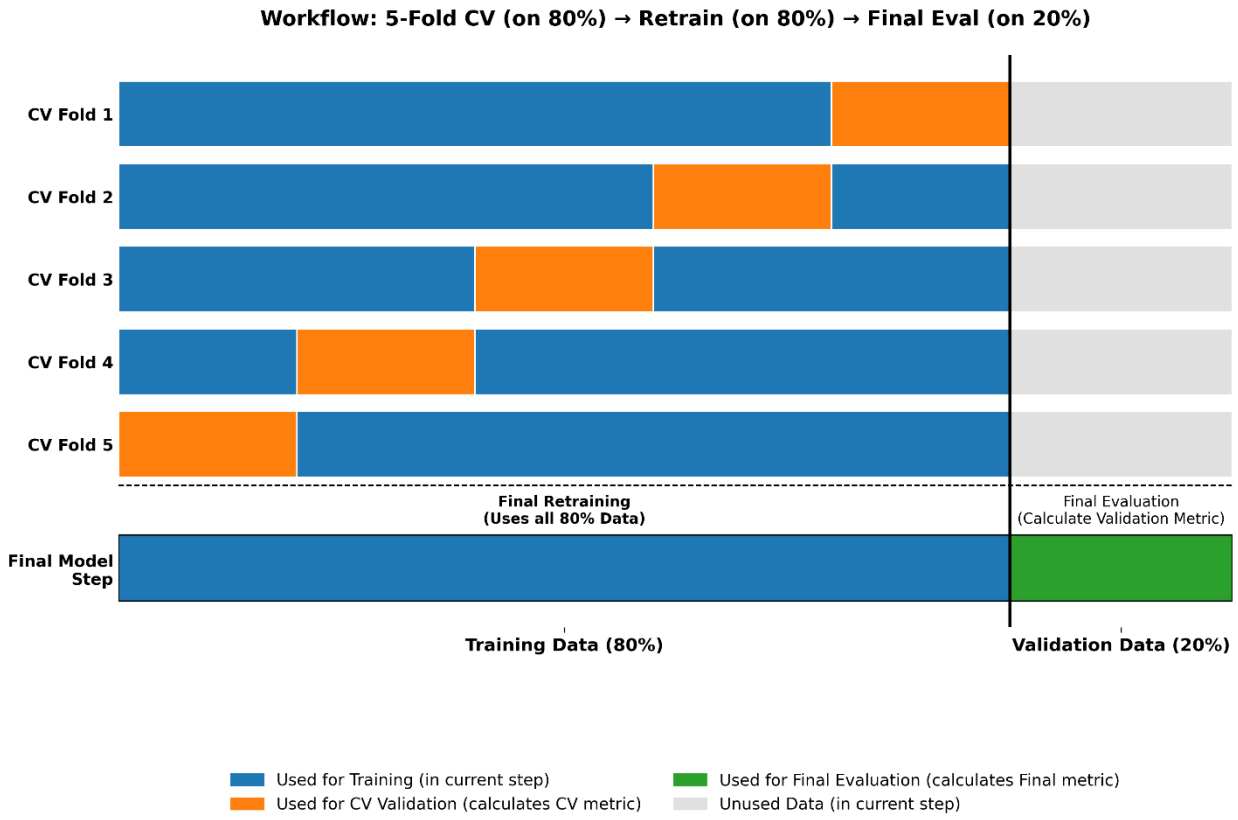


Figure 4.4 - *Schematic representation of the experimental workflow, illustrating 5-fold cross-validation performed on the 80% training set followed by final model evaluation on the 20% independent validation set.*

4.2.5 Feature selection

To reduce dimensionality and computational load, the Minimum Redundancy Maximum Relevance (mRMR) algorithm was applied to a combined feature set of 2,278 descriptors (composition + correlation + embeddings) [210]. mRMR selects features that are maximally relevant to the target variable while minimizing redundancy among selected features. Seven feature subsets were evaluated—containing 10, 50, 100, 500, 1,000, 1,500, and 2,000 features—and their performance was assessed across all cell lines. Feature importance was additionally evaluated using Pearson correlation with CNRCI values.

4.2.6 Model development

We developed models using three different strategies. First, the fine-tuned DNABERT-2 model was directly used as a standalone classifier to predict localization based on raw sequence input. Second, a hybrid approach combined DNABERT-2 embeddings (from both pre-trained and fine-tuned models) with classical machine-learning algorithms such as SVM, Random Forest, Gradient Boosting, Logistic Regression, and others. Third, we trained machine-learning models using only composition and correlation-based features. All the combinations were tested to evaluate the strengths of each feature type and determine which modeling strategy achieved the highest accuracy across all cell lines.

4.2.7 Model evaluation metrics

Model performance was assessed using several binary classification metrics to ensure a balanced evaluation. These included accuracy, precision, sensitivity, specificity, F1-score, Matthew's correlation coefficient (MCC), and area under the ROC curve (AUC). MCC was particularly important because it provides a balanced measure even when class sizes differ, making it a strong indicator of real-world usefulness. AUC measures the model's ability to distinguish cytoplasmic from nuclear lncRNAs across different thresholds. These metrics provided a quantitative measure of how well each model and feature set performed.

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP + TN}{TP + TN + FN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

$$F1 - score = \frac{TP}{TP + (0.5 \times (FN + FP))}$$

4.3 Results

This study attempts to design a localization prediction method that is cell-line specific using a diverse range of approaches. An overview of the average performance for all the approaches used in the study is depicted in Figure 4.5.

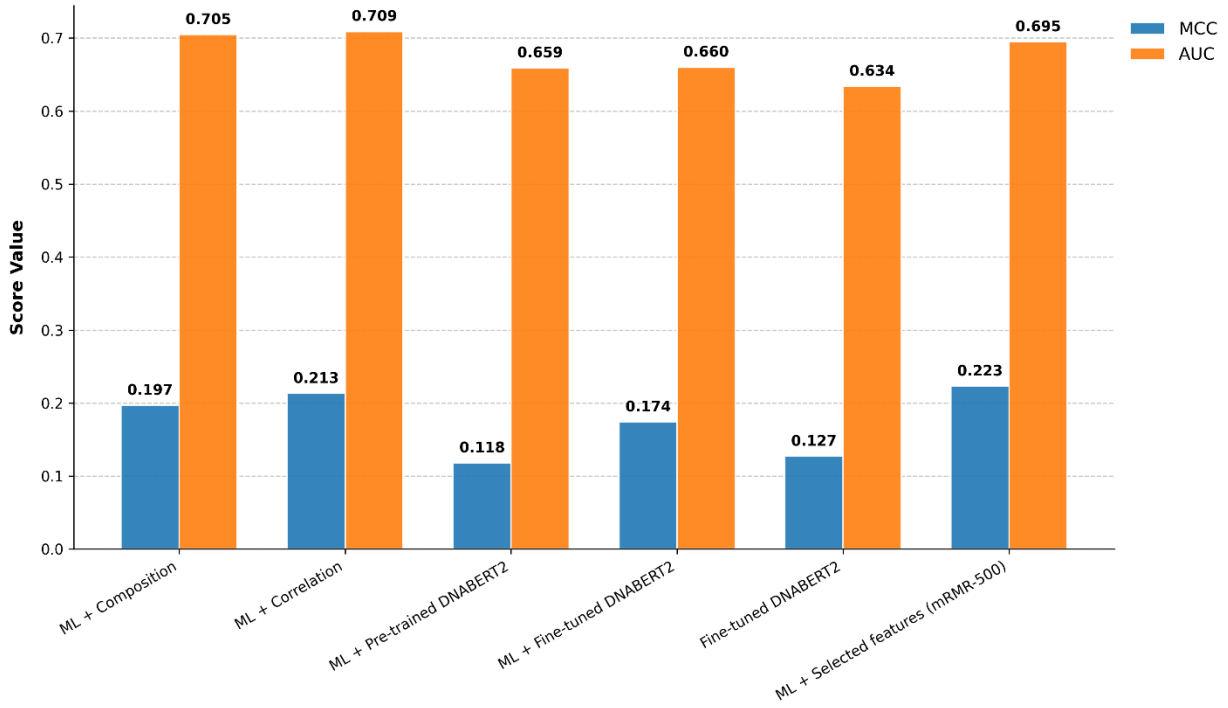


Figure 4.5 - *Comparison of performance for various methods used in this study*

4.3.1 Functional enrichment analysis

We examined sequence features associated with cytoplasmic versus nuclear labels across the 15 training cell lines. Short k-mer enrichment analyses showed that CG dinucleotides and CG-containing trinucleotides (like GCG, CGC, CGG, CCG and CGA) were consistently higher in

cytoplasm-labeled lncRNAs (FDR < 0.05 in up to 11 cell lines), whereas TG-, CT- and related T-rich k-mers were enriched among nuclear transcripts, reinforcing the cytosine- versus thymine-biased composition signals linked to CNRCI. Simultaneous comparisons of the dinucleotide physicochemical indices underlying the correlation descriptors (p1–p12) further indicated compartment-associated differences in physicochemical properties - most notably Roll, Twist and stacking energy versus enthalpy, tilt and entropy - whose magnitude and direction varied by cell line, supporting the need for cell-context-specific models. Literature-inspired motif proxies highlighted that high CG-density was linked with cytoplasmic labels, while classic AG-rich or long C-rich nuclear-retention signals were rather weak and inconsistent. This implied that bulk localization is dependent on multiple factors rather than being governed by a single motif.

GO and KEGG enrichment analysis was performed using RNAenrich to characterize the biological differences between cytoplasmic and nuclear lncRNA sets [211–213]. Among significantly enriched GO terms (adjusted p-value < 0.05), 2,511 were shared between compartments. Cytoplasmic lncRNAs uniquely enriched 397 terms, including "response to interferon-beta" (GO:0035456), "positive regulation of apoptotic process" (GO:0043065), and "RNA splicing" (GO:0008380). KEGG analysis associated cytoplasmic lncRNAs with Ferroptosis (hsa04216) and Autophagy (hsa04140), consistent with cytoplasmic stress and degradation functions. Nuclear lncRNAs uniquely enriched 254 terms, including "eukaryotic 48S preinitiation complex" (GO:0033290) and "MLL1/2 complex" (GO:0044665), pointing to roles in transcriptional regulation and chromatin remodeling. These compartment-specific enrichments confirm that the biological distinctions between cytoplasmic and nuclear lncRNAs are real and substantial, validating the classification objective.

4.3.2 Feature importance

Correlation of each feature with CNRCI values was used to assess which descriptors best distinguish cytoplasmic from nuclear lncRNAs. A positive correlation indicates the feature favors cytoplasmic localization; a negative correlation favors nuclear. Across composition-based features, cytosine-rich k-mers were consistently positively correlated with cytoplasmic localization, while thymine-containing k-mers showed negative correlations supporting nuclear localization. In order to estimate how much these patterns vary across cell lines, the difference between maximum and

minimum correlation values per feature was computed across all 15 lines. Figures 4.6 and 4.7 show heatmaps of the most variable features. The variation visible in these heatmaps highlights that localization-associated sequence signals are partly cell-line-specific, consistent with the overall motivation for building cell-line-specific models.

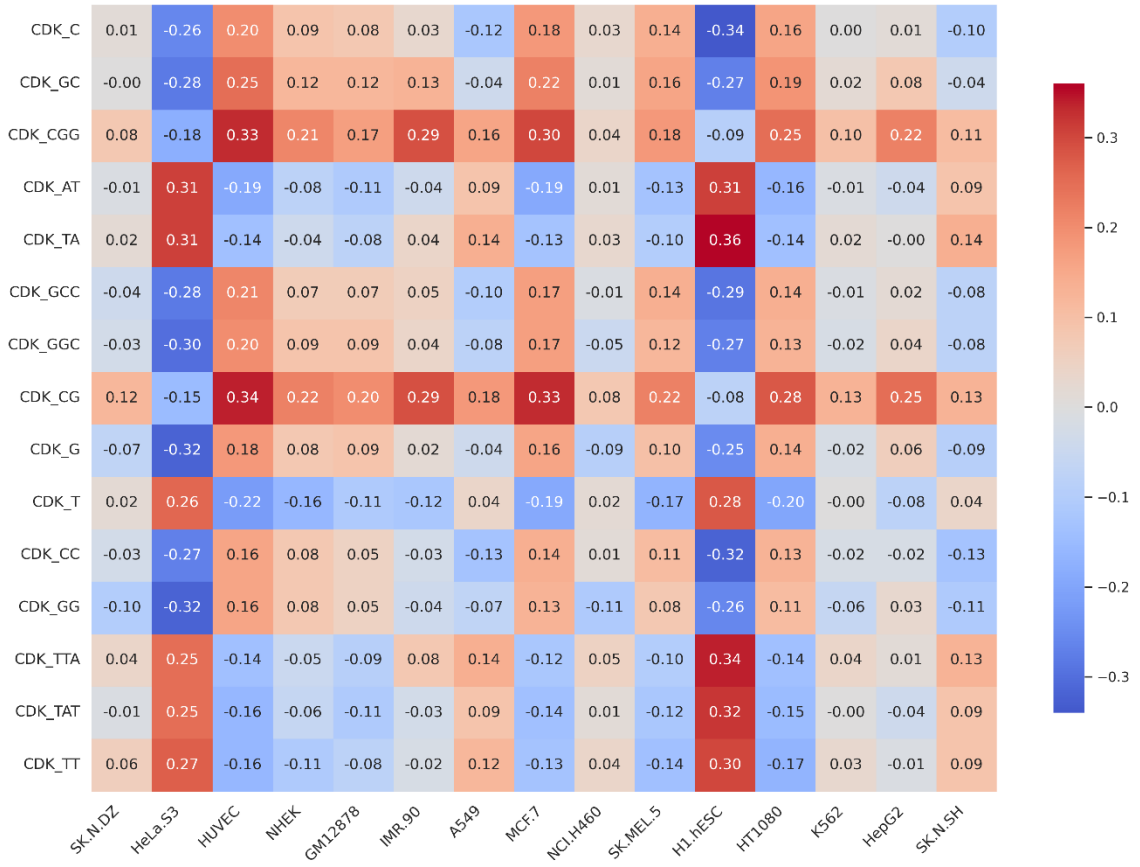


Figure 4.6 - *Top composition-based features selected for their high correlation variance with CNRCI values across* 15 *cell lines.*

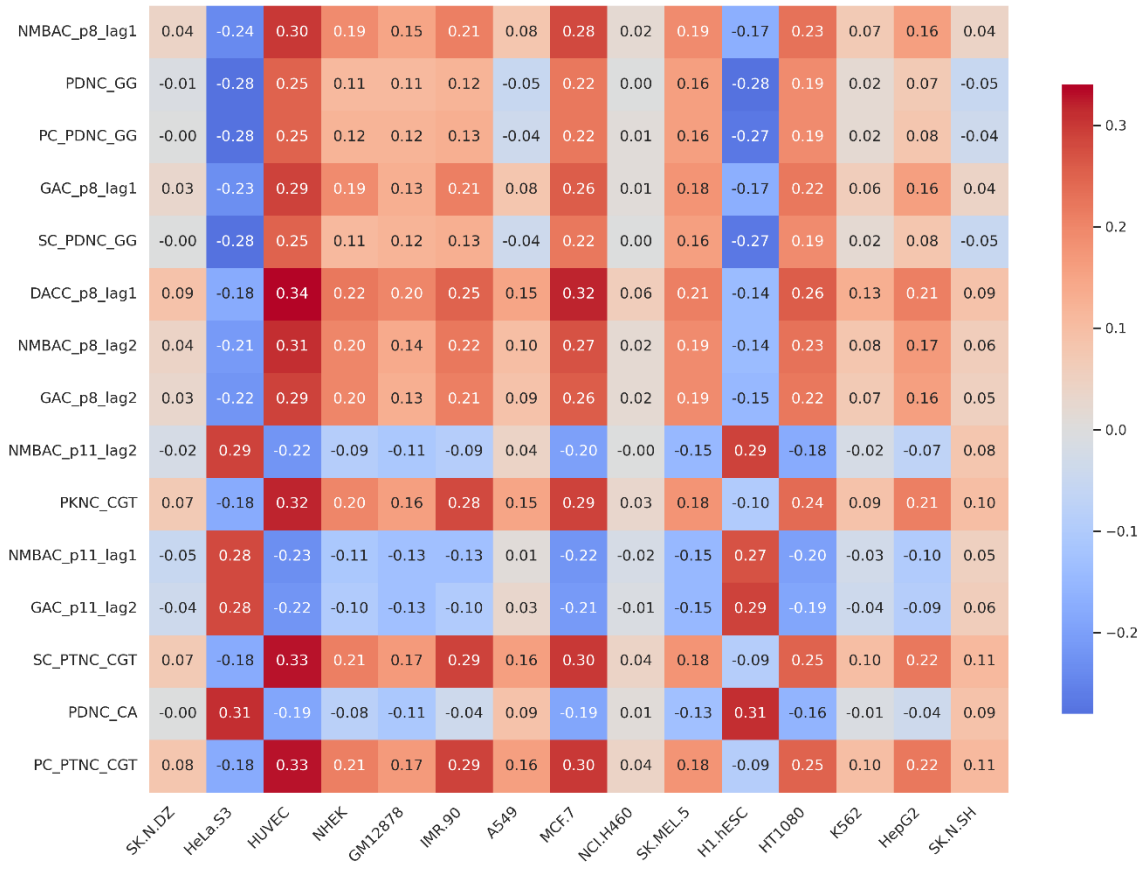


Figure 4.7 - *Top correlation-based features selected for their high correlation variance with CNRCI values across 15 cell lines*

4.3.3 Model based on composition and correlation features

Across the 15 cell lines, composition-based features combined with ML classifiers achieved an average AUC of 0.7049 and MCC of 0.197. The average AUC and MCC of correlation-based features were 0.7089 and 0.213, respectively, which was a marginal improvement. Tables 4.1 and 4.2 provide a summary of these findings. Each cell line has a different top-performing machine learning technique; Random Forest, Gradient Boosting, and GaussianNB classifiers were commonly found in both feature types. Performance varied somewhat by cell line; NCI.H460 and SK.MEL.5 were more difficult to use, indicating differences in dataset size and class balance, whereas cell lines such as A549 and HeLa.S3 displayed higher AUC values.

Table 4.1 – *Performance of the best-performing machine learning models across cell lines on the independent validation dataset using composition-based features, reported in terms of sensitivity, specificity, accuracy, MCC, F*1*-score, and AUC.*

Cell-lines	Feature used	ML Model used	Sensitivity	Specificity	Accuracy	MCC	F1-score	AUC
A549	RDK	RandomForest Classifier	0.600	0.825	0.736	0.438	0.643	0.761
H1.hESC	CDK	RandomForest Classifier	0.500	0.785	0.671	0.297	0.548	0.708
HeLa.S3	PKNC	GaussianNB	0.826	0.593	0.631	0.310	0.422	0.779
HepG2	RDK	GaussianNB	0.429	0.847	0.733	0.292	0.466	0.718
HT1080	CDK	MLPClassifier	0.517	0.769	0.659	0.296	0.569	0.728
HUVEC	PKNC	SVC	0.032	0.964	0.706	-0.011	0.056	0.756
MCF.7	CDK	GaussianNB	0.452	0.809	0.721	0.259	0.444	0.724
NCI.H460	PKNC	SVC	0.000	1.000	0.826	0.000	0.000	0.723
NHEK	CDK	SVC	0.000	1.000	0.742	0.000	0.000	0.643
SK.MEL.5	CDK	XGBClassifier	0.063	0.950	0.763	0.023	0.100	0.676
SK.N.DZ	CDK	MLPClassifier	0.545	0.692	0.635	0.237	0.537	0.679
SK.N.SH	PDNC	QuadraticDiscriminantAnalysis	0.547	0.732	0.683	0.259	0.476	0.703
GM12878	PKNC	GradientBoostingClassifier	0.133	0.939	0.752	0.115	0.200	0.687
K562	ALL_COMP	DecisionTree Classifier	0.452	0.788	0.692	0.243	0.458	0.620
IMR.90	PKNC	AdaBoostClassifier	0.448	0.735	0.603	0.192	0.510	0.668
Average			0.370	0.829	0.704	0.197	0.362	0.705

Note: Reprinted from [214] *under the Creative Commons Attribution License (CC BY).*

Table 4.2 – *Performance of the best-performing machine learning models across cell lines on the independent validation dataset using correlation-based features, reported in terms of sensitivity, specificity, accuracy, MCC, F*1*-score, and AUC.*

Cell-lines	Feature used	ML Model used	Sensitivity	Specificity	Accuracy	MCC	F1-score	AUC
A549	PC_PDNC	GradientBoostingClassifier	0.567	0.803	0.709	0.381	0.607	0.757
H1.hESC	PDNC	RandomForestClassifier	0.495	0.811	0.685	0.324	0.556	0.720
HeLa.S3	PKNC	GaussianNB	0.783	0.585	0.617	0.272	0.400	0.779
HepG2	MAC	GaussianNB	0.679	0.567	0.597	0.218	0.478	0.715
HT1080	SC_PTNC	GaussianProcessClassifier	0.583	0.769	0.688	0.359	0.619	0.738
HUVEC	PDNC	AdaBoostClassifier	0.302	0.897	0.732	0.243	0.384	0.757
MCF.7	SC_PDNC	SVC	0.000	1.000	0.754	0.000	0.000	0.753
NCI.H460	PDNC	SVC	0.000	1.000	0.826	0.000	0.000	0.683
NHEK	PKNC	AdaBoostClassifier	0.231	0.866	0.702	0.116	0.286	0.635
SK.MEL.5	PDNC	GradientBoostingClassifier	0.063	0.933	0.750	-0.007	0.095	0.654
SK.N.DZ	TAC	SVC	0.424	0.865	0.694	0.327	0.519	0.739
SK.N.SH	PC_PTNC	GaussianNB	0.750	0.531	0.588	0.248	0.490	0.689
GM12878	PC_PDNC	XGBClassifier	0.350	0.899	0.771	0.288	0.416	0.715
K562	DACC	AdaBoostClassifier	0.405	0.808	0.692	0.221	0.430	0.642
IMR.90	PC_PTNC	AdaBoostClassifier	0.621	0.588	0.603	0.208	0.590	0.655
Average			0.417	0.795	0.694	0.213	0.391	0.709

Note: Reprinted from [214] *under the Creative Commons Attribution License (CC BY).*

4.3.4 Models based on embeddings from DNABERT-2

When embeddings from the pre-trained DNABERT-2 model were used as features, ML classifiers achieved an average AUC of 0.659 and MCC of 0.118. Using fine-tuned embeddings improved this slightly to an average AUC of 0.660 and MCC of 0.174. Detailed results for each cell line are given in Tables 4.8 and 4.9. The fine-tuned end-to-end DNABERT-2 model gave an average AUC of 0.634 and MCC of 0.127 (Table 4.10). In several cell lines, the fine-tuned model assigned most sequences to a single class, resulting in zero sensitivity, which indicates that the model was prone to class collapse on imbalanced data without further calibration.

Table 4.3 - *Performance of the best-performing machine learning models across cell lines on the independent validation dataset using embeddings derived from the pre-trained DNABERT-2 model, evaluated in terms of sensitivity, specificity, accuracy, MCC, F1-score, and AUC.*

Cell-lines	ML model used	Sensitivity	Specificity	Accuracy	MCC	F1-score	AUC
A549	SVC linear	0.500	0.759	0.656	0.267	0.536	0.661
H1.hESC	MLP Classifier	0.588	0.638	0.618	0.223	0.552	0.666
HeLa.S3	XGBoost Classifier	0.087	0.992	0.844	0.201	0.154	0.699
HepG2	MLP Classifier	0.000	1.000	0.728	0.000	0.000	0.706
HT1080	MLP Classifier	0.817	0.513	0.645	0.338	0.667	0.735
HUVEC	Random Forest Classifier	0.032	0.982	0.719	0.041	0.059	0.690
MCF.7	MLP Classifier	0.012	0.984	0.745	-0.013	0.022	0.657
NCI.H460	Gradient Boosting Classifier	0.063	1.000	0.837	0.228	0.118	0.599

NHEK	Gaussian Naive Bayes Classifier	0.538	0.705	0.662	0.223	0.452	0.660
SK.MEL.5	KNN	0.063	0.967	0.776	0.061	0.105	0.638
SK.N.DZ	Gradient Boosting Classifier	0.364	0.808	0.635	0.191	0.436	0.678
SK.N.SH	MLP Classifier	0.469	0.709	0.646	0.166	0.411	0.618
GM12878	Logistic Regression	0.117	0.955	0.760	0.125	0.184	0.655
K562	Logistic Regression	0.071	0.923	0.678	-0.009	0.113	0.609
IMR.90	Random Forest Classifier	0.345	0.824	0.603	0.193	0.444	0.699
Average		0.271	0.851	0.704	0.149	0.284	0.665

Note: Reprinted from [214] *under the Creative Commons Attribution License (CC BY).*

Table 4.4 - *Performance of the best-performing machine learning models across cell lines on the independent validation dataset using embeddings derived from the fine-tuned DNABERT-2 model, evaluated in terms of sensitivity, specificity, accuracy, MCC, F1-score, and AUC.*

Cell-lines	ML model used	Sensitivity	Specificity	Accuracy	MCC	F1-score	AUC
A549	SVC_linear	0.489	0.708	0.621	0.200	0.506	0.660
H1.hESC	SVC_linear	0.422	0.801	0.650	0.241	0.490	0.687
HeLa.S3	MLP Classifier	0.000	0.992	0.830	-0.037	0.000	0.732
HepG2	Gaussian Naive Bayes Classifier	0.571	0.767	0.714	0.321	0.520	0.708
HT1080	SVC_linear	0.450	0.756	0.623	0.217	0.509	0.700
HUVEC	MLP Classifier	0.206	0.903	0.711	0.147	0.283	0.726

MCF.7	Gaussian Naive Bayes Classifier	0.512	0.739	0.683	0.232	0.443	0.700
NCI.H460	XGBoost Classifier	0.000	0.987	0.815	-0.048	0.000	0.535
NHEK	AdaBoost Classifier	0.359	0.821	0.702	0.189	0.384	0.600
SK.MEL.5	SVC_radial	0.000	1.000	0.789	0.000	0.000	0.539
SK.N.DZ	Gradient Boosting Classifier	0.515	0.692	0.624	0.207	0.515	0.717
SK.N.SH	AdaBoost Classifier	0.156	0.866	0.679	0.028	0.204	0.587
GM12878	Gradient Boosting Classifier	0.183	0.939	0.764	0.182	0.265	0.683
K562	Gradient Boosting Classifier	0.143	0.894	0.678	0.052	0.203	0.590
IMR.90	Gradient Boosting Classifier	0.552	0.706	0.635	0.261	0.582	0.671
Average		0.304	0.838	0.701	0.146	0.327	0.656

Note: Reprinted from [214] *under the Creative Commons Attribution License (CC BY).*

4.3.5 Fine-tuned DNABERT-2 model

We tested DNABERT-2 as a standalone predictor by fine-tuning it directly on our training dataset. Due to computational constraints, cross-validation could not be applied during training. When evaluated on the validation dataset, the model achieved a modest performance, with AUC and MCC values of 0.634 and 0.127 respectively, which is lower than those obtained using classical ML with simple features. Performance of the fine-tuned DNABERT-2 model is provided in Table 4.5. The fine-tuned DNABERT-2 model was not able to distinguish cytoplasmic from nuclear lncRNAs consistently. This suggests that transformer-based models may require more extensive

domain-specific adaptation or larger, higher-quality training datasets to perform well for this problem.

