

**T.C.  
REPUBLIC OF TURKEY  
HACETTEPE UNIVERSITY  
GRADUATE SCHOOL OF HEALTH SCIENCES**

**IMAGE, SEQUENCE, AND INTERACTOME BASED  
PREDICTION OF SUBCELLULAR LOCALIZATION OF  
PROTEINS**

**Ecem KUŞCUOĞLU**

**Program of Bioinformatics  
MASTER'S THESIS**

**ANKARA  
2024**



**T.C.  
REPUBLIC OF TURKEY  
HACETTEPE UNIVERSITY  
GRADUATE SCHOOL OF HEALTH SCIENCES**

**IMAGE, SEQUENCE, AND INTERACTOME BASED  
PREDICTION OF SUBCELLULAR LOCALIZATION OF  
PROTEINS**

**Ecem KUŞCUOĞLU**

**Program of Bioinformatics  
MASTER'S THESIS**

**ADVISOR OF THE THESIS  
Prof. Dr. Tunca DOĞAN**

**ANKARA**

**2024**

**IMAGE, SEQUENCE, AND INTERACTOME BASED PREDICTION OF  
SUBCELLULAR LOCALIZATION OF PROTEINS**

**Ecem Kuşcuođlu**

**Supervisor: Prof. Tunca Dođan, PhD**

This thesis study has been approved and accepted as a Master dissertation in “Bioinformatics Program” by the assessment committee, whose members are listed below, on January 02,2024.

**Chairman of the Committee:** Assoc. Prof. Ercüment ÇİÇEK, PhD  
Bilkent University

**Advisor of the Dissertation:** Prof. Tunca DOĐAN, PhD  
Hacettepe University

**Member:** Assist. Prof. Gülşah Merve KILINÇ, PhD  
Hacettepe University

This dissertation has been approved by the above committee in conformity to the related issues of Hacettepe University Graduate Education and Examination Regulation.

Prof. Müge YEMİŐCİ ÖZKAN, MD, PhD

**Director**

## YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan “**Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge**” kapsamında tezimin aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H.Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir. <sup>(1)</sup>
- Enstitü / Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren ... ay ertelenmiştir. <sup>(2)</sup>
- Tezimle ilgili gizlilik kararı verilmiştir. <sup>(3)</sup>

26/01/2024

Ecem Kuşcuoğlu

i

<sup>1</sup>“*Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge*”

(1) *Madde 6. 1. Lisansüstü teze ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.*

(2) *Madde 6. 2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç imkanı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulunun gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.*

(3) *Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, tezin yapıldığı kurum tarafından verilir \*. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, ilgili kurum ve kuruluşun önerisi ile enstitü veya fakültenin uygun görüşü üzerine üniversite yönetim kurulu tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir. Madde 7.2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir*

\* Tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu tarafından karar verilir.

## **ETHICAL DECLARATION**

In this thesis study, I declare that all the information and documents have been obtained in the base of the academic rules and all audio-visual and written information and results have been presented according to the rules of scientific ethics. I did not do any distortion in the data set. In the case of using other works, related studies have been fully cited in accordance with the scientific standards. I also declare that my thesis study is original except cited references. It was produced by me in consultation with supervisor Prof. Dr. Tunca DOĞAN and written according to the rules of thesis writing of Hacettepe University Institute of Health Sciences.

Ecem Kuşcuođlu

## ACKNOWLEDGEMENT

I want to convey my deep appreciation to my family, whose unwavering support, boundless love, selfless sacrifices, and constant encouragement have been the bedrock of my journey. Their presence has been a guiding light, and I am truly grateful for their steadfast belief in me.

I express my gratitude to Prof. Dr. Tunca Doğan for his essential guidance and unwavering support, significantly shaping my academic journey. His mentorship has been crucial and deeply inspiring.

I'm thankful for my colleagues and friends. They've been there for me during tough times, providing great support.

I want to give a heartfelt thanks to Tuna and Banu. They've been there for me, even staying up late nights and being true companions. Their support when everything seemed tough has been my rock.

In remembrance of Zehra, I want to express my deepest gratitude. Your memory will forever hold a special place in my heart. Your influence has left an indelible mark on my journey, and I am thankful for the time we shared.

To everyone who has played a role in my academic and personal development, I extend my sincere thanks. Your collective impact has been immeasurable, and I am grateful for the privilege of having you in my life.

## ABSTRACT

**Kuşcuoğlu, E. Image, Sequence, And Interactome Based Prediction of Subcellular Localization of Proteins, Hacettepe University Graduate School of Health Sciences Bioinformatics Program Master's Thesis, Ankara, 2024.**

Knowledge of subcellular localization (SL) of proteins is essential for drug development, systems biology, proteomics, and functional genomics. Due to the high costs associated with experimental studies, it has become crucial to develop computational systems to accurately predict proteins' SLs. With different modes of biological data (e.g., biomolecular sequences, biomedical images, unstructured text, etc.) becoming readily available to ordinary scientists, it is possible to leverage complementary types of data to increase both the performance and coverage of predictions. In this study, we propose HoliLoc, a new method for predicting protein SLs via multi-modal deep learning. Our approach makes use of three different types of data (i.e., 2D confocal microscopy images, amino acid sequences, and protein-protein interactions – PPIs) to predict SLs of proteins in a multi-label manner for 22 different cell compartments using protein language models, graph embeddings and convolutional and feed forward neural networks. The system was trained in an end-to-end manner, and the performances were calculated on the unseen hold-out test dataset. The average test performance of individual models (each using a single data type) was 0.18 (macro F1-score) and 0.55 (accuracy), whereas for HoliLoc (the fusion of 3 modalities) it was observed to be 0.26 (F1-score) and 0.60 (accuracy), indicating the effectiveness of the multi-modal learning approach proposed. According to our comparison against state-of-the-art SL predictors, HoliLoc displays highly competitive performance. HoliLoc is distributed as an open-access programmatic tool, which is anticipated to benefit life science researchers by reducing the cost and time required for wet-lab experiments by accurately predicting the SLs of the protein of interest in advance.

**Key words:** Protein subcellular localization prediction, deep learning, protein research



## ÖZET

**Kuşcuoğlu, E. Proteinlerin Subselüler Yerleşimlerinin Görüntü, Sekans ve İnteraktom Verisi Tabanlı Tahmini, Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü Biyoinformatik Programı Yüksek Lisans Tezi, Ankara, 2024.**

Proteinlerin subselüler yerleşimleri (SY) bilgisi, ilaç geliştirme, sistem biyolojisi, proteomik ve fonksiyonel genomik alanlarında önemlidir. Deneysel çalışmalarla ilişkilendirilen yüksek maliyetler nedeniyle, proteinlerin SY'lerini doğru bir şekilde tahmin edecek için hesaplamalı sistemleri geliştirmek gerekli hale gelmiştir. Farklı biyolojik veri türlerinin (örneğin, biyomoleküler diziler, biyomedikal görüntüler, yapılandırılmamış metinler vb.) araştırmacılar için kolayca erişilebilir hale gelmesi, tahminlerin hem performansını hem de kapsamını artırmak için tamamlayıcı veri türlerinden yararlanma olasılığı sunmuştur. Bu çalışmada, protein SY'lerini çok modlu derin öğrenme ile tahmin etmek için HoliLoc adlı yeni bir yöntem önerilmiştir. Yaklaşımımız, protein dil modellerini, çizge öğrenme tekniklerini ve evrişimli ve ileri beslemeli sinir ağlarını kullanarak, 22 farklı kompartıman için proteinlerin SY'lerini tahmin etmek için üç farklı veri türünden yararlanır (2D konfokal mikroskopi görüntüleri, amino asit dizileri ve protein-protein etkileşimleri). Sistem, uçtan uca bir şekilde eğitildi ve performanslar daha önce görülmeyen test veri setinde hesaplandı. Her biri tek bir veri tipini kullanan bireysel modellerin test performansı ortalama 0.18 makro F1 puanı ve 0.55 doğrulukta iken, HoliLoc'un (3 modalitenin birleşimi) gözlemlenen ortalama test performansı 0.26 makro F1 puanı ve 0.60 doğruluk olarak tespit edildi. Bu sonuçlar, önerilen çoklu modlu öğrenme yaklaşımının başarısını göstermektedir. Literatürde mevcut SY tahmincilerine karşı yaptığımız karşılaştırmaya göre, HoliLoc oldukça rekabetçi bir performans sergilemektedir. HoliLoc, yaşam bilimleri araştırmacılarına, ilgilendikleri proteinin subselüler yerleşimlerini doğru bir şekilde tahmin ederek laboratuvar deneyleri için gereken maliyeti ve zamanı azaltacak açık erişimli bir programlama aracı olarak sunulmaktadır.

**Anahtar kelimeler:** Protein subselüler yerleşimlerinin tahmini, derin öğrenme, protein bilimi

## TABLE OF CONTENTS

APPROVAL PAGE	iii
YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI	iv
ETHICAL DECLARATION	v
ACKNOWLEDGMENTS	vi
ABSTRACT	vii
ÖZET	viii
TABLE OF CONTENTS	ix
LIST OF SYMBOLS AND ABBREVIATIONS	xii
LIST OF FIGURES	xiii
LIST OF TABLES	xv
<b>1. INTRODUCTION</b>	<b>1</b>
1.1. Problem Definition and Motivation	2
1.2. Aim and Scope	3
1.3. Structure of the Thesis	4
<b>2. BACKGROUND INFORMATION</b>	<b>6</b>
2.1. Protein Subcellular Localization	6
2.2. Immunofluorescence	6
2.3. Human Protein Atlas Subcellular Section	8
2.4. Interactome Data and IntAct	8
2.5. Protein Sequences, Function and The UniProtKB	9
2.6. Gene Ontology (GO) Database	10
2.7. Machine Learning and Deep Learning	11
2.7.1. Feedforward Neural Network (FFN)	12
2.7.2. Convolutional Neural Network (CNN)	13
2.7.3. Graph Learning	15
2.7.4. Node2Vec	16

2.7.5 Sequence Embeddings	17
2.8. Related Work	18
2.8.1. MULocDeep	18
2.8.2. DeepLoc	19
2.8.3. MuSIC	20
2.8.4. SLPred	21
2.8.5. Human Protein Atlas Image Classification Kaggle Challenge	22
2.8.6. Team 1 (bestfitting)	23
2.8.7. Team 2 (WAIR)	24
2.8.8. Team 3 (Pudae)	25
2.9. Fusion Strategies Using Deep Learning	26
<b>3. MATERIALS AND METHODS</b>	28
3.1. HoliLoc Dataset Construction and Splitting	28
3.1.1 Image Data	29
3.1.2. Interactome (PPI) Data	30
3.1.3. Sequence Data	31
3.1.4. Arrangement of Location Classes	31
3.2. Data Preprocessing	35
3.2.1. Image Data Preprocessing	35
3.2.2. Sequence Data Preprocess	37
3.2.3. Interactome (PPI) Data Preprocess	38
3.3. Classification Models	40
3.3.1. Image Model	40
3.3.2. Sequence Model	43
3.3.3. Interactome Model (PPI)	45
3.3.4. Model Fusion (HoliLoc)	47

<b>4. RESULTS</b>	49
4.1. Evaluation of Performance	49
4.2. Comparison of HoliLoc and Individual Feature Based Models	50
4.2.1. Comparison of HoliLoc and Individual Feature Based Models on Single Location Prediction Setting	54
4.2.2. Comparison of HoliLoc and Individual Feature Based Models on Multi Location Prediction Setting	50
4.3. Comparison With HPA Kaggle Challenge	60
4.4. Use Case	62
<b>5. DISCUSSION</b>	65
<b>6. CONCLUSION</b>	73
<b>7. REFERENCES</b>	76
<b>8. APPENDIX</b>	80
<b>EK-1: Tez Çalışması ile İlgili Etik Kurul İzinleri</b>	
<b>EK-2: Tez Çalışması Orijinallik Raporu</b>	
<b>9. CURRICULUM VITAE</b>	

**LIST OF SYMBOLS AND ABBREVIATIONS**

<b>AI</b>	Artificial Intelligence
<b>BFS</b>	Breadth-first Sampling
<b>CNN</b>	Convolutional neural network
<b>DFS</b>	Depth-first Sampling
<b>DL</b>	Deep Learning
<b>DT</b>	Decision Tree
<b>ER</b>	Endoplasmic Reticulum
<b>FFN</b>	Feedforward neural network
<b>FFN</b>	Feed Forward Neural Network
<b>GO</b>	Gene Ontology
<b>HPA</b>	Human Protein Atlas
<b>HPA</b>	Human Protein Atlas
<b>IF</b>	Immunofluorescence
<b>IF</b>	Immunofluorescence
<b>K-NN</b>	K-Nearest Neighbor Algorithm
<b>LM</b>	Language Models
<b>LSTM</b>	Long Short-Term Memory
<b>MTOC</b>	Microtubule Organizing Center
<b>NLP</b>	Natural Language Processing
<b>PM</b>	Plasma Membrane
<b>PPI</b>	Protein Protein interactions
<b>SVM</b>	Support Vector Machine
<b>UniProtKB</b>	Universal Protein Knowledge Base
<b>UniRef</b>	UniProt Reference Clusters

## LIST OF FIGURES

<b>Figures</b>	<b>Pages</b>	
1.1.	Representation of the subcellular locations of a eukaryotic cell	1
1.2.	The workflow of the proposed protein subcellular localization prediction method: HoliLoc	2
2.1.	Direct immunofluorescence	7
2.2.	Indirect immunofluorescence	8
2.3.	Architecture of a FFN composed of three layers	13
2.4.	Architecture of a CNN.	15
2.5.	Illustration of graph representation learning input and output.	16
2.6.	BFS and DFS search strategies	17
2.7.	MULocDeep workflow and neural network architecture	19
2.8.	Overview of DeepLoc	20
2.9.	Overview of multi scale integrated map of the cell: MuSIC	21
2.10.	Schematic representation of the subcellular localization predictor SLPred	22
2.11.	Model architecture of team 1: bestfitting	24
2.12.	Model architecture of team 2: WAIR	25
2.13.	Model architecture of team 3: pudae	26
2.14.	Fusion strategies using deep learning	27
3.1.	Heatmap illustrating the shared subcellular locations among various cell lines c1 and c2. Each cell in the heatmap represents the percentage of protein localization similarity between corresponding cell lines.	30

<b>3.2.</b>	Bar graph displaying the number of proteins associated with each subcellular location before arrangement of location classes	32
<b>3.3.</b>	Bar graph displaying the number of proteins associated with each subcellular location after arrangement of location classes	32
<b>3.4.</b>	The bar graph showing the distribution of train and test data with the x-axis representing subcellular locations and the y-axis indicating the corresponding sample sizes in both the training and test datasets.	33
<b>3.5.</b>	Data distribution among 22 subcellular locations	35
<b>3.6.</b>	Overview of image data preprocess	36
<b>3.7.</b>	Visual examples of HoliLoc input protein image data	37
<b>3.8.</b>	Overview of sequence data preprocess	38
<b>3.9.</b>	Overview of interactome (PPI) data preprocess	39
<b>3.10.</b>	Image model structure	42
<b>3.11.</b>	Sequence model structure	44
<b>3.12.</b>	Interactome (PPI) model structure	46
<b>3.13.</b>	HoliLoc model structure	48
<b>4.1.</b>	Performance comparison of HoliLoc and individual models for endoplasmic reticulum (ER) prediction	53
<b>4.2.</b>	Performance comparison of HoliLoc and individual feature-based models' macro F1 scores in the single-location models per subcellular location and multi-location prediction setting	54
<b>4.3.</b>	Comparison of average prediction performances of HoliLoc and individual feature-based models in the multi-location prediction setting.	56
<b>4.4.</b>	Confocal microscopy images of target protein P68431, obtained from Human Protein Atlas database in which green is target protein, blue is nucleus, red is microtubules and yellow is ER. A: all channels are visible, B: red and green channels are visible, C: blue and green channels are visible, D: yellow and green channels are visible, E: Only green channel is visible.	64

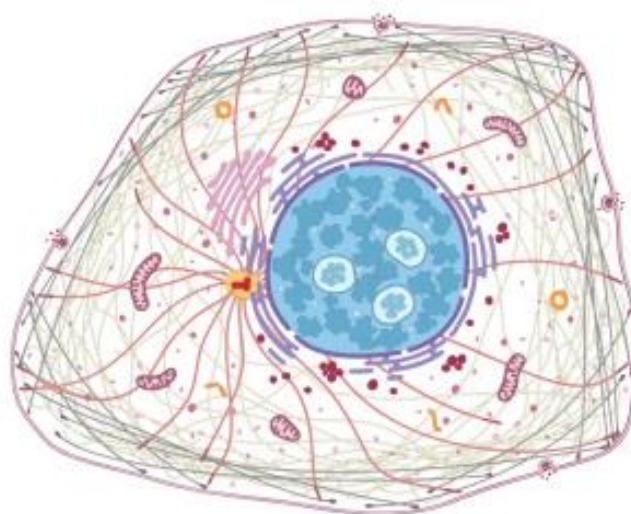
## LIST OF TABLES

<b>Tables</b>		<b>Pages</b>
<b>2.1.</b>	Models and their performance for top ranking and selected teams of HPA Kaggle Challenge	23
<b>3.1.</b>	HoliLoc's multi label human protein data summary	34
<b>4.1.</b>	Performance Comparison of HoliLoc and Individual Feature-based Models in the Single-location Prediction Settings	52
<b>4.2.</b>	F1 score performance comparison of HoliLoc, individual feature-based and class-wise score distribution for the top ten teams of HPA Kaggle Challenge. Comparing mean HPA kaggle challenge results and HoliLoc highest performance results are shown bold.	56
<b>4.3.</b>	Comparative analysis of Image, Sequence, PPI, and HoliLoc multi-location prediction models' average 10-Fold F1 scores across subcellular locations, bold entries signify top-performing models within each subcellular location.	59



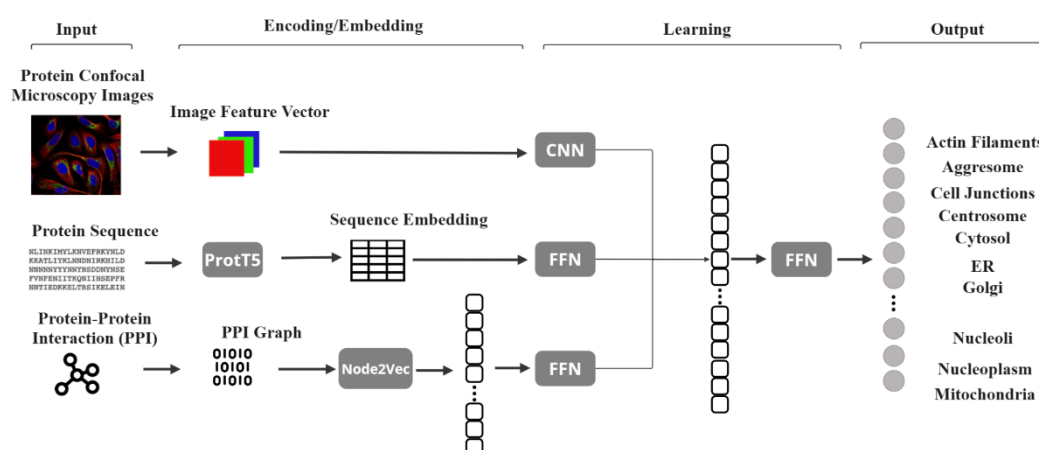
## 1. INTRODUCTION

Regions of eukaryotic cells differ both morphologically and functionally. To function properly, proteins need to be at the appropriate location. The compartments where proteins are found restrict the functionalities since they create a physiological environment (Figure 1.1). If a protein localization is improper, its function will be damaged and may result in a variety of disorders, including cancer, metabolic problems, and neurodegenerative diseases (1). Understanding where proteins are located inside the cell not only improves our understanding of the activities of certain proteins but also clarifies how cells are structured in general. For domains like drug detection, systems biology, and proteomics, understanding protein function is crucial knowledge. Both *in vitro* and *in vivo* methods can be used to identify proteins' subcellular localization. However, experimental techniques such as mass spectrometry and fluorescence-tagging methods are expensive and time-consuming. These methods, however, might be costly and result in unreliable data.