Table 4.5 - *Performance of the fine-tuned DNABERT-2 model used directly for classification across cell lines on the validation dataset*

Cell-lines	Sensitivity	Specificity	Accuracy	MCC	F1-score	AUC
A549	0.500	0.861	0.718	0.393	0.584	0.762
H1.hESC	0.520	0.713	0.636	0.235	0.533	0.644
HeLa.S3	0.000	1.000	0.837	0.000	0.000	0.527
HepG2	0.000	1.000	0.728	0.000	0.000	0.698
HT1080	0.567	0.731	0.659	0.301	0.591	0.676
HUVEC	0.159	0.976	0.750	0.251	0.260	0.710
MCF.7	0.000	1.000	0.754	0.000	0.000	0.562
NCI.H460	0.000	1.000	0.826	0.000	0.000	0.576
NHEK	0.000	1.000	0.742	0.000	0.000	0.564
SK.MEL.5	0.000	1.000	0.789	0.000	0.000	0.468
SK.N.DZ	0.515	0.885	0.741	0.439	0.607	0.791
SK.N.SH	0.000	1.000	0.737	0.000	0.000	0.589
GM12878	0.000	1.000	0.767	0.000	0.000	0.718
K562	0.190	0.885	0.685	0.099	0.258	0.570
IMR.90	0.655	0.529	0.587	0.185	0.594	0.649
Average	0.207	0.905	0.730	0.127	0.228	0.634

Note: Reprinted from [214] *under the Creative Commons Attribution License (CC BY).*

4.3.6 Model based on features selected by mRMR algorithm

To check whether reducing the number of features could further improve performance, we applied the mRMR method to select smaller feature subsets. However, using mRMR did not lead to any clear improvement in the model results. The best-performing models already showed strong performance with the full or selected feature sets, suggesting that these features contain useful and complementary information rather than unnecessary redundancy. In several cases, mRMR-based filtering appeared to remove weak but helpful signals, leading to similar or slightly poorer

sensitivity–specificity trade-offs without improving AUC or MCC. Although mRMR reduced the number of features and simplified model training, this did not translate into better prediction on unseen data. Therefore, mRMR feature selection was not used in the final models, which were trained using the feature sets that gave the most reliable performance across cell lines. Detailed performance results for these final models are provided in Table 4.6.

Table 4.6 - *Impact of the number of features (k) selected using minimum Redundancy–Maximum Relevance (mRMR) on classification performance across cell lines, with AUC reported on the independent validation dataset.*

Cell lines	k =10	k = 50	k = 100	k = 500	k = 1000	k = 1500	k = 2000
A549	0.756	0.75	0.749	0.741	0.739	0.738	0.738
GM12878	0.725	0.706	0.706	0.727	0.742	0.734	0.719
H1.hESC	0.677	0.687	0.694	0.71	0.712	0.716	0.717
HT1080	0.71	0.73	0.728	0.731	0.745	0.749	0.742
HUVEC	0.733	0.744	0.762	0.771	0.76	0.748	0.764
HeLa.S3	0.669	0.728	0.756	0.734	0.718	0.776	0.779
HepG2	0.736	0.722	0.725	0.72	0.691	0.673	0.685
IMR.90	0.725	0.651	0.647	0.621	0.636	0.602	0.663
K562	0.615	0.665	0.645	0.628	0.585	0.581	0.586
MCF.7	0.706	0.728	0.73	0.701	0.715	0.721	0.711
NCI.H460	0.535	0.514	0.533	0.638	0.553	0.593	0.568
NHEK	0.64	0.634	0.626	0.609	0.603	0.632	0.65
SK.MEL.5	0.672	0.652	0.552	0.725	0.597	0.619	0.575
SK.N.DZ	0.671	0.664	0.675	0.704	0.705	0.715	0.707
SK.N.SH	0.683	0.684	0.685	0.669	0.683	0.684	0.682
Average	0.683	0.684	0.681	0.695	0.679	0.685	0.686

Note: Reprinted from [214] *under the Creative Commons Attribution License (CC BY).*

4.3.7 Performance comparison of CytoLNCpred and existing state-of-the-art classifiers

CytoLNCpred was benchmarked against lncLocator 2.0 and TACOS on the same held-out validation datasets. lncLocator 2.0 achieved an average AUC of 0.523 across all 15 cell lines.

TACOS, which only covers 10 of the 15 cell lines, averaged an AUC of 0.676. CytoLNCpred using composition features achieved an average AUC of 0.705, and using correlation features achieved 0.709. CytoLNCpred consistently outperformed both existing tools across nearly every cell line, with especially notable improvements in NCI.H460, MCF.7, and SK.N.DZ where TACOS had no predictions and lncLocator 2.0 was close to random.

To further investigate the importance of cell-line-specific prediction, we also evaluated GraphLncLoc [87], a state-of-the-art sequence-based predictor of general lncRNA subcellular localization, on the same validation datasets. Unlike CytoLNCpred, GraphLncLoc is designed to predict global localization classes and is not trained to capture cell-line-specific localization patterns. Consequently, it achieved a substantially lower average AUC of 0.432 across the 15 cell lines, with individual cell-line AUCs ranging from 0.345 to 0.612. Figure 4.8 gives a comparative overview of the existing methods against CytoLNCpred on the basis of AUC.

The reduced performance highlights the limited transferability of general localization models to the task of predicting cell-line-specific cytoplasmic enrichment. Since lncRNA localization is highly context dependent and varies across cellular environments, these findings underscore the importance of developing dedicated cell-line-specific predictors. Collectively, these results demonstrate that incorporating cell-line-specific localization information enables CytoLNCpred to consistently achieve superior predictive performance compared with existing general-purpose localization prediction tools.

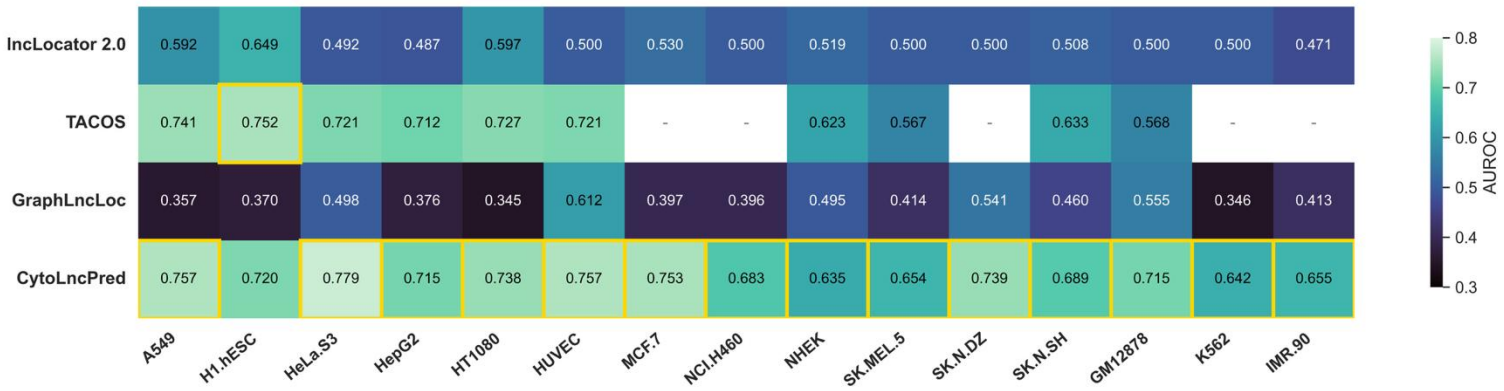


Figure 4.8 – *Comparison of existing tools (lncLocator 2.0, TACOS, GraphLncLoc) with CytoLNCpred based on AUC*

4.4 Discussion and Conclusion

CytoLNCpred was designed to address concrete problems with existing tools—contaminated datasets, exclusion of borderline-localized lncRNAs, and lack of cell-line specificity. The cleaner dataset construction strategy contributed directly to the performance improvements over lncLocator 2.0 and TACOS. Among all approaches tested, correlation-based features combined with machine learning classifiers consistently achieved the highest predictive performance. Although recent advances in genomics have emphasized transformer-based language models, our results suggest that carefully engineered correlation-based descriptors remain more effective for predicting cell line-specific lncRNA localization. This is likely because nucleocytoplasmic localization is determined by relatively subtle sequence signals, where higher-order nucleotide dependencies provide greater discriminatory power than simple nucleotide composition or general-purpose sequence embeddings. Consistent with this observation, negative correlation scores were predominantly associated with nuclear-localized transcripts, whereas positive scores were enriched among cytoplasmic lncRNAs, demonstrating that correlation-based features effectively capture biologically relevant localization signatures. In contrast, global sequence properties such as transcript length and GC content exhibited only modest and inconsistent differences between nuclear and cytoplasmic lncRNAs, with their influence varying across individual cell lines. These findings indicate that broad compositional characteristics alone are insufficient to explain localization and further justify the use of correlation-derived sequence descriptors.

The enrichment of CG dinucleotides and CG-containing trinucleotides among cytoplasm-labelled lncRNAs, together with the enrichment of TG- and related T-rich motifs among nuclear transcripts, suggests that short compositional features may play a role in determining localization. Differences in dinucleotide physicochemical properties, such as helix structure and base-stacking stability, suggest that RNA localization depends not only on nucleotide composition but also on structural and thermodynamic characteristics. These associations also varied across cell lines, and classical nuclear-retention elements, such as AG-rich or extended C-Rich motifs, showed only weak and inconsistent relationships with the localization labels. Collectively, these findings imply that subcellular localization is multifactorial and context-dependent rather than governed by a single motif class. These findings also generate testable hypotheses, such as whether CG-rich

cytoplasmic and T-rich nuclear lncRNAs interact with different RNA-binding proteins, and whether changes in the abundance or activity of these proteins cause cell line-specific differences in lncRNA localization.

The analyses also reinforce the central role of cellular context in determining lncRNA localization. The same lncRNA frequently exhibited different localization labels across cell lines, indicating that localization is governed not only by intrinsic sequence information but also by cell-specific regulatory processes, including RNA-binding protein availability, nuclear retention, RNA export, ribonucleoprotein complex assembly, and alternative splicing. This context dependency provides a biological explanation for why cell line-specific prediction models outperform generalized approaches. For example, *LINC00852* was predominantly nuclear across most cell lines but strongly cytoplasmic in NCI-H460, consistent with reports that it functions in the cytoplasm of lung carcinoma cells through interaction with S100A9 and activation of the MAPK pathway, while remaining nuclear in osteosarcoma where it regulates AXL transcription. Similarly, *SNHG3* was predominantly cytoplasmic in H1.hESC, HepG2, and GM12878 but nuclear in MCF-7 and NHEK, consistent with its reported cytoplasmic role as a competing endogenous RNA in colorectal cancer. Considerable heterogeneity was also observed in the proportion of cytoplasmic lncRNAs across the fifteen cell lines, ranging from approximately 16% to 47%. HeLa.S3 and NCI-H460 were among the most nucleus-enriched cell lines, whereas IMR-90 displayed a comparatively balanced distribution. These observations indicate that model evaluation should be performed separately for each cell line, as pooled performance metrics may obscure biologically meaningful differences arising from distinct localization distributions. Furthermore, grouping cell lines according to localization characteristics revealed trends that partially reflected cellular lineage, with epithelial and carcinoma-derived cell lines generally exhibiting greater nuclear enrichment, hematopoietic cell lines containing relatively few cytoplasmic lncRNAs, and fibroblast IMR-90 together with human embryonic stem cell H1 showing higher cytoplasmic prevalence. The presence of notable exceptions, however, emphasizes that lineage alone does not fully determine localization patterns and highlights the complexity of cell-specific regulatory mechanisms.

The localization predictions from CytoLNCpred also have direct therapeutic implications. Antisense oligonucleotides (ASOs) are generally more effective against nuclear lncRNAs, whereas siRNAs preferentially target cytoplasmic transcripts. Consequently, CytoLNCpred predictions can

guide early therapeutic modality selection, with cytoplasmic predictions favouring siRNA-based approaches and nuclear predictions supporting ASO development. Likewise, CRISPR-based strategies can be selected according to localization, with nuclear lncRNAs being more suitable for CRISPRi or Cas9-mediated regulation and cytoplasmic transcripts representing more appropriate targets for RNA-targeting CRISPR-Cas13 systems. CytoLNCpred therefore serves as a valuable decision-support tool during the initial stages of RNA therapeutic design.

Despite demonstrating improved performance over existing methods, the overall predictive accuracy of CytoLNCpred indicates that cell line-specific lncRNA localization prediction remains a challenging computational task. The observed prediction errors can be attributed to both biological complexity and technical limitations. From a biological perspective, subcellular localization is governed by a combination of sequence-dependent and sequence-independent mechanisms. While the primary nucleotide sequence contains localization-associated signals, additional determinants including RNA secondary and tertiary structure, RNA-binding protein interactions, alternative splicing, epitranscriptomic modifications, and the availability of cellular interaction partners collectively influence the final localization of a transcript. These regulatory processes vary substantially across cell types and physiological conditions and are not captured by sequence-based models, thereby limiting their predictive capacity. Furthermore, the binary classification labels used in this study were derived from continuous CNRCI values obtained from RNA-seq experiments. Consequently, transcripts with CNRCI values close to zero may exhibit ambiguous localization patterns, introducing label uncertainty that can adversely affect model training and evaluation.

Despite outperforming existing methods, the overall predictive performance of CytoLNCpred indicates that cell-line-specific lncRNA localization remains a challenging computational problem. The observed prediction errors arise from both biological complexity and technical limitations. From a biological perspective, subcellular localization is regulated by a combination of sequence-dependent and sequence-independent mechanisms. Although the primary nucleotide sequence encodes localization-associated signals, additional determinants, including RNA secondary and tertiary structure, RNA-binding protein interactions, alternative splicing, epitranscriptomic modifications, and the availability of interacting molecular partners, collectively govern the spatial distribution of lncRNAs. These regulatory mechanisms vary substantially across cell types and

physiological conditions and are not captured by sequence-based models, thereby limiting their predictive capability. Furthermore, the binary labels used in this study were derived from continuous CNRCI values obtained from bulk RNA-seq experiments. Consequently, transcripts with CNRCI values close to zero may exhibit ambiguous localization patterns, introducing label uncertainty that can adversely affect model training and evaluation.

Several technical factors further contributed to the moderate predictive performance observed across the evaluated models. The relatively small number of experimentally validated lncRNAs available for individual cell lines limits the ability of machine learning algorithms, particularly transformer-based architectures, to learn robust and generalizable representations, as reflected by the comparatively lower performance of DNABERT-2. Variations in class distribution among cell lines also reduced model sensitivity and, in certain cases, resulted in a bias toward predicting the majority class. In addition, the present study relies on lncAtlas, which is based on ENCODE poly(A)+ fractionation RNA-seq data from 15 cell lines generated under uniform experimental conditions and has not been updated since 2017. Consequently, the dataset is restricted to GENCODE-annotated lncRNAs and represents steady-state localization under basal cellular conditions, without capturing the dynamic redistribution of lncRNAs in response to cellular stress, differentiation, or other perturbations. Moreover, the current framework relies exclusively on sequence-derived features and does not incorporate complementary biological information, such as RNA secondary structure, RNA-binding protein interaction profiles, transcript abundance, epitranscriptomic modifications, or chromatin accessibility. Future improvements should therefore focus on integrating multimodal biological data, expanding training datasets using more recent and diverse RNA-seq resources, incorporating single-cell and perturbation-based localization datasets, and developing strategies to effectively model long lncRNA transcripts, thereby improving both predictive accuracy and biological interpretability.

CytoLNCpred is a cell-line-specific predictor of cytoplasmic lncRNA localization across 15 human cell lines. This method utilizes correlation-based features combined with classical ML algorithms and consistently outperform both large language model–based approaches and existing state-of-the-art tools. We used cleaner training datasets, tested multiple modelling strategies, and applied rigorous validation to achieve this. The method is publicly available as a web server and

standalone package, providing a practical resource for functional annotation of lncRNAs and selection of RNA-based therapeutic strategies.

5

Prediction of lncRNA – protein interaction

5.1 Introduction

RNA molecules longer than 200 nucleotides that do not encode proteins yet play crucial regulatory roles in cells are known as long non-coding RNAs, or lncRNAs [54]. Many of these processes include direct binding to proteins, creating ribonucleoprotein complexes in which the biological result is decided by both the RNA and protein components [215,216]. Examples that are well-established include *MALAT1* controlling alternative splicing by attaching to SR proteins and XIST enlisting the Polycomb Repressive Complex 2 (PRC2) to silence the X chromosome [217,218]. These associations show how important lncRNA–protein interactions (LPIs) contribute to nuclear organization, gene regulation, chromatin remodeling, and cell signaling.

Potential therapeutic targets, disrupted LPIs are implicated in a number of illnesses, such as cancer and neurological disorders. Mapping these relationships empirically is still lacking, though. Although they take a lot of time, money, and specialized knowledge, techniques like RNA immunoprecipitation (RIP) and crosslinking methods like PAR-CLIP and iCLIP can accurately identify RNA–protein interaction sites throughout the transcriptome. Large-scale screening is therefore not feasible [219–221].

Researchers have created computational algorithms for predicting LPIs in order to maneuver around these restrictions. Earlier techniques included conventional machine learning that was trained on manually generated sequence features such as physicochemical characteristics, projected RNA structure, and k-mer composition [104,222,223]. Subsequent methods used network data, inferring likely associations from protein interaction networks and gene expression patterns [106,224]. Although the majority of these tools were developed using outdated NPInter database versions and are no longer maintained or accessible to the public, combining many prediction models produced some slight improvements [113].

Deep learning techniques such as graph transformers, graph autoencoders, and graph convolutional networks have been used to address this issue in recent work. But almost all of the current tools have an important drawback: they usually are unable to evaluate new user-provided sequences and rely on small, out-of-date training datasets. This paper introduces lncrnaPI, a complete prediction method trained on the largest curated LPI dataset currently accessible from NPInter v5.0, as a solution to these problems. This strategy incorporates embeddings from big language models,

convolutional deep learning, alignment-based techniques, and classical machine learning. Both a web server and a stand-alone Python tool for large-scale LPI prediction are freely accessible.

5.2 Materials and Methods

The pipeline consists of four primary steps: dataset construction and curation, multi-scale feature generation, model development across alignment-based, machine learning, and deep learning frameworks, and systematic performance evaluation. A graphical overview of the complete workflow is presented in Figure 5.1.

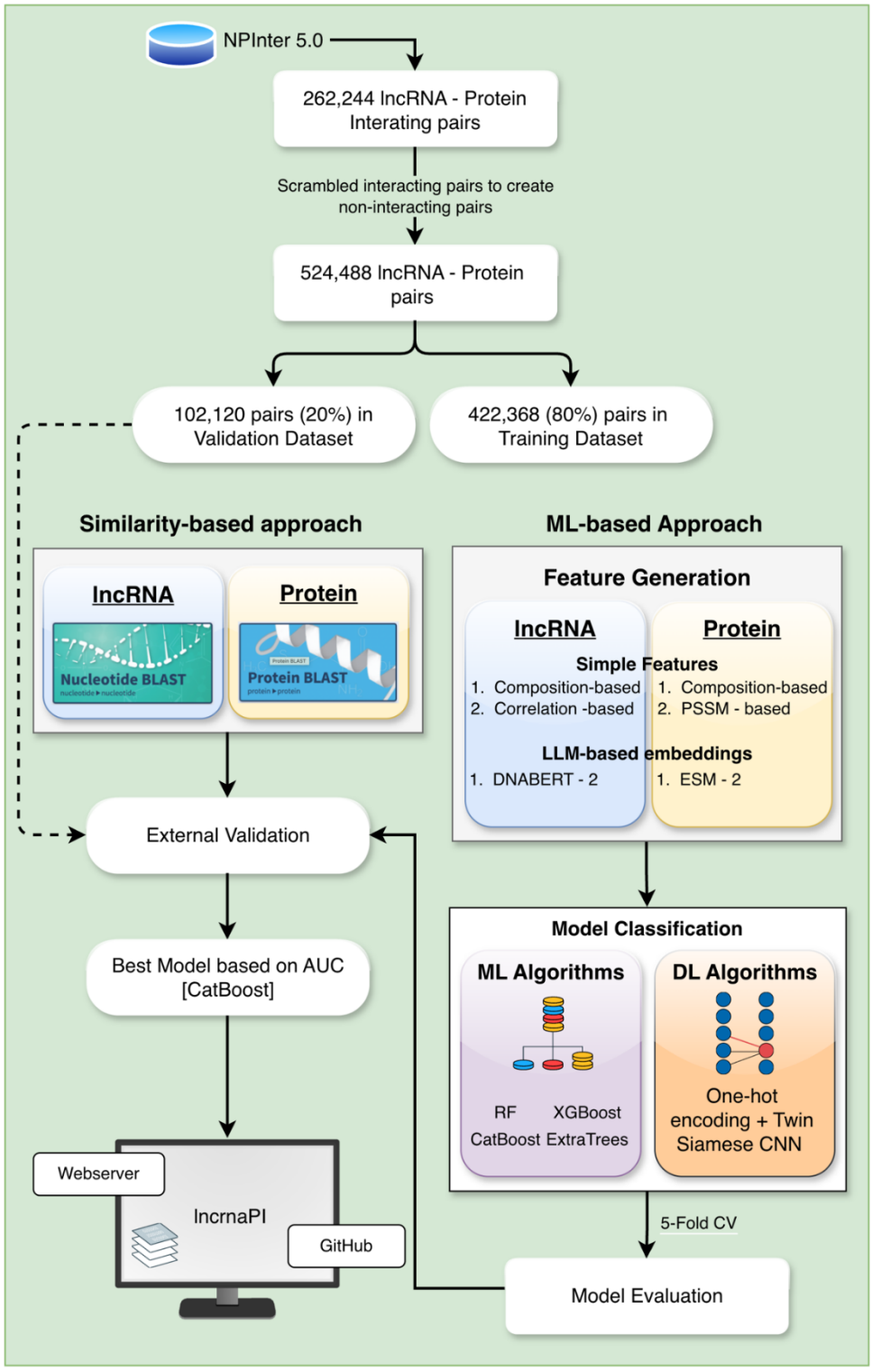


Figure 5.1 - *Overall methodological workflow of lncrnaPI*

5.2.1 Dataset Curation

We constructed a high-quality, non-redundant dataset by extracting human lncRNA–protein interactions from NPInter v5.0 [102]. NPInter integrates experimentally supported interactions from multiple sources, including high-throughput CLIP-based studies (e.g., eCLIP, PAR-CLIP, CLIP-seq and iCLIP), the ENCODE project, and manually literature-curated interaction databases. For this study, interactions originating from high-throughput datasets (including records annotated as High-throughput data combined with LncPro prediction), ENCODE, and literature-curated sources were retained. The distribution of interaction sources and the supporting experimental evidence are summarized in Figure 5.2. This resulted in 488,982 human lncRNA–protein interactions from the database's ~2.5 million experimentally supported ncRNA–biomolecule interactions. lncRNA sequences between 200 and 20,000 nucleotides in length were retained, while protein sequences containing only the 20 standard amino acids were included. After applying these sequence quality filters, the dataset was reduced to 370,242 lncRNA–protein interactions. Sequence redundancy was subsequently removed using CD-HIT at an 80% sequence identity threshold for lncRNAs and a 95% threshold for proteins, yielding a final positive dataset of 262,244 unique interacting lncRNA–protein pairs [225].

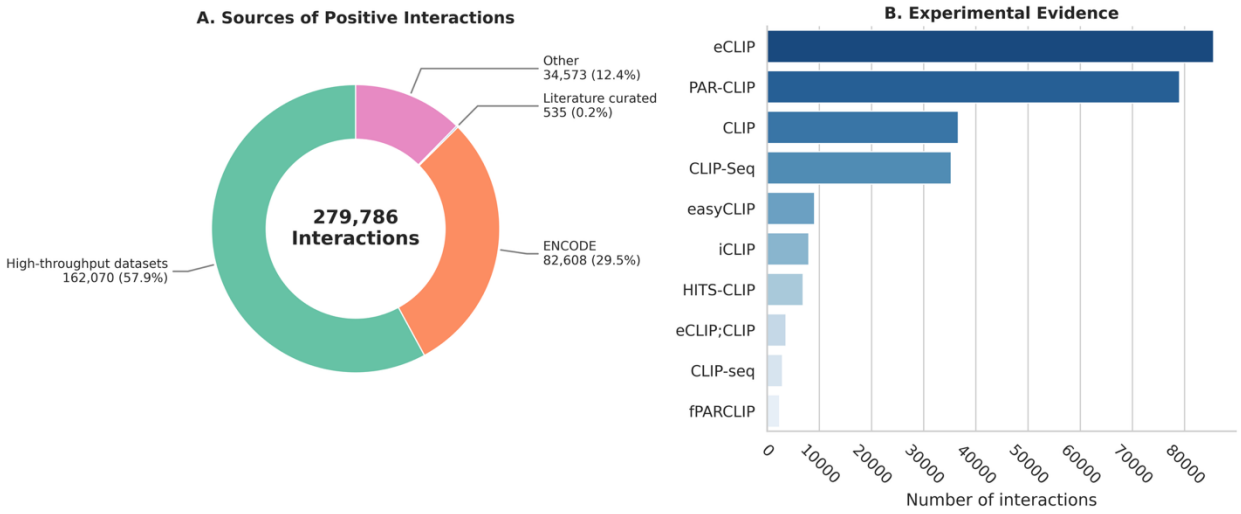


Figure 5.2 - *Data sources and experimental evidence supporting the positive lncRNA–protein interaction dataset. Panel (A) shows the distribution of positive interactions by source, including*

high-throughput datasets, ENCODE, and literature-curated records. Panel (B) summarizes the most frequent experimental techniques supporting these interactions.

To construct the negative dataset, an equal number of non-interacting lncRNA–protein pairs were generated by randomly shuffling the positive interaction pairs while excluding all known positive interactions to minimize the inclusion of false negatives. The resulting balanced dataset comprised 524,488 lncRNA–protein pairs (262,244 positive and 262,244 negative interactions). Finally, the dataset was partitioned by stratified sampling into an 80% training set and a 20% independent validation set to preserve the class distribution. The overall dataset construction workflow is summarized in Figure 5.3.

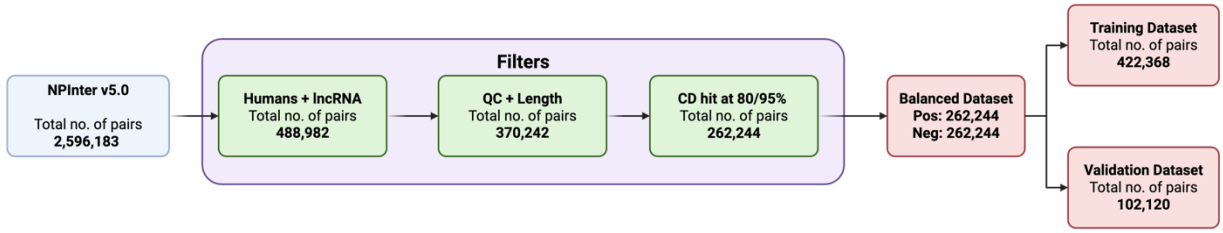


Figure 5.3 - *Schematic workflow of the data collection and preprocessing pipeline for the construction of the balanced human lncRNA–protein interaction dataset.*

5.2.2 Feature Generation

We created numerical representations of lncRNA and protein sequences using multiple feature extraction methods to capture different aspects of sequences that influence molecular binding.

For compositional features, we used the Nfeature toolkit for lncRNAs and the Pfeature toolkit for proteins [208,226]. These tools calculate basic descriptors like nucleotide and amino acid frequencies, along with various physicochemical and structural properties. We tested all of the available descriptors and found that Dinucleotide-based Auto-Cross Covariance (DACC) were particularly effective for lncRNAs because it captures both distant and local nucleotide patterns associated with RNA folding. Protein sequences were encoded using Amphiphilic Pseudo-Amino Acid Composition (APAAC) and Shannon Entropy of Physicochemical Properties (SPC). SPC

provides a measure of compositional complexity by quantifying the variety of physicochemical properties along a sequence. APAAC measures the distribution and sequence-order effects of hydrophobic and hydrophilic residues, which are important factors in determining interaction interfaces and binding surfaces. Position-Specific Scoring Matrices (PSSMs) were created utilizing PSI-BLAST searches against the SwissProt database using POSSUM in order to include evolutionary information [227,228]. Functionally constrained sites are highlighted by PSSMs, which encode residue substitution probabilities inferred from evolutionary conservation.

Large language models were used to generate contextual embeddings in addition to manually created descriptors. For lncRNAs, sequence representations were obtained using DNABERT-2, a transformer model pretrained on genomic sequences across multiple species [209]. For proteins, mean-pooled embeddings were extracted from three ESM-2 variants (t30, t33, and t48), each trained on extensive protein sequence corpus and capable of capturing implicit structural and functional features encoded in primary sequence data [229]. For the convolutional neural network, we transformed sequences to one-hot encoded matrices, normalizing lncRNA sequences to 2,500 places and protein sequences to 1,500 locations using padding or truncation. Table 5.1 outlines the primary feature categories employed in this study.

Table 5.1 - *Key features utilized for numerical representation of lncRNA and protein sequences, categorized by sequence type and biological significance.*

Selected Feature	Sequence Type	Source / Tool	Biological Significance
DACC	lncRNA	Nfeature	Models local and long-range nucleotide correlations influencing RNA secondary structure.
APAAC	Protein	Pfeature	Captures the amphiphilic residue distribution essential for protein-binding interfaces.
SPC	Protein	Pfeature	Quantifies the informational diversity and complexity of physical properties along the protein.
PSSM	Protein	POSSUM	Identifies evolutionarily conserved residues that serve as functionalf interaction hotspots.
Embedding	lncRNA	DNABERT-2	Extracts high-level semantic patterns and structural motifs from genomic sequences.
Embedding	Protein	ESM-2 (150M)	Represents latent structural folds and functional properties via deep representation learning.

5.2.3 Alignment-Based Prediction

A sequence homology baseline was used to quantify the role of direct sequence similarity in LPI prediction and to interpret the performance of machine learning algorithms in comparison to a simpler reference [230]. The training sets of lncRNA and protein sequences were used to create separate BLAST databases. Validation sequences were compared to these databases using two prediction techniques: a nearest-neighbour strategy, which assigned the interaction label of the highest-scoring training pair to each query, and a consensus voting model, which determined the predicted label by majority vote across all statistically significant BLAST hits.

5.2.4 Machine Learning Models

Four tree-based ensemble classifiers were trained using the feature representations: Random Forest, Extra Trees, XGBoost, and CatBoost. Random Forest and Extra Trees were set up with 100 estimators and a fixed random seed of 42 to ensure consistency between experimental runs. Given the size of the dataset, techniques that support native parallel execution were preferred over single-threaded operations, which were computationally slower. The XGBoost classifier was implemented with log-loss as the optimisation criterion. All models were trained on the training set with stratified five-fold cross-validation, and their final performance was reported on the independent held-out validation partition.

5.2.5 Deep Learning Models

We constructed a two-tower Convolutional Neural Network (CNN) that learns directly from sequence data without the need for manually created features. It analyzes lncRNA and protein sequences via distinct but parallel pathways. The lncRNA route utilizes two successive convolutional blocks to inputs with the shape (2500, 4). The first block applies dropout at 0.4, max-pooling with a pool size of 4, and employs 128 filters with a kernel size of 16. The same pooling and dropout procedures are used in the second block, which employs 256 filters with a kernel size of 10. The protein pathway has a similar design, employing 128 filters with kernel size

10 in the first block and 256 filters with kernel size 8 in the second block, with the same pooling and dropout parameters. The inputs are shaped as (1500, 20).

Post processing, a single combined representation was created by flattening and combining the outputs from both branches. Then, using ReLU activation algorithms, this combined representation moved through two completely linked layers with 256 and 128 neurons, respectively. To prevent overfitting, we included a dropout layer with a rate of 0.5 after the first fully connected layer. The final prediction comes from a single output neuron with sigmoid activation for binary classification. We trained the model using the Adam optimizer with a learning rate of 0.001, binary cross-entropy as the loss function, and monitored AUC as the main performance measure during training.

5.2.6 Evaluation Metrics

Standard binary classification metrics as discussed in Section 4.2.7, such as accuracy, precision, sensitivity (recall), specificity, F1-score, Matthews Correlation Coefficient (MCC), and Area Under the Receiver Operating Characteristic Curve (AUC), were used to assess the model's performance. For model comparison, MCC and AUC were given top priority. AUC evaluates a classifier's capacity to differentiate between classes without regard to a decision threshold, but MCC offers a fair evaluation by taking into account every component of the confusion matrix and maintains stability in the face of class imbalance. The mean across five stratified cross-validation folds was used to estimate training performance. An independent held-out test set was used for the final model evaluation in order to evaluate generalization to unknown data.

5.2.7 Webserver Implementation and Standalone Application

lncrnaPI was made available in two formats to meet the needs of diverse users. Users can upload lncRNA and protein sequences and retrieve interaction predictions directly through a browser using a web-based interface (https://webs.iiitd.edu.in/raghava/lncrnapi). No local installation or command-line usage is required. This tool allows people without a lot of computing experience to do rapid analysis. Through the Python Package Index (PyPI), lncrnaPI is also accessible as a

Python package for workflow integration and large-scale applications. This version can be integrated into automated bioinformatics pipelines and offers programmatic access. When combined, these deployment methods facilitate high-throughput genome-scale analysis as well as routine experimental application.

5.3 Results

5.3.1 Overall Performance Evaluation

We systematically compared alignment-based, machine learning, and deep learning models on the independent held-out validation set, using accuracy, MCC, and AUC as the primary evaluation metrics. The alignment-based method performed notably worse than both machine learning and deep learning approaches, suggesting that raw sequence similarity alone is not sufficient to reliably predict LPIs. Machine learning and deep learning models both achieved considerably higher accuracy and discriminative performance, indicating that the sequence signals underlying lncRNA–protein interaction specificity are complex and non-homologous in nature, and therefore require more sophisticated modeling approaches to detect effectively.

5.3.2 Performance of Alignment-Based Method

Evaluation of the alignment-based baselines revealed a fundamental limitation of homology-driven prediction for the present dataset. Of the 102,120 pairs comprising the validation set, the large majority did not return any significant BLAST hits against the training database, reflecting the substantial sequence divergence between training and validation partitions following redundancy removal by CD-HIT. Among the subset of pairs for which alignments were obtained, the nearest-neighbour approach correctly classified 13,630 of 17,777 matchable pairs, corresponding to an accuracy of 76.67%. The consensus voting model achieved a marginally higher accuracy of 79.71% by correctly classifying 13,659 of 17,136 pairs. Although sequence homology carries predictive value in cases where close training-set relatives exist, the absence of BLAST hits for the majority of validation pairs, combined with an accuracy approximately 15 percentage points below the machine learning models, confirms that alignment-based approaches

are inadequate as primary prediction tools for this task. These findings establish that LPI prediction depends on complex physicochemical and structural sequence patterns not captured by sequence similarity alone.

5.3.3 Compositional Analysis

In order to understand which sequence-based patterns explain lncRNA–protein interactions in our dataset, we examined all 524,488 training and validation pairs (1,482 proteins and 33,054 lncRNAs). The interaction network was observed to be slightly skewed. A small group of RNA-binding proteins - including FUS, WDR33, HNRNPA2B1, IGF2BP1/3, RBM6, DHX36, UPF1, FBL, SFPQ and NONO, accounted for most positive pairs, each with positive rates above 95%. On the lncRNA side, hubs included well-known transcripts such as *MALAT1, XIST, GAS5, RMRP* and several SNHG RNAs. Highly interacting proteins had lower leucine composition and higher in glycine, asparagine, tyrosine and glutamine than proteins having few lncRNA partners. Short peptide motifs supported the same pattern: RGG, GYG, GRG, YGG and related Gly/Arg sequences were strongly enriched in these hub proteins, in line with known RGG-box and disordered RNA-binding regions. For lncRNAs, base composition (A/T/G/C) was only a weak signal, whereas longer transcripts were more favored among interacting lncRNAs.

Gene Ontology analysis of the top hub proteins, compared with all proteins in the dataset, showed enrichment for proteins involved in miRNA and pre-miRNA processing, RISC complex assembly, alternative splicing, and related mRNA-processing pathways (FDR < 0.05). Overall, our labelled dataset and the prediction model is driven mainly by RNA-binding protein sequence features and nuclear RNA-processing pathways, rather than by broad composition differences across all proteins or lncRNAs. This leads to a simple biological hypothesis that high-confidence predictions should often involve Gly/Arg-rich RNA-binding proteins and longer nuclear lncRNAs linked to splicing or miRNA pathways, while predictions for non-hub proteins may highlight less-studied interactions outside these dominant patterns.

We also investigated the intrinsic sequence properties that may govern lncRNA-protein interactions. We performed a detailed compositional analysis of both amino acids and nucleotides. This analysis revealed a strong predictive signal within the amino acid composition (AAC) of the

protein sequences, as shown in Figure 5.4. We found that specific amino acids are significantly enriched or depleted in the interacting protein set. The most discriminative amino acid was Leucine (L), which was significantly depleted in positive (interacting) pairs (mean composition: 6.34%) compared to negative pairs (9.33%), yielding a standalone predictive accuracy of 71.6%. Conversely, Glycine (G) was the most enriched amino acid in positive pairs (11.55% vs. 7.30%), achieving 64.9% accuracy. Other amino acids, including Cysteine (C), Glutamine (Q), and Tyrosine (Y), also showed strong biases with accuracies over 60%. In contrast, the analysis of nucleotide composition in the lncRNA sequences showed only marginal differences between positive and negative sets, with predictive accuracies ranging between 50-51%. This suggests that while simple nucleotide frequency is not a strong driver of interaction specificity, the amino acid composition of the protein partner is a key determinant and provides a strong, predictive signal.

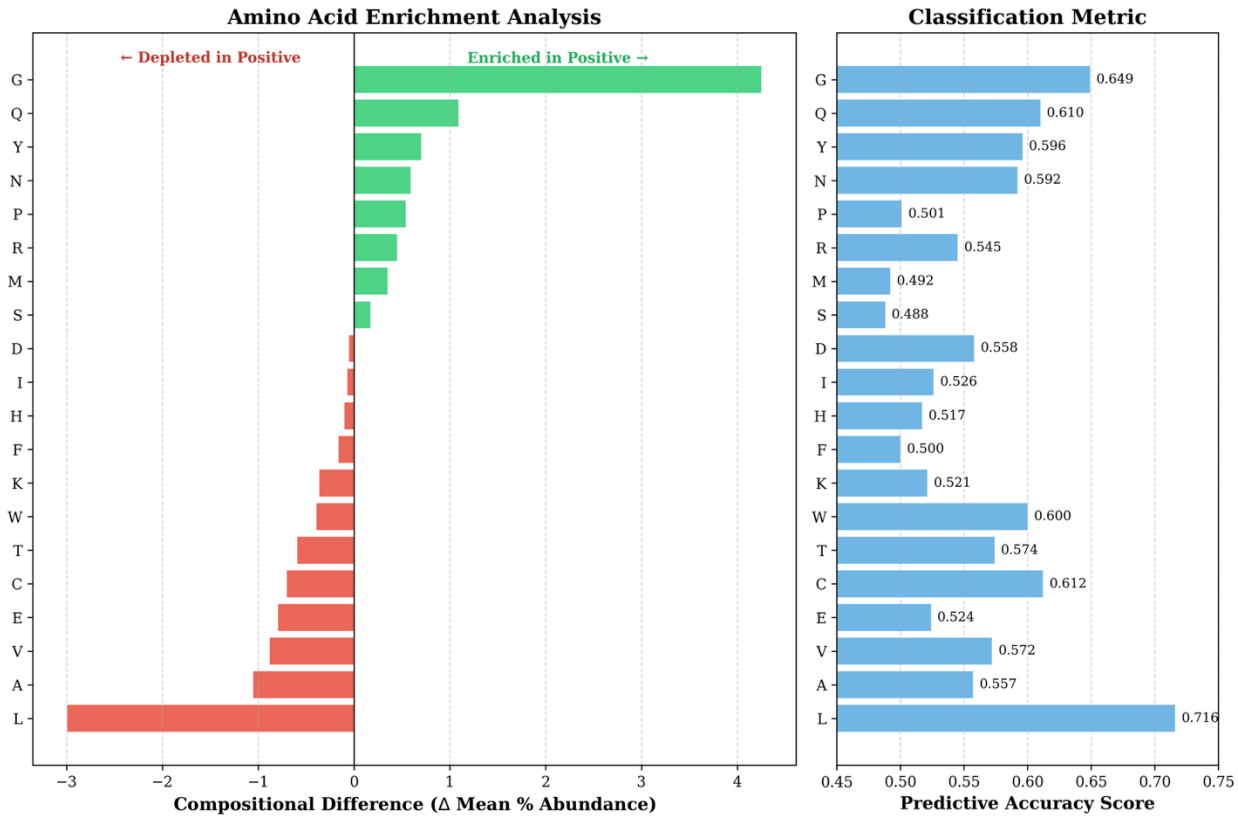


Figure 5.4 - *Amino acid compositional bias and predictive performance. (Left) Difference in mean abundance (Δ%) of amino acids between interacting and non-interacting pairs; positive values*

indicate enrichment and negative values indicate depletion. (Right) Predictive accuracy for individual residues, ranked by enrichment magnitude.

5.3.4 Performance of Machine Learning Models

Machine learning models were trained using CatBoost, Random Forest, Extra Trees, and XGBoost across multiple features that were generated. Among these classifiers, CatBoost consistently achieved the highest performance, indicating its suitability for this prediction task. Among traditional sequence-derived descriptors, the CatBoost model trained on the DACC–SPC feature combination achieved a validation AUC of 0.988 and MCC of 0.911, while the DACC–APAAC combination produced equivalent performance. Incorporation of LLM-derived embeddings—specifically DNABERT-2 for lncRNA and ESM-2-t30 for protein sequences—yielded a marginal but consistent improvement, with the corresponding CatBoost model achieving the highest recorded validation AUC of 0.989 and MCC of 0.915. The complete performance summary for top-performing models is reported in Table 5.2.

Table 5.2 - *Model performance metrics for individual lncRNA-protein feature sets on training and validation data using CatBoost models*

lncRNA features	Protein features	Training				Validation			
		Accuracy	F1_Score	MCC	AUC	Accuracy	F1_Score	MCC	AUC
DACC	SPC	0.954	0.956	0.910	0.988	0.955	0.955	0.911	0.988
DACC	APAAC	0.954	0.956	0.910	0.988	0.954	0.954	0.910	0.988
DACC	PSSM	0.953	0.955	0.907	0.987	0.953	0.954	0.908	0.988
ALL_COMP	PSSM	0.953	0.954	0.907	0.987	0.953	0.953	0.907	0.987
One hot encoded vectors		-	-	-	-	0.953	0.953	0.908	0.987
DNABERT2	ESM2.t30	0.929	0.958	0.914	0.989	0.932	0.957	0.915	0.989

[#] DACC - Dinucleotide based auto cross correlation; SPC - Shannon Entropy of Physicochemical Property; APAAC - Amphiphilic Pseudo Amino-acid Composition; PSSM - Position-Specific Scoring Matrix; ESM-2 t30 - esm2_t30_150M_UR50D

5.3.5 Deep Learning Model Performance

The two-tower CNN, trained end-to-end on one-hot encoded sequence representations in the absence of any handcrafted feature input, demonstrated strong and consistent predictive performance on the independent validation set. The model achieved an MCC of 0.908 and AUC of 0.987, values that are closely comparable to those obtained by the best-performing CatBoost configurations trained on traditional features. The CatBoost model trained on LLM-derived embeddings delivered a nominal performance advantage but CNN architecture's strength is that it eliminates the need for any feature engineering or external tool dependencies. The convolutional filters learn discriminative sequence motifs and interaction-relevant patterns directly from the training data. They providing a fully automated and extensible modelling approach that does not require modification when applied to novel sequence types.

5.3.6 Comparison with Existing Tools

We surveyed 32 previously published LPI prediction tools to position our approach within the existing literature. Most of these tools were no longer functional, had broken or discontinued web servers, or did not support user-provided sequence input. Of all tools reviewed, only LPI-BLS, developed by Fan et al., offered a working standalone implementation that could be used for direct comparison.

We evaluated LPI-BLS on our validation dataset, where it performed poorly, achieving an accuracy of 0.513, MCC of −0.003, and AUC of 0.512. These values are statistically equivalent to random guessing. Closer inspection of the predictions showed that the tool assigned a non-interacting label to nearly every input pair, as evidenced by an extremely low sensitivity of 0.012. This failure likely stems from the tool being trained on an older and much smaller dataset from NPInter v2.0, which has not generalized well to the broader and more diverse set of interactions captured in NPInter v5.0. Figure 5.5 compares the performance of all evaluated models against LPI-BLS.

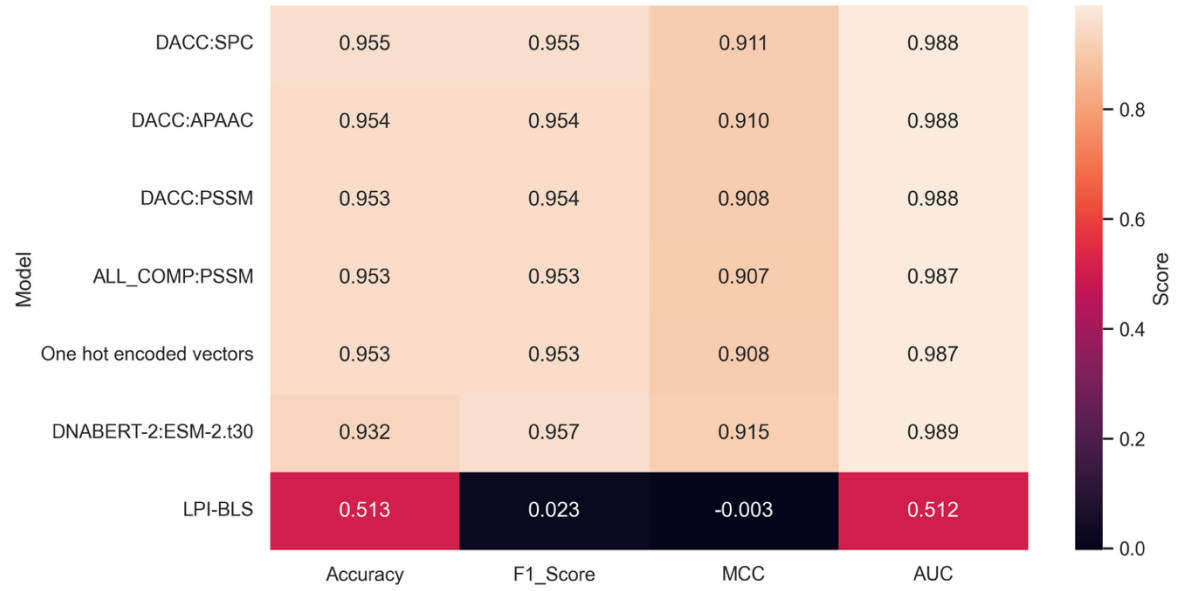


Figure 5.5 - *Comparative analysis of validation metrics across seven model architectures, demonstrating the performance superiority of DACC and Transformer-based models*

5.3.7 High-Throughput Application and Speed Optimization

The best-performing models in this study, those using large language model embeddings or features from the Pfeature and Nfeature pipelines, are computationally expensive at proteome or transcriptome scale due to the time and hardware demands of embedding inference and feature extraction. To address this, we developed a lightweight feature representation that maintains strong predictive performance while significantly reducing the computational burden. A CatBoost model trained on a compact descriptor comprising the composition of ten selected amino acids (G, Q, Y, N, R, L, C, W, and V) and the four standard nucleotides (A, T, G, and C) achieved a validation AUC of 0.980. Under this streamlined approach, prediction of interactions between a single lncRNA and the entire human proteome can be completed in approximately 30 seconds, rendering the tool suitable for routine high-throughput screening applications without meaningful loss in predictive accuracy.

5.4 Discussion and Conclusion

The current study's findings show that accurate lncRNA–protein interaction prediction requires modeling techniques that can capture intricate, non-homologous sequence characteristics that traditional alignment-based approaches are unable to capture. Due to the low sequence similarity between training and validation sets after redundancy removal, the alignment-based baselines were unable to produce predictions for most validation pairs. Even in the few instances where alignments were obtained, accuracy was roughly 15% lower than that of the machine learning models. The biological expectation that lncRNA–protein binding specificity results from structural compatibility, evolutionarily refined interface properties, and physicochemical complementarity rather than from the direct sequence homology that characterizes, for instance, conserved protein domains is consistent with this performance failure.

Our sequence-based analysis suggests that predictive performance in this dataset is closely linked to biological structure in the curated interactome. Interaction partners are not evenly distributed: a small set of RNA-binding proteins, including FUS, WDR33, HNRNPA2B1, IGF2BP1/3, UPF1, FBL, SFPQ and NONO, dominate positive labels. These proteins show sequence features typical of RNA binders, namely reduced leucine content, enrichment of glycine, asparagine, tyrosine and glutamine, and frequent RGG-like tripeptide motifs. On the lncRNA side, longer transcripts such as *MALAT1, XIST, GAS5* and SNHG family members are more often found in positive pairs, while simple nucleotide composition contributes much less. Overall, these results indicate that the labelled dataset mainly reflects an RBP-centred binding profile rather than a single universal RNA or protein motif.

This interpretation has direct implications for how predictions should be used. High-confidence pairs are most likely to involve Gly/Arg-rich RNA-binding proteins and longer nuclear lncRNAs linked to splicing, miRNA processing or related post-transcriptional pathways, as supported by Gene Ontology enrichment of hub proteins. Such predictions can therefore generate useful experimental hypotheses, for example prioritising RIP or CLIP validation for under-studied lncRNAs assigned to known RBPs, or exploring candidate partners of non-hub proteins that fall outside the dominant CLIP-enriched programmes. At the same time, because positive labels come from curated and CLIP-rich sources and negatives are generated by partner shuffling, these patterns should be read as features that distinguish membership in the annotated interactome, not as proof of a specific binding site or regulatory outcome. Even with this limitation, linking

performance to key proteins, lncRNAs, sequence motifs and pathways strengthens the biological value of the work beyond accuracy alone.

A clear hierarchy in the predictive value of different feature representations was observed across the evaluated models. Basic compositional descriptors—particularly amino acid frequency profiles capturing leucine depletion and glycine enrichment—provided a useful and interpretable first-order discriminative signal, with individual amino acid classifiers reaching accuracies of up to 71.6%. However, the highest performance was consistently achieved by models trained on LLM-derived embeddings from DNABERT-2 and ESM-2, which encode contextual and evolutionary sequence information through transformer-based pre-training on large-scale genomic and proteomic corpora. The improvement of the embedding-based CatBoost model (AUC 0.989, MCC 0.915) over the best traditional feature-based model (AUC 0.988, MCC 0.911) reflects the additional representational capacity of contextual embeddings relative to fixed-length handcrafted descriptors, particularly in capturing long-range dependencies and structural preferences within sequences.