**Figure 1.1.** Representation of the subcellular locations of a eukaryotic cell. Adapted from (2).

To reduce cost, it is important to use prior in-silico tools to narrow down the research area. In accordance with this need, within the scope of this thesis study it is aimed to perform a new method for predicting protein subcellular localization via multi-modal deep learning. For this purpose, in this thesis, 3 different types of data are used: 2D confocal microscopy images (Human Protein Atlas), amino acid sequences (UniProt), and protein-protein interactions - PPIs (IntAct) to predict proteins' subcellular localization in a multi-label manner for 22 different cell compartments (Figure 1.2). HoliLoc takes protein data from 3 different modalities as input, encodes and embeds them, conducts learning separately on each data type, uses joint fusion for multi-modality, and transforms them into probabilities for 22 locations via another feed forward neural network (FFN). The system was trained in an end-to-end manner, and the performances were calculated on the unseen hold-out test dataset, which is significantly different from the training dataset.



**Figure 1.2.** The workflow of the proposed protein subcellular localization prediction method: HoliLoc.

### 1.1. Problem Definition and Motivation

Understanding the subcellular locations of proteins is essential for investigations in systems biology, proteomics, drug development, and protein function. Out of the 20,394 reviewed human proteins, 7,348 have localization annotations with experimental verification, according to UniProt (version 2020\_05)

(3). Additionally, there is very little data on sub-organellar compartment localization. The wetlab experimental approaches to studying SL of proteins are costly, and also labour and time-intensive. Due to this, new methods are required to aid researchers in this endeavour. Approaches based on artificial intelligence have evolved nowadays, and the biological data variety has grown, which presents a huge potential to advance protein SL-related work. However, existing SL prediction methods in the literature lack the necessary accuracy and coverage to be utilised in prospective studies in protein science. Therefore, it has become crucial to develop cutting-edge deep learning techniques that utilise the available data. Recently, the approach of predicting biological properties of proteins by combining various forms or types of data from genomics, proteomics and other omic types started to draw interest. To conclude, there is a current need to develop new computational approaches to accurately predict the SL of proteins with low resource requirements considering time and funds, using available large-scale and complex biological data and suitable algorithms from data science and artificial intelligence.

## **1.2. Aim and Scope**

The aim of this thesis is firstly to investigate the effect of utilising multiple sources of data on protein subcellular localization prediction task and secondly to create a multi-modal deep learning model that would provide high-performance protein subcellular location predictions. The main objectives are:

1. To construct a novel dataset combining protein confocal microscopy images, sequence embeddings, and protein-protein interaction (PPI) embeddings.
2. To utilise diverse deep learning methods together: convolutional neural networks (CNN), protein language models, graph learning, and feed forward neural networks (FFN).
3. To conduct multi-modular model fusion and observe the model effect of holistic data integration.
4. To predict protein subcellular localization in a multi-label manner for 22 locations.

5. To compare the effect of holistic data integration and multimodality on the most well-known Kaggle challenge benchmark dataset.
6. To prepare an easy to use programmatic tool for the researchers who are investigating unknown localizations of proteins.

To achieve this aim, image, amino acid sequence, and interactome/ protein-protein interaction (PPI) data are utilised together for human proteins, and deep learning models are constructed by taking simplicity into account to mainly observe the data diversity impact. Hence, three different data types are obtained from the public biological data sources and organised according to the scope in which all proteins belong to human and containing only one protein from a UniRef50 cluster to ensure each protein's uniqueness and mitigate the inclusion of closely related proteins that can lead to overfitting. The model architectures were specifically selected as simply as possible to observe the effect of a holistic data application approach. Performance assessment of HoliLoc was conducted as an inter-studies in which it was compared with the state-of-the-art protein subcellular location predictors as well as an intra-study in which individual models (each using a single data type) were compared with HoliLoc (modularly fusing individual models) in both the single-location and multi-location prediction settings. Therefore, this study allows us to see the effect of using different data types together in a challenging task for 22 different subcellular locations with extremely unbalanced data in a multi-class and multi-label manner.

### **1.3. Structure of the Thesis**

The “Introduction” section gives information about the protein subcellular localization, problem definition, aim and scope of this thesis. “Background Information” section describes more comprehensively protein subcellular localization and how it is interpreted in laboratories and introduces main databases used in this thesis as well as general deep learning information and state of art solutions in computational subcellular localization prediction and fusion strategies using deep learning. In the “Materials and Methods” section data preparation, dataset construction, data preprocessing, splitting and model structure overview is explained. There are detailed performance assessments and comparisons with the Kaggle

challenge in the “Results” section, as well as inter-study performance comparisons. Additionally, a use case study is presented in this section. The thesis' findings are analysed in the “Discussion” section in relation to the objectives of the study. Finally, potential future work and limitations are provided together with an overall summary in the “Conclusion” section.

## **2. BACKGROUND INFORMATION**

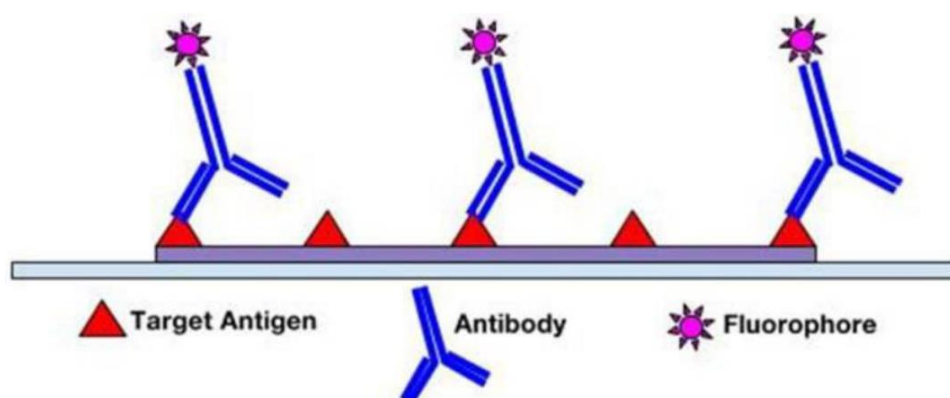
### **2.1. Protein Subcellular Localization**

Protein subcellular localization describes the particular compartment or area inside a cell where a protein is most frequently detected or localised. It is essential to how the protein interacts and works inside the cell. Many different compartments, including the nucleus, cytoplasm, mitochondria, endoplasmic reticulum, Golgi apparatus, lysosomes, peroxisomes, and plasma membrane, can be where proteins are located (2). Protein's amino acid sequence contains signals or specific sequences that determine where in the cell it will be found. These signals serve as molecular tags that point the protein in the direction of the correct cellular location (3). Specific protein complexes or machinery that promote the transport of the protein to the appropriate compartment can recognize the targeting sequences. Protein subcellular localization can be investigated using a variety of approaches, such as immunofluorescence microscopy, live-cell imaging, and fractionation procedures (4). In immunofluorescence microscopy, proteins are marked with fluorescent antibodies that attach to the target protein with high specificity, making it possible to see where the protein is located within the cell. Fluorescent protein tags that are genetically fused to the protein of interest are employed in live-cell imaging techniques. This makes it possible to see the protein's dynamic mobility in real time within the cell. Using fractionation procedures, cellular compartments are divided, and proteins from each fraction are isolated to ascertain their unique localization (5). Understanding protein function, cellular processes, and signalling cascades requires a precise understanding of protein subcellular distribution. This enables researchers to understand how proteins interact with other molecules and cellular organelles to understand their roles in diverse cellular processes and disorders.

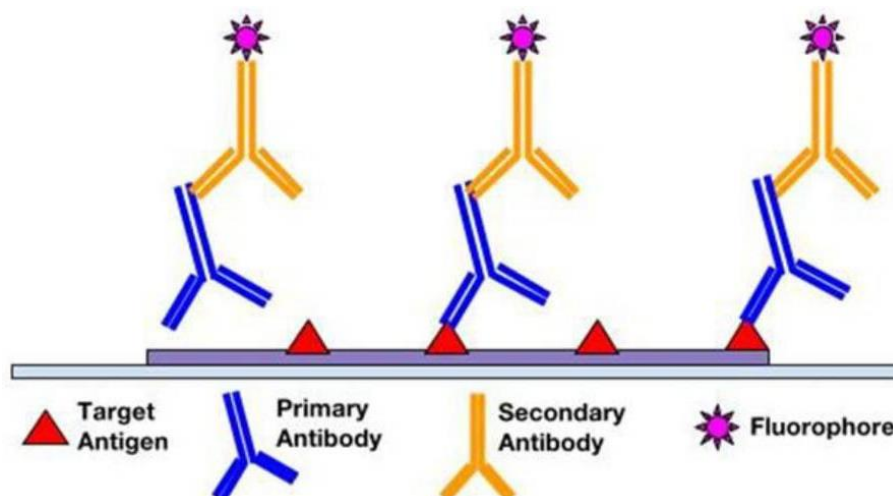
### **2.2. Immunofluorescence**

Immunofluorescence (IF) is a technique for locating and visualizing proteins or other antigens in cells. Using antibodies that uniquely recognize the desired target

of interest. The reporter fluorophore-conjugated antibodies make possible fluorescence detection under confocal fluorescence microscopy (Figure 2.1). Signal amplification, specific targeting, resolution, and analytical capabilities are the main advantages of the IF method. Different fluorophores sensitive to specific targets can be stained simultaneously which makes it ideal for co-localization studies. Cultured cells, cell suspensions, tissue samples and entire organisms are available for IF. IF can be direct or indirect depending on the usage of a secondary antibody (Figure 2.2). Indirect IF uses a secondary antibody, has more flexible usage and results in greater signal detection (6). IF is widely used in the field of protein subcellular localization. IF is used to systematically map the subcellular location of the human proteins in the Human Protein Atlas project (7). Target proteins are shown in green, nucleus in blue, microtubules in red and ER in yellow (8).



**Figure 2.1.** Direct immunofluorescence. Adapted from (8).



**Figure 2.2.** Indirect immunofluorescence. Adapted from. (8).

### 2.3. Human Protein Atlas Subcellular Section

Human Protein Atlas (HPA), a project that aims to map all the human proteins in cells, tissues, and organs using various technologies. The website (<https://www.proteinatlas.org/humanproteome/subcellular>) provides open access to data and images for researchers and the public. The database has twelve sections, each focusing on a different aspect of the human proteome, such as tissue expression, subcellular localization, immune cells, metabolic pathways, and 3D structures. The database also describes the methods and sources used to generate and analyse the data. The website is a valuable resource for understanding human biology and disease. Currently, the subcellular part of the HPA shows where proteins are found in different parts of the cells selected from 37 cell-lines. It uses IF and confocal microscopy to see the proteins in up to three different cell-lines to show how proteins from 13147 genes (65% of the human protein-coding genes) are in one or more of 35 different organelles and subcellular structures (2).

### 2.4. Interactome Data and IntAct

Molecular interactions are non-bonding interactions, noncovalent forces, and intermolecular forces among molecules (9). Interactome is all molecular interactions



in cells, especially protein-protein interaction which is the most prominent branch of this field (10). This information is important to understand the relationship between biological substances and molecular mechanisms of the cellular apparatuses.

Organelles and suborganelles create the ideal conditions for their particular functions and are the structural units of eukaryotic proteins (11). Proteins must naturally be located in the same subcellular location in order to interact (12). Since protein-protein interactions (PPIs) information is very crucial for protein subcellular location prediction, in this thesis, interaction information of human proteins is utilised. There are many publicly available databases for interactome data. Some well-known examples are STRING, MINT, IntAct, BioGRID and in this thesis the IntAct database is utilised since it is one of the most comprehensive databases in terms of the number of PPIs for 123071 proteins (13). The IntAct portal is a web-based resource for molecular interaction data, curated from literature or user submissions (<https://www.ebi.ac.uk/intact/home>). The IntAct portal provides a free, open-source database system and analysis tools for molecular interaction data. Users have access to the data in a variety of formats for browsing, downloading, and searching. IntAct is a member of the IMEx Consortium, an international collaboration between a group of major public interaction data providers who have agreed on sharing curation effort and practices (14).

## **2.5. Protein Sequences, Function and The UniProtKB**

Amino acids link together with covalent peptide bonds, forming a long chain that ultimately constructs proteins. For this reason, polypeptides are another name for proteins. Despite the fact that nature contains hundreds of different amino acids, only 20 amino acids are required to synthesise all of the proteins found in the human body and the majority of other living forms (15). The 3 primary classes of amino acids are hydrophobic, polar, and charged, according to their chemical content. Since they tend to repel water, hydrophobic amino acids, which include alanine, valine, leucine, isoleucine, methionine, proline, phenylalanine, and tryptophan are frequently found inside proteins and help to stabilise their structural integrity. Serine, threonine, cysteine, tyrosine, asparagine, and glutamine are examples of polar amino acids that

have hydrophilic qualities and can interact with water molecules. Charged amino acids are essential for the formation of ionic interactions within proteins. Examples of charged amino acids include lysine, arginine, histidine, aspartic acid, and glutamic acid. Furthermore, glycine is a special kind of amino acid that is regarded as neutral but, because of its small size and distinct structure, can display both hydrophobic and polar properties. There are known to be tens of thousands of distinct proteins, each with a unique amino acid sequence (16). These amino acids work together to give proteins their varied and complex structures, which allow them to perform vital biological tasks. The sequence itself and how this chain is folded are 2 of the main features of protein functionality. There are many mechanisms affecting protein localization in the cell; however, the most important ones are pre-sequences, which are found at the N- or C-terminus of a protein sequence and internal signals, which are found in the middle of the protein sequence. Those sequences are used as part of the cell signalling mechanism and bring proteins to their destination (3). Universal Protein Resource Knowledge Base (UniProtKB) is a protein database that contains information on the sequence and function of proteins (<https://www.uniprot.org/>). The European Bioinformatics Institute, The Protein Information Resource, and the Swiss Institute of Bioinformatics collaborate to create UniProt. Over 100 individuals are engaged in various roles within all three institutes, including software development, support, and database curation. UniProt offers a central repository of reliable, extensive, publicly available information on protein sequences and functional annotation. It also provides tools and services for searching, analysing, and downloading protein data. UniProtKB also provides comprehensive information on proteins, such as their interactions, and subcellular locations. UniRef clusters similar protein sequences based on sequence identity, and UniRef information can be obtained from UniProtKB, which provides clustered sets of all protein sequences from the UniProtKB and selected UniProt archive records to obtain complete coverage of sequence space at resolutions of 100%, 90%, and 50% identity (17).

## **2.6. Gene Ontology (GO) Database**

Ontology is a representation of a community's shared background knowledge (18). A framework and a set of concepts for defining the functions of gene products from all organisms are provided by the Gene Ontology (GO). The definition of "function," however, is more complicated than it first appears. In the context of molecular biology, function describes particular and coordinated actions for a specific purpose. Molecular function, cellular component, and biological process are the three different ways that gene functions can be explained that are considered by the GO. The gene product in this representation is responsible for carrying out a molecular-level activity (molecular function) at a particular location in relation to the cell (cellular component). This molecular process is part of a larger biological objective (biological process) that is made up of several molecular-level processes. A single macromolecular machine can perform a molecular function by directly interacting physically with other molecules. The biological process is the largest of those aspects in the GO, and the one with the highest diversity (19). Computational representation of the functions fulfilled by proteins and non-coding RNA molecules produced by genes from a wide range of organisms, including bacteria and humans, can be found in the Gene Ontology knowledgebase (<https://geneontology.org/>). Currently, the GO consists of over 7,648,957 annotations of which 2,862,941 belong to biological processes, 2,442,438 belong to molecular function and 2,343,578 belong to cellular components. 951,061 of all annotations experimentally supported annotations from over 175,927 published papers (20).