The deep learning CNN architecture merits particular consideration because it achieves near-equivalent performance to the LLM-based models—MCC 0.908, AUC 0.987—without any feature extraction pipeline, relying instead on convolutional filters trained directly on one-hot encoded sequence inputs. This result demonstrates that the interaction signal is sufficiently structured within the raw nucleotide and amino acid sequence data that an end-to-end convolutional model can learn to detect it without external biological annotations. The CNN thus provides an important and practically valuable alternative for deployment scenarios in which LLM inference infrastructure is unavailable, or when generalization to sequence types outside the pre-training distribution of the LLMs is required.

A clear performance gap between what the field currently delivers and what can be achieved with updated methodologies was discovered when comparing our models to existing tools. Only LPI-BLS remained operational out of the 32 tools assessed, and its near-random performance on our validation set (AUC 0.512) is probably due to its dependence on an old training dataset. This highlights a larger issue in computational biology: tools developed using outdated database versions lose their usefulness as databases grow, particularly in the absence of a system to retrain

or reassess them against updated standards. These issues led us to utilize NPInter v5.0 for training our models as it had undergone a massive update from the previous versions.

More than 524,000 balanced lncRNA–protein interaction pairs made up the training dataset, which allowed the models to pick up subtle sequence-derived patterns not found in smaller datasets frequently employed in earlier research. Comparable performance across cross-validation and independent test evaluations demonstrates that the models did not overfit the training data while maintaining predictive stability on unseen samples. This consistency across assessment methods demonstrates the method's dependability.

Despite the strong predictive performance of our models, we must acknowledge a fundamental limitation regarding the construction of our negative dataset. In the absence of a large-scale, experimentally validated repository of non-interacting lncRNA-protein pairs, we generated negative samples by randomly shuffling the protein partners within our curated positive set. While this approach successfully created a balanced dataset and ensured that the resulting synthetic pairs were not present in the original positive dataset, it inherently assumes that these random pairings do not interact in vivo. Consequently, there is a distinct possibility that some of these computationally generated "non-interactions" may actually represent undiscovered, true biological interactions. The presence of such latent false negatives in the training data could inadvertently introduce noise, potentially biasing the model's learning process or slightly overestimating real-world performance metrics. Future iterations of LPI prediction tools, including lncrnaPI, would greatly benefit from the integration of experimentally verified true negative datasets as they become available in the literature, which would further refine model accuracy and biological reliability.

Performance could be further improved through several future directions. Fine-tuning large language models on lncRNA and RNA-binding protein sequences may yield embeddings that better capture interaction-specific signals. Although AI-based protein structure prediction methods such as AlphaFold provide valuable structural representations, this study intentionally relied on sequence-derived features to develop a robust and broadly applicable framework independent of predicted structures. Furthermore, while protein structure prediction has advanced considerably, accurate tertiary structure prediction for lncRNAs remains challenging, limiting the reliable incorporation of RNA structural features. As RNA structure prediction methods mature, integrating

high-confidence protein and RNA structural information may further enhance predictive performance. The model could also benefit from the inclusion of interaction data from additional species, while the computationally efficient composition-based framework remains well suited for rapid large-scale screening.

All things considered, this chapter presents a scalable and systematically verified methodology for predicting lncRNA-protein interactions. The method outperforms existing methods in terms of predictive performance by merging large-scale curated data, gradient-boosting classifiers, and language model-derived embeddings. This framework is appropriate for both focused hypothesis-driven research and genome-wide interaction analysis.

6

lncRNA-based diagnostic biomarkers for PDAC

6.1 Introduction

One of the deadliest types of cancer is pancreatic ductal adenocarcinoma (PDAC). Patients with PDAC have a five-year survival rate of less than 10%, and these numbers have not improved recently [231]. PDAC is a biologically aggressive tumor that can disseminate to distant organs early in the disease, creates an immunosuppressive stroma, and is inherently resistant to chemotherapy and targeted agents [232]. Once PDAC has spread outside of the pancreas, these characteristics make treatment very challenging. The absence of reliable biomarkers for early detection is a significant contributing factor to this poor prognosis. Since PDAC doesn't show any symptoms in its early stages, most patients don't receive a diagnosis until the disease has spread locally or metastasized, at which point surgical removal is no longer an option.

The diagnostic tools available today are insufficient for population-level early detection. The most widely used clinical biomarker is the serum glycoprotein CA19-9 [233]. The most popular biomarker in clinical practice, serum glycoprotein CA19-9, has poor sensitivity and specificity, is elevated in a number of benign inflammatory conditions, such as chronic pancreatitis, and is completely absent in Lewis-negative people, who make up 10–15% of the population [134]. Cross-sectional imaging modalities—computed tomography and magnetic resonance imaging—are indispensable for disease staging but are unsuitable for large-scale screening owing to cost, limited accessibility, and reduced sensitivity for sub-centimetre lesions [234]. These diagnostic limitations have generated an interest in minimally invasive liquid biopsy approaches capable of detecting disease-specific molecular signatures in blood or other accessible body fluids.

Long non-coding RNAs (lncRNAs) have shown great potential as candidate liquid biopsy biomarkers for PDAC, supported by their tissue- and condition-specific expression patterns, their documented involvement in tumour initiation, progression, and microenvironmental remodelling, and the biological stability of RNA enclosed within extracellular vesicles (EVs) in the circulation. The feasibility of a transcriptomics-based diagnostic approach was demonstrated by Yu et al. (2020), who applied a support vector machine classifier to plasma EV-derived RNA-sequencing data from GSE133684 and identified an eight-gene biomarker panel achieving a validation AUC of 0.939 [138]. This result established a proof of concept for EV transcriptomic profiling as a non-invasive diagnostic modality, while simultaneously providing a publicly available and well-

characterised benchmark dataset against which subsequent methodological advances could be evaluated.

In this study, we address these challenges by developing a robust machine learning framework for PDAC prediction using plasma EV-derived RNA-seq data. We integrate multiple complementary feature selection strategies to identify compact gene panels and systematically evaluate their predictive performance across a diverse set of classifiers. Our results demonstrate that small, information-rich gene signatures can achieve high diagnostic accuracy and generalizability, highlighting the potential of EV transcriptomics as a non-invasive platform for improving early PDAC detection.

6.2 Materials and Methods

Dataset collection and standardisation, multi-strategy feature selection, machine learning model building, and performance evaluation were the four consecutive steps that made up the analytical framework. Figure 6.1 shows a schematic representation of the entire pipeline.

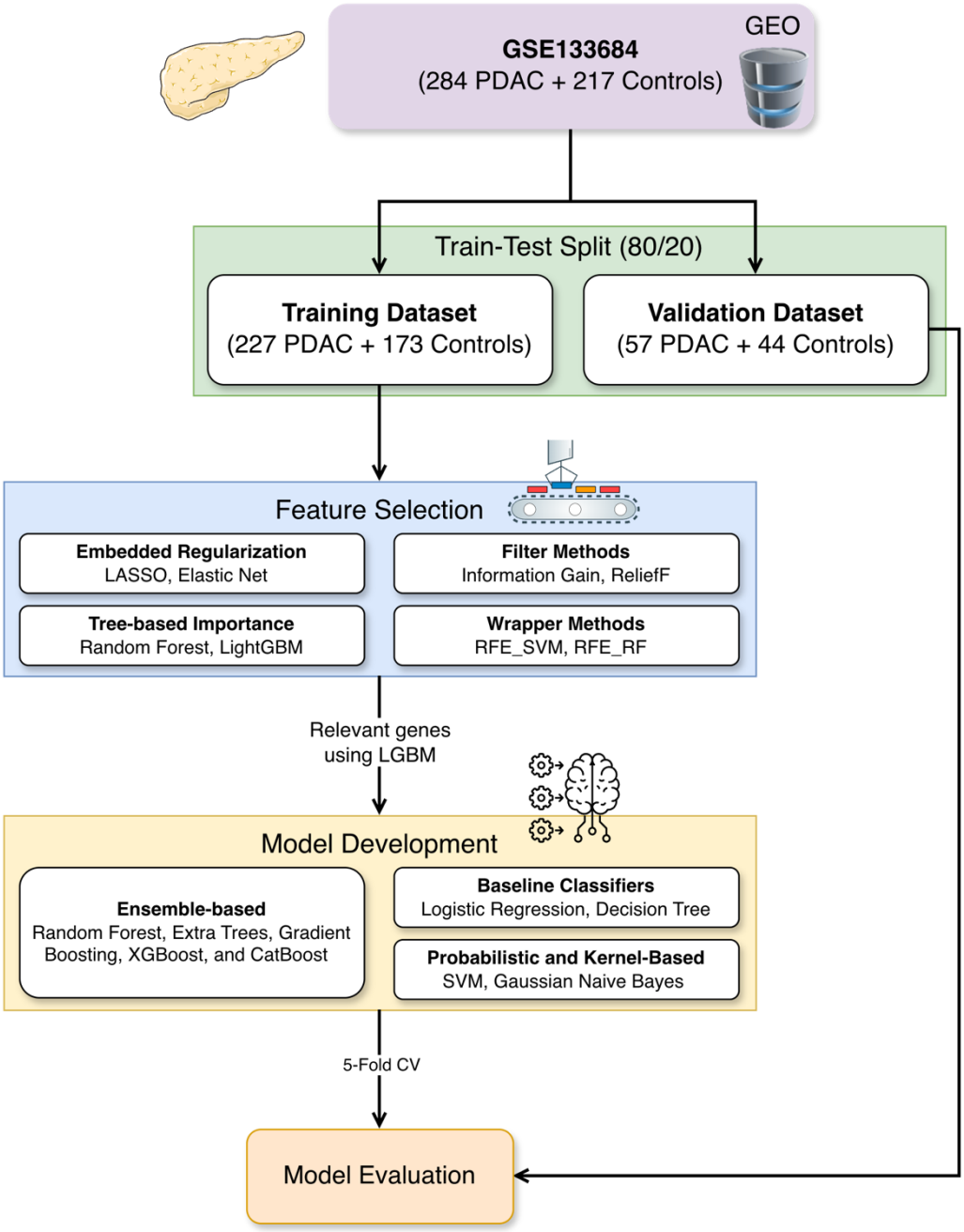


Figure 6.1 - *An overview of the methodology used in PDACpred*

6.2.1 Dataset Curation

For this study, a publicly available plasma extracellular vesicle (EV) RNA-sequencing dataset (accession number GSE133684) was obtained from the Gene Expression Omnibus (GEO). 117

healthy controls (HC), 100 patients with chronic pancreatitis (CP), and 284 patients with pancreatic ductal adenocarcinoma (PDAC) comprised the cohort's 501 samples. For 12,943 lncRNA transcripts, Transcripts Per Million (TPM) values for gene expression were provided. To ensure comparability with previous analyses, the TPM values were used exactly as submitted, without any additional normalization.

PDAC samples (n = 284) were classified as the positive class for binary classification, whereas CP and HC samples were combined to create the negative class (n = 217). This classification was created to differentiate cancer from both non-cancerous conditions because chronic pancreatic inflammation can present with overlapping symptoms and elevated CA19-9 values. The entire dataset was then separated using stratified 80:20 splitting into a training set (n = 400) and an independent held-out validation set (n = 101). Z-score normalization was then used to standardize the gene expression data. Scaling parameters were only obtained from the training set and then applied to the validation set in order to avoid data leaking. All feature selection and model training processes were based on these standardized matrices.

6.2.2 Feature Selection Strategy

A multi-strategy feature selection system was built for handling the enormous complexity of the standardized training dataset, which included 54,148 gene transcripts, and to separate compact, high-value biomarker signatures. Eight algorithms were chosen, each of which reflects a different concept for characterizing the link between characteristics and the target class, because no single method has been demonstrated to consistently outperform others across high-dimensional biological datasets. Each algorithm produced gene subsets of nine predetermined sizes: 5, 8, 10, 15, 20, 25, 30, 40, and 50 genes.

Since sparsity-based methods incorporate feature selection into the model training procedure, they were employed first. By using L1 regularization, LASSO eliminated non-informative gene coefficients from the model by reducing them to zero. Elastic Net, which combines L1 and L2 penalty terms, was also used because it can handle correlated gene groups, which is a common trend in transcriptomic data.

Non-linear feature interactions were then taken into consideration using ensemble-based ranking techniques. While LightGBM provided a different gradient-boosted ranking utilizing both gain and split-based criteria, Random Forest importance scores were calculated based on mean reductions in node impurity.

The third category of feature selection involved statistical and instance-based filtering approaches. The dependency between gene expression and disease status was measured using Mutual Information, which captured both linear and non-linear connections without depending on a particular classifier. In parallel, the ReliefF method (implemented in the skrebate package) evaluated genes depending on how well they distinguished samples from different groups within their local neighborhood in feature space. In the last step, model performance-driven wrapper-based selection was used. Two classifiers were used for Recursive Feature Elimination (RFE): a linear Support Vector Machine and Random Forest. This method produced increasingly fine-tuned feature subsets by iteratively eliminating genes according to their contribution to predicted performance.

A varied collection of candidate biomarker panels reflecting statistical significance, local discriminative capacity, ensemble behaviour, sparsity-based selection, and classifier-guided optimisation was produced by combining these eight feature selection techniques. All generated gene subsets, regardless of methodology or panel size, were meticulously recorded in a structured JSON format to enable consistent benchmarking and comparative analysis.

6.2.3 Machine Learning Model Development

Each of the 72 gene panels was assessed using nine different classification algorithms in order to look at the ways in which various feature selection techniques interact with model architecture. The ensemble-based techniques Random Forest, Extra Trees, Gradient Boosting, XGBoost, and CatBoost were chosen because they can effectively handle high-dimensional biological data and capture non-linear feature interactions [235–238]. These methods work especially well with transcriptomic datasets, which frequently contain noise and complex dependencies. Logistic regression was included as a linear reference model for comparison, offering a clear standard by which non-linear approaches could be evaluated. In order to model non-linear decision boundaries

in high-dimensional space, a Support Vector Machine with a radial basis function (RBF) kernel was integrated [239]. A standalone learner's performance was also assessed using a single Decision Tree, and a probabilistic classification framework based on conditional independence assumptions was represented by Gaussian Naive Bayes.

Python's scikit-learn, XGBoost, and CatBoost libraries were used to implement each model. To guarantee reproducibility and prevent performance inflation brought on by extensive parameter optimization, the default hyperparameters were kept and a fixed random seed (42) was applied.

6.2.4 Model Training

Each classifier was assessed on each selected gene panel in order to identify the best model and biomarker set. The evaluation was carried out using a two-step, structured process to evaluate training stability and true generalization performance. Initially, a five-fold stratified cross-validation was performed for the training dataset (n = 401). Eighty percent of the samples were used to train the model in each fold, and the remaining twenty percent were used for evaluation. All five fold's performance was averaged to produce a reliable estimate. This approach ensured that the results would show consistent behavior across multiple partitions and be independent of a single data split. After cross-validation, each model was retrained using the complete standardized training dataset (n = 401). After that, the trained model was assessed using a separate held-out validation set (n = 100) that had not been used during the feature selection or model training phases. Performance on this validation set demonstrated how well the model generalized to data that had never been seen before.

6.2.5 Performance evaluation

For both training and validation, model performance was comprehensively measured using a standard suite of binary classification metrics (Section 4.2.7): Accuracy, Precision, Sensitivity (Recall), Specificity, F1-Score, Matthews Correlation Coefficient (MCC), and the Area Under the Receiver Operating Characteristic Curve (AUC). The final training and validation performance

results for every combination were then compared to identify the most robust and high-performing model.

6.3 Results

The present study aimed to whether the systematic application of multi-strategy feature selection and diverse classifier benchmarking could yield diagnostic models for PDAC that improve upon previously reported performance on the GSE133684 dataset.

6.3.1 Transcriptomic Heterogeneity and Feature Variance Analysis

Before using supervised models, we conducted a preliminary unsupervised analysis to investigate the transcriptome makeup of the dataset and determine whether any disease-related trends were apparent. A heatmap (Figure 6.2) displaying distinct blocks of co-expressed genes was created by hierarchical clustering of the 50 most variable genes using Ward's linkage and Euclidean distance; however, the sample-level dendrogram showed no discernible difference between PDAC and non-PDAC samples. Throughout the dendrogram, data from both positive and negative classes were mixed, indicating that biological noise, inter-individual variations, and other physiological factors unrelated to cancer are the primary causes of transcriptome variance rather than disease state. This conclusion was corroborated by principal component analysis of the raw and z-score-standardized expression matrices, which revealed significant overlap between the PDAC and non-PDAC samples along the first two principal components. These findings demonstrate that the extremely variable gene expression used in unsupervised approaches is insufficient to consistently distinguish between diagnostic classifications in this dataset. Supervised machine learning techniques are required because of the intricacy and interconnectedness of expression profiles across classes.

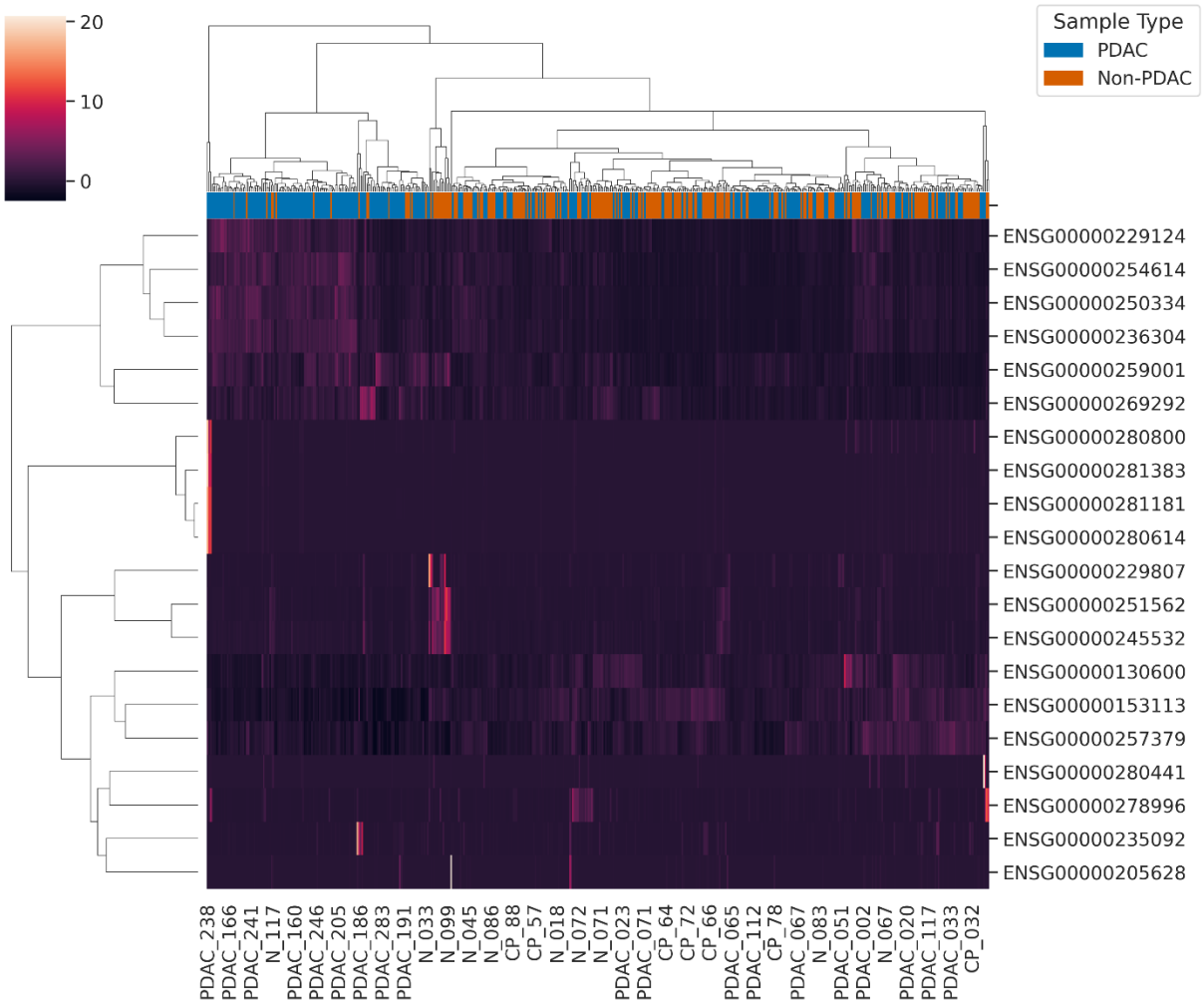


Figure 6.2 - *Hierarchical clustering heatmap of the 20 most variable genes, with Ward's linkage and Euclidean distance*

6.3.2 Effect of Feature Set Size on Model Performance

By evaluating the CatBoost classifier on LGBM-ranked gene subsets ranging from 5 to 50 genes, we investigated the impact of feature panel size on prediction performance. As the number of genes increased from 5 to 20, performance improved gradually. At 20 features, with an MCC of 0.779 and an AUC of 0.954, validation performance peaked. This pattern implies that a core collection of genes carrying complementary, non-redundant information is successfully identified by the LGBM ranking in order to differentiate PDAC from non-PDAC samples. Figure 6.3 displays performance across all panel sizes.

Consistent improvements were not obtained when genes were added above the 20-feature limit. At 25 and 40 features, MCC fell to 0.718, and at 50 features, it only partially recovered (MCC 0.778). This decrease is in line with the addition of less significant genes, which increase input dimensionality without enhancing generalization by adding transcriptome noise and redundancy instead of helpful diagnostic signal. These findings demonstrated that the 20-gene LGBM-selected panel was the best setup, minimizing the chance of overfitting from excessively high-dimensional inputs while maintaining a balance between diagnostic accuracy and a small feature set.

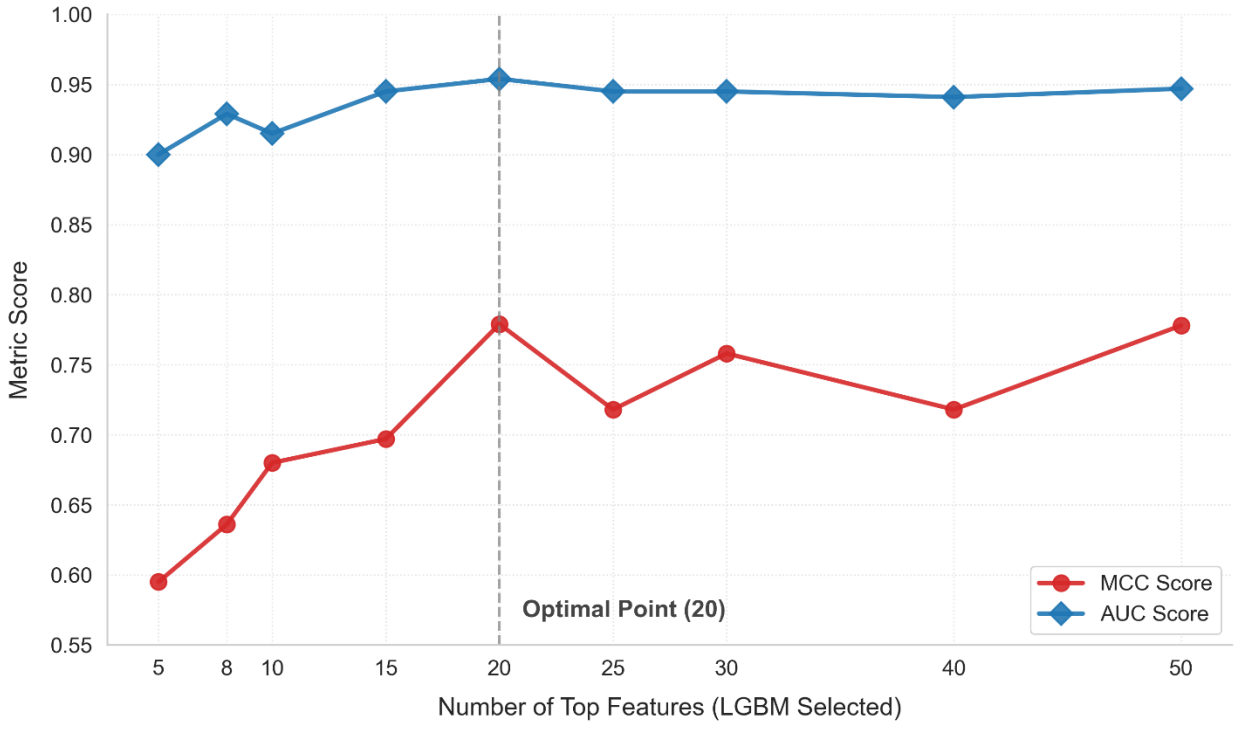


Figure 6.3 - *Performance of the CatBoost classifier across incremental feature subsets*

6.3.3 Biological Significance of Selected Genes

We investigated the biological significance of the 20-gene signature chosen by LGBM and discovered a group of transcripts with known functions in PDAC biology, confirming the biological plausibility of the model. In line with its established function in promoting the epithelial-mesenchymal transition and metastasis in pancreatic cancer, *MALAT1* was one of the most highly

scored characteristics [240]. A mechanism implicated in PDAC invasion, MET receptor signaling, is regulated by the lncRNA *MACC1-OT1* [122]. Immune evasion mechanisms in the PDAC tumor microenvironment are associated with *PCED1B-AS1* and *LRRC32-AS1* [241]. Among the genes that code for proteins, *CLDN1*—a tight-junction protein whose overexpression in pancreatic ductal cells has been linked to enhanced tumor aggressiveness and metabolic alterations—and CD44—a hallmark of cancer stem cells and tumor-initiating populations—were given high significance ratings [242,243]. The DNA damage response and transcriptional alterations typical of genomically unstable PDAC are captured by *ATM* and *ETV4*, whereas the read-through transcript *ARPC4-TTLL3* represents cytoskeletal remodeling linked to tumor cell motility [244–246].

It's interesting to note that a number of unannotated lncRNAs with no known biological function were consistently placed among the top predictors and significantly improved model performance across cross-validation folds. These transcripts are excellent candidates for further functional research because of their frequent selection in spite of the fact that they lack recognized functional roles. This shows that these transcripts contain genuine disease-associated signals in the EV transcriptome that have not yet been experimentally identified.

6.3.4 Comparison of Machine Learning Classifiers

With the highest validation MCC of 0.779 and AUC of 0.954 on the 20-gene LGBM panel, CatBoost demonstrated the most consistently strong performance on lncRNA-based feature subsets across all classifier and panel combinations. Gradient Boosting (MCC 0.758) and XGBoost (MCC 0.760) came in second and third, respectively, among various gradient-boosting techniques, indicating that boosting-based ensemble approaches typically do well on this classification problem. It was evident that non-linear ensemble methods performed better than simpler models and linear classifiers. The diagnostic signal in lncRNA expression profiles rely on complex, non-additive feature interactions that linear decision boundaries are unable to adequately represent, as Support Vector Machines only managed an MCC of 0.578 while ExtraTrees reached 0.678. A comparative overview of classifier performance on the ideal 20-gene panel is shown in Figure 6.4.

CatBoost performance across all nine panel sizes is reported in Table 6.1. The results show a consistent pattern of improving diagnostic performance as panel size increases from 5 to 20 genes,

where all five key metrics—sensitivity, specificity, accuracy, MCC, and F1-score—simultaneously reached their peak values. Beyond 20 genes, performance fluctuated and MCC declined overall, consistent with the inclusion of lower-importance transcripts that reduce the signal-to-noise ratio of the feature set.

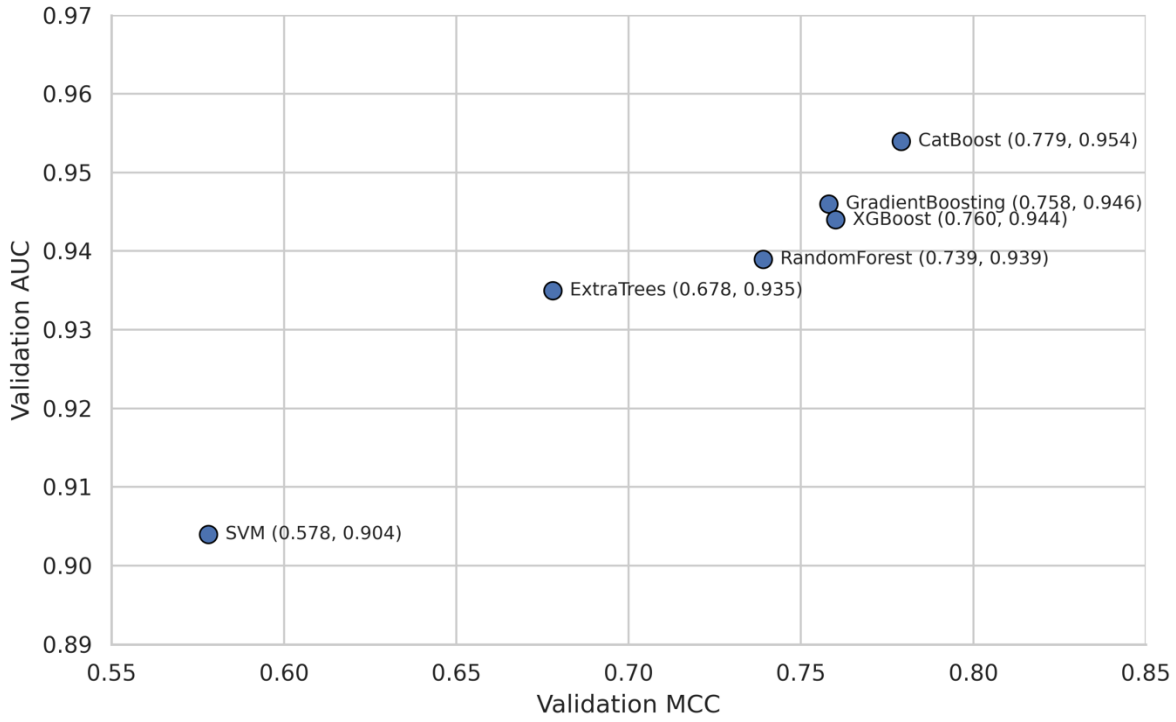


Figure 6.4 - *Scatterplot indicating the performance of different machine learning algorithms for a 20-gene biomarker panel selected by LGBM*

Table 6.1 - *Validation performance of the CatBoost classifier across LGBM-selected feature subsets of increasing size.*

No. of Features	Sensitivity	Specificity	Accuracy	MCC	F1 Score	AUC
5	0.877	0.705	0.802	0.595	0.833	0.900
8	0.877	0.750	0.822	0.636	0.847	0.929
10	0.930	0.727	0.842	0.680	0.869	0.915
15	0.877	0.818	0.851	0.697	0.870	0.945
20	0.895	0.886	0.891	0.779	0.903	0.954

25	0.877	0.841	0.861	0.718	0.877	0.945
30	0.895	0.864	0.881	0.758	0.895	0.945
40	0.912	0.795	0.861	0.718	0.881	0.941
50	0.930	0.841	0.891	0.778	0.906	0.947

6.3.5 Performance of mRNA-based models

In order to establish a baseline performance reference for the lncRNA-based classifiers equivalent models were trained on feature subsets selected from the full mRNA expression matrix rather than from lncRNA-annotated transcripts exclusively. Models trained on mRNA-derived features demonstrated consistently better performance across all panel sizes compared to their lncRNA-features. The top-performing configuration—CatBoost with 20 LGBM-selected mRNA features—achieved a validation accuracy of 0.960, sensitivity of 0.947, specificity of 0.977, MCC of 0.921, and AUC of 0.992. At the reduced panel size of eight genes, the mRNA-based CatBoost model retained strong diagnostic performance, with a validation accuracy of 0.921 and MCC of 0.843. The comparative performance trajectories of lncRNA-based and mRNA-based models across all panel sizes are shown in Figure 6.5.

The performance advantage of mRNA-based models over lncRNA-only models indicates that the mRNA transcriptome encodes additional diagnostic information beyond what is captured by lncRNA-annotated features alone. The lncRNA-only models nonetheless demonstrated clinically meaningful performance levels, with the peak validation AUC of 0.954 substantially exceeding the published baseline on this dataset. These results collectively suggest that the diagnostic ceiling achievable with compact gene panels is higher when the feature space is not restricted to a single RNA biotype, and they provide a quantitative rationale for exploring hybrid lncRNA–mRNA panels in future studies.

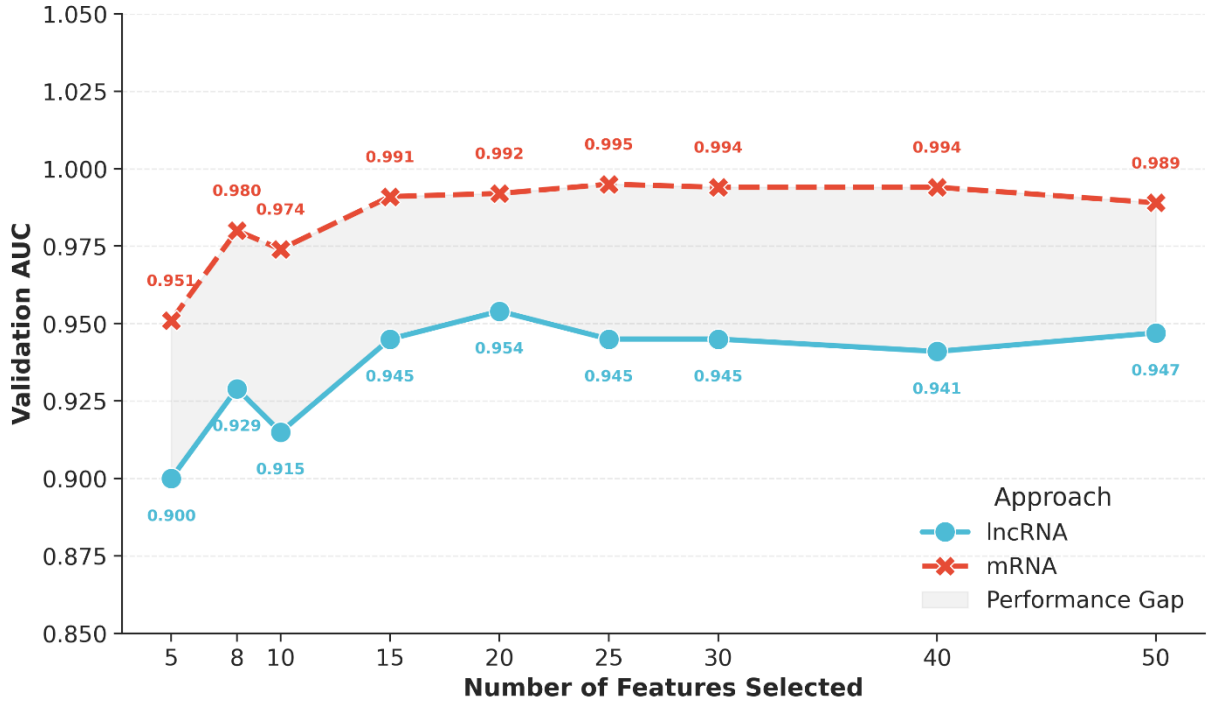


Figure 6.5 - *Comparative analysis of diagnostic performance (Validation AUC) between lncRNA and mRNA-based models across incremental feature subset sizes*

6.3.6 Benchmarking Against Published Classifiers

To contextualise the performance of the present framework within the existing literature, our best-performing models were compared against the classifiers originally reported by Yu et al. (2020) for the GSE133684 dataset. Using an equivalent eight-gene panel, the SVM classifier employed in the original study achieved a validation AUC of 0.0.936. Our CatBoost model trained on the LGBM-top-8 lncRNA panel attained a validation AUC of 0.929 and MCC of 0.636, which is comparable to the performance achieved by Yu et al. However, the mRNA panel achieved significantly higher performance using the 5 and 8 gene biomarker panels, with AUC of 0.951 and 0.980, respectively. However, it was observed that a larger lncRNA biomarker panel (20 gene) was able to outperform the existing method. The comparative results are summarised in Table 6.2.

Table 6.2 – Comparative performance of our study's best models against published classifiers

Model	Biomarker type	No. of genes in biomarker panel	Sensitivity	Specificity	Accuracy	MCC	F1-score	AUC
CatBoost	lncRNA	5	0.877	0.705	0.802	0.595	0.833	0.900
CatBoost	mRNA	5	0.877	0.886	0.881	0.760	0.893	0.951
CatBoost	lncRNA	8	0.877	0.750	0.822	0.636	0.847	0.929
CatBoost	mRNA	8	0.895	0.955	0.921	0.843	0.927	0.980
CatBoost	lncRNA	20	0.895	0.886	0.891	0.779	0.903	0.954
Yu et al. 2020	RNA	8	0.937	0.916	0.927	-	-	0.936
He et al. 2024	RNA	4	0.720	0.890	-	-	-	0.848

6.4 Discussion and Conclusion

Applying a structured feature selection strategy together with systematic evaluation of multiple classifiers led to improved models for detecting pancreatic ductal adenocarcinoma (PDAC) from plasma extracellular vesicle (EV) RNA-seq profiles. These improvements were achieved through analytical refinement rather than expansion of the patient cohort. Validation performance, measured using MCC, increased as the number of selected genes expanded from 5 to 20, indicating that additional genes contributed meaningful and complementary signal. When panel size exceeded 20 genes, performance no longer improved and began to decline, reflecting the impact of increased dimensionality and noise in a dataset with moderate sample size. The 20-gene panel therefore represents a data-supported balance between information gain and model stability.

Inspection of the highest-performing panels revealed biologically relevant markers. *MALAT1* consistently ranked among the strongest predictors, in line with its documented role in PDAC progression. Other recurrent features included lncRNAs such as *MACC1-OT1, PCED1B-AS1,* and *LRRC32-AS1*, along with protein-coding genes including *CLDN1* and *CD44*. The repeated selection of these transcripts suggests that the optimal panels capture signals derived from tumor biology as well as immune and stromal components, consistent with the heterogeneous molecular landscape of PDAC.

Systematic comparison of mRNA-based and lncRNA-only models revealed that the broader mRNA transcriptome encodes diagnostic information not fully captured within the lncRNA subset alone, with mRNA models outperforming lncRNA counterparts by approximately 3–4 AUC points

at the 20-gene level. Nevertheless, lncRNA-only models achieved clinically meaningful performance substantially exceeding the published baseline, affirming that lncRNAs constitute an independent source of diagnostic signal rather than mere surrogates for protein-coding gene activity. Several functionally unannotated lncRNAs were retained with high relevance scores across independent cross-validation cycles, suggesting that purely data-driven approaches can implicate biologically meaningful transcripts not yet characterised by hypothesis-driven inquiry. Of particular clinical relevance, the 20-gene lncRNA model achieved a specificity of 0.886 against a negative class that explicitly included chronic pancreatitis patients. The major issue in using CA19-9 levels for PDAC diagnostics is that it is generally elevated in the case of any chronic inflammatory processes.

Several methodological limitations in this study still need to be addressed. The GSE133684 cohort, while among the largest publicly available EV RNA-sequencing datasets for PDAC at 501 samples, remains modest relative to the scales conventionally required for clinical biomarker validation, and the held-out validation set of 101 samples cannot substitute for prospective, multi-centre external validation. Classifier benchmarking relied on default hyperparameter configurations to ensure equitable inter-algorithm comparison, leaving open the possibility that targeted optimisation would yield further performance gains. Feature selection and model evaluation were performed on a single train–validation split, which limits assessment of panel stability. Incorporating bootstrap-based resampling or repeated cross-validation would provide stronger evidence for the reproducibility of the identified gene signatures. In addition, the absence of validation in an independent external cohort restricts conclusions regarding generalizability beyond the analyzed dataset.

This chapter presents a rigorously evaluated computational framework for developing compact lncRNA- and mRNA-based biomarker panels for non-invasive PDAC detection using plasma EV RNA-sequencing data. By integrating eight complementary feature selection strategies and benchmarking nine classifiers across 72 candidate panels, the analysis demonstrates that smaller, information-rich signatures outperform larger panels on independent validation. Gradient-boosting methods, in particular, extracted greater diagnostic signal than approaches previously applied to the same dataset. The final 20-gene lncRNA panel (MCC 0.779, AUC 0.954) and eight-gene mRNA panel (MCC 0.843, AUC 0.980) represent the primary outcomes of this work and provide

a framework that can be extended to EV transcriptomic datasets from other cancer types and future prospective validation studies.

7

Expression-based cancer biomarkers using LLM

7.1 Introduction

Pancreatic ductal adenocarcinoma (PDAC) accounts for more than 90% of all pancreatic cancer cases [247]. It is characterized by a dense desmoplastic stroma, pronounced inflammatory infiltration, early metastatic spread, and marked resistance to both cytotoxic chemotherapy and targeted therapies [248]. The five-year survival rate remains below 8%, making PDAC one of the leading causes of cancer-related mortality worldwide [249]. This poor outcome is largely attributable to the absence of specific symptoms at early stages, resulting in most patients being diagnosed when the disease is already advanced and surgical resection is no longer feasible [249]. Epidemiological trends further indicate that PDAC incidence is expected to rise substantially, driven by population ageing, increasing obesity rates, and the growing prevalence of type 2 diabetes, all established risk factors.

Clinical diagnosis primarily depends on cross-sectional imaging modalities, including computed tomography (CT), magnetic resonance imaging (MRI), and endoscopic ultrasound (EUS) [250]. These approaches are typically complemented by measurement of the serum biomarker carbohydrate antigen 19-9 (CA19-9) [251]. Despite their clinical utility, these approaches exhibit well-documented limitations: imaging modalities lack the sensitivity required to detect sub-centimetre lesions and are subject to operator variability and high cost, while CA19-9 demonstrates suboptimal sensitivity and specificity and is absent entirely in Lewis-negative individuals [252]. As a consequence, only 15–20% of patients present with surgically resectable disease at diagnosis. These limitations have motivated sustained research into minimally invasive liquid biopsy approaches, with particular emphasis on transcriptomic markers and extracellular vesicle-associated RNA as sources of disease-specific molecular information.

Computational approaches to transcriptomic biomarker identification have predominantly employed traditional machine learning algorithms, including random forests, LASSO regression, and support vector machines. A representative example is the study by Yu et al. (2020), who applied an SVM classifier trained on TPM-normalised gene expression values to identify an eight-gene biomarker signature achieving an AUC of 0.939 on an external validation cohort [138]. While such methods have demonstrated the feasibility of EV transcriptomics-based PDAC detection, they treat gene expression data as static numerical feature vectors and may be insufficient for

capturing the complex, higher-order interaction patterns distributed across thousands of transcripts in a high-dimensional expression matrix.

Large language models (LLMs) represent a computationally powerful class of deep learning architectures that have demonstrated transformative capacity for pattern recognition in sequential data, primarily in natural language processing through masked language modelling on large unlabelled corpora. Their application to biological sequence analysis—through models such as ProtBERT, PeptideBERT, and the ESM-2 family—has produced pre-trained representations of substantial utility for protein structure and function prediction. The applicability of these architectures to numerical gene expression data, however, has not previously been explored.

The present chapter describes a novel methodological framework in which continuous gene expression values are systematically converted into discrete amino acid sequences, thereby enabling standard pre-trained protein language models to be fine-tuned on transcriptomic data without modification to their underlying architecture. Alignment-based classification strategies employing PSI-BLAST and MERCI motif discovery are evaluated as baselines and ensemble components. To the best of our knowledge, this chapter constitutes the first application of protein language models to the mining of cancer patient transcriptomic profiles.

7.2 Material and Methods

The analysis comprised five sequential stages: dataset preparation, expression-to-probability transformation, pseudo-peptide sequence generation, LLM fine-tuning, and performance evaluation. Alignment-based and ensemble approaches were developed as complementary strategies. An overview of the complete methodology is presented in Figure 7.1.

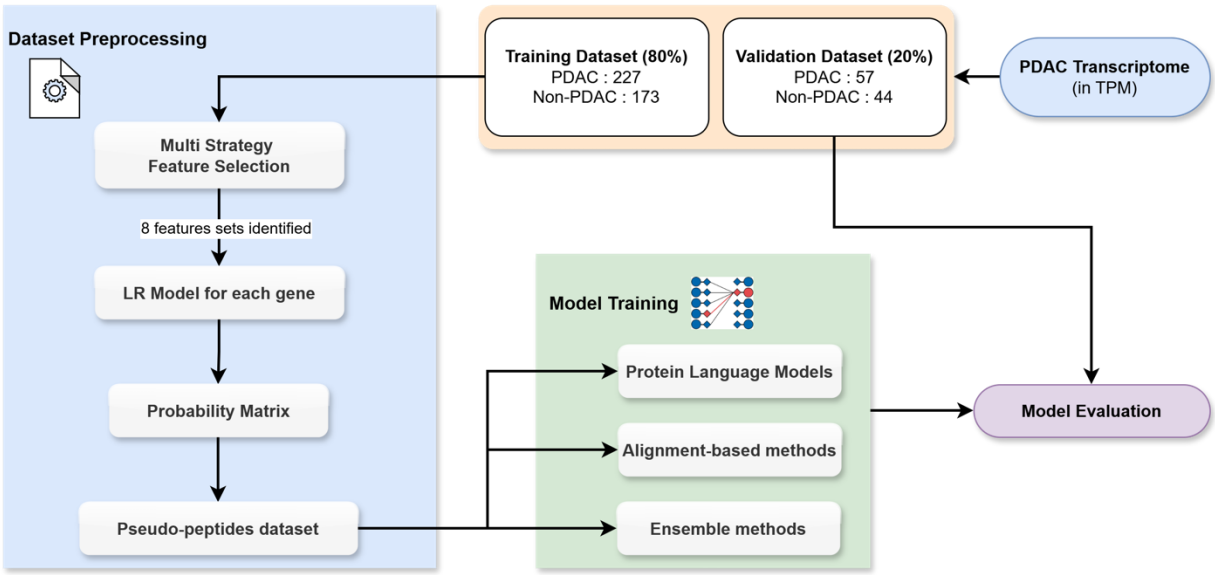


Figure 7.1 **-** *An overview of the methodology followed in this study*

7.2.1 Dataset Description

The transcriptomic dataset employed in this study was obtained from the Gene Expression Omnibus (GEO) repository under accession number GSE133684, consistent with prior work in this thesis and with the original study by Yu et al. (2020) [138]. Expression data were provided in pre-processed format as Transcripts Per Million (TPM) values across 12,943 lncRNA genes for a cohort of 501 patient samples, comprising 284 patients with PDAC, 100 patients with chronic pancreatitis (CP), and 117 healthy controls (HC). For the binary classification task, PDAC samples constituted the positive class (n = 284) and the pooled CP and HC cohorts formed the negative class (n = 217). The full dataset was partitioned 80:20 into a training set (n = 401) and an independent held-out validation set (n = 100) using stratified sampling to preserve class proportions across both subsets.

7.2.2 Feature Selection Strategies

Dimensionality reduction of the initial 12,943-feature matrix was performed using eight complementary feature selection algorithms, applied after z-score standardisation of all features to account for the wide variation in expression magnitude across genes. Gene subsets at nine panel

sizes (5, 8, 10, 15, 20, 25, 30, 40, and 50 characteristics) were produced by applying each algorithm separately. This produced a structured library of candidate gene panels for methodical assessment.

Four methodological categories were used to group the selection algorithms. LASSO (L1 regularisation via LassoCV), which encourages model sparsity by pushing the coefficients of non-informative features to zero, and Elastic Net (ElasticNetCV), which combines L1 and L2 penalties to account for linked gene groups, were examples of sparsity-based embedding techniques. Among the ensemble-based significance techniques were LightGBM, which provided feature rankings based on split frequency and gain, and Random Forest, which generated important scores based on mean drops in node impurity. Among the filter techniques were ReliefF (skrebate), which assessed feature quality using nearest-neighbor comparisons in the feature space, and Mutual Information (mutual_info_classif), which quantified the statistical dependence between each gene and the binary class label. One of the wrapper approaches was Recursive Feature Elimination, which iteratively eliminated the least informative feature at each stage using a Random Forest and a linear SVM as base estimators. The scikit-learn, lightgbm, and skrebate Python libraries were utilized in all implementations, and the resulting gene panels were stored in JSON format.

7.2.3 Conversion of Gene Expression Profiles into Pseudo-Peptide Sequences

To enable the application of pre-trained protein language models to numerical transcriptomic data, a two-stage transformation pipeline was implemented to convert continuous TPM expression values into discrete amino acid sequence strings. The complete transformation process is illustrated in Figure 7.2.

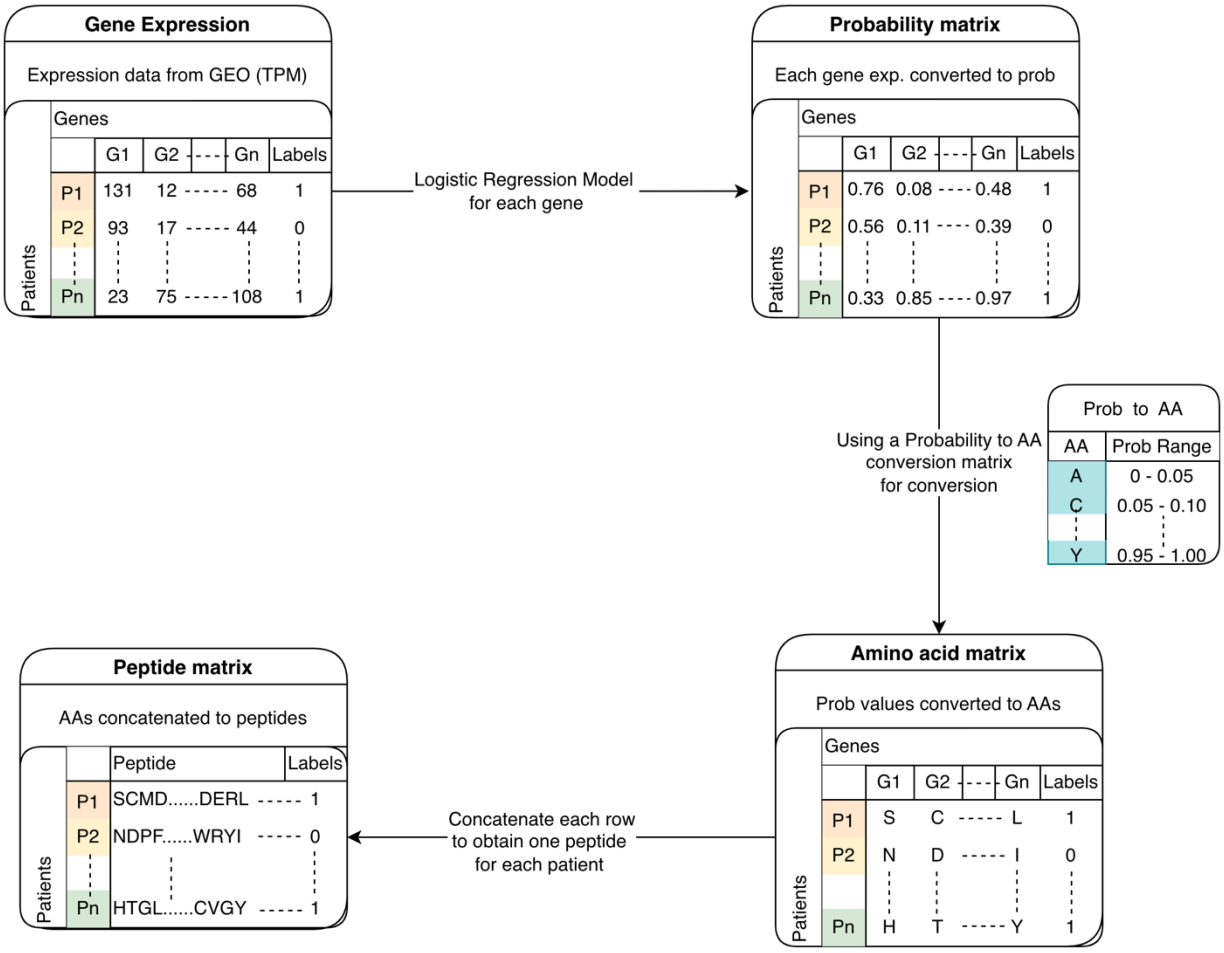


Figure 7.2 *- A graphical representation of the process used for the conversion of gene expression to peptide sequences*

In the first stage, the TPM expression matrix was converted to a probability matrix of identical dimensions (501 samples × 12,943 genes). For each of the 12,943 genes, an individual logistic regression (LR) model was trained on the 401-sample training set to predict the binary PDAC classification outcome from the TPM values of that gene alone. The fitted model was then applied to all 501 samples—training and validation cohorts—to generate a PDAC probability score for each sample–gene combination. This procedure was repeated independently for every gene, producing a probability matrix in which each entry encodes the marginal likelihood of a sample being PDAC-positive based on single-gene expression.