## **2.7. Machine Learning and Deep Learning**

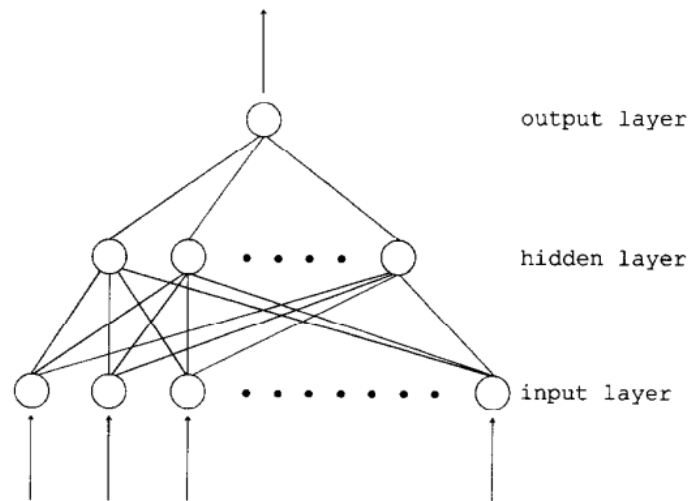
Artificial Intelligence (AI) includes any methodology that allows computers to imitate human behaviour and go above human decision-making abilities to solve complex problems autonomously or with minimal human involvement (21). Machine learning (ML) is a subclass of artificial intelligence that is capable of self-learning. Without human intervention, ML models gain experience and become more intelligent during training. This is achieved through the utilisation of algorithms for statistical learning, enabling the models to autonomously learn and improve without the need for human assistance. Deep learning (DL), on the other hand, relies on large

datasets or a lot of information provided as input to learn from experience. While shallow neural networks only have a maximum of two layers between the input and output of the neural network, deep neural networks have multiple layers for multiple data transformations between the input and output (22). Typically, deep neural networks have multiple hidden layers and advanced neurons, which are mathematical representations of connected processing units (23). Some well-known machine learning algorithms are support vector machine (SVM), K-nearest neighbour algorithm (K-NN) and decision tree (DT) algorithms. K-NN algorithm predicts the relationship between the unknown data and the known data for a given dataset. Then, it imputes the new data to the existing category that most closely matches it. The goal of the SVM algorithm is to create the best hyperplane, or decision limit, that divides n-dimensional space into distinct classes and makes it simple to assign a different point to the appropriate category. Recursively dividing the data into progressively smaller subsets according to the feature values is how the DT algorithm works. The algorithm selects the feature at each node that divides the data into groups with distinct target values the best (24). For the majority of applications where text, image, video, speech, and audio data needs to be processed, deep neural networks perform better than shallow ML algorithms because DL is especially helpful in domains with large and high-dimensional data (25). Rather than employing a straightforward activation function, they might make use of sophisticated operations like convolutions or numerous activations in a single neuron. These features enable raw input data to be fed into deep neural networks, which then automatically identify the representation required for the associated learning task (23).

### **2.7.1. Feedforward Neural Network (FFN)**

Neural units in a feedforward neural network (FFN) are arranged in layers. The layers that are between are known as hidden layers, and the topmost layer is referred to as the output layer. Information travels through one or more hidden layers and the input layer in this architecture, then proceeds in a unidirectional fashion to the output layer (Figure 2.3). Activation functions are used by each neuron to process input data, and through training, the network adjusts weights and biases to learn the

relationships between input and output. FFN is very adaptable and can be used for a variety of tasks including classification and regression (26).

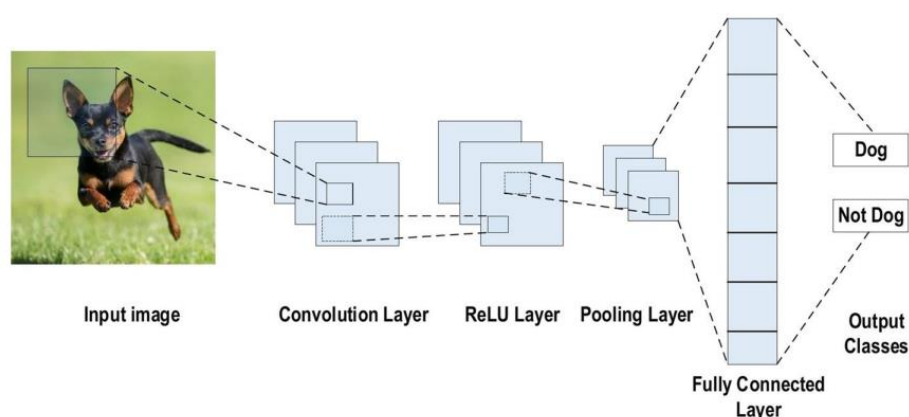


**Figure 2.3.** Architecture of a FFN composed of three layers. Adapted from (26).

### 2.7.2. Convolutional Neural Network (CNN)

An example of an artificial neural network created especially for computer vision applications is the convolutional neural network (CNN), which is used for image recognition and classification. Because CNNs can automatically and adaptively learn the spatial hierarchies of features from input data, they have been shown to be very effective in these tasks. The convolutional layer is the fundamental component of CNNs. In convolution, input images are convolved with filters or kernels to extract features. When a  $f \times f$  filter is used to convolve a  $N \times N$  image, the same feature is learned across the board. Following each operation, the window slides, and the feature maps learn the features. Using shared weights and biases, the feature maps capture the image's local receptive field (27). Convolution, pooling, fully connected, and nonlinearity are the four layers that make up CNN (Figure 2.4). Following convolution, the output is made simpler by the pooling layers, reducing the amount of computation required and the number of parameters by creating downsampling representations. The fully connected layer, also referred to as the

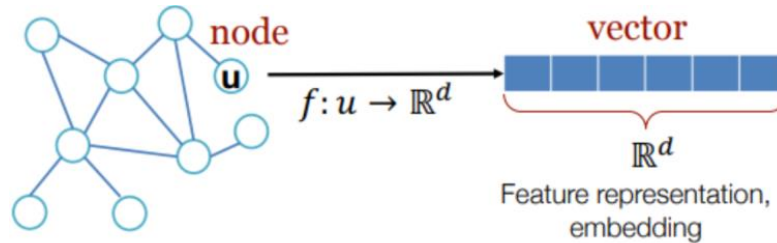
convolutional output layer, gets input from the convolutional output layer or final pooling and flattens it before sending it to the next layer. In a neural network, every neuron node takes the output value from the neuron in the layer above as its input value and sends it to the neuron in the layer below. The value of the input attribute will be directly passed to the following layer by the input layer neuron node. The input of the lower node and the output of the upper node in a multilayer neural network are functionally related. The activation function is the name given to this function (28). ReLU, Sigmoid, Tanh, and Softmax functions are commonly used by CNNs to add non-linearity to the network; hence, it can learn intricate patterns (29). After several convolutional and pooling layers, CNNs usually end with one or more fully connected layers, which perform the final classification based on the features learned in the preceding layers. Data is flattened into a vector before being passed from convolutional and pooling layers to fully connected layers. This is required because one-dimensional input is required for fully connected layers. A regularisation method called dropout is employed to prevent overfitting, which temporarily removes units from a neural network and improves the network's ability to generalise to new, untested data (30). The output layer, which is the final layer in the CNN architecture, is where the final classification is accomplished. The output layer of the CNN model uses loss functions to determine the expected error that was generated throughout the training set of data. The selection of hyperparameters significantly impacts the performance of CNN. The overall performance of CNN can be impacted by even minor adjustments to the hyper-parameter values. Because of this, selecting parameters carefully is a very important factor that needs to be taken into account when developing an optimization scheme (29).



**Figure 2.4.** Architecture of a CNN. Adapted from (29).

### 2.7.3. Graph Learning

In computer science, a graph is a fundamental data structure made up of nodes and edges. Nodes stand for objects or points, and edges indicate the connections or relationships between these objects. A node is essentially a basic unit, and an edge is a connection or channel of communication that represents the relationships between two nodes (Figure 2.5). Graph representation makes it possible to efficiently store and retrieve the relational knowledge of interacting entities. Data analysis of graphs can offer important new information about community detection, behaviour analysis, as well as additional practical uses like clustering, link prediction, and node classification. The goal of graph representation learning, also known as graph embedding, is to map each node to a vector while maintaining the distance characteristics between nodes. Graph embedding has been studied since the early 1900s. Since then, numerous approaches have been proposed. Matrix factorization, random walk, and dimensionality reduction are some of the well-known graph embedding techniques. A graph can be traced by creating multiple paths by starting random walks from random initial nodes. These paths show the connected edges' context. By passing through nearby edges, one can explore the graph and gather both local and global structural information because of the randomness of these walks. Subsequently, random sampling paths are subjected to probability models such as skip-gram and bag-of-words to acquire node representations (31).

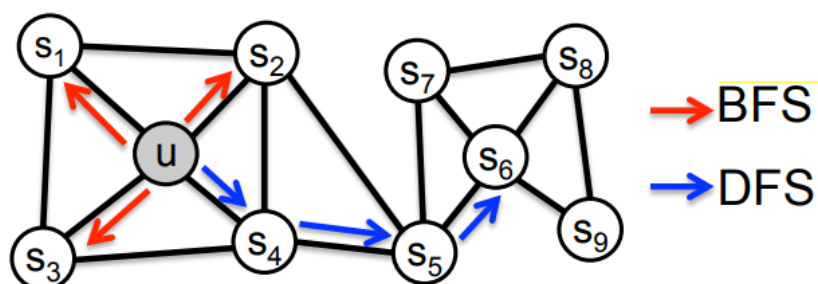


**Figure 2.5.** Illustration of graph representation learning input and output. Adapted from (31).

#### 2.7.4. Node2Vec

One of the most well-known random walk-based methods for scalable feature learning in networks is Node2Vec. Node2Vec discovers continuous vector representations for nodes by utilising stochastic traversal strategies to explore the network. To create neighbourhood sets, two extreme sampling techniques are typically used. Depth-First Sampling (DFS), in which the neighbourhood is made up of nodes that are successively sampled at increasing distances from the source node, and Breadth-First Sampling (BFS), in which the neighbourhood is limited to nodes that are the source's immediate neighbours (Figure 2.6). The flexible neighbourhood sampling strategy of Node2Vec enables seamless interpolation between BFS and DFS accomplished by creating a versatile biased random walk method that can investigate neighbourhoods in both a BFS and a DFS manner (32). Hence, Node2Vec generates embeddings that function as rich feature representations, facilitating enhanced performance and generalisation across various graph-based tasks.





**Figure 2.6.** BFS and DFS search strategies. Adapted from (32).

### 2.7.5 Sequence Embeddings

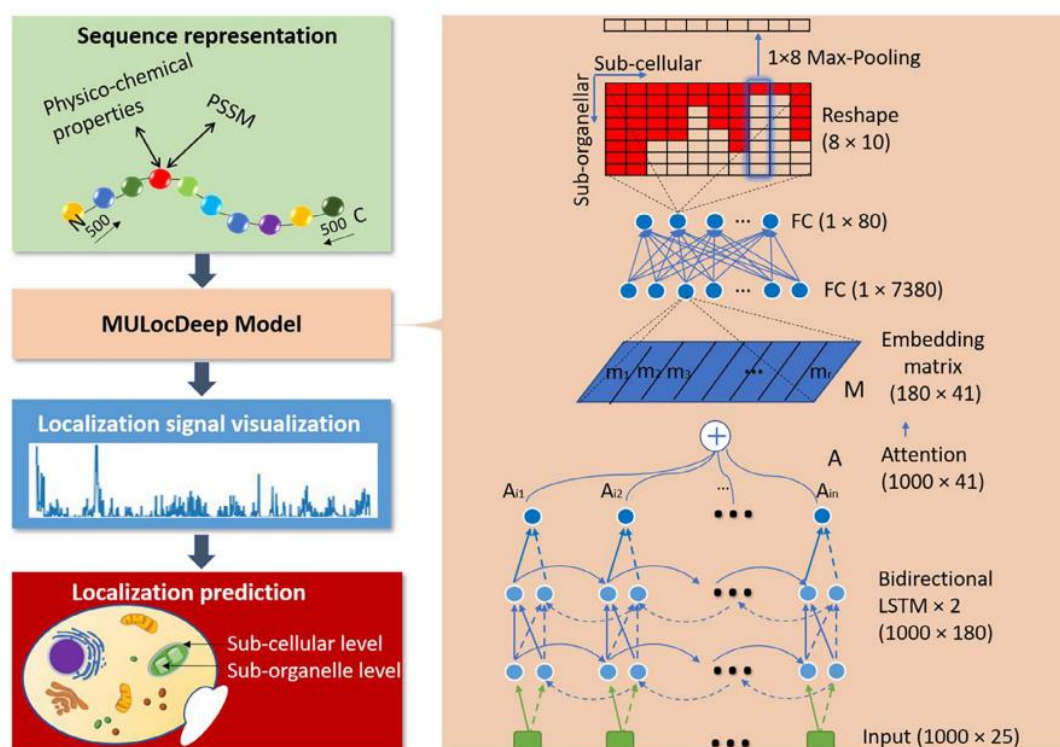
Natural Language Processing (NLP) gives computers the ability to understand text and speak words similar to what humans can. This led to the adaptation of language models (LMs) to encode implicit language encoded in protein sequences. Protein LMs exhibit great promise in producing descriptive representations (embeddings) for proteins based solely on their sequences faster compared to earlier methods, while maintaining comparable or enhanced predictive capabilities. Many protein LMs that have been trained by researchers are likely to shed light on various facets of the protein language. Protein sequences are made up of character strings that stand for individual amino acids. LMs are trained to represent language by reconstructing corrupted text or by anticipating the next word in a sentence based on its preceding context. LM representations, or embeddings, have been a source for various techniques (33). Predictions based on embeddings are faster than those based on evolutionary information, but they are typically less accurate (36). Some of the very well-known LMs are T5 (34), Electra (35), BERT (36), Albert (37), Transformer-XL (38) and XLNet (39). The de-facto standard for transfer learning in natural language processing, BERT was the first bidirectional model in NLP to attempt to reconstruct corrupted tokens. To increase the number of attention heads, Albert reduced the complexity of BERT by forcing hard parameter sharing between its attention layers. By training two networks, a discriminator and a generator, Electra attempts to increase the pre-training task's sampling efficiency. As

an alternative to simply reconstructing corrupted input tokens, the discriminator (Electra) identifies which tokens were masked, while the generator BERT reconstructs masked tokens as well. A transformer model is a type of neural network that tracks relationships in sequential data to learn meaning and context. T5 makes use of the original transformer architecture for sequence translation, which is made up of a decoder that creates a translation to a target language based on the encoder's embedding and an encoder that projects a source language into an embedding space. For every attention head, T5 learns a positional encoding that is shared by all layers. Protein predictions for datasets could be made more quickly and affordably by using sequence embeddings as the input to relatively small-size CNN/FNN for secondary structure, localization, and classification predictions (40).

## **2.8. Related Work**

### **2.8.1. MULocDeep**

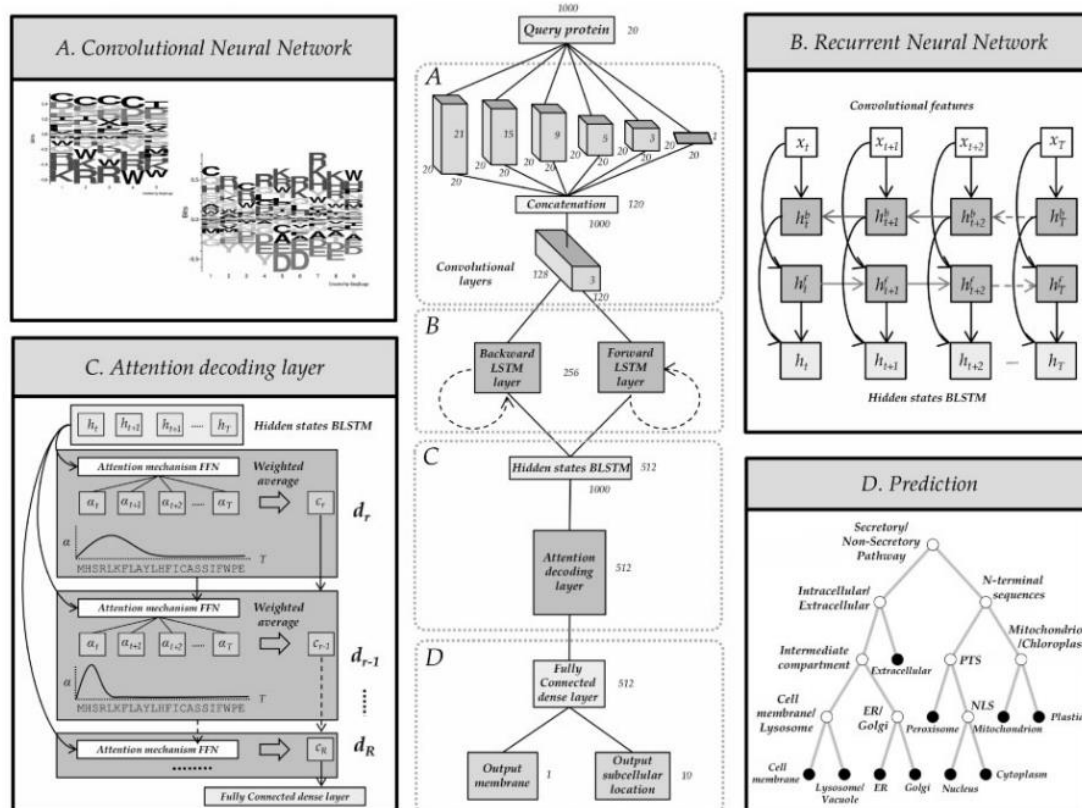
MULocDeep is a deep learning-based protein localization prediction framework that predicts the localization of proteins in both subcellular and suborganellar manner. Main MULocDeep model is the bidirectional Long Short Term Memory (LSTM) which handles protein sequence information, and the multihead self-attention to assign weights to each amino acid of a sequence for interpretation. It also provides a web server for users to submit protein sequences and visualise the results. A dataset containing eukaryotic species' proteins in 44 suborganellar compartments in 10 subcellular localizations with experimental evidence from the UniProt database was collected, which they called the UniLoc dataset. It is trained and tested with the UniLoc dataset. MULocDeep performs better than other major methods at both subcellular and suborganellar levels in most cases. MULocDeep also identified some known and novel localization signals from the attention weights, which could provide insights into the mechanism of protein sorting and localization (3).