In the second stage, each probability value was mapped to a specific amino acid residue according to a deterministic binning scheme spanning the unit interval in increments of 0.05: probabilities in [0.00, 0.05) were encoded as Alanine (A), [0.05, 0.10) as Cysteine (C), and so forth through the 20 standard amino acids, with [0.95, 1.00] encoded as Tyrosine (Y). For each patient sample, the amino acid characters derived from the probability values of the selected feature genes were concatenated in gene-rank order to form a single pseudo-peptide sequence of length equal to the number of selected features. This procedure was applied to all 501 samples, producing a dataset of sample-specific pseudo-peptide sequences that encapsulate the underlying transcriptomic signature of each patient in a format directly interpretable by protein language models.

7.2.4 Selection and Fine-Tuning of Large Language Models

Five pre-trained protein and peptide language models were selected for evaluation: PeptideBERT, ProtBERT, and three variants of the ESM-2 architecture (t6_8M_UR50D, t12_35M_UR50D, and t33_650M_UR50D), spanning a range of model capacities from 8 million to 650 million parameters [229,253,254]. PeptideBERT was specifically designed for peptide property prediction and is pre-trained on curated peptide sequence datasets; ProtBERT is a BERT-based transformer pre-trained on a large corpus of protein sequences from UniRef databases; and the ESM-2 models are evolutionary-scale protein language models pre-trained on the UR50D dataset using masked language modelling objectives. Each model was fine-tuned on the pseudo-peptide sequences derived from the 401-sample training cohort, optimising the model parameters for the binary PDAC classification task. Predictive performance was subsequently assessed on the 100-sample independent validation set using accuracy, precision, recall, F1-score, MCC, and AUC.

7.2.5 Alignment-Based and Motif-Based Baseline Methods

Two alignment-based classification strategies were implemented as baselines and ensemble components. For PSI-BLAST-based classification, a reference database was constructed from the pseudo-peptide sequences of the 401-sample training cohort [230]. Validation sequences were queried against this database, and the class label of the single highest-scoring training sequence was assigned to each query. The search was conducted across a range of expectation value (E-

value) thresholds spanning from 10^{-7} to 10^{3} to characterise the coverage-accuracy trade-off at each peptide length.

For motif-based classification, the MERCI tool was used to identify class-discriminative sequence motifs within the training data [255]. Motifs identified as specific to the PDAC class were designated positive predictors, while motifs specific to the non-PDAC class served as negative predictors. Validation sequences were classified by the presence or absence of these motifs; detection of a PDAC-specific motif resulted in a positive classification, and detection of a non-PDAC-specific motif resulted in a negative classification. The minimum positive-class hit threshold parameter (fp) and the number of motifs (k) were varied to characterise the performance-coverage trade-off.

7.2.6 Ensemble Model Construction and Performance Evaluation

Ensemble models were constructed by combining the probabilistic outputs of each fine-tuned LLM with the binary signals generated by the alignment-based methods. The integration mechanism adjusted LLM-generated probability scores based on alignment evidence: detection of a class-specific MERCI motif or a statistically significant PSI-BLAST hit to a training sequence of the positive class increased the LLM probability score by 0.5, while detection of a negative-class signal decreased it by 0.5 correspondingly. This additive adjustment transformed the final output into a calibrated probability metric reflecting both the LLM's learned sequence representation and explicit alignment-based evidence of class membership. Model performance was assessed across all LLM architectures, peptide lengths, and alignment parameter settings using AUC and MCC as primary metrics, supplemented by sensitivity, specificity, accuracy, and F1-score.

7.3 Results

7.3.1 Identification of Discriminative Transcriptomic Biomarkers

Feature selection was performed on the probability-transformed expression matrix using the eight complementary algorithms described in Section 7.2.2. The LGBM importance ranking, which demonstrated the strongest alignment between feature ranks and model performance across

downstream evaluations, identified a primary signature composed of the lncRNAs *MALAT1, MACC1-OT1, PCED1B-AS1,* and *LRRC32-AS1,* alongside the uncharacterised transcripts ENSG00000279738 and ENSG00000263244.

Each of these transcripts has documented mechanistic relevance to PDAC biology. *MALAT1* promotes epithelial-mesenchymal transition and metastatic dissemination through modulation of the miR-217/KRAS signalling axis and regulation of alternative splicing. *MACC1-OT1* functions as an oncogenic regulatory element within the *MACC1* signalling network, driving metabolic reprogramming and invasive phenotypes. *PCED1B-AS1* facilitates tumour progression by modulating the miR-411-3p/HIF-1 axis, influencing hypoxia-induced malignant behaviour and cell invasion. *LRRC32-AS1*, an antisense transcript associated with TGF-β signalling regulation, contributes to the maintenance of an immunosuppressive microenvironmental niche within the PDAC stroma. The recurrent selection of uncharacterised non-coding transcripts (ENSG00000279738 and ENSG00000263244) with high importance scores across multiple feature selection methods suggests these elements represent novel disease-associated regulatory transcripts warranting experimental follow-up.

7.3.2 Predictive Performance of Large Language Models

All five LLM architectures were evaluated across pseudo-peptide lengths of 5 to 50 residues. Table 7.1 lists the top-performing independent model for every peptide length. A consistent pattern of design suitability as a function of sequence length was found when the performance data from various peptide lengths were analysed. ESM2_t33, which reached a peak AUC of 0.905 and MCC of 0.665 at a length of 5, showed greater discriminative capacity for shorter pseudo-peptides (lengths 5–20). ProtBERT fared better than the ESM-2 versions for intermediate and longer sequences, attaining the maximum overall AUC of 0.905 and MCC of 0.678 at a peptide length of 25. This discrepancy implies that ProtBERT's bidirectional attention mechanism is better suited to capturing discriminative patterns across longer pseudo-peptide sequences, whereas the inductive biases of ESM-2 architectures—which are optimized for local motif recognition in their pre-training—are more effective for shorter fragments. Overall, the LLM-based approach's diagnostic performance was consistent and repeatable across the tested parameter space, with AUC values ranging from 0.856 to 0.905 for all LLM designs and peptide lengths.

Table 7.1 - *Performance metrics of the top-performing standalone Large Language Model architecture for each pseudo-peptide length.*

Peptide Length	Model	Sensitivity	Specificity	Accuracy	MCC	F1-Score	AUC
5	esm2_t33	0.947	0.682	0.832	0.665	0.864	0.905
8	esm2_t33	0.877	0.705	0.802	0.595	0.833	0.897
10	esm2_t33	0.877	0.727	0.812	0.616	0.840	0.892
15	esm2_t33	0.860	0.659	0.772	0.534	0.810	0.878
20	esm2_t33	0.895	0.705	0.812	0.617	0.843	0.878
25	protbert	0.982	0.636	0.832	0.678	0.868	0.905
30	protbert	0.860	0.773	0.822	0.636	0.845	0.900
40	esm2_t6	0.860	0.705	0.792	0.575	0.824	0.856
50	protbert	0.930	0.341	0.673	0.343	0.763	0.815

7.3.3 Performance of Alignment-Based Approaches

The diagnostic value of MERCI motif identification and PSI-BLAST sequence similarity search was evaluated for pseudo-peptide lengths ranging from 8 to 50 residues. Tables 7.2 and 7.3 present the findings, respectively. For peptide lengths of 8 residues and longer, PSI-BLAST alignments were available; no hits were found for 5-residue sequences, suggesting that this length is below the practical threshold for sequence-based homology searches. Accuracy varied significantly with sequence length and E-value threshold among the lengths for which alignments were available. A peptide length of 40 residues had the highest alignment accuracy (95.5% correct classification at E-value 10^{-5}), followed by a peptide length of 30 (93.3% at E-value 10^{-5}). Significantly, while sequence length increases database search specificity, larger sequences needed stricter e-value limits to maintain classification accuracy.

Table 7.2 - *PSI-BLAST similarity search performance by peptide length using the e-value threshold that maximized accuracy for each length.*

Peptide length	e-value	Total hits	Correct hits	Incorrect hits	% of correct hits
8	1	29	25	4	86.207

10	1.00E+02	95	71	24	74.737
15	1.00E -02	17	13	4	76.471
20	1.00E -02	27	24	3	88.889
25	1.00E -04	17	15	2	88.235
30	1.00E -05	15	14	1	93.333
40	1.00E -05	22	21	1	95.455
50	1.00E -05	29	24	5	82.759

PSI-BLAST and MERCI motif discovery showed complementary performance profiles. Shorter pseudo-peptide sequences had the highest classification accuracy: 8-residue peptides had 97.5% accuracy with fp = 5, while 10-residue peptides had 100% accuracy with a minimum positive-class hit threshold (fp) of 10. Only 11 and 40 validation sequences were categorized, respectively, indicating that coverage was constrained under these conditions. In contrast to the length-40 maximum seen for PSI-BLAST, performance decreased as peptide length increased beyond 20 residues, falling to 76.7% accuracy for length-40 peptides. When combined, these findings show that similarity-based alignment works better for longer sequences (lengths ≥30) and motif-based methods are most dependable for shorter pseudo-peptides (lengths ≤20), indicating a complimentary diagnostic utility between the two techniques.

Table 7.3 - *Diagnostic accuracy of MERCI-based motif discovery across varying peptide lengths, using the minimum positive hit threshold (fp) that yielded maximum accuracy.*

Peptide Length	fp	Total hits	Correct hits	Incorrect hits	% of correct hits
8	5	40	39	1	97.500
10	10	11	11	0	100
15	10	18	17	1	94.444
20	10	21	19	2	90.476
25	10	21	19	2	90.476
30	10	19	18	1	94.737
40	10	43	33	10	76.744
50	10	30	25	5	83.333

7.3.4 Ensemble Model Performance

Ensemble architectures integrating LLM probability scores with alignment-based signals—MERCI motifs (Table 7.4) and PSI-BLAST similarities (Table 7.5)—were evaluated across all peptide lengths (Figure 7.3). Contrary to the hypothesis that explicit alignment evidence would augment LLM performance, the ensemble models consistently failed to improve upon their standalone LLM counterparts. The highest ensemble AUC was 0.897 (LLM + PSI-BLAST, length 30), compared to 0.905 for the standalone LLM at length 25. MCC values were similarly lower for all ensemble configurations relative to the corresponding standalone models. These results indicate that the discriminative features learned by the LLMs during fine-tuning on the pseudo-peptide training set are sufficient for optimal classification, and that the addition of alignment-based signals introduces noise rather than complementary information. This finding supports the interpretation that LLMs trained on pseudo-peptides internalise sequence-level patterns that relevant.

Table 7.4 - *Diagnostic performance of the Ensemble (LLM + MERCI) architecture across varying peptide lengths, utilizing the motif discovery hit thresholds (fp) that yielded maximum accuracy.*

Peptide length	Model	fp	Sensitivity	Specificity	Accuracy	MCC	F1-Score	AUC
10	protbert	10	0.836	0.618	0.762	0.461	0.824	0.856
15	esm_t33	10	0.806	0.647	0.752	0.45	0.812	0.833
20	esm_t6	10	0.866	0.735	0.822	0.601	0.866	0.853
25	protbert	10	0.925	0.676	0.842	0.635	0.886	0.877
30	protbert	10	0.806	0.794	0.802	0.58	0.844	0.880
40	esm_t33	5	0.806	0.647	0.752	0.45	0.812	0.815
50	esm_t33	5	0.791	0.765	0.782	0.537	0.828	0.847

Table 7.5 - *Diagnostic performance of the Ensemble (LLM + PSI-BLAST) architecture across varying peptide lengths, utilizing the sequence alignment e-value thresholds that yielded maximum accuracy.*

Peptide Length	Model	e-value	Sensitivity	Specificity	Accuracy	MCC	F1-Score	AUC
8	esm_t33	1	0.860	0.705	0.792	0.575	0.824	0.881
10	esm_t33	1.00E+02	0.842	0.636	0.753	0.492	0.793	0.886
15	esm_t33	1.00E-02	0.860	0.636	0.762	0.514	0.803	0.835
20	esm_t6	1.00E-02	0.912	0.659	0.802	0.600	0.839	0.874
25	protbert	1.00E-04	0.983	0.591	0.812	0.642	0.855	0.876
30	protbert	1.00E-05	0.877	0.750	0.822	0.636	0.848	0.897
40	esm_t6	1.00E-05	0.860	0.705	0.792	0.575	0.824	0.873
50	protbert	1.00E-05	0.772	0.750	0.762	0.520	0.786	0.852

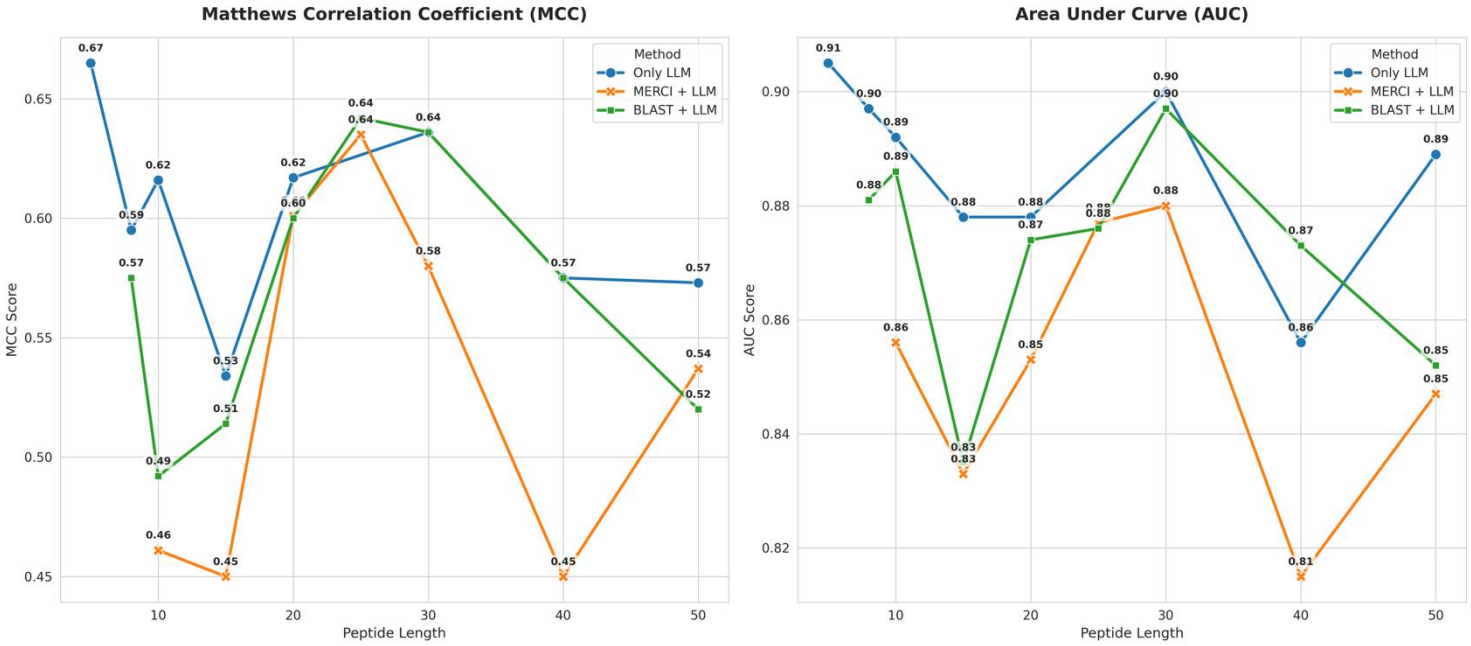


Figure 7.3 - *Comparison of Matthews Correlation Coefficient (MCC) and Area Under the Curve (AUC) metrics for standalone and ensemble Large Language Models across varying peptide lengths.*

7.3.5 Benchmark Comparison

The LLM-based models were benchmarked against two published PDAC transcriptomic classifiers derived from the same GSE133684 dataset. The eight-gene SVM model of Yu et al.

(2020) achieved a reported AUC of 0.936, and a four-lncRNA signature reported by Wang et al. (2024) achieved an AUC of 0.848. The best-performing standalone LLM in the present study—ProtBERT at peptide length 25—achieved an AUC of 0.905, surpassing the four-lncRNA baseline by a margin of 0.057. However, it was not able to improve upon the eight-gene biomarker panel by Yu et al. This comparison demonstrates that LLM-based transcriptomic mining has potential and can be an alternative to conventional machine learning classifiers trained on handcrafted numerical feature representations.

7.4 Discussion and Conclusion

This chapter presents the first application of protein language models for cancer patient classification using transcriptomic expression data, demonstrating a proof of concept for large language model (LLM)–based transcriptomic analysis. The core methodological innovation is a two-stage transformation pipeline that converts continuous gene expression probability values into discrete amino acid–like pseudo-peptide sequences. This representation enables pretrained protein language models, including ProtBERT and members of the ESM-2 family, to process transcriptomic profiles without requiring architectural modification.

Standalone LLM-based classifiers achieved a maximum AUC of 0.905 and MCC of 0.678, indicating that the pseudo-peptide encoding preserves diagnostically relevant information. Performance differences between ProtBERT and ESM-2 were dependent on pseudo-peptide length. ProtBERT, with its bidirectional self-attention mechanism, performed more effectively on longer sequences where global dependencies are informative. In contrast, ESM-2, which emphasizes local positional representations, showed more stable performance at shorter sequence lengths. These findings suggest that pseudo-peptide length—determined by the number of selected genes—is a critical factor in model selection and optimization.

A theoretically significant finding is that ensemble models combining LLM outputs with PSI-BLAST sequence similarity and MERCI motif signals consistently underperformed their standalone LLM counterparts, despite both alignment and motif components demonstrating independent classification accuracy on subsets of the validation data. This outcome indicates that fine-tuning allows the LLM to internalise the same discriminative patterns that explicit alignment

and motif approaches detect, rendering their separate integration redundant. Unlike alignment-based methods constrained to matching against a fixed training database, or motif approaches limited to predefined pattern strings, the fine-tuned LLM learns a continuous, context-sensitive embedding space that generalises beyond specific training instances. This finding parallels observations in natural language processing and protein structure prediction, where ensemble augmentation of fine-tuned transformer models with rule-based components has similarly failed to yield performance improvements when training data is sufficient for the model to learn relevant representations autonomously.

Several conceptual and practical limitations need to be considered here. The pseudo-peptide sequences generated by this pipeline bear no biological correspondence to natural proteins, as amino acid identity at each position is determined by logistic regression probabilities for individual genes rather than by evolutionary selection or structural constraints. However, we are still unsure to what extent these pseudo-peptides encode physicochemical and evolutionary properties of natural sequences. Several practical limitations should be acknowledged. The independent validation set comprised 100 samples, which restricts the precision of clinical confidence interval estimates. Pseudo-peptide length was limited to 50 residues, although performance trends suggest that longer sequences may yield further improvements. The computational cost associated with LLM fine-tuning and inference also limits immediate clinical scalability and would require optimization through model compression or distillation strategies. In addition, the logistic regression–based probability transformation step introduces potential overfitting risk when training cohorts are small relative to the number of genes; future implementations could incorporate stronger regularization or calibration procedures to reduce this risk.

In summary, this chapter establishes the feasibility of applying pretrained protein language models to transcriptomic cancer classification through a structured expression-to-sequence transformation pipeline. This approach represents a distinct methodological alternative to conventional machine learning strategies explored in earlier chapters. Standalone ProtBERT and ESM-2 models achieved competitive diagnostic performance relative to established multi-gene SVM classifiers without extensive feature engineering. The lack of improvement from alignment-based ensemble augmentation further indicates that fine-tuned language models capture relevant discriminative patterns directly from the transformed sequences. The proposed framework is adaptable to other

expression-based datasets and disease contexts and provides a foundation for future development using domain-adapted language models and external prospective validation.

8

Summary

A significant portion of the human transcriptome is made up of long non-coding RNAs (lncRNAs), which are essential for controlling chromatin structure, gene expression, and cellular signaling. Because of their strong context specificity, modest expression levels, and limited experimental characterization, lncRNAs still lack functional annotation despite their significance. Large volumes of transcriptome data have been produced by developments in high-throughput sequencing, but strong computational techniques are needed to extract significant biological insights from these datasets. In order to tackle these issues, this thesis methodically examines lncRNA activity from three interrelated perspectives: subcellular localization, lncRNA–protein interactions, and the application of lncRNAs as cancer diagnostic biomarkers.

Subcellular localization is a key factor in determining the function of long noncoding RNAs (lncRNAs), since the cellular compartment in which a lncRNA localizes greatly influences its molecular partners and regulatory roles. While cytoplasmic lncRNAs affect mRNA stability, translation, and signaling pathways, nuclear-localized lncRNAs are frequently implicated in transcriptional regulation, chromatin remodeling, and epigenetic control. Experimental techniques for lncRNA localization determination, however, are costly, technically complex, and restricted to a few cell types, leading to fragmented and occasionally conflicting datasets. This study initially looked at the localization resources that were already available and identified some of their main drawbacks, such as out-of-date annotations, uneven labeling, and the inclusion of transcripts that were incorrectly classified. A strong computational methodology for predicting the subcellular location of lncRNA across several human cell lines was created based on this investigation. This study demonstrated that correlation-driven features combined with classical machine-learning models consistently outperformed large language model embeddings using carefully curated datasets and a combination of composition-based, correlation-based, and transformer-derived features. By addressing the inadequacies of current methods, feature selection produced a reliable cell-line-specific localization predictor that significantly enhanced prediction stability and interpretability.

In addition to their location, lncRNAs mainly regulate transcription, splicing, chromatin structure, and signaling pathways via interacting with proteins to form dynamic ribonucleoprotein complexes. For large-scale study, computational prediction is crucial since experimental identification of lncRNA–protein interactions (LPIs) is time-consuming and only covers a portion

of the situation. This thesis addressed a significant drawback of current tools trained on out-of-date interaction data by developing a thorough LPI prediction framework utilizing the most recent version of the NPInter database. Large language model embeddings, deep learning architectures, alignment-based techniques, and conventional machine-learning models were among the modeling strategies that were assessed. It was evident from the results that alignment-based approaches did not perform well, proving that LPIs are not based on mere sequence similarity. The best results, however, were obtained by models that combined gradient boosting classifiers with large language model embeddings, such as DNABERT-2 for lncRNAs and ESM-2 for proteins. Additionally, a Siamese-style deep learning model that was trained directly on raw sequences demonstrated competitive performance. The final framework was deployed as a standalone application and a web server, which is significant since it allows the research community to predict lncRNA–protein interactions in a scalable and accessible manner.

With a focus on pancreatic ductal adenocarcinoma (PDAC), one of the deadliest diseases because of its late diagnosis and few available treatments, the translational potential of long non-coding was investigated through their use as diagnostic biomarkers. Poor sensitivity and specificity plague current diagnostic methods, such as imaging and the blood biomarker CA19-9, particularly in early-stage disease. While blood-based transcriptome profiling, especially with extracellular vesicle RNA, provides a less invasive option, it is still difficult to extract reliable indicators from high-dimensional expression data. In this study, concise lncRNA signatures for PDAC detection were found by analyzing plasma EV RNA-seq data using a systematic feature-selection and benchmarking approach. The outcomes demonstrated that small, rigorously chosen lncRNA panels continuously performed better than larger gene sets, attaining high diagnostic accuracy and robustness across validation techniques. The most stable and dependable classifier was CatBoost, and the chosen lncRNAs demonstrated strong discriminatory strength between PDAC and chronic pancreatitis, a significant clinical confounder, as well as obvious biological relevance.

Concurrently, this thesis investigated a radically different approach to transcriptome analysis through the application of huge language models to gene-expression data. The conversion of gene-expression profiles into peptide-like sequences made it possible to apply pretrained protein language models like ESM-2 and ProtBERT. Without the use of manually created numerical characteristics, this transformation made it possible for language models to identify intricate, non-

linear patterns in expression data. The findings showed that large language models can perform diagnostically on par with conventional machine-learning techniques, especially when incorporating intermediate-length pseudo-peptides. The robustness of these transformed sequences was enhanced by ensemble models that combined several techniques, whereas traditional alignment- and motif-based methods performed poorly. This study offers a proof-of-concept for the use of language models as a versatile and expandable transcriptome mining framework.

This thesis offers a thorough computational analysis of lncRNA biology, with an emphasis on both basic functional annotation and therapeutically relevant applications. It emphasizes the significance of feature selection, model selection, and data quality in extracting biologically significant conclusions from complex transcriptome data. This work illustrates how complementary computational solutions can be customized to various biological concerns by combining cutting-edge deep-learning and language-model-based techniques with traditional machine learning. In addition to improving our knowledge of lncRNA function, the frameworks and tools created here offer useful resources for the larger scientific community. All of these contributions provide a solid basis for further research in precision diagnostics, interaction mapping, and lncRNA-driven functional genomics.

References

[1] Taft RJ, Pheasant M, Mattick JS. The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays 2007;29:288–99. https://doi.org/10.1002/bies.20544.

[2] Gall JG. Chromosome structure and the C-value paradox. J Cell Biol 1981;91:3s–14s. https://doi.org/10.1083/jcb.91.3.3s.

[3] Kung JTY, Colognori D, Lee JT. Long noncoding RNAs: past, present, and future. Genetics 2013;193:651–69. https://doi.org/10.1534/genetics.112.146704.

[4] Brannan CI, Dees EC, Ingram RS, et al. The product of the H19 gene may function as an RNA. Mol Cell Biol 1990;10:28–36. https://doi.org/10.1128/mcb.10.1.28-36.1990.

[5] Lee JT. Gracefully ageing at 50, X-chromosome inactivation becomes a paradigm for RNA and chromatin control. Nat Rev Mol Cell Biol 2011;12:815–26. https://doi.org/10.1038/nrm3231.

[6] Brown CJ, Hendrich BD, Rupert JL, et al. The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell 1992;71:527–42. https://doi.org/10.1016/0092-8674(92)90520-m.

[7] Zhao J, Sun BK, Erwin JA, et al. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science 2008;322:750–6. https://doi.org/10.1126/science.1163045.

[8] Kapusta A, Feschotte C. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. Trends Genet 2014;30:439–52. https://doi.org/10.1016/j.tig.2014.08.004.

[9] Statello L, Guo C-J, Chen L-L, et al. Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol 2021;22:96–118. https://doi.org/10.1038/s41580-020-00315-9.

[10] Ulitsky I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nat Rev Genet 2016;17:601–14. https://doi.org/10.1038/nrg.2016.85.

[11] Sarropoulos I, Marin R, Cardoso-Moreira M, et al. Developmental dynamics of lncRNAs across mammalian organs and species. Nature 2019;571:510–4. https://doi.org/10.1038/s41586-019-1341-x.

[12] Darbellay F, Necsulea A. Comparative transcriptomics analyses across species, organs, and developmental stages reveal functionally constrained lncRNAs. Mol Biol Evol 2020;37:240–59. https://doi.org/10.1093/molbev/msz212.

[13] Bartolomei MS, Zemel S, Tilghman SM. Parental imprinting of the mouse H19 gene. Nature 1991;351:153–5. https://doi.org/10.1038/351153a0.

[14] Eren K, Taktakoğlu N, Pirim I. DNA sequencing methods: From past to present. Eurasian J Med 2022;54:47–56. https://doi.org/10.5152/eurasianjmed.2022.22280.

[15] Collins FS, Morgan M, Patrinos A. The Human Genome Project: lessons from large-scale biology. Science 2003;300:286–90. https://doi.org/10.1126/science.1084564.

[16] Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921. https://doi.org/10.1038/35057062.

[17] Carninci P, Kasukawa T, Katayama S, et al. The transcriptional landscape of the mammalian genome. Science 2005;309:1559–63. https://doi.org/10.1126/science.1112014.

[18] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74. https://doi.org/10.1038/nature11247.

[19] Chen L-L. Towards higher-resolution and in vivo understanding of lncRNA biogenesis and function. Nat Methods 2022;19:1152–5. https://doi.org/10.1038/s41592-022-01626-9.

[20] Schlackow M, Nojima T, Gomes T, et al. Distinctive patterns of transcription and RNA processing for human lincRNAs. Mol Cell 2017;65:25–38. https://doi.org/10.1016/j.molcel.2016.11.029.

[21] Wu H, Yang L, Chen L-L. The diversity of long noncoding RNAs and their generation. Trends Genet 2017;33:540–52. https://doi.org/10.1016/j.tig.2017.05.004.

[22] Wilusz JE, JnBaptiste CK, Lu LY, et al. A triple helix stabilizes the 3’ ends of long noncoding RNAs that lack poly(A) tails. Genes Dev 2012;26:2392–407. https://doi.org/10.1101/gad.204438.112.

[23] Yin Q-F, Yang L, Zhang Y, et al. Long noncoding RNAs with snoRNA ends. Mol Cell 2012;48:219–30. https://doi.org/10.1016/j.molcel.2012.07.033.

[24] Memczak S, Jens M, Elefsinioti A, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature 2013;495:333–8. https://doi.org/10.1038/nature11928.

[25] Hansen TB, Jensen TI, Clausen BH, et al. Natural RNA circles function as efficient microRNA sponges. Nature 2013;495:384–8. https://doi.org/10.1038/nature11993.

[26] Yan P, Luo S, Lu JY, et al. Cis- and trans-acting lncRNAs in pluripotency and reprogramming. Curr Opin Genet Dev 2017;46:170–8. https://doi.org/10.1016/j.gde.2017.07.009.

[27] Wang KC, Chang HY. Molecular mechanisms of long noncoding RNAs. Mol Cell 2011;43:904–14. https://doi.org/10.1016/j.molcel.2011.08.018.

[28] Davidovich C, Cech TR. The recruitment of chromatin modifiers by long noncoding RNAs: lessons from PRC2. RNA 2015;21:2007–22. https://doi.org/10.1261/rna.053918.115.

[29] Arunkumar G. LncRNAs: the good, the bad, and the unknown. Biochem Cell Biol 2024;102:9–27. https://doi.org/10.1139/bcb-2023-0155.

[30] Zhang X, Wang W, Zhu W, et al. Mechanisms and functions of long non-coding RNAs at multiple regulatory levels. Int J Mol Sci 2019;20:5573. https://doi.org/10.3390/ijms20225573.

[31] Warwick T, Brandes RP, Leisegang MS. Computational methods to study DNA:DNA:RNA triplex formation by lncRNAs. Noncoding RNA 2023;9:10. https://doi.org/10.3390/ncrna9010010.

[32] Engreitz JM, Pandya-Jones A, McDonel P, et al. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science 2013;341:1237973. https://doi.org/10.1126/science.1237973.

[33] Li X, Fu X-D. Chromatin-associated RNAs as facilitators of functional genomic interactions. Nat Rev Genet 2019;20:503–19. https://doi.org/10.1038/s41576-019-0135-1.

[34] Luo H, Zhu G, Xu J, et al. HOTTIP lncRNA promotes hematopoietic stem cell self-renewal leading to AML-like disease in mice. Cancer Cell 2019;36:645-659.e8. https://doi.org/10.1016/j.ccell.2019.10.011.

[35] Graf J, Kretz M. From structure to function: Route to understanding lncRNA mechanism. Bioessays 2020;42:e2000027. https://doi.org/10.1002/bies.202000027.

[36] Niu X, Zhang L, Wu Y, et al. Biomolecular condensates: Formation mechanisms, biological functions, and therapeutic targets. MedComm 2023;4:e223. https://doi.org/10.1002/mco2.223.

[37] Nakagawa S, Yamazaki T, Mannen T, et al. ArcRNAs and the formation of nuclear bodies. Mamm Genome 2022;33:382–401. https://doi.org/10.1007/s00335-021-09881-5.

[38] Ku D, Yang Y, Kim Y. RNA-associated nuclear condensates: Where the nucleus keeps its RNAs in check. Mol Cells 2025;48:100240. https://doi.org/10.1016/j.mocell.2025.100240.

[39] Qian Y-Y, Li K, Liu Q-Y, et al. Long non-coding RNA PTENP1 interacts with miR-193a-3p to suppress cell migration and invasion through the PTEN pathway in hepatocellular carcinoma. Oncotarget 2017;8:107859–69. https://doi.org/10.18632/oncotarget.22305.

[40] Fotuhi SN, Khalaj-Kondori M, Hoseinpour Feizi MA, et al. Long non-coding RNA BACE1-AS may serve as an Alzheimer's disease blood-based biomarker. J Mol Neurosci 2019;69:351–9. https://doi.org/10.1007/s12031-019-01364-2.

[41] Kretz M. TINCR, staufen1, and cellular differentiation. RNA Biol 2013;10:1597–601. https://doi.org/10.4161/rna.26249.

[42] Karakas D, Ozpolat B. The role of LncRNAs in translation. Noncoding RNA 2021;7:16. https://doi.org/10.3390/ncrna7010016.

[43] Gupta RA, Shah N, Wang KC, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 2010;464:1071–6. https://doi.org/10.1038/nature08975.

[44] Li S, Mei Z, Hu H-B, et al. The lncRNA MALAT1 contributes to non-small cell lung cancer development via modulating miR-124/STAT3 axis. J Cell Physiol 2018;233:6679–88. https://doi.org/10.1002/jcp.26325.

[45] Wang Y, Liu X-J, Yao X-D. Function of PCA3 in prostate tissue and clinical research progress on developing a PCA3 score. Chin J Cancer Res 2014;26:493–500. https://doi.org/10.3978/j.issn.1000-9604.2014.08.08.

[46] Chung DW, Rudnicki DD, Yu L, et al. A natural antisense transcript at the Huntington's disease repeat locus regulates HTT expression. Hum Mol Genet 2011;20:3467–77. https://doi.org/10.1093/hmg/ddr263.

[47] Hu Y, Hu J. Diagnostic value of circulating lncRNA ANRIL and its correlation with coronary artery disease parameters. Braz J Med Biol Res 2019;52:e8309. https://doi.org/10.1590/1414-431X20198309.

[48] Yan B, Yao J, Liu J-Y, et al. lncRNA-MIAT regulates microvascular dysfunction by functioning as a competing endogenous RNA. Circ Res 2015;116:1143–56. https://doi.org/10.1161/CIRCRESAHA.116.305510.

[49] Zhang J, Chen M, Chen J, et al. Long non-coding RNA MIAT acts as a biomarker in diabetic retinopathy by absorbing miR-29b and regulating cell apoptosis. Biosci Rep 2017;37:BSR20170036. https://doi.org/10.1042/BSR20170036.

[50] Frankish A, Diekhans M, Ferreira A-M, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 2019;47:D766–73. https://doi.org/10.1093/nar/gky955.

[51] Zhao L, Wang J, Li Y, et al. NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. Nucleic Acids Res 2021;49:D165–71. https://doi.org/10.1093/nar/gkaa1046.

[52] Chen L, Zhang Y-H, Pan X, et al. Tissue expression difference between mRNAs and lncRNAs. Int J Mol Sci 2018;19:3416. https://doi.org/10.3390/ijms19113416.

[53] Mattick JS, Amaral PP, Carninci P, et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat Rev Mol Cell Biol 2023;24:430–47. https://doi.org/10.1038/s41580-022-00566-8.

[54] Peng W-X, Koirala P, Mo Y-Y. LncRNA-mediated regulation of cell signaling in cancer. Oncogene 2017;36:5661–7. https://doi.org/10.1038/onc.2017.184.

[55] Cantile M, Di Bonito M, Cerrone M, et al. Long non-coding RNA HOTAIR in breast cancer therapy. Cancers (Basel) 2020;12:1197. https://doi.org/10.3390/cancers12051197.

[56] Toker J, Iorgulescu JB, Ling AL, et al. Clinical importance of the lncRNA NEAT1 in cancer patients treated with immune checkpoint inhibitors. Clin Cancer Res 2023;29:2226–38. https://doi.org/10.1158/1078-0432.CCR-22-3714.

[57] Wu H, Chen S, Li A, et al. LncRNA expression profiles in systemic lupus erythematosus and rheumatoid arthritis: Emerging biomarkers and therapeutic targets. Front Immunol 2021;12:792884. https://doi.org/10.3389/fimmu.2021.792884.

[58] Lodde V, Murgia G, Simula ER, et al. Long noncoding RNAs and circular RNAs in autoimmune diseases. Biomolecules 2020;10:1044. https://doi.org/10.3390/biom10071044.

[59] Wu G-C, Pan H-F, Leng R-X, et al. Emerging role of long noncoding RNAs in autoimmune diseases. Autoimmun Rev 2015;14:798–805. https://doi.org/10.1016/j.autrev.2015.05.004.

[60] Zhang H, Zhang N, Wu W, et al. Machine learning-based tumor-infiltrating immune cell-associated lncRNAs for predicting prognosis and immunotherapy response in patients with glioblastoma. Brief Bioinform 2022;23. https://doi.org/10.1093/bib/bbac386.

[61] Qiu X, Tian Y, Xu J, et al. Development and validation of an immune-related long non-coding RNA prognostic model in glioma. J Cancer 2021;12:4264–76. https://doi.org/10.7150/jca.53831.

[62] Yu W, Ma Y, Hou W, et al. Identification of immune-related lncRNA prognostic signature and molecular subtypes for glioblastoma. Front Immunol 2021;12:706936. https://doi.org/10.3389/fimmu.2021.706936.

[63] Zhao Y, Teng H, Yao F, et al. Challenges and strategies in ascribing functions to long noncoding RNAs. Cancers (Basel) 2020;12:1458. https://doi.org/10.3390/cancers12061458.

[64] Wang G, Lee-Yow Y, Chang HY. Approaches to probe and perturb long noncoding RNA functions in diseases. Curr Opin Genet Dev 2024;85:102158. https://doi.org/10.1016/j.gde.2024.102158.

[65] Yan C, Zhang Z, Bao S, et al. Computational methods and applications for identifying disease-associated lncRNAs as potential biomarkers and therapeutic targets. Mol Ther Nucleic Acids 2020;21:156–71. https://doi.org/10.1016/j.omtn.2020.05.018.

[66] Miller JR, Yi W, Adjeroh DA. Evaluation of machine learning models that predict lncRNA subcellular localization. NAR Genom Bioinform 2024;6:lqae125. https://doi.org/10.1093/nargab/lqae125.

[67] Cao Z, Pan X, Yang Y, et al. The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier. Bioinformatics 2018;34:2185–94. https://doi.org/10.1093/bioinformatics/bty085.

[68] Philip M, Chen T, Tyagi S. A survey of current resources to study lncRNA-protein interactions. Noncoding RNA 2021;7:33. https://doi.org/10.3390/ncrna7020033.

[69] Mo Y, Adu-Amankwaah J, Qin W, et al. Unlocking the predictive potential of long non-coding RNAs: a machine learning approach for precise cancer patient prognosis. Ann Med 2023;55:2279748. https://doi.org/10.1080/07853890.2023.2279748.

[70] Bridges MC, Daulagala AC, Kourtidis A. LNCcation: lncRNA localization and function. J Cell Biol 2021;220. https://doi.org/10.1083/jcb.202009045.

[71] Brockdorff N, Bowness JS, Wei G. Progress toward understanding chromosome silencing by Xist RNA. Genes Dev 2020;34:733–44. https://doi.org/10.1101/gad.337196.120.

[72] Hutchinson JN, Ensminger AW, Clemson CM, et al. A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. BMC Genomics 2007;8:39. https://doi.org/10.1186/1471-2164-8-39.

[73] Clemson CM, Hutchinson JN, Sara SA, et al. An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol Cell 2009;33:717–26. https://doi.org/10.1016/j.molcel.2009.01.026.

[74] Miao H, Wang L, Zhan H, et al. A long noncoding RNA distributed in both nucleus and cytoplasm operates in the PYCARD-regulated apoptosis by coordinating the epigenetic and translational regulation. PLoS Genet 2019;15:e1008144. https://doi.org/10.1371/journal.pgen.1008144.

[75] Poloni J de F, Oliveira FHS de, Feltes BC. Localization is the key to action: regulatory peculiarities of lncRNAs. Front Genet 2024;15:1478352. https://doi.org/10.3389/fgene.2024.1478352.

[76] Mas-Ponte D, Carlevaro-Fita J, Palumbo E, et al. LncATLAS database for subcellular localization of long noncoding RNAs. RNA 2017;23:1080–7. https://doi.org/10.1261/rna.060814.117.

[77] Cui T, Dou Y, Tan P, et al. RNALocate v2.0: an updated resource for RNA subcellular localization with increased coverage and annotation. Nucleic Acids Res 2022;50:D333–9. https://doi.org/10.1093/nar/gkab825.

[78] Su Z-D, Huang Y, Zhang Z-Y, et al. iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. Bioinformatics 2018;34:4196–204. https://doi.org/10.1093/bioinformatics/bty508.

[79] Gudenas BL, Wang L. Prediction of LncRNA subcellular localization with deep learning from sequence features. Sci Rep 2018;8:16385. https://doi.org/10.1038/s41598-018-34708-w.

[80] Ahmad A, Lin H, Shatabda S. Locate-R: Subcellular localization of long non-coding RNAs using nucleotide compositions. Genomics 2020;112:2583–9. https://doi.org/10.1016/j.ygeno.2020.02.011.

[81] Feng S, Liang Y, Du W, et al. LncLocation: Efficient subcellular location prediction of long non-coding RNA-based multi-source heterogeneous feature fusion. Int J Mol Sci 2020;21:7271. https://doi.org/10.3390/ijms21197271.

[82] Zeng M, Wu Y, Li Y, et al. LncLocFormer: a Transformer-based deep learning model for multi-label lncRNA subcellular localization prediction by using localization-specific attention mechanism. Bioinformatics 2023;39. https://doi.org/10.1093/bioinformatics/btad752.

[83] Li M, Zhao B, Li Y, et al. SGCL-LncLoc: An interpretable deep learning model for improving lncRNA subcellular localization prediction with supervised graph contrastive learning. Big Data Min Anal 2024;7:765–80. https://doi.org/10.26599/bdma.2024.9020002.

[84] Zeng M, Wu Y, Lu C, et al. DeepLncLoc: a deep learning framework for long non-coding RNA subcellular localization prediction based on subsequence embedding. Brief Bioinform 2022;23. https://doi.org/10.1093/bib/bbab360.

[85] Cai J, Wang T, Deng X, et al. GM-lncLoc: LncRNAs subcellular localization prediction based on graph neural network with meta-learning. BMC Genomics 2023;24:52. https://doi.org/10.1186/s12864-022-09034-1.

[86] Li M, Zhao B, Yin R, et al. GraphLncLoc: long non-coding RNA subcellular localization prediction using graph convolutional networks based on sequence to graph transformation. Brief Bioinform 2023;24. https://doi.org/10.1093/bib/bbac565.

[87] Zhang T, Tan P, Wang L, et al. RNALocate: a resource for RNA subcellular localizations. Nucleic Acids Res 2017;45:D135–8. https://doi.org/10.1093/nar/gkw728.

[88] Fan Y, Chen M, Zhu Q. LncLocPred: Predicting LncRNA subcellular localization using multiple sequence feature information. IEEE Access 2020;8:124702–11. https://doi.org/10.1109/access.2020.3007317.

[89] Lin Y, Pan X, Shen H-B. lncLocator 2.0: a cell-line-specific subcellular localization predictor for long non-coding RNAs with interpretable deep learning. Bioinformatics 2021;37:2308–16. https://doi.org/10.1093/bioinformatics/btab127.

[90] Jeon Y-J, Hasan MM, Park HW, et al. TACOS: a novel approach for accurate prediction of cell-specific long noncoding RNAs subcellular localization. Brief Bioinform 2022;23. https://doi.org/10.1093/bib/bbac243.

[91] Derrien T, Johnson R, Bussotti G, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 2012;22:1775–89. https://doi.org/10.1101/gr.132159.111.

[92] Quinn JJ, Chang HY. Unique features of long non-coding RNA biogenesis and function. Nat Rev Genet 2016;17:47–62. https://doi.org/10.1038/nrg.2015.10.

[93] Bhan A, Mandal SS. LncRNA HOTAIR: A master regulator of chromatin dynamics and cancer. Biochim Biophys Acta 2015;1856:151–64. https://doi.org/10.1016/j.bbcan.2015.07.001.

[94] Zhao J, Ohsumi TK, Kung JT, et al. Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol Cell 2010;40:939–53. https://doi.org/10.1016/j.molcel.2010.12.011.

[95] Darnell JC, Mele A, Hung KYS, et al. Immunoprecipitation and SDS-PAGE for cross-linking immunoprecipitation (CLIP). Cold Spring Harb Protoc 2018;2018:db.prot097956. https://doi.org/10.1101/pdb.prot097956.

[96] Simon MD. Capture hybridization analysis of RNA targets (CHART). Curr Protoc Mol Biol 2013;Chapter 21:Unit 21.25. https://doi.org/10.1002/0471142727.mb2125s101.

[97] Castello A, Horos R, Strein C, et al. Comprehensive identification of RNA-binding proteins by RNA interactome capture. Methods Mol Biol 2016;1358:131–9. https://doi.org/10.1007/978-1-4939-3067-8_8.

[98] Li J-H, Liu S, Zhou H, et al. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res 2014;42:D92-7. https://doi.org/10.1093/nar/gkt1248.

[99] Kang J, Tang Q, He J, et al. RNAInter v4.0: RNA interactome repository with redefined confidence scoring system and improved accessibility. Nucleic Acids Res 2022;50:D326–32. https://doi.org/10.1093/nar/gkab997.

[100] Hu B, Yang Y-CT, Huang Y, et al. POSTAR: a platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. Nucleic Acids Res 2017;45:D104–14. https://doi.org/10.1093/nar/gkw888.

[101] Zheng Y, Luo H, Teng X, et al. NPInter v5.0: ncRNA interaction database in a new era. Nucleic Acids Res 2023;51:D232–9. https://doi.org/10.1093/nar/gkac1002.

[102] Junge A, Refsgaard JC, Garde C, et al. RAIN: RNA-protein association and interaction networks. Database (Oxford) 2017;2017:baw167. https://doi.org/10.1093/database/baw167.

[103] Mann CM, Muppirala UK, Dobbs D. Computational prediction of RNA-protein interactions. Methods Mol Biol 2017;1543:169–85. https://doi.org/10.1007/978-1-4939-6716-2_8.

[104] Suresh V, Liu L, Adjeroh D, et al. RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information. Nucleic Acids Res 2015;43:1370–9. https://doi.org/10.1093/nar/gkv020.

[105] Muppirala UK, Honavar VG, Dobbs D. Predicting RNA-protein interactions using only sequence information. BMC Bioinformatics 2011;12:489. https://doi.org/10.1186/1471-2105-12-489.

[106] Li A, Ge M, Zhang Y, et al. Predicting long noncoding RNA and protein interactions using Heterogeneous Network Model. Biomed Res Int 2015;2015:671950. https://doi.org/10.1155/2015/671950.

[107] Fan X-N, Zhang S-W. LPI-BLS: Predicting lncRNA–protein interactions with a broad learning system-based stacked ensemble classifier. Neurocomputing 2019;370:88–93. https://doi.org/10.1016/j.neucom.2019.08.084.

[108] Yi H-C, You Z-H, Wang M-N, et al. RPI-SE: a stacking ensemble learning framework for ncRNA-protein interactions prediction using sequence information. BMC Bioinformatics 2020;21:60. https://doi.org/10.1186/s12859-020-3406-0.

[109] Zhou L, Duan Q, Tian X, et al. LPI-HyADBS: a hybrid framework for lncRNA-protein interaction prediction integrating feature selection and classification. BMC Bioinformatics 2021;22:568. https://doi.org/10.1186/s12859-021-04485-x.

[110] Ren Z-H, Yu C-Q, Li L-P, et al. SAWRPI: A stacking ensemble framework with adaptive weight for predicting ncRNA-protein interactions using sequence information. Front Genet 2022;13:839540. https://doi.org/10.3389/fgene.2022.839540.

[111] Wu Q, Liu Y, Chen S, et al. Attention holistic processing multi-channel graph transformer with graph residual connections for predicting lncRNA–Protein interactions. Knowl Based Syst 2025;310:112957. https://doi.org/10.1016/j.knosys.2025.112957.

[112] Yuan J, Wu W, Xie C, et al. NPInter v2.0: an updated database of ncRNA interactions. Nucleic Acids Res 2014;42:D104-8. https://doi.org/10.1093/nar/gkt1057.

[113] Lu Q, Ren S, Lu M, et al. Computational prediction of associations between long non-coding RNAs and proteins. BMC Genomics 2013;14:651. https://doi.org/10.1186/1471-2164-14-651.

[114] Huang L, Jiao S, Yang S, et al. LGFC-CNN: Prediction of lncRNA-protein interactions by using multiple types of features through deep learning. Genes (Basel) 2021;12:1689. https://doi.org/10.3390/genes12111689.

[115] Zhang X, Liu M, Li Z, et al. Fusion of multi-source relationships and topology to infer lncRNA-protein interactions. Mol Ther Nucleic Acids 2024;35:102187. https://doi.org/10.1016/j.omtn.2024.102187.

[116] Do H, Kim W. Roles of oncogenic long non-coding RNAs in cancer development. Genomics Inform 2018;16:e18. https://doi.org/10.5808/GI.2018.16.4.e18.

[117] Luo H, Jing H, Chen W. An extensive overview of the role of lncRNAs generated from immune cells in the etiology of cancer. Int Immunopharmacol 2024;133:112063. https://doi.org/10.1016/j.intimp.2024.112063.

[118] Deng Y, Luo S, Zhang X, et al. A pan-cancer atlas of cancer hallmark-associated candidate driver lncRNAs. Mol Oncol 2018;12:1980–2005. https://doi.org/10.1002/1878-0261.12381.

[119] Deng S, Wang J, Zhang L, et al. LncRNA HOTAIR promotes cancer stem-like cells properties by sponging miR-34a to activate the JAK2/STAT3 pathway in pancreatic Ductal Adenocarcinoma. Onco Targets Ther 2021;14:1883–93. https://doi.org/10.2147/OTT.S286666.

[120] Hu Q, Ye Y, Chan L-C, et al. Oncogenic lncRNA downregulates cancer cell antigen presentation and intrinsic tumor suppression. Nat Immunol 2019;20:835–51. https://doi.org/10.1038/s41590-019-0400-7.

[121] Ghafouri-Fard S, Fathi M, Zhai T, et al. LncRNAs: Novel biomarkers for pancreatic cancer. Biomolecules 2021;11:1665. https://doi.org/10.3390/biom11111665.

[122] Kozłowska-Masłoń J, Guglas K, Paszkowska A, et al. Radio-lncRNAs: Biological function and potential use as biomarkers for personalized oncology. J Pers Med 2022;12:1605. https://doi.org/10.3390/jpm12101605.

[123] Kashyap D, Sharma R, Goel N, et al. Coding roles of long non-coding RNAs in breast cancer: Emerging molecular diagnostic biomarkers and potential therapeutic targets with special reference to chemotherapy resistance. Front Genet 2022;13:993687. https://doi.org/10.3389/fgene.2022.993687.

[124] Lin Y, Zhao W, Pu R, et al. Long non-coding RNAs as diagnostic and prognostic biomarkers for colorectal cancer (Review). Oncol Lett 2024;28:486. https://doi.org/10.3892/ol.2024.14619.

[125] Derlatka K, Kulczycka M, Prendecka-Wróbel M, et al. An emerging role of long noncoding RNAs as novel biomarkers for breast cancer metastasis. Appl Sci (Basel) 2024;14:6667. https://doi.org/10.3390/app14156667.

[126] Bosso M, Haddad D, Al Madhoun A, et al. Targeting the metabolic paradigms in cancer and diabetes. Biomedicines 2024;12:211. https://doi.org/10.3390/biomedicines12010211.

[127] Orth M, Metzger P, Gerum S, et al. Pancreatic ductal adenocarcinoma: biological hallmarks, current status, and future perspectives of combined modality treatment approaches. Radiat Oncol 2019;14:141. https://doi.org/10.1186/s13014-019-1345-6.

[128] Sarantis P, Koustas E, Papadimitropoulou A, et al. Pancreatic ductal adenocarcinoma: Treatment hurdles, tumor microenvironment and immunotherapy. World J Gastrointest Oncol 2020;12:173–81. https://doi.org/10.4251/wjgo.v12.i2.173.