**Figure 2.7.** MULocDeep workflow and neural network architecture. Adapted from (3).

### 2.8.2. DeepLoc

DeepLoc is a deep learning method for predicting protein subcellular localization from amino acid sequences that uses a recurrent neural network (RNN) with an attention mechanism to identify protein regions that are important for localization. The method also uses a CNN to detect protein motifs and a hierarchical tree to model the protein sorting pathways. The method is trained and tested on a new dataset extracted from UniProt, where proteins have experimental evidence for their subcellular locations. The method is available as a web server and a code example. The method predicts where proteins are in eukaryotic cells based on their amino acid sequences. It uses CNNs to find short motifs and RNNs to scan the sequence in both directions. It also uses an attention layer to focus on the important 20 regions and a dense layer to output the location. The method has two outputs: 1 for membrane-bound and 10 for different locations (41).

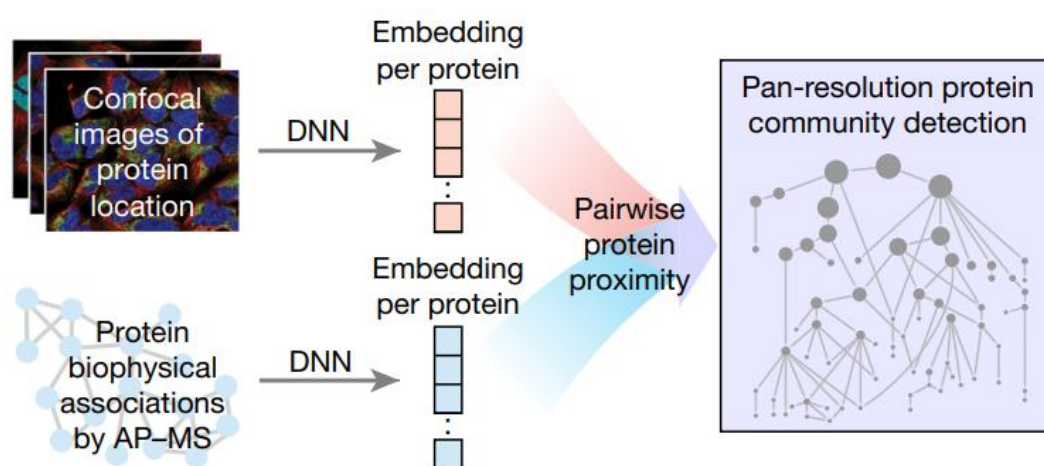


**Figure 2.8.** Overview of DeepLoc. Adapted from (41).

### 2.8.3. MuSIC

MuSIC is a method to map the cell structure at different scales by combining images and interactions of proteins. It uses deep neural networks to create embeddings for proteins based on their immunofluorescence and affinity purification data, then adjusts and merges the embeddings to show subcellular systems at various scales. A matched dataset of immunofluorescence images from HPA and affinity purification-mass spectroscopy data from BioPlex is used. MuSIC offers a new way to study cell organisation and function across multiple scales. It also explores the benefits and challenges of using different kinds of data to map the cell structure and proposes future directions to include more data types and deal with cellular diversity and dynamics. For image embedding, it uses DenseNet7, a convolutional neural

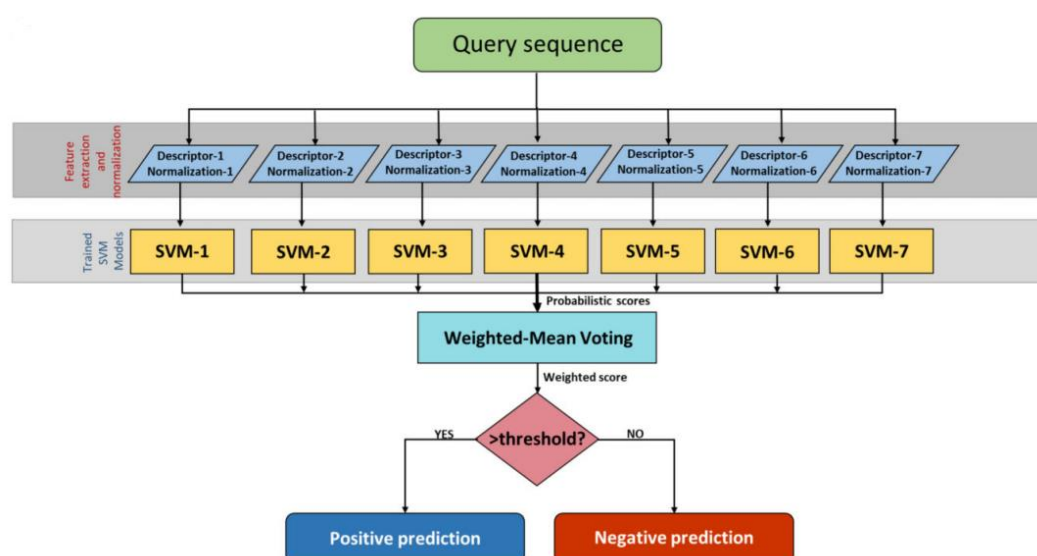
network that performs better than cellular markers in capturing protein locations. Also, it uses a node2vec neural network to embed each protein based on its extended affinity purification and mass spectroscopy interaction neighbourhood (42).



**Figure 2.9.** Overview of multi scale integrated map of the cell: MuSIC. Adapted from (42).

#### 2.8.4. SLPred

SLPred is a sequence-based multi-view and multi-label protein subcellular location prediction ensemble machine-learning system which can predict all possible subcellular localizations of a protein for 9 main subcellular locations (cytoplasm, nucleus, cell membrane, mitochondrion, endoplasmic reticulum, secreted, Golgi apparatus, lysosome, and peroxisome). 9 independent protein sequence-based machine learning models using SVM produce a binary prediction for each subcellular location and threshold applied, and the weighted mean of votes coming from those models is given as the protein's all possible subcellular localizations (43).



**Figure 2.10.** Schematic representation of the subcellular localization predictor SLPred. Adapted from (43).

### 2.8.5. Human Protein Atlas Image Classification Kaggle Challenge

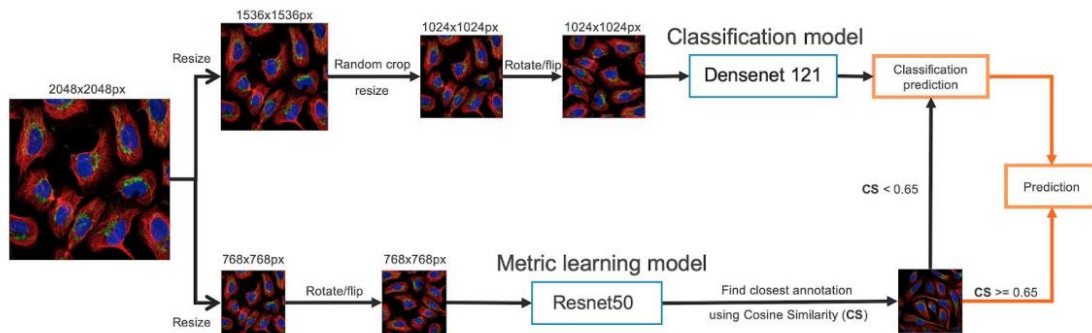
This competition was held with the aim of finding deep learning solutions for classifying protein subcellular localization patterns in fluorescence microscopy images from the Human Protein Atlas (HPA) project. The competition faced two main challenges: the multi-label problem, where each image can have multiple labels, and the class imbalance problem, where some labels are much more frequent than others. Participants were permitted to use any external data, including the approximately 78,000 images that are publicly accessible on the HPA Cell Atlas (HPAv18), and 42,774 non-public images (44). The performance was measured with macro F1. The score for human experts is 0.71 while the winning team has macro F1 scores 0.59. (Table 2.1). Even though the best model was unable to perform at a human expert level, the challenge still provides opportunities for the advancement of cell biology.

**Table 2.1.** Models and their performance for top ranking and selected teams of HPA Kaggle Challenge. Adapted from (44).

Rank	Team Name	Score
1	Team 1: bestfitting	0.593
2	Team 2: WAIR	0.571
3	Team 3: pudae	0.570
4	Team 4: Wienerschnitzelgemeinschaft	0.567
5	Team 5: vpp	0.566
8	Team 8: One More Layer (Of Stacking)	0.563
10	Team 10: conv is all u need	0.557
16	Team 16: NTU_MiRA	0.553
39	Team 39: Random Walk	0.540

### 2.8.6. Team 1 (bestfitting)

The labels of each sample were predicted using a CNN multi-label classification model, and the closest sample was found for comparison using a metric learning model. The HPAv18 dataset and competition data were divided into training and validation sets using multi-label stratification. Focal loss over the validation set was used to estimate the model's performance. A Densenet121 serves as the model's backbone. The final CNN feature map's GlobalMaxPool and GlobalAvgPool layers were concatenated before being fed to two fully connected layers, which determined the probability of each class. Augmentation implemented. The model was trained using a combined loss function consisting of focal loss, Lovasz loss, and log loss. The ratio of labels in the training set was used to threshold the output (44).

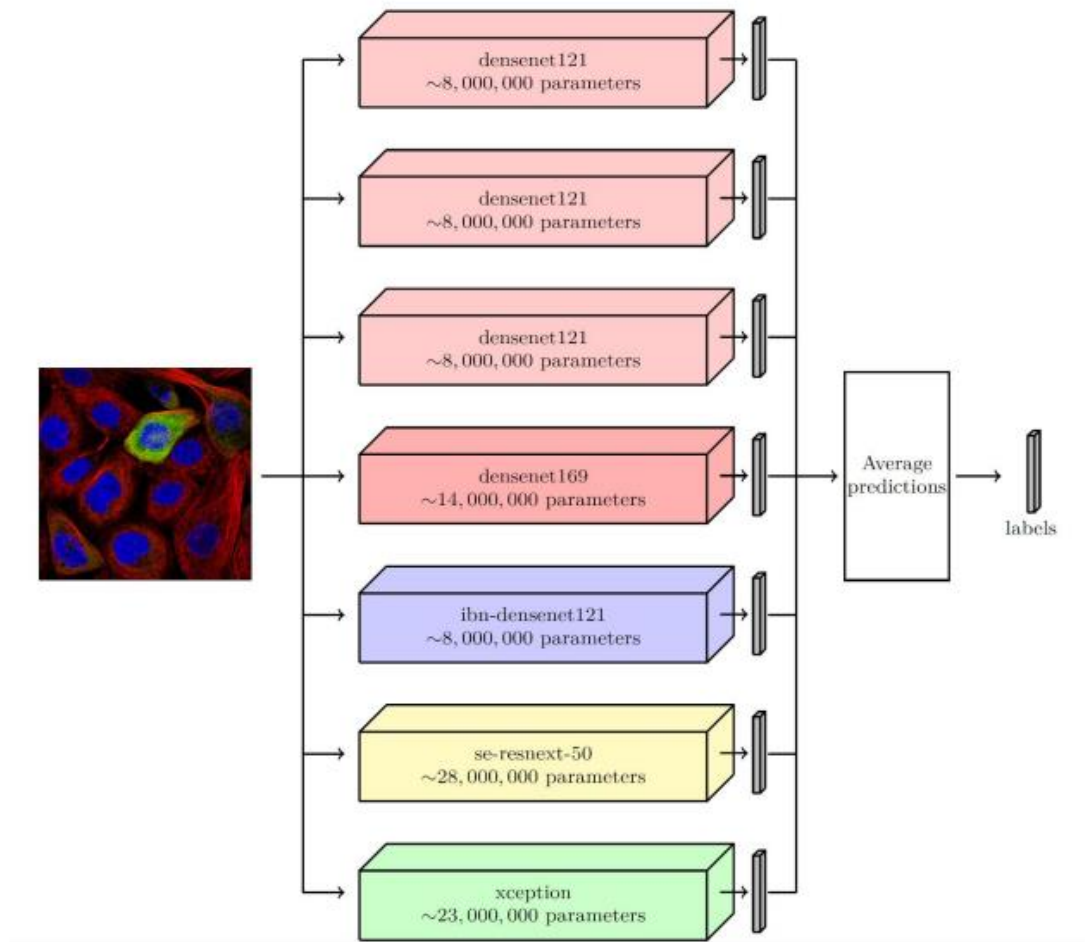


**Figure 2.11.** Model architecture of team 1: bestfitting. Adapted from (44).

### 2.8.7. Team 2 (WAIR)

The Kaggle challenge data and the external HPA $\nu$ 18 dataset were used to train and assess an ensemble of CNN models with a variety of architectures. To address the multi-label classification problem, 7 end-to-end models were developed. Throughout the seven models, a total of 5 distinct architectures were used: 3 instances of densenet121, densenet169, ibn-densenet121, se-resnext50, and Xception 3,4,5,6. The backbone models were pre-trained ImageNet models. Each model's final fully connected layer was swapped out for a fully connected layer consisting of 28 neurons, and each neuron's sigmoid activation function was connected to produce the final predicted probability for each image class (44).

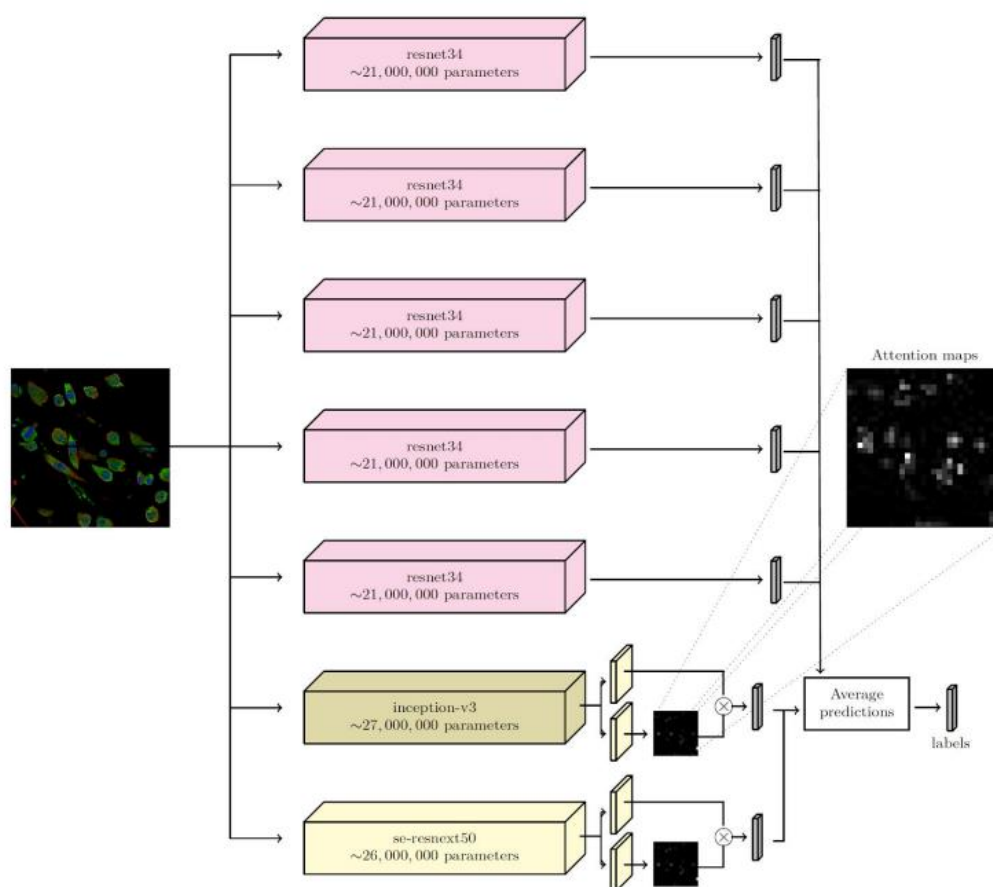




**Figure 2.12.** Model architecture of team 2: WAIR. Adapted from (44).

### 2.8.8. Team 3 (Pudae)

Both attention-gated CNN models and an ensemble of regular models were employed. The ensemble consisted of seven different models: one attention-gated inceptionv3, one attention-gated se-resnet50 and five resnet34 models, which served as the model's framework. The output from each model is averaged to produce the final forecast. Focal loss was used to address the class imbalance, negating the need for other techniques like undersampling and oversampling (44).