[129] Vincent A, Herman J, Schulick R, et al. Pancreatic cancer. Lancet 2011;378:607–20. https://doi.org/10.1016/S0140-6736(10)62307-0.

[130] Katz MHG, Pisters PWT, Evans DB, et al. Borderline resectable pancreatic cancer: the importance of this emerging stage of disease. J Am Coll Surg 2008;206:833–46; discussion 846-8. https://doi.org/10.1016/j.jamcollsurg.2007.12.020.

[131] Lal A, Christians K, Evans DB. Management of borderline resectable pancreatic cancer. Surg Oncol Clin N Am 2010;19:359–70. https://doi.org/10.1016/j.soc.2009.11.006.

[132] Lopez NE, Prendergast C, Lowy AM. Borderline resectable pancreatic cancer: definitions and management. World J Gastroenterol 2014;20:10740–51. https://doi.org/10.3748/wjg.v20.i31.10740.

[133] Ballehaninna UK, Chamberlain RS. The clinical utility of serum CA 19-9 in the diagnosis, prognosis and management of pancreatic adenocarcinoma: An evidence based appraisal. J Gastrointest Oncol 2012;3:105–19. https://doi.org/10.3978/j.issn.2078-6891.2011.021.

[134] Ramya Devi KT, Karthik D, Mahendran T, et al. Long noncoding RNAs: role and contribution in pancreatic cancer. Transcription 2021;12:12–27. https://doi.org/10.1080/21541264.2021.1922071.

[135] Strum S, Evdokimova V, Radvanyi L, et al. Extracellular vesicles and their applications in tumor diagnostics and immunotherapy. Cells 2024;13:2031. https://doi.org/10.3390/cells13232031.

[136] Xue J, Qin S, Ren N, et al. Extracellular vesicle biomarkers in circulation for the diagnosis of gastric cancer: A systematic review and meta-analysis. Oncol Lett 2023;26:423. https://doi.org/10.3892/ol.2023.14009.

[137] Yu S, Li Y, Liao Z, et al. Plasma extracellular vesicle long RNA profiling identifies a diagnostic signature for the detection of pancreatic ductal adenocarcinoma. Gut 2020;69:540–50. https://doi.org/10.1136/gutjnl-2019-318860.

[138] Reese M, Flammang I, Yang Z, et al. Potential of exosomal microRNA-200b as liquid biopsy marker in pancreatic ductal adenocarcinoma. Cancers (Basel) 2020;12:197. https://doi.org/10.3390/cancers12010197.

[139] He X, Chen L, Di Y, et al. Plasma-derived exosomal long noncoding RNAs of pancreatic cancer patients as novel blood-based biomarkers of disease. BMC Cancer 2024;24:961. https://doi.org/10.1186/s12885-024-12755-z.

[140] Takahashi K, Ota Y, Kogure T, et al. Circulating extracellular vesicle-encapsulated HULC is a potential biomarker for human pancreatic cancer. Cancer Sci 2020;111:98–111. https://doi.org/10.1111/cas.14232.

[141] Dykes IM, Emanueli C. Transcriptional and post-transcriptional gene regulation by long non-coding RNA. Genomics Proteomics Bioinformatics 2017;15:177–86. https://doi.org/10.1016/j.gpb.2016.12.005.

[142] Noh JH, Kim KM, McClusky WG, et al. Cytoplasmic functions of long noncoding RNAs. Wiley Interdiscip Rev RNA 2018;9:e1471. https://doi.org/10.1002/wrna.1471.

[143] Raj A, Tyagi S. Detection of individual endogenous RNA transcripts in situ using multiple singly labeled probes. Methods Enzymol 2010;472:365–86. https://doi.org/10.1016/S0076-6879(10)72004-8.

[144] Eng C-HL, Lawson M, Zhu Q, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. Nature 2019;568:235–9. https://doi.org/10.1038/s41586-019-1049-y.

[145] Levesque MJ, Ginart P, Wei Y, et al. Visualizing SNVs to quantify allele-specific expression in single cells. Nat Methods 2013;10:865–7. https://doi.org/10.1038/nmeth.2589.

[146] Mellis IA, Gupte R, Raj A, et al. Visualizing adenosine-to-inosine RNA editing in single mammalian cells. Nat Methods 2017;14:801–4. https://doi.org/10.1038/nmeth.4332.

[147] Chen KH, Boettiger AN, Moffitt JR, et al. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 2015;348:aaa6090. https://doi.org/10.1126/science.aaa6090.

[148] Kishi JY, Lapan SW, Beliveau BJ, et al. SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues. Nat Methods 2019;16:533–44. https://doi.org/10.1038/s41592-019-0404-0.

[149] Shah S, Lubeck E, Zhou W, et al. SeqFISH accurately detects transcripts in single cells and reveals robust spatial organization in the hippocampus. Neuron 2017;94:752-758.e1. https://doi.org/10.1016/j.neuron.2017.05.008.

[150] Wang X, Allen WE, Wright MA, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science 2018;361:eaat5691. https://doi.org/10.1126/science.aat5691.

[151] Lee JH, Daugharthy ER, Scheiman J, et al. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. Nat Protoc 2015;10:442–58. https://doi.org/10.1038/nprot.2014.191.

[152] Gyllborg D, Langseth CM, Qian X, et al. Hybridization-based in situ sequencing (HybISS) for spatially resolved transcriptomics in human and mouse brain tissue. Nucleic Acids Res 2020;48:e112. https://doi.org/10.1093/nar/gkaa792.

[153] Abudayyeh OO, Gootenberg JS, Essletzbichler P, et al. RNA targeting with CRISPR-Cas13. Nature 2017;550:280–4. https://doi.org/10.1038/nature24049.

[154] Gardini A. Global run-On sequencing (GRO-seq). Methods Mol Biol 2017;1468:111–20. https://doi.org/10.1007/978-1-4939-4035-6_9.

[155] Shiraki T, Kondo S, Katayama S, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A 2003;100:15776–81. https://doi.org/10.1073/pnas.2136655100.

[156] German MA, Luo S, Schroth G, et al. Construction of Parallel Analysis of RNA Ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. Nat Protoc 2009;4:356–62. https://doi.org/10.1038/nprot.2009.8.

[157] Keene JD, Komisarow JM, Friedersdorf MB. RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. Nat Protoc 2006;1:302–7. https://doi.org/10.1038/nprot.2006.47.

[158] Smola MJ, Rice GM, Busan S, et al. Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. Nat Protoc 2015;10:1643–69. https://doi.org/10.1038/nprot.2015.103.

[159] Wan Y, Qu K, Ouyang Z, et al. Genome-wide mapping of RNA structure using nuclease digestion and high-throughput sequencing. Nat Protoc 2013;8:849–69. https://doi.org/10.1038/nprot.2013.045.

[160] Jain R, Devine T, George AD, et al. RIP-Chip analysis: RNA-Binding Protein Immunoprecipitation-Microarray (Chip) Profiling. Methods Mol Biol 2011;703:247–63. https://doi.org/10.1007/978-1-59745-248-9_17.

[161] Femino AM, Fay FS, Fogarty K, et al. Visualization of single RNA transcripts in situ. Science 1998;280:585–90. https://doi.org/10.1126/science.280.5363.585.

[162] Lubeck E, Cai L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. Nat Methods 2012;9:743–8. https://doi.org/10.1038/nmeth.2069.

[163] Tsanov N, Samacoits A, Chouaib R, et al. smiFISH and FISH-quant - a flexible single RNA detection approach with super-resolution capability. Nucleic Acids Res 2016;44:e165. https://doi.org/10.1093/nar/gkw784.

[164] Symmons O, Chang M, Mellis IA, et al. Allele-specific RNA imaging shows that allelic imbalances can arise in tissues through transcriptional bursting. PLoS Genet 2019;15:e1007874. https://doi.org/10.1371/journal.pgen.1007874.

[165] Rouhanifard SH, Mellis IA, Dunagin M, et al. ClampFISH detects individual nucleic acid molecules using click chemistry-based amplification. Nat Biotechnol 2018;37:84–9. https://doi.org/10.1038/nbt.4286.

[166] Xia C, Fan J, Emanuel G, et al. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. Proc Natl Acad Sci U S A 2019;116:19490–9. https://doi.org/10.1073/pnas.1912459116.

[167] Heinrich S, Sidler CL, Azzalin CM, et al. Stem-loop RNA labeling can affect nuclear and cytoplasmic mRNA processing. RNA 2017;23:134–41. https://doi.org/10.1261/rna.057786.116.

[168] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10:57–63. https://doi.org/10.1038/nrg2484.

[169] Schena M, Shalon D, Davis RW, et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270:467–70. https://doi.org/10.1126/science.270.5235.467.

[170] Bertone P, Stolc V, Royce TE, et al. Global identification of human transcribed sequences with genome tiling arrays. Science 2004;306:2242–6. https://doi.org/10.1126/science.1103388.

[171] Velculescu VE, Zhang L, Vogelstein B, et al. Serial analysis of gene expression. Science 1995;270:484–7. https://doi.org/10.1126/science.270.5235.484.

[172] Pelechano V, Wei W, Steinmetz LM. Extensive transcriptional heterogeneity revealed by isoform profiling. Nature 2013;497:127–31. https://doi.org/10.1038/nature12121.

[173] Kaewsapsak P, Shechner DM, Mallard W, et al. Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking. Elife 2017;6. https://doi.org/10.7554/eLife.29224.

[174] Tilgner H, Knowles DG, Johnson R, et al. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res 2012;22:1616–25. https://doi.org/10.1101/gr.134445.111.

[175] Choudhury S, Rathore AS, Raghava GPS. Compilation of resources on subcellular localization of lncRNA. Front RNA Res 2024;2. https://doi.org/10.3389/frnar.2024.1419979.

[176] Abugessaisa I, Ramilowski JA, Lizio M, et al. FANTOM enters 20th year: expansion of transcriptomic atlases and functional annotation of non-coding RNAs. Nucleic Acids Res 2021;49:D892–8. https://doi.org/10.1093/nar/gkaa1054.

[177] GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45:580–5. https://doi.org/10.1038/ng.2653.

[178] Li Z, Liu L, Jiang S, et al. LncExpDB: an expression database of human long non-coding RNAs. Nucleic Acids Res 2021;49:D962–8. https://doi.org/10.1093/nar/gkaa850.

[179] Li Z, Liu L, Feng C, et al. LncBook 2.0: integrating human long non-coding RNAs with multi-omics annotations. Nucleic Acids Res 2023;51:D186–91. https://doi.org/10.1093/nar/gkac999.

[180] Volders P-J, Anckaert J, Verheggen K, et al. LNCipedia 5: towards a reference set of human long non-coding RNAs. Nucleic Acids Res 2019;47:D135–9. https://doi.org/10.1093/nar/gky1031.

[181] Szcześniak MW, Wanowska E. CANTATAdb 3.0: An updated repository of plant long non-coding RNAs. Plant Cell Physiol 2024;65:1486–93. https://doi.org/10.1093/pcp/pcae081.

[182] RNAcentral Consortium. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. Nucleic Acids Res 2021;49:D212–20. https://doi.org/10.1093/nar/gkaa921.

[183] Varabyou A, Sommer MJ, Erdogdu B, et al. CHESS 3: an improved, comprehensive catalog of human genes and transcripts based on large-scale expression data, phylogenetic analysis, and protein structure. Genome Biol 2023;24:249. https://doi.org/10.1186/s13059-023-03088-4.

[184] Wen X, Gao L, Guo X, et al. lncSLdb: a resource for long non-coding RNA subcellular localization. Database (Oxford) 2018;2018:1–6. https://doi.org/10.1093/database/bay085.

[185] Lin X, Lu Y, Zhang C, et al. LncRNADisease v3.0: an updated database of long non-coding RNA-associated diseases. Nucleic Acids Res 2024;52:D1365–9. https://doi.org/10.1093/nar/gkad828.

[186] Gao Y, Shang S, Guo S, et al. Lnc2Cancer 3.0: an updated resource for experimentally supported lncRNA/circRNA cancer associations and web tools based on RNA-seq and scRNA-seq data. Nucleic Acids Res 2021;49:D1251–8. https://doi.org/10.1093/nar/gkaa1006.

[187] Frankish A, Diekhans M, Jungreis I, et al. GENCODE 2021. Nucleic Acids Res 2021;49:D916–23. https://doi.org/10.1093/nar/gkaa1087.

[188] ENCODE Project Consortium, Moore JE, Purcaro MJ, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 2020;583:699–710. https://doi.org/10.1038/s41586-020-2493-4.

[189] Li J, Han L, Roebuck P, et al. TANRIC: An interactive open platform to explore the function of lncRNAs in cancer. Cancer Res 2015;75:3728–37. https://doi.org/10.1158/0008-5472.CAN-15-0273.

[190] Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res 2013;41:D991-5. https://doi.org/10.1093/nar/gks1193.

[191] Wu L, Wang L, Hu S, et al. RNALocate v3.0: Advancing the repository of RNA subcellular localization with dynamic analysis and prediction. Nucleic Acids Res 2025;53:D284–92. https://doi.org/10.1093/nar/gkae872.

[192] Zhang S, Qiao H. KD-KLNMF: Identification of lncRNAs subcellular localization with multiple features and nonnegative matrix factorization. Anal Biochem 2020;610:113995. https://doi.org/10.1016/j.ab.2020.113995.

[193] Sang L, Yang L, Ge Q, et al. Subcellular distribution, localization, and function of noncoding RNAs. Wiley Interdiscip Rev RNA 2022;13:e1729. https://doi.org/10.1002/wrna.1729.

[194] Clark MB, Johnston RL, Inostroza-Ponta M, et al. Genome-wide analysis of long noncoding RNA stability. Genome Res 2012;22:885–98. https://doi.org/10.1101/gr.131037.111.

[195] Saxena A, Carninci P. Long non-coding RNA modifies chromatin: epigenetic silencing by long non-coding RNAs. Bioessays 2011;33:830–9. https://doi.org/10.1002/bies.201100084.

[196] Gong C, Maquat LE. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. Nature 2011;470:284–8. https://doi.org/10.1038/nature09701.

[197] Du Z, Sun T, Hacisuleyman E, et al. Integrative analyses reveal a long noncoding RNA-mediated sponge regulatory network in prostate cancer. Nat Commun 2016;7:10982. https://doi.org/10.1038/ncomms10982.

[198] Miyagawa R, Tano K, Mizuno R, et al. Identification of cis- and trans-acting factors involved in the localization of MALAT-1 noncoding RNA to nuclear speckles. RNA 2012;18:738–51. https://doi.org/10.1261/rna.028639.111.

[199] Lubelsky Y, Ulitsky I. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. Nature 2018;555:107–11. https://doi.org/10.1038/nature25757.

[200] Yin Y, Lu JY, Zhang X, et al. U1 snRNP regulates chromatin retention of noncoding RNAs. Nature 2020;580:147–50. https://doi.org/10.1038/s41586-020-2105-3.

[201] Porto FW, Daulatabad SV, Janga SC. Long non-coding RNA expression levels modulate cell-type-specific splicing patterns by altering their interaction landscape with RNA-binding proteins. Genes (Basel) 2019;10:E593. https://doi.org/10.3390/genes10080593.

[202] Guo C-J, Xu G, Chen L-L. Mechanisms of long noncoding RNA nuclear retention. Trends Biochem Sci 2020;45:947–60. https://doi.org/10.1016/j.tibs.2020.07.001.

[203] Carlevaro-Fita J, Polidori T, Das M, et al. Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs. Genome Res 2019;29:208–22. https://doi.org/10.1101/gr.229922.117.

[204] He R-Z, Jiang J, Luo D-X. The functions of N6-methyladenosine modification in lncRNAs. Genes Dis 2020;7:598–605. https://doi.org/10.1016/j.gendis.2020.03.005.

[205] Davis CA, Hitz BC, Sloan CA, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res 2018;46:D794–801. https://doi.org/10.1093/nar/gkx1081.

[206] Cunningham F, Allen JE, Allen J, et al. Ensembl 2022. Nucleic Acids Res 2022;50:D988–95. https://doi.org/10.1093/nar/gkab1049.

[207] Girgis HZ. MeShClust v3.0: high-quality clustering of DNA sequences using the mean shift algorithm and alignment-free identity scores. BMC Genomics 2022;23:423. https://doi.org/10.1186/s12864-022-08619-0.

[208] Mathur M, Patiyal S, Dhall A, et al. Nfeature: A platform for computing features of nucleotide sequences 2021. https://doi.org/10.1101/2021.12.14.472723.

[209] Zhou Z, Ji Y, Li W, et al. DNABERT-2: Efficient foundation model and benchmark for multi-species genome. arXiv [q-BioGN] 2023. https://doi.org/10.48550/ARXIV.2306.15006.

[210] Zhao Z, Anand R, Wang M. Maximum Relevance and Minimum Redundancy feature selection methods for a marketing machine learning platform. arXiv [StatML] 2019. https://doi.org/10.48550/ARXIV.1908.05376.

[211] Zhang S, Amahong K, Zhang Y, et al. RNAenrich: a web server for non-coding RNA enrichment. Bioinformatics 2023;39. https://doi.org/10.1093/bioinformatics/btad421.

[212] Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res 2015;43:D1049-56. https://doi.org/10.1093/nar/gku1179.

[213] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30. https://doi.org/10.1093/nar/28.1.27.

[214] Choudhury S, Mehta NK, Raghava GPS. CytoLNCpred-a computational method for predicting cytoplasm associated long non-coding RNAs in 15 cell-lines. Front Bioinform 2025;5:1585794. https://doi.org/10.3389/fbinf.2025.1585794.

[215] Chu C, Spitale RC, Chang HY. Technologies to probe functions and mechanisms of long noncoding RNAs. Nat Struct Mol Biol 2015;22:29–35. https://doi.org/10.1038/nsmb.2921.

[216] Ribeiro DM, Zanzoni A, Cipriano A, et al. Protein complex scaffolding predicted as a prevalent function of long non-coding RNAs. Nucleic Acids Res 2018;46:917–28. https://doi.org/10.1093/nar/gkx1169.

[217] Tripathi V, Ellis JD, Shen Z, et al. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol Cell 2010;39:925–38. https://doi.org/10.1016/j.molcel.2010.08.011.

[218] Bousard A, Raposo AC, Żylicz JJ, et al. The role of Xist-mediated Polycomb recruitment in the initiation of X-chromosome inactivation. EMBO Rep 2019;20:e48019. https://doi.org/10.15252/embr.201948019.

[219] Gagliardi M, Matarazzo MR. RIP: RNA immunoprecipitation. Methods Mol Biol 2016;1480:73–86. https://doi.org/10.1007/978-1-4939-6380-5_7.

[220] Danan C, Manickavel S, Hafner M. PAR-CLIP: A method for transcriptome-wide identification of RNA binding protein interaction sites. Methods Mol Biol 2016;1358:153–73. https://doi.org/10.1007/978-1-4939-3067-8_10.

[221] Huppertz I, Attig J, D'Ambrogio A, et al. iCLIP: protein-RNA interactions at nucleotide resolution. Methods 2014;65:274–87. https://doi.org/10.1016/j.ymeth.2013.10.011.

[222] Hu H, Zhang L, Ai H, et al. HLPI-Ensemble: Prediction of human lncRNA-protein interactions based on ensemble strategy. RNA Biol 2018;15:797–806. https://doi.org/10.1080/15476286.2018.1457935.

[223] Li Y, Sun H, Feng S, et al. Capsule-LPI: a LncRNA-protein interaction predicting tool based on a capsule network. BMC Bioinformatics 2021;22:246. https://doi.org/10.1186/s12859-021-04171-y.

[224] Xiao Y, Zhang J, Deng L. Prediction of lncRNA-protein interactions using HeteSim scores based on heterogeneous networks. Sci Rep 2017;7:3664. https://doi.org/10.1038/s41598-017-03986-1.

[225] Huang Y, Niu B, Gao Y, et al. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 2010;26:680–2. https://doi.org/10.1093/bioinformatics/btq003.

[226] Pande A, Patiyal S, Lathwal A, et al. Pfeature: A tool for computing wide range of protein features and building prediction models. J Comput Biol 2023;30:204–22. https://doi.org/10.1089/cmb.2022.0241.

[227] Wang J, Yang B, Revote J, et al. POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. Bioinformatics 2017;33:2756–8. https://doi.org/10.1093/bioinformatics/btx302.

[228] UniProt Consortium. UniProt: The universal protein knowledgebase in 2023. Nucleic Acids Res 2023;51:D523–31. https://doi.org/10.1093/nar/gkac1052.

[229] Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379:1123–30. https://doi.org/10.1126/science.ade2574.

[230] Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. BMC Bioinformatics 2009;10:421. https://doi.org/10.1186/1471-2105-10-421.

[231] Becker AE, Hernandez YG, Frucht H, et al. Pancreatic ductal adenocarcinoma: risk factors, screening, and early detection. World J Gastroenterol 2014;20:11182–98. https://doi.org/10.3748/wjg.v20.i32.11182.

[232] Bilimoria KY, Bentrem DJ, Ko CY, et al. Validation of the 6th edition AJCC Pancreatic Cancer Staging System: report from the National Cancer Database. Cancer 2007;110:738–44. https://doi.org/10.1002/cncr.22852.

[233] Kim J, Bamlet WR, Oberg AL, et al. Detection of early pancreatic ductal adenocarcinoma with thrombospondin-2 and CA19-9 blood markers. Sci Transl Med 2017;9:eaah5583. https://doi.org/10.1126/scitranslmed.aah5583.

[234] Cao K, Xia Y, Yao J, et al. Large-scale pancreatic cancer detection via non-contrast CT and deep learning. Nat Med 2023;29:3033–43. https://doi.org/10.1038/s41591-023-02640-w.

[235] Breiman L. Random forests. Mach Learn 2001;45:5–32. https://doi.org/10.1023/a:1010933404324.

[236] Ji L, Yi Z, Pu X. Fingerprint classification by SPCNN and combined LVQ networks. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg; 2006, p. 395–8. https://doi.org/10.1007/11881070_55.

[237] Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features. arXiv [CsLG] 2017. https://doi.org/10.48550/ARXIV.1706.09516.

[238] Chen T, Guestrin C. XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM; 2016. https://doi.org/10.1145/2939672.2939785.

[239] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97. https://doi.org/10.1023/a:1022627411411.

[240] Li L, Chen H, Gao Y, et al. Long noncoding RNA MALAT1 promotes aggressive pancreatic cancer proliferation and metastasis via the stimulation of autophagy. Mol Cancer Ther 2016;15:2232–43. https://doi.org/10.1158/1535-7163.MCT-16-0008.

[241] Zhang Y, Ma H, Chen C. Long non-coding RNA PCED1B-AS1 promotes pancreatic ductal adenocarcinoma progression by regulating the miR-411-3p/HIF-1α axis. Oncol Rep 2021;46. https://doi.org/10.3892/or.2021.8085.

[242] Kyuno D, Asano H, Okumura R, et al. The role of claudin-1 in enhancing pancreatic cancer aggressiveness and drug resistance via metabolic pathway modulation. Cancers (Basel) 2025;17:1469. https://doi.org/10.3390/cancers17091469.

[243] Zhao S, Chen C, Chang K, et al. CD44 expression level and isoform contributes to pancreatic cancer cell plasticity, invasiveness, and response to therapy. Clin Cancer Res 2016;22:5592–604. https://doi.org/10.1158/1078-0432.CCR-15-3115.

[244] Armstrong SA, Schultz CW, Azimi-Sadjadi A, et al. ATM dysfunction in pancreatic adenocarcinoma and associated therapeutic implications. Mol Cancer Ther 2019;18:1899–908. https://doi.org/10.1158/1535-7163.MCT-19-0208.

[245] Gao X, Jiang M, Chu Y, et al. ETV4 promotes pancreatic ductal adenocarcinoma metastasis through activation of the CXCL13/CXCR5 signaling axis. Cancer Lett 2022;524:42–56. https://doi.org/10.1016/j.canlet.2021.09.026.

[246] Kashiwaya K, Nakagawa H, Hosokawa M, et al. Involvement of the tubulin tyrosine ligase-like family member 4 polyglutamylase in PELP1 polyglutamylation and chromatin remodeling in pancreatic cancer cells. Cancer Res 2010;70:4024–33. https://doi.org/10.1158/0008-5472.CAN-09-4444.

[247] Halbrook CJ, Lyssiotis CA, Pasca di Magliano M, et al. Pancreatic cancer: Advances and challenges. Cell 2023;186:1729–54. https://doi.org/10.1016/j.cell.2023.02.014.

[248] Kane S, Engelhart A, Guadagno J, et al. Pancreatic ductal adenocarcinoma: Characteristics of tumor microenvironment and barriers to treatment. J Adv Pract Oncol 2020;11:693–8. https://doi.org/10.6004/jadpro.2020.11.7.4.

[249] Nakaoka K, Ohno E, Kawabe N, et al. Current status of the diagnosis of early-stage pancreatic ductal adenocarcinoma. Diagnostics (Basel) 2023;13:215. https://doi.org/10.3390/diagnostics13020215.

[250] Winter K, Talar-Wojnarowska R, Dąbrowski A, et al. Diagnostic and therapeutic recommendations in pancreatic ductal adenocarcinoma. Recommendations of the Working Group of the Polish Pancreatic Club. Prz Gastroenterol 2019;14:1–18. https://doi.org/10.5114/pg.2019.83422.

[251] Poruk KE, Gay DZ, Brown K, et al. The clinical utility of CA 19-9 in pancreatic adenocarcinoma: diagnostic and prognostic updates. Curr Mol Med 2013;13:340–51. https://doi.org/10.2174/1566524011313030003.

[252] Lee T, Teng TZJ, Shelat VG. Carbohydrate antigen 19-9 - tumor marker: Past, present, and future. World J Gastrointest Surg 2020;12:468–90. https://doi.org/10.4240/wjgs.v12.i12.468.

[253] Guntuboina C, Das A, Mollaei P, et al. PeptideBERT: A language model based on transformers for peptide property prediction. J Phys Chem Lett 2023;14:10427–34. https://doi.org/10.1021/acs.jpclett.3c02398.

[254] Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: Toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 2022;44:7112–27. https://doi.org/10.1109/tpami.2021.3095381.

[255] Vens C, Rosso M-N, Danchin EGJ. Identifying discriminative classification-based motifs in biological sequences. Bioinformatics 2011;27:1231–8. https://doi.org/10.1093/bioinformatics/btr110.