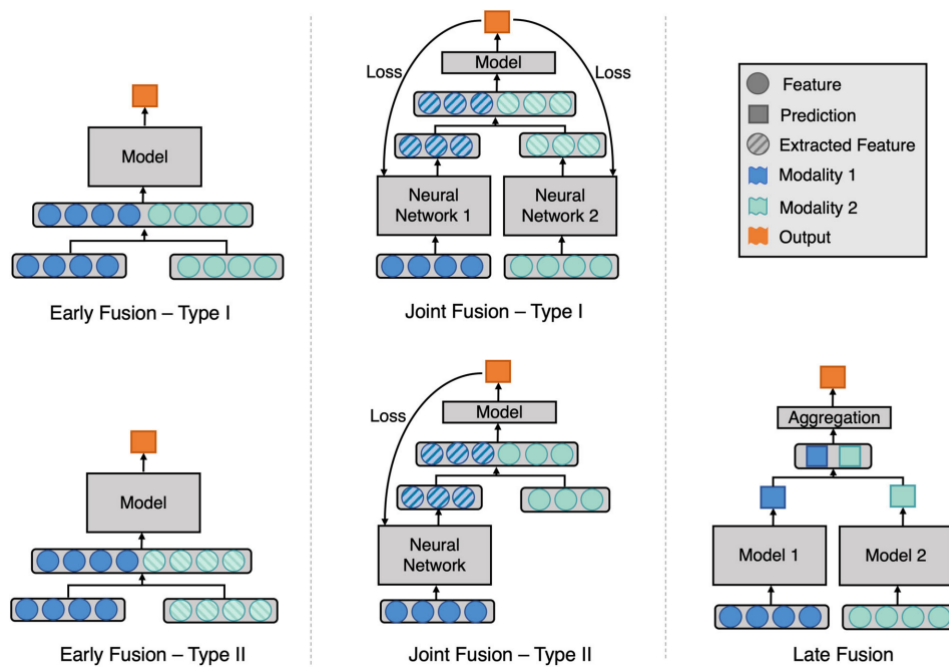


**Figure 2.13.** Model architecture of team 3: pudae. Adapted from (44).

## 2.9. Fusion Strategies Using Deep Learning

Data fusion involves combining information from various sources to enhance machine learning models' performance by extracting complementary and more comprehensive data, in contrast to relying on a single data modality. This trend is evident in the latest medical imaging literature, where the “fusion paradigm” integrates both electronic health records and pixel data to address complex tasks beyond the capabilities of individual modalities. 3 primary fusion strategies are early, joint, and late fusion. Early fusion, also known as feature-level fusion, integrates various input modalities into a feature vector before training a ML model. This can be achieved through methods like pooling, concatenation or using a gated unit. Early

fusion type-1 fuses original features, while early fusion type-2 incorporates features extracted from manual processes such as imaging analysis software, or learned representations from other neural networks, including predicted probabilities. Joint fusion, or intermediate fusion, merges learned feature representations from intermediate neural network layers with features from other modalities as input to a final model. Joint fusion type-1 involves extracting feature representations from all modalities. However, not all input features require the feature extraction step for joint fusion in type-2. Late fusion refers to using predictions from multiple models to make a final decision, also known as decision-level fusion. Different modalities train separate models, and an aggregation function combines their predictions. Examples of aggregation functions include averaging, weighted voting, majority voting or a meta-classifier based on each model's predictions. The choice of the aggregation function is empirical, varying based on the application and input modalities (45).



**Figure 2.14.** Fusion strategies using deep learning. Adapted from (45).

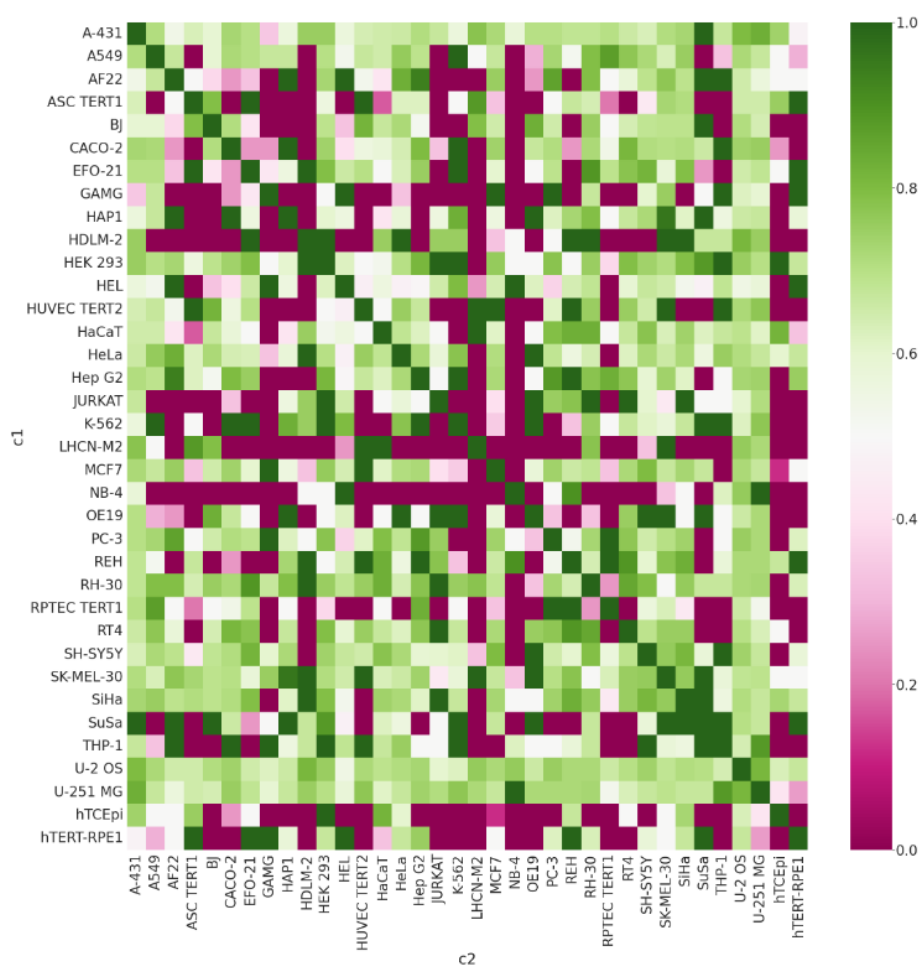
### 3. MATERIALS AND METHODS

#### 3.1. HoliLoc Dataset Construction and Splitting

HoliLoc's in-house dataset is constructed by using 3 different types of data: amino acid sequences extracted from UniProt, 2D confocal microscopy images from the HPA, and PPI data obtained from IntAct. It is notable that, due to data volume constraints, this integration is restricted to human proteins. Subcellular location data is sourced from the Human Protein Atlas, which contributes location insights to the dataset. After constructing the main HoliLoc dataset splitting was conducted just once to ensure all data preprocessing was conducted on separate train and test data. The final dataset, consisting of 9182 proteins, was partitioned, with 10% reserved for testing, while the remaining 90% was retained for training to ensure reliable model evaluation (Table 3.1). In the final step of dataset construction, GO-based cellular component annotations obtained from the HPA are compared with cellular component annotations directly from the GO database for each protein. If all cellular component annotations for a protein obtained from the HPA are present in the GO database's cellular component annotations, it is assumed that the GO database has covered the protein's HPA information. However, if at least one cellular component annotation from the HPA is not found in the GO database's cellular component annotations, the protein is considered uncovered. When the percentage is calculated across all HoliLoc proteins, it is observed that GO cellular component annotations cover 50% of the proteins in terms of HPA cellular component annotations. Therefore, we've chosen to utilise HPA as our source of subcellular localization information in HoliLoc's train and test datasets. This decision aligns with the HPA Kaggle challenge conditions, which is a crucial subject of inter-study comparisons and ensuring consistency, while avoiding potential issues arising from coverage discrepancies.

### 3.1.1 Image Data

The first step of HoliLoc's data construction was image data collection. HoliLoc's image data is obtained from the Human Protein Atlas version 21.0 Release date: 2021.11.18. The subcellular section of HPA in XML format and parsed with the Python ElementTree library to obtain desired information, including location, GO ID, cell-line, URL of IF images, and converted to a data frame with the Pandas library. In the first stage, a mixed and large data set was obtained. This data is composed of unique, 12761 proteins, from 2 different organisms and 36 cell-lines. The HPA data provides multiple options of cell-lines for each protein. However, all cell-line combinations were not possible for each protein; hence, location similarity of cell-lines investigated. Similarities between the cell-lines on how many locations they shared are investigated and reported with a heatmap (Figure 3.1). Overall similarity among cell-lines calculated by determining the percentage of shared location information relative to all possible combinations of shared location information. The average protein localization similarity was found to be 0.66 among all cell-lines. Hence all cell-lines treated the same and only one cell-line was selected per protein, which has the highest number of location annotations. Also, more than 99% of the data belongs to human. Hence, only human proteins are filtered. The dataset initially comprised data from the HPA with 12761 unique Uniprot IDs, totalling 82439 protein-cell-line pairs, including duplications. Following a cleaning and simplification process, the dataset was refined to include 12759 unique proteins.



**Figure 3.1.** Heatmap illustrating the shared subcellular locations among various cell lines c1 and c2. Each cell in the heatmap represents the percentage of protein localization similarity between corresponding cell lines.

### 3.1.2. Interactome (PPI) Data

Interactome (PPI) data addition to HoliLoc data was the second step. From the IntAct interactome information obtained as tab delimited text file for 19236 unique human proteins. This file was merged with the image data and 10201 unique proteins obtained. Hence, a data frame with the IF image URL and PPI information was obtained. In this step a UniRef investigation was also conducted. UniRef50 information was obtained from UniProtKB

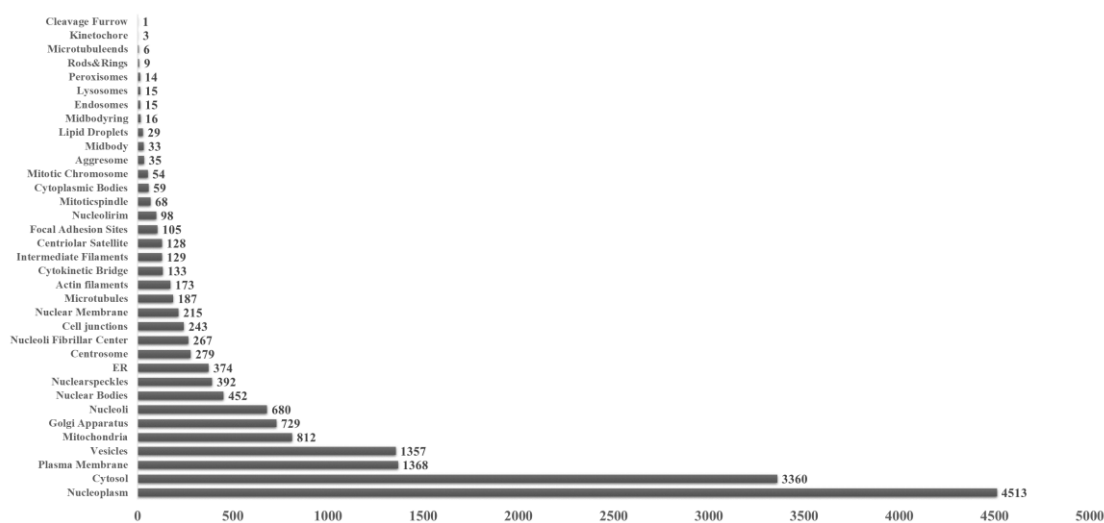
(<https://www.uniprot.org/help/downloads>) and only 1 member of each cluster was selected randomly. With this procedure, 9187 unique proteins remained.

### **3.1.3. Sequence Data**

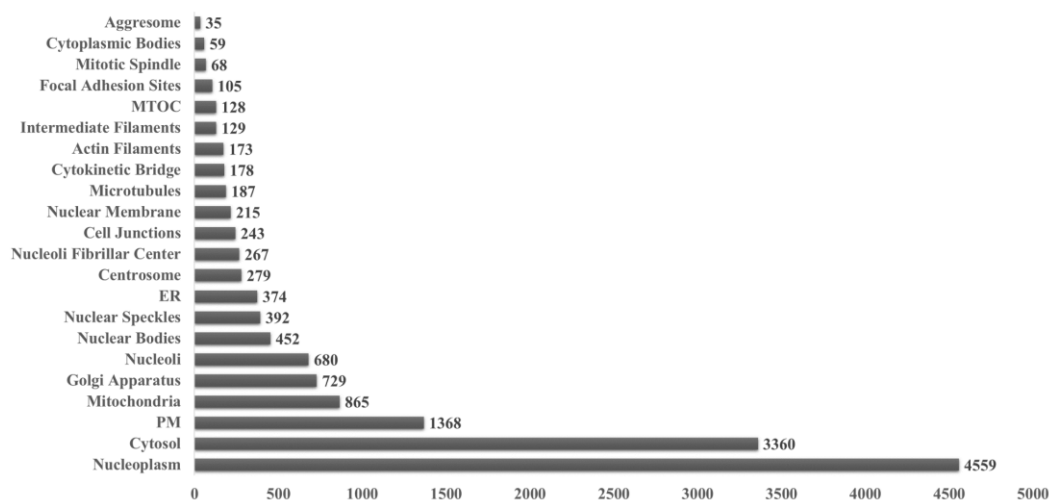
Sequence data addition is the third step of data construction. In this step amino acid sequence embeddings produced with the ProtT5 protein language model (protein level embeddings) were obtained directly from the UniProtKB protein embeddings section (<https://www.uniprot.org/help/embeddings>) with the instructions UniProt provided. Those sequence embeddings were merged with the image and interactome data of HoliLoc. This merging leads to the overall unique protein data count of HoliLoc 9182, which is the final HoliLoc dataset to be divided to train and test.

### **3.1.4. Arrangement of Location Classes**

Vesicles and Kinetochore are dropped from the study because of low sample size and being unable to group with others. The rest is grouped according to their biological relevance and the HPA Kaggle Challenge Study's approach. After grouping, locations with less than 30 proteins were also dropped. (endosomes, lipid-droplets, lysosomes, microtubule-ends, peroxisomes, rod-rings). Hence, the location classes count dropped from 35 (Figure 3.2) to 22. (Figure 3.3). Approximately 10% of each location is allocated to the test dataset, while the rest forms the training data. This approach ensures a balanced representation of locations in both datasets. (Figure 3.4).

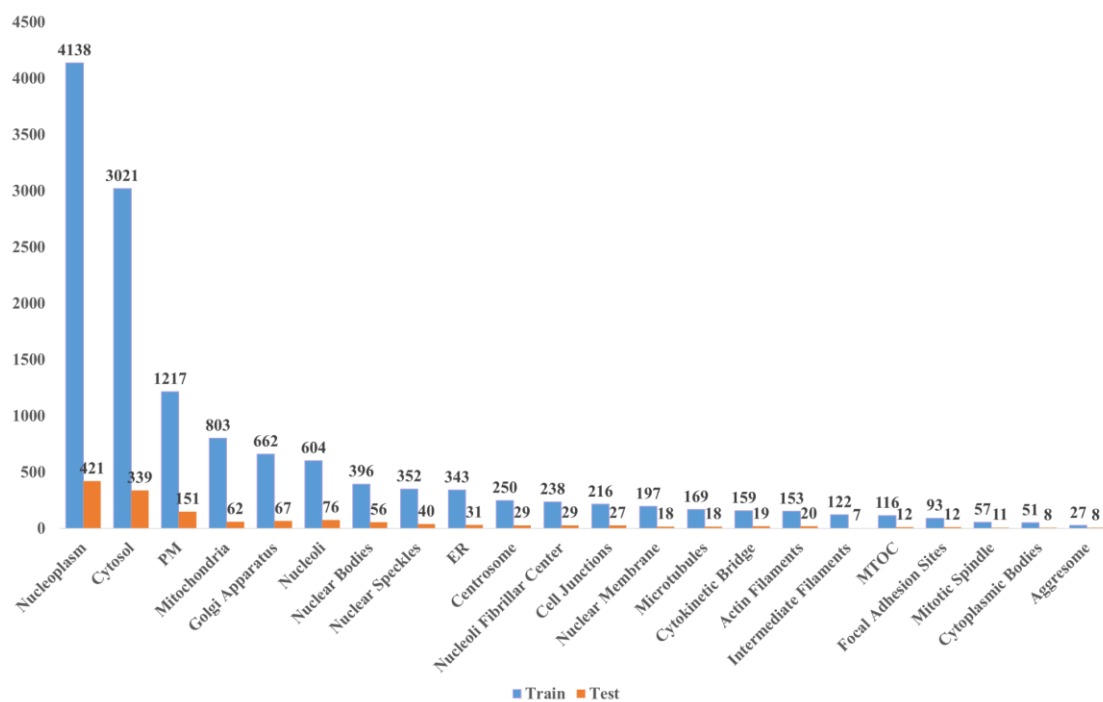


**Figure 3.2.** Bar graph displaying the number of proteins associated with each subcellular location before arrangement of location classes.



**Figure 3.3.** Bar graph displaying the number of proteins associated with each subcellular location after arrangement of location classes.

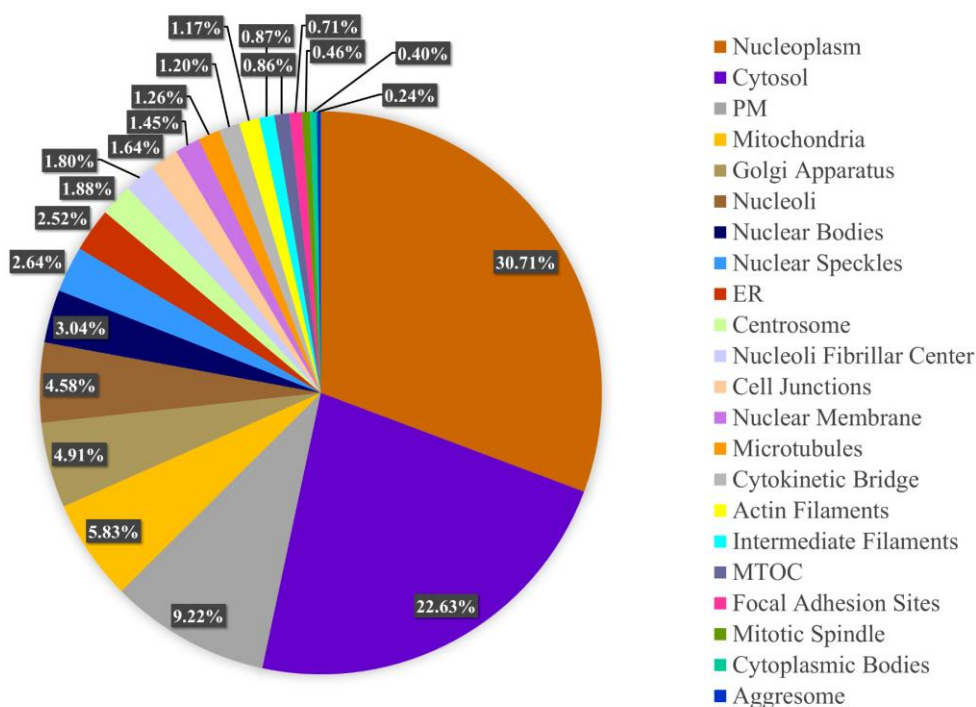




**Figure 3.4.** The bar graph showing the distribution of train and test data with the x-axis representing subcellular locations and the y-axis indicating the corresponding sample sizes in both the training and test datasets.

**Table 3.1.** HoliLoc's multi label human protein data summary.

	<b>Train</b>	<b>Test</b>
<b>Nucleoplasm</b>	4138	421
<b>Cytosol</b>	3021	339
<b>Plasma Membrane (PM)</b>	1217	151
<b>Mitochondria</b>	803	62
<b>Golgi Apparatus</b>	662	67
<b>Nucleoli</b>	604	76
<b>Nuclear Bodies</b>	396	56
<b>Nuclear Speckles</b>	352	40
<b>Endoplasmic Reticulum (ER)</b>	343	31
<b>Centrosome</b>	250	29
<b>Nucleoli Fibrillar Center</b>	238	29
<b>Cell Junctions</b>	216	27
<b>Nuclear Membrane</b>	197	18
<b>Microtubules</b>	169	18
<b>Cytokinetic Bridge</b>	159	19
<b>Actin Filaments</b>	153	20
<b>Intermediate Filaments</b>	122	7
<b>Microtubule Organizing Center (MTOC)</b>	116	12
<b>Focal Adhesion Sites</b>	93	12
<b>Mitotic Spindle</b>	57	11
<b>Cytoplasmic Bodies</b>	51	8
<b>Aggresome</b>	27	8
<b>Total Unique Protein Count</b>	8459	723



**Figure 3.5.** Data distribution among 22 subcellular locations.

## 3.2. Data Preprocessing

The data preprocess is necessary to transform raw data into a format that can be effectively used by HoliLoc models. Different data preprocessing strategies are applied to each data type according to their own needs.

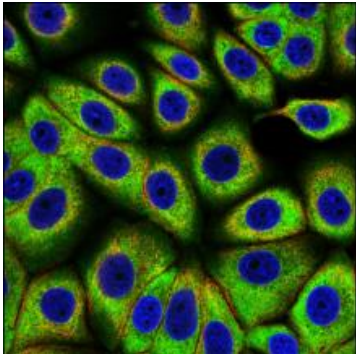
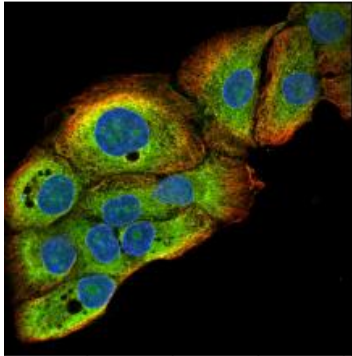
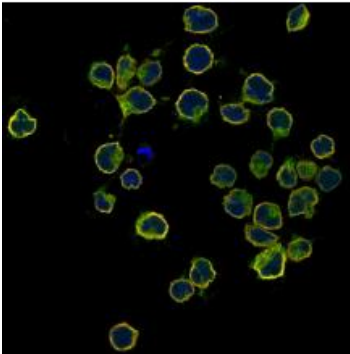
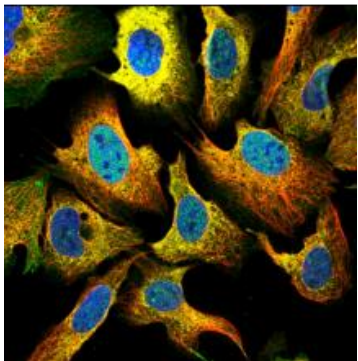
### 3.2.1. Image Data Preprocessing

The preprocess operations of image data were conducted with the OpenCV library. The first preprocess step is converting IF images of proteins from BGR to RGB and resizing these immunofluorescence images to a standardised dimension of 224x224 pixels. Deep learning frameworks like TensorFlow, which is used in this thesis, typically expect images to be in RGB (Red-Green-Blue) colour format rather than BGR (Blue-Green-Red), which is a common colour format used by many image processing libraries, including OpenCV. This standardisation of dimensions not only

ensures consistent data input for subsequent analytical procedures but also aids in reducing computational complexity. Subsequently, the resized images were transformed into numerical arrays. Each pixel within these images is represented as a series of numerical values, enabling the application of deep learning techniques. Secondly, normalisation of these arrays was carried out by dividing each numerical value by 255. The process of dividing pixel values by 255 is a vital step in the standardisation of digital images. In a typical 8-bit colour image, pixel values range from 0 (no colour) to 255 (full colour intensity) for each colour channel (red, green, blue). Dividing by 255 scales these values to a normalised range between 0 and 1, which ensures consistent data input and prevents numerical instability in mathematical operations, enhances interpretability, and improves the performance of deep learning algorithms that rely on consistent, scaled input data. This transformation, therefore, is an essential preprocessing step in making images convenient for deep learning models. Finally, arrays constructed from images that have a consistent size of  $224 \times 224 \times 3$ , where the '3' defines the three primary colour channels (Figure 3.6). This multichannel representation allows comprehensive exploration of the IF image data and permits the simultaneous investigation of protein SL across various cellular structures and regions (Figure 3.7).



**Figure 3.6.** Overview of image data preprocess.

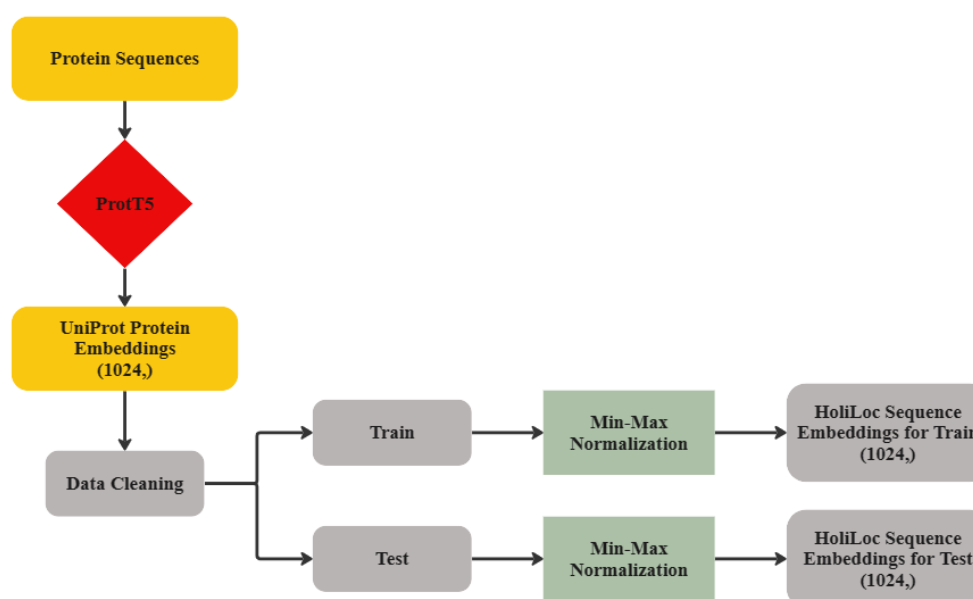
	
<p>Uniprot: P61513            Cell-Line: MCF7            Subcellular Locations: Cytosol, ER</p>	<p>Uniprot: Q96HY6            Cell-Line: HaCaT            Subcellular Locations: ER, Nucleoli</p>
	
<p>Uniprot: O60383            Cell-Line: JURKAT            Subcellular Locations: Cytosol, Golgi</p>	<p>Uniprot: Q96C12            Cell-Line: MCF7            Subcellular Locations: Nucleoplasm,            Focal Adhesion Sites, Cytosol</p>

**Figure 3.7.** Visual examples of HoliLoc input protein image data.

### 3.2.2. Sequence Data Preprocess

Sequence embedding information was obtained from the Uniprot Protein Embeddings Section (<https://www.uniprot.org/help/embeddings>) with the

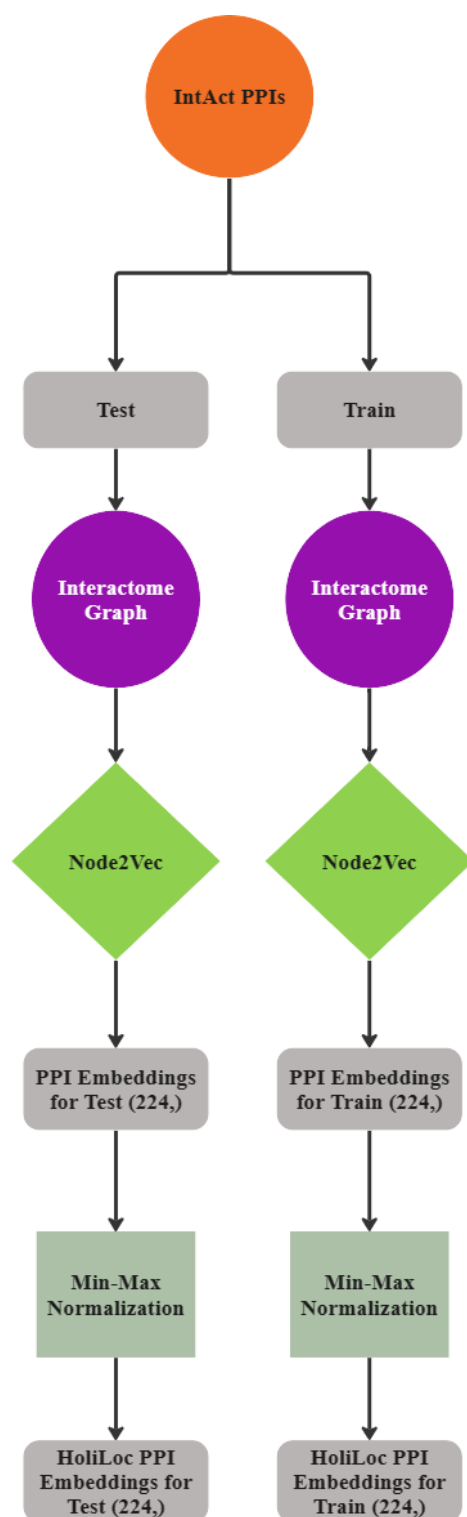
instructions of Uniprot, where raw embeddings of protein sequences are present and obtained with the ProtT5 protein LM in size 1024 (46). Sequence embeddings were merged with the main HoliLoc data by mapping to related proteins and train test split was conducted for HoliLoc data. Min-max normalization, scaling numerical features to a specified range between 0 and 1, applied to train and test data separately to prevent data leakage, and final HoliLoc sequence embeddings in size 1024 obtained (Figure 3.8).



**Figure 3.8.** Overview of sequence data preprocess.

### 3.2.3. Interactome (PPI) Data Preprocess

PPI data obtained from IntAct was merged with the final HoliLoc data by mapping to related proteins. After the train test split process of HoliLoc, PPI information was converted to PPI graphs and given into the node2vec algorithm separately, and embeddings in size 224 obtained then min-max normalisation was then applied to embeddings of train and test data separately to prevent data leakage. Finally normalised PPI embeddings obtained in size 224 to be given into FFN (Figure 3.9).



**Figure 3.9.** Overview of interactome (PPI) data preprocess.

### 3.3. Classification Models

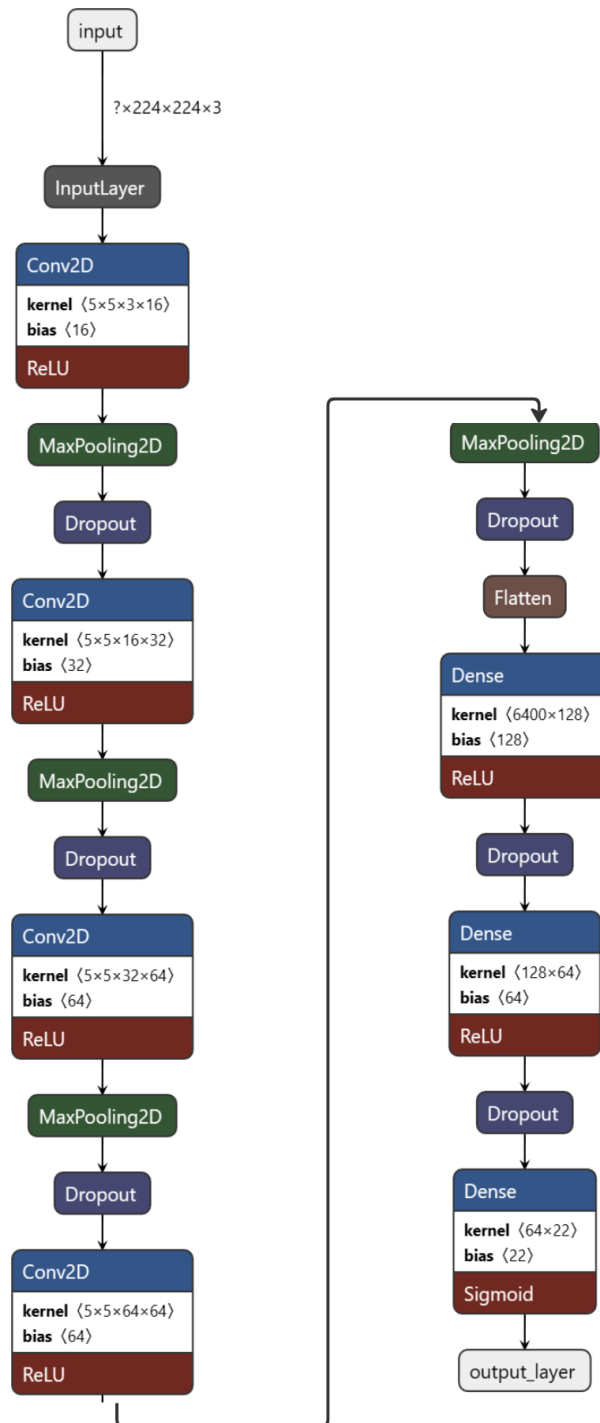
HoliLoc leverages a diverse range of data modalities to enhance predictive accuracy and provide a comprehensive understanding of protein SL. Our approach outlined three key modalities: image, sequence, and PPI. In this thesis, the deep learning models were constructed using the TensorFlow framework with the Keras. Models trained on GPU and data preprocess conducted on CPU with features, HP Z8 G4 Workstation, 2 x HP Intel Xeon Gold 5215 2.50GHz CPU -40 cores-, HP 128GB memory, 2 x NVIDIA GeForce RTX2080, HP Z Turbo Drive G2 512GB SSD, HP 1TB 7200rpm SATA HDD. For hyperparameter optimization, hyperband algorithm from Keras Tuner was employed for each model, image, sequence, PPI and HoliLoc separately. The primary objective was to maximise validation accuracy, and the optimization process, guided by early stopping with a patience of 10 epochs, iteratively explored hyperparameter configurations over a maximum of 30 epochs. Investigated parameters include dense layer units and dropout rates. The Hyperband algorithm efficiently searches through a large hyperparameter space while using resources effectively and sampling many configurations randomly. This helps in exploring a diverse range of hyperparameter combinations. This approach is a fine choice with limited computational sources.

#### 3.3.1. Image Model

Image feature vectors were employed as input for CNN, which has a total of 20 layers, which include convolutional layers, pooling layers, dropout layers, flattening layers, and dense layers. The convolutional layers, which played an initial role in capturing fine details within the IF images, utilise various filter sizes (16, 32, and 64) with a kernel size of (5, 5) and ReLU activation. MaxPooling2D layers follow each convolutional layer, employing a pool size of (2, 2) for down-sampling and retaining critical image features, which reduces computational expense while simultaneously enabling the model to recognize features in various regions of the image. Dropout layers with rates of 0.3 and 0.5, along with a specified seed value, are incorporated for regularisation. The model flattens the output before progressing



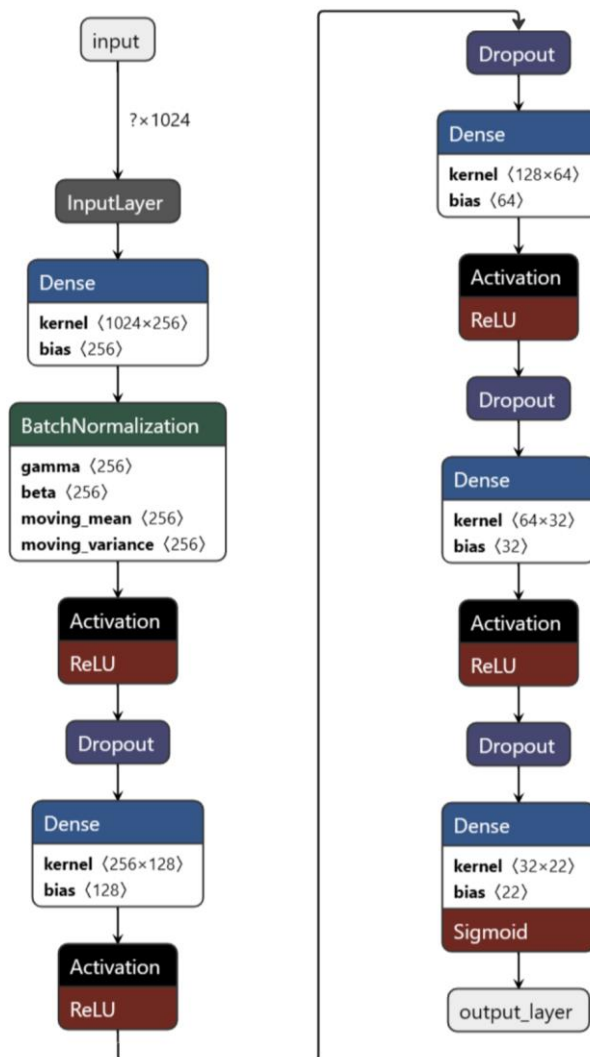
through dense layers, including two with 128 and 64 units, respectively, both employing ReLU activation. Dropout layers with rates of 0.3 are applied after each dense layer in which high-level abstractions and relationships between the detected features are recognized. The final dense layer, designated as the output layer, consists of 22 units with sigmoid activation, suitable for multi-label classification. The architecture is designed for image classification tasks on input data with dimensions (224, 224, 3). Sigmoid activation allows each output unit to independently produce values in the range [0, 1], which aligns well with multilabel classification, where each class can be associated with multiple labels; in this case, it is appropriate for proteins' presence in multiple locations. The model is compiled using the Adam optimizer with a binary cross-entropy loss.



**Figure 3.10.** Image model structure.

### 3.3.2. Sequence Model

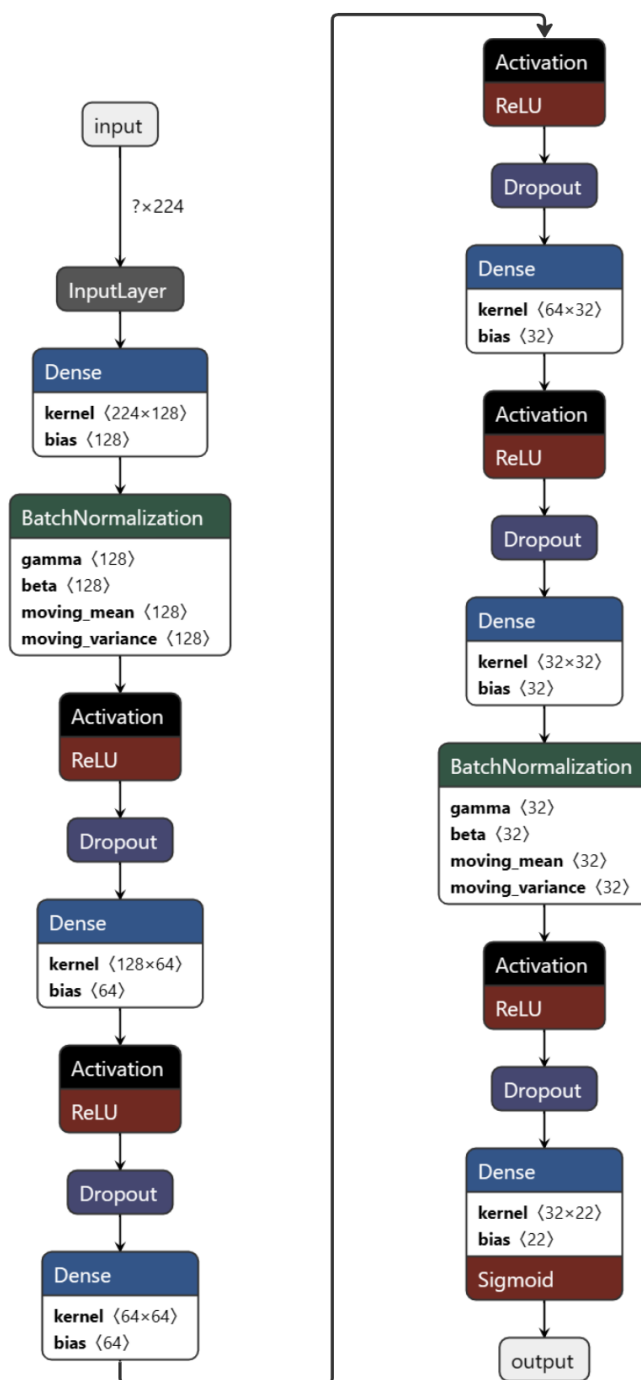
The protein sequence embeddings are given into the FFN model, which is composed of 16 layers in total, which include dense layers, batch normalisation layers, activation layers (ReLU), and dropout layers. Each is tailored to optimise its performance for the classification task. To enhance model stability, batch normalisation is applied, followed by the ReLU activation function to introduce non-linearity. Additionally, dropout regularisation is applied to cope with overfitting while increasing the model's robustness (Figure 3.11). The output layer consists of 22 units with sigmoid activation, designed for multi-label classification tasks. The model is designed for sequence embeddings with an input shape of (1024,). The model is compiled using the Adam optimizer with a binary cross-entropy loss.



**Figure 3.11.** Sequence model structure.

### 3.3.3. Interactome Model (PPI)

Interactome FFN is composed of 20 layers. The model includes dense layers with varying units (128, 64, 64, 32, 32), batch normalisation layers, activation layers with ReLU, and dropout layers with different dropout rates (0.4, 0.5, 0.1, 0.1, 0.1). Each of these layers collectively enables the model to understand complicated patterns within protein interactions. The final dense layer, comprising 22 units with a sigmoid activation, serves as the output layer for our classification task (Figure 3.12). The architecture is designed for input data with dimensions (224,). The model is compiled using the Adam optimizer with a binary cross-entropy loss.



**Figure 3.12.** Interactome (PPI) model structure.

### 3.3.4. Model Fusion (HoliLoc)

To harness the synergistic potential of three distinct modalities, HoliLoc employs a technique known as joint fusion, also referred to as intermediate fusion. This process centres around the combination of feature representations learned from intermediate layers of neural networks with data from other modalities. The primary objective is to harmonise the variations in dimensionality and information content across these diverse modalities. The model incorporates three individual modules—image, sequence, and PPI that are fused to construct a powerful multi-modal neural network. This feature vector is subsequently fed into a FFN, consisting of 17 layers in which 6 dense layers, batch normalisation, activation, and dropout layers exist. The final output layer utilises sigmoid activation for multi label classification with 22 classes (Figure 3.13). The model is compiled using the Adam optimizer and binary cross-entropy loss. The model's architecture consists of a total of 4,663,606 parameters, with 4,654,390 being trainable and an additional 9,216 non-trainable.

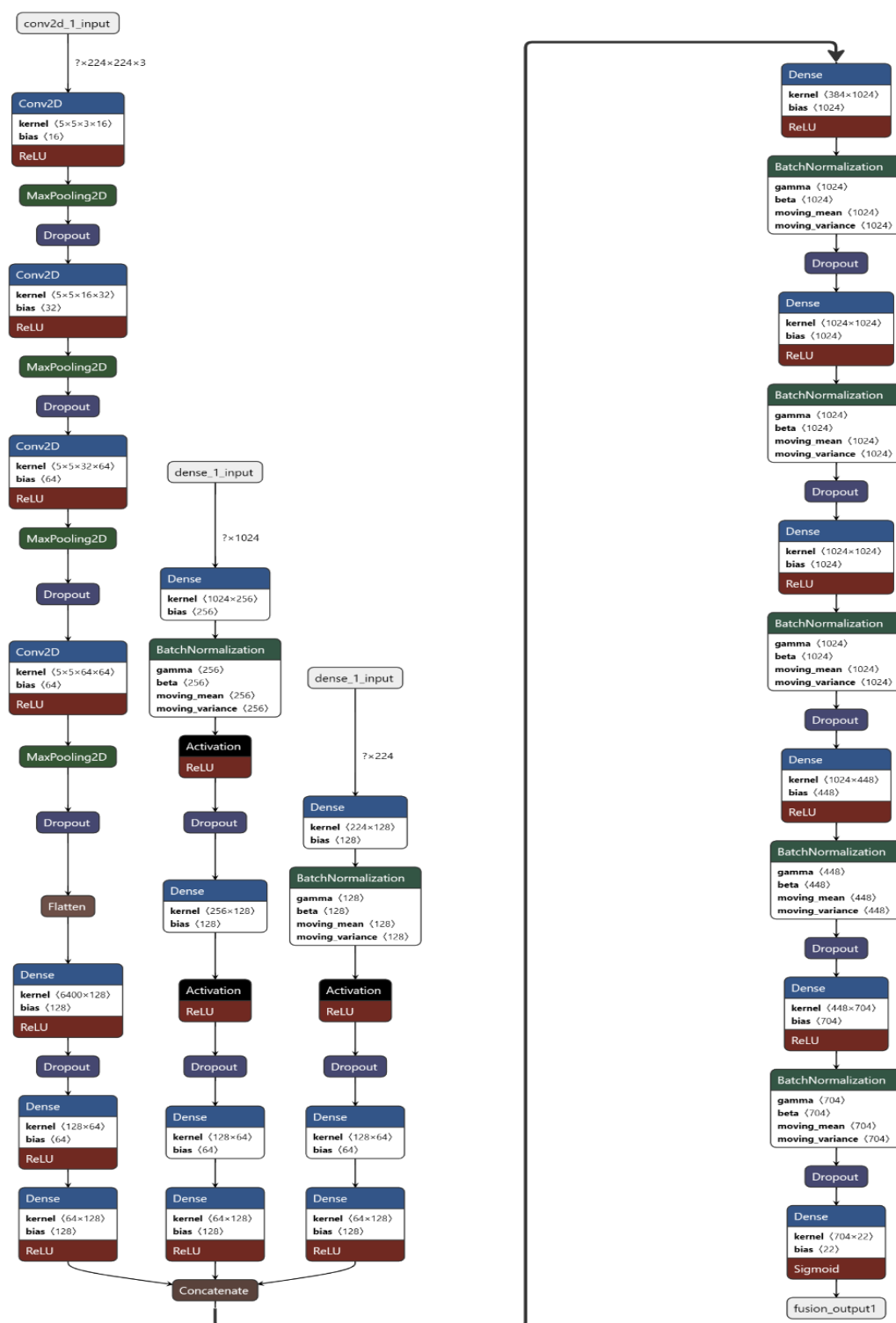


Figure 3.13. HoliLoc model structure.



## 4. RESULTS

### 4.1. Evaluation of Performance

The performance evaluation of HoliLoc follows a thorough methodology, utilising the macro-F1 score to assess its predictive power in 22 subcellular classes. This metric offers a comprehensive understanding of the model's efficacy because it treats each class label equally. To understand HoliLoc's intra-study capabilities, comparisons among individual feature based models and HoliLoc were made in multi-location and single-location prediction settings. Furthermore, analysis of HoliLoc's performance is made by adding the accuracy, recall, and precision metrics of single-location models. Weighting metrics according to the sample size metrics makes possible the more transparent evaluation despite an unbalanced data distribution. Beyond the intra-study performance, the comparative analysis that compared the single-location models from HoliLoc with the top 10 teams' scores from the HPA Kaggle challenge for each SL was conducted for the inter-study investigation. Macro F1 score, precision, recall, and accuracy can be represented by the following equations (4.1, 4.2, 4.3, 4.4, 4.5, 4.6).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4.1.)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4.2.)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4.3.)$$

$$P = \frac{1}{T} \sum_{t \in T} P_t \quad (4.4.)$$

$$R = \frac{1}{T} \sum_{t \in T} R_t \quad (4.5.)$$

$$\text{Macro F1} = \frac{2PR}{P + R} \quad (4.6.)$$

In those equations, TP is true positive, TN is true negative, FP is false positive, FN is false negative,  $t$  is the class label among all labels  $T$ ,  $P$  is average precision over all labels, and  $R$  is average recall over all labels.

## 4.2. Comparison of HoliLoc and Individual Feature Based Models

HoliLoc is trained separately in both multi-location and single-location settings. In the multi-location setting, the model generates predictions for all 22 SLs simultaneously in a multi-label format. While, in the single-location setting, the same training procedure is applied independently to each SL. This leads to the development of 22 distinct HoliLoc models, each producing a binary output specific to its corresponding SL. In summary, for the multi-location setting, a unified HoliLoc model is created by training and fusing individual feature-based models (image, sequence, and interactome models). This results in a total of 4 models. On the other hand, in the single-location setting, individual feature-based models, and HoliLoc are trained for each of the 22 SLs. Hence, a total of 88 models are developed for single-location settings.

### 4.2.1. Comparison of HoliLoc and Individual Feature Based Models on Single Location Prediction Setting

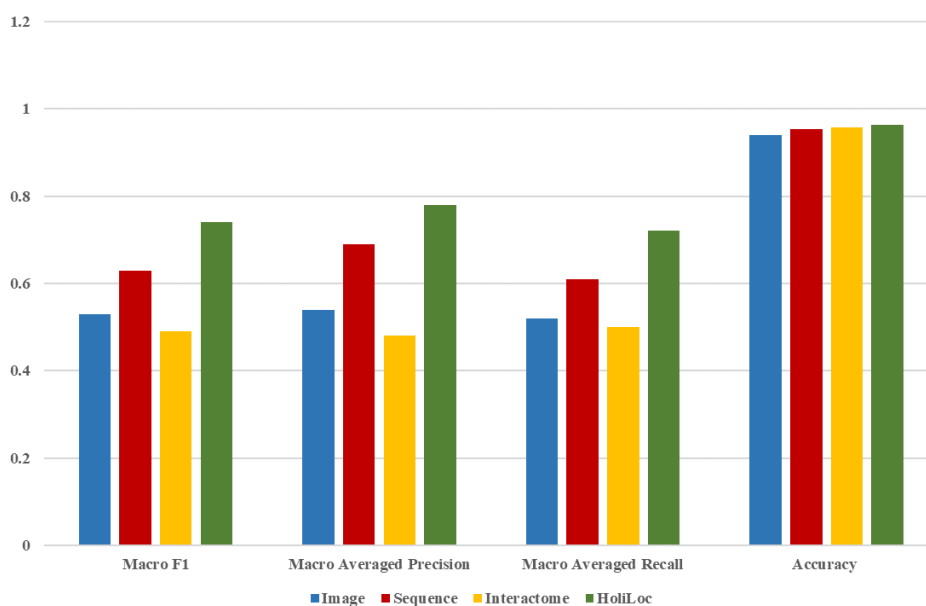
In the context of single-location prediction, HoliLoc demonstrates its most substantial performance improvement in PM, ER, mitochondria, and nucleoplasm. A comparison of average macro F1 scores between individual feature-based models and HoliLoc reveals the following order: PM (0.54 to 0.79), ER (0.55 to 0.74), mitochondria (0.59 to 0.79), and nucleoplasm (0.65 to 0.83). Despite the relatively

small sample size, the notable enhancement in ER's performance with HoliLoc draws attention, highlighting the effectiveness of HoliLoc in significantly improving predictive success for these subcellular localizations in a single-location setting (Figure 4.1). Despite their relatively large sample sizes within the dataset, nuclear speckles, nuclear bodies, and nucleoli fail to demonstrate significant performance improvement with HoliLoc. A comparison of average macro F1 scores between individual feature-based models' average and HoliLoc shows moderate increase for nuclear speckles (0.52 to 0.55), nuclear bodies (0.53 to 0.55), and nucleoli (0.52 to 0.54). These findings suggest that HoliLoc's impact on predictive success for nuclear speckles, nuclear bodies, and nucleoli in the single-location setting is limited, indicating potential complexities or challenges associated with these specific subcellular localizations. In single-location prediction settings, the comparative analysis presented in Table 4.1 reveals the notable superiority of HoliLoc over single feature-based models across various subcellular locations. Specifically, when evaluating accuracy, HoliLoc exhibits enhanced performance in 9 out of 22 locations, including centrosome, cytokinetic bridge, cytoplasmic bodies, cytosol, ER, intermediate filaments, mitochondria, nucleoli fibrillar centre, and PM. Moreover, HoliLoc exceeds single feature-based models in terms of recall and precision for 17 out of the 22 subcellular locations. Furthermore, for both average and weighted average all accuracy, recall and precision get better scores compared with individual feature-based models. The statistical analyses, conducted through the Wilcoxon signed-rank test on the F1 score values, establish HoliLoc's significant superiority over individual feature-based models. Upon examining macro F1 scores, significant improvements were observed and found to be statistically significant across all model comparisons. HoliLoc outperformed image feature based model ( $p = 1.80e-04$ ), sequence feature based model ( $p = 8.73e-05$ ), and PPI feature based model ( $p = 2.99e-04$ ), as well as the average of individual feature based models ( $p = 4.77e-07$ ). These results provide compelling evidence, affirming HoliLoc's consistent outperformance of individual feature-based models. The graphical representation of results in Figure 4.2 serves as visual support for these findings, further establishing HoliLoc as an advanced and promising model for single-location prediction scenarios in subcellular localization tasks.

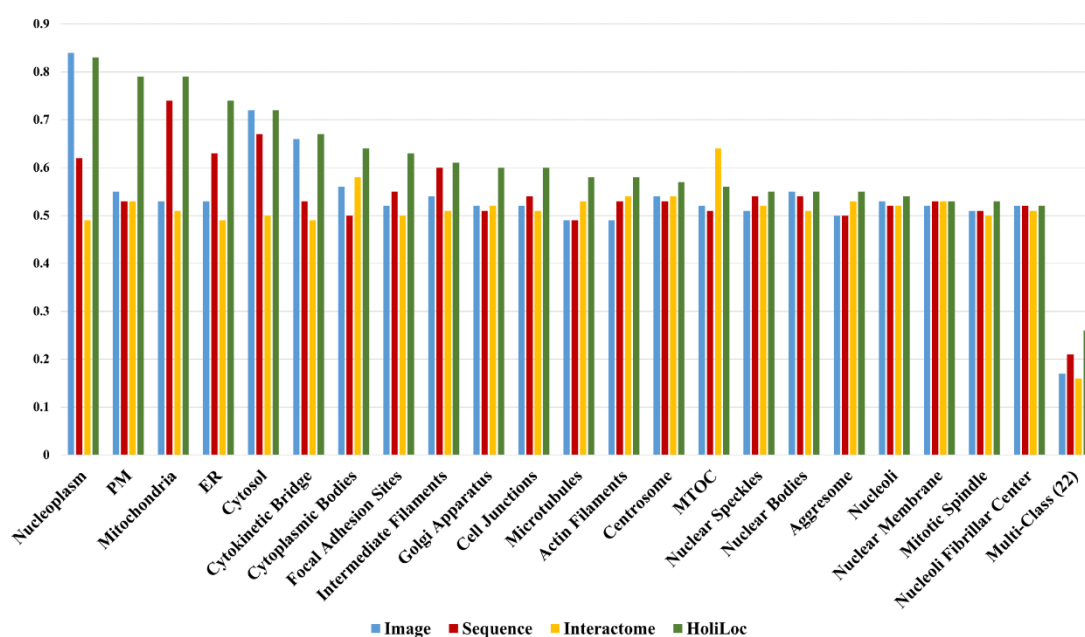
**Table 4.1.** Performance comparison of HoliLoc and individual feature-based models in the single-location prediction settings. The highest performance results are shown in bold font.

Model	Image			Sequence			PPI			HoliLoc		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Actin Filaments	<b>0.97</b>	0.49	0.50	0.92	0.52	0.55	0.95	0.53	0.54	0.92	<b>0.56</b>	<b>0.64</b>
Aggresome	0.93	0.51	0.53	<b>0.99</b>	0.49	0.50	0.97	0.52	0.55	0.96	<b>0.54</b>	<b>0.61</b>
Cell Junctions	<b>0.96</b>	0.48	0.50	0.95	0.56	0.53	0.95	0.52	0.51	0.89	<b>0.58</b>	<b>0.71</b>
Centrosome	0.92	0.54	<b>0.55</b>	0.95	0.55	0.53	0.95	0.57	0.53	<b>0.96</b>	<b>0.65</b>	<b>0.55</b>
Cytokinetic Bridge	0.93	0.56	<b>0.61</b>	0.91	0.54	0.60	0.96	0.52	0.52	<b>0.98</b>	<b>0.99</b>	<b>0.61</b>
Cytoplasmic Bodies	0.98	0.56	0.56	<b>0.99</b>	0.49	0.50	<b>0.99</b>	0.62	0.56	<b>0.99</b>	<b>0.66</b>	<b>0.62</b>
Cytosol	<b>0.72</b>	<b>0.72</b>	<b>0.72</b>	0.67	0.67	0.67	0.51	0.50	0.50	<b>0.72</b>	<b>0.72</b>	<b>0.72</b>
Endoplasmic Reticulum	0.94	0.54	0.52	0.95	0.69	0.61	<b>0.96</b>	0.48	0.50	<b>0.96</b>	<b>0.78</b>	<b>0.72</b>
Focal Adhesion Sites	0.93	0.51	0.55	<b>0.98</b>	0.56	0.54	<b>0.98</b>	0.49	0.50	0.96	<b>0.60</b>	<b>0.69</b>
Golgi Apparatus	<b>0.91</b>	<b>0.67</b>	0.52	0.77	0.51	0.52	0.89	0.56	0.52	0.88	0.61	<b>0.59</b>
Intermediate Filaments	0.97	0.53	0.56	<b>0.99</b>	0.66	0.57	0.90	0.52	<b>0.67</b>	<b>0.99</b>	<b>0.75</b>	0.57
Microtubules	<b>0.98</b>	0.49	0.50	<b>0.98</b>	0.49	0.50	0.94	0.52	0.54	0.97	<b>0.69</b>	<b>0.55</b>
Mitochondria	0.85	0.53	0.53	0.93	0.77	0.71	0.87	0.52	0.51	<b>0.94</b>	<b>0.87</b>	<b>0.74</b>
Mitotic Spindle	0.94	0.51	0.52	<b>0.98</b>	0.49	0.50	<b>0.98</b>	0.49	0.50	0.97	<b>0.53</b>	<b>0.54</b>
MTOC	0.96	0.52	0.53	0.91	0.51	0.55	<b>0.99</b>	<b>0.99</b>	<b>0.58</b>	0.96	0.55	0.57
Nuclear Bodies	0.84	0.54	<b>0.57</b>	0.88	<b>0.55</b>	0.54	<b>0.89</b>	0.51	0.51	0.86	<b>0.55</b>	0.56
Nuclear Membrane	0.95	0.52	0.52	<b>0.97</b>	<b>0.57</b>	0.52	0.90	0.52	<b>0.57</b>	0.96	0.54	0.52
Nuclear Speckles	<b>0.90</b>	0.51	0.51	0.89	0.53	0.54	0.89	0.52	0.52	0.85	<b>0.54</b>	<b>0.58</b>
Nucleoli	0.79	0.53	0.53	<b>0.80</b>	0.52	0.52	0.79	0.52	0.52	0.79	<b>0.54</b>	<b>0.55</b>

<b>Nucleoli Fibrillar Center</b>	0.89	0.51	<b>0.53</b>	<b>0.91</b>	<b>0.52</b>	<b>0.53</b>	0.92	0.51	0.51	<b>0.96</b>	0.73	0.52
<b>Nucleoplasm</b>	<b>0.84</b>	<b>0.84</b>	<b>0.83</b>	0.62	0.65	0.65	0.51	0.49	0.49	0.83	0.83	<b>0.83</b>
<b>Plasma Membrane</b>	0.69	0.55	0.55	0.72	0.54	0.53	0.66	0.52	0.53	<b>0.85</b>	<b>0.78</b>	<b>0.81</b>
<b>Average</b>	0.90	0.55	0.56	0.89	0.56	0.55	0.88	0.54	0.53	<b>0.92</b>	<b>0.66</b>	<b>0.63</b>
<b>Weighted Average</b>	0.82	0.67	0.66	0.74	0.61	0.61	0.67	0.51	0.51	<b>0.83</b>	<b>0.73</b>	<b>0.72</b>



**Figure 4.1.** Performance comparison of HoliLoc and individual models for endoplasmic reticulum (ER) single location setting prediction.



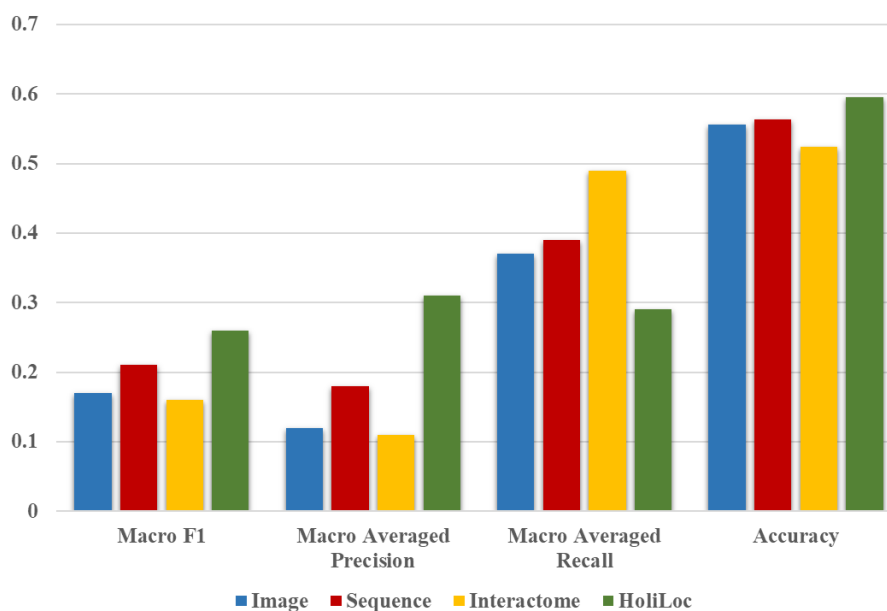
**Figure 4.2.** Performance comparison of HoliLoc and individual feature-based models' macro F1 scores in the single-location models per subcellular location and multi-location prediction setting.

#### 4.2.2. Comparison of HoliLoc and Individual Feature Based Models on Multi Location Prediction Setting

In the context of predicting multiple locations, an improvement in macro F1 performance is demonstrated by HoliLoc compared to individual feature-based models. The average macro F1 score for individual feature-based models is 0.18, while a higher score of 0.26 is achieved by HoliLoc (Figure 4.3). The Wilcoxon signed-rank test, conducted on the F1 score values, reveals HoliLoc's superiority over individual feature-based models in multi-location prediction setting. HoliLoc outperforms the image feature-based model ( $p = 4.8e-03$ ), the sequence feature-based model ( $p = 1.22e-02$ ), the PPI feature-based model ( $p = 6.06e-03$ ), and the average of individual feature-based models ( $p = 1.9e-03$ ). Figure 4.3 visually supports these statistical findings. Additionally, a difference is observed between average recall (0.42) and average precision (0.14) for individual feature-based models. In contrast, a more balanced relationship is maintained by HoliLoc between recall (0.29) and precision (0.31). This suggests that not only is overall performance enhanced by

HoliLoc, but it also achieves a more harmonized trade-off between recall and precision compared to individual feature-based models. In specific subcellular localizations, the most significant impact of HoliLoc is observed in ER, mitochondria, and mitotic spindle SLs. The average macro F1 scores for individual feature-based models at these locations, along with HoliLoc, follow this order: ER (0.17, 0.46), mitochondria (0.23, 0.55), and mitotic spindle (0.18, 0.53) (Table 4.2). Despite their relatively small sample sizes, attention is drawn to both ER and mitotic spindle SLs, as illustrated in Figure 3.5, emphasizing the significance of HoliLoc's impact on cellular localization, especially in instances where the data may be limited in terms of sample size. However, in the multi-location setting, a relatively modest impact is observed for cytosol and nucleoplasm SLs, despite their being two of the most abundant SLs in the dataset (Figure 3.5). The average F1 scores for individual feature-based models are notably high, measuring 0.64 for cytosol and 0.75 for nucleoplasm. In comparison, F1 scores of 0.69 for cytosol and 0.83 for nucleoplasm are achieved by HoliLoc. These results suggest that while HoliLoc may not outperform individual feature-based models for cytosol and nucleoplasm SLs, competitive performance is still maintained, considering the high baseline set by the individual models. The inconsistency in performance across different subcellular localizations underlines the complexity of the multi-location setting and the impact of HoliLoc on various cellular structures. Considering the 0 F1 scores in Table 4.2 for the multi-location prediction models, a comprehensive assessment was undertaken with 10-fold analysis. The dataset underwent randomised division into training and testing sets, with approximate ratios consistently maintained over 10 iterations. Each iteration independently evaluated with the F1 score for the multi-location prediction model, contributing to the averaged results in Table 4.3. This analysis was conducted to evaluate the models' performance across different data combinations. The results revealed enhancements for some SLs, such as cytokinetic bridge, focal adhesion sites, intermediate filaments, mitotic spindle, MTOC, nuclear membrane, and nucleoli fibrillar center, showing performance increment from 0 F1 scores reported in Table 4.2. However, some SLs remained with 0 F1 score for some models. For example, the image model had 0 F1 scores for cytoplasmic bodies, the

sequence model for aggresome, and the HoliLoc model for both aggresome and cytoplasmic bodies. The PPI model demonstrated a notable absence of 0 F1 values.



**Figure 4.3.** Comparison of average prediction performances of HoliLoc and individual feature-based models in the multi-location prediction setting.

**Table 4.2.** F1 score performance comparison of HoliLoc, individual feature-based and class-wise score distribution for the top ten teams of HPA Kaggle Challenge. Adapted from (44). Comparing mean HPA Kaggle challenge results and HoliLoc highest performance results are shown bold.

Subcellular Location	HPA Kaggle Challenge						HoliLoc Multi Location Prediction				HoliLoc Single Location Prediction			
	Mean	Minimum	25% Percentile	50% Percentile	75% Percentile	Maximum	Image	Sequence	PPI	HoliLoc	Image	Sequence	PPI	HoliLoc
Actin Filaments	0.58	0.55	0.56	0.58	0.60	<b>0.62</b>	0.18	0.14	0.12	0.15	0.49	0.53	0.54	0.58



Aggresome	0.64	0.60	0.64	0.64	0.65	<b>0.69</b>	0.00	0.00	0.00	0.00	0.50	0.50	0.53	0.55
Cell Junctions	0.57	0.54	0.56	0.56	0.57	<b>0.63</b>	0.11	0.02	0.23	0.09	0.49	0.54	0.51	0.60
Centrosome	0.50	0.48	0.49	0.50	0.51	0.53	0.13	0.10	0.23	0.05	<b>0.54</b>	0.53	<b>0.54</b>	<b>0.57</b>
Cytokinetic Bridge	0.39	0.36	0.37	0.39	0.40	0.44	0.00	0.19	0.05	0.24	<b>0.57</b>	<b>0.55</b>	<b>0.52</b>	<b>0.67</b>
Cytoplasmic Bodies	0.41	0.33	0.38	0.42	0.43	0.44	0.00	0.00	0.00	0.00	<b>0.56</b>	<b>0.50</b>	<b>0.58</b>	<b>0.64</b>
Cytosol	0.59	0.58	0.58	0.60	0.60	0.60	<b>0.64</b>	<b>0.64</b>	<b>0.64</b>	<b>0.69</b>	<b>0.72</b>	<b>0.67</b>	0.50	<b>0.72</b>
Endoplasmic Reticulum	0.54	0.44	0.50	0.54	0.56	0.65	0.18	0.26	0.08	0.46	0.53	0.63	0.49	<b>0.74</b>
Focal Adhesion Sites	0.59	0.56	0.58	0.59	0.60	0.62	0.16	0.08	0.00	0.21	0.52	0.55	0.50	<b>0.63</b>
Golgi Apparatus	0.69	0.67	0.68	0.69	0.70	<b>0.72</b>	0.18	0.21	0.17	0.25	0.52	0.51	0.52	0.60
Intermediate Filaments	0.70	0.67	0.68	0.69	0.71	<b>0.75</b>	0.00	0.07	0.00	0.12	0.54	0.60	0.51	0.61

Microtubules	0.76	0.72	0.74	0.75	0.77	<b>0.78</b>	0.03	0.31	0.07	0.25	0.49	0.49	0.53	0.58
Mitochondria	0.67	0.65	0.65	0.67	0.68	0.71	0.21	0.31	0.16	0.55	0.53	<b>0.74</b>	0.51	<b>0.79</b>
Mitotic Spindle	0.40	0.35	0.39	0.39	0.41	0.46	0.00	<b>0.53</b>	0.00	<b>0.53</b>	<b>0.51</b>	<b>0.50</b>	<b>0.50</b>	<b>0.53</b>
MTOC	0.42	0.39	0.40	0.41	0.44	0.47	0.00	0.00	0.17	0.15	<b>0.52</b>	<b>0.51</b>	<b>0.64</b>	<b>0.56</b>
Nuclear Bodies	0.56	0.54	0.55	0.56	0.58	<b>0.59</b>	0.18	0.16	0.15	0.22	0.55	0.54	0.51	0.55
Nuclear Membrane	0.74	0.73	0.73	0.74	0.76	<b>0.76</b>	0.06	0.00	0.02	0.00	0.52	0.53	0.53	0.53
Nuclear Speckles	0.64	0.60	0.62	0.64	0.66	<b>0.67</b>	0.16	0.13	0.11	0.18	0.51	0.54	0.52	0.55
Nucleoli	0.73	0.69	0.72	0.73	0.74	<b>0.76</b>	0.20	0.20	0.19	0.26	0.53	0.52	0.52	0.54
Nucleoli Fibrillar Center	0.66	0.63	0.65	0.66	0.68	<b>0.70</b>	0.03	0.07	0.02	0.00	0.52	0.52	0.51	0.52

Nucleoplasm	0.79	0.76	0.78	0.79	0.79	0.80	0.76	0.74	0.74	<b>0.83</b>	<b>0.84</b>	0.62	0.49	<b>0.83</b>
Plasma Membrane	0.59	0.58	0.58	0.59	0.60	0.62	0.43	0.36	0.35	0.54	0.55	0.53	0.53	<b>0.79</b>
Average	0.60	0.56	0.58	0.60	0.61	<b>0.64</b>	0.17	0.21	0.16	0.26	0.55	0.55	0.52	0.62

**Table 4.3.** Comparative analysis of Image, Sequence, PPI, and HoliLoc multi-location prediction models' average 10-Fold F1 scores across subcellular locations, bold entries signify top-performing models within each subcellular location.

	Image	Sequence	PPI	HoliLoc
<b>Actin Filaments</b>	0.15	0.07	0.06	<b>0.21</b>
<b>Aggresome</b>	<b>0.02</b>	0.00	<b>0.02</b>	0.00
<b>Cell Junctions</b>	0.11	0.10	0.07	<b>0.20</b>
<b>Centrosome</b>	0.06	0.11	0.09	<b>0.12</b>
<b>Cytokinetic Bridge</b>	0.04	0.11	0.06	<b>0.10</b>
<b>Cytoplasmic Bodies</b>	0.00	<b>0.02</b>	<b>0.02</b>	0.00
<b>Cytosol</b>	0.63	0.63	0.63	<b>0.66</b>
<b>Endoplasmic Reticulum</b>	0.15	0.25	0.08	<b>0.47</b>
<b>Focal Adhesion Sites</b>	0.04	0.05	0.02	<b>0.17</b>
<b>Golgi Apparatus</b>	0.18	0.19	0.17	<b>0.26</b>
<b>Intermediate Filaments</b>	0.05	0.07	0.04	<b>0.13</b>
<b>Microtubules</b>	0.10	0.15	0.05	<b>0.16</b>
<b>Mitochondria</b>	0.20	0.23	0.18	<b>0.44</b>

<b>Mitotic Spindle</b>	0.05	<b>0.16</b>	0.03	<b>0.16</b>
<b>MTOC</b>	0.01	0.04	0.04	<b>0.09</b>
<b>Nuclear Bodies</b>	0.17	0.15	0.13	<b>0.20</b>
<b>Nuclear Membrane</b>	0.06	0.07	0.07	<b>0.08</b>
<b>Nuclear Speckles</b>	0.12	0.12	0.09	<b>0.14</b>
<b>Nucleoli</b>	0.21	0.21	0.19	<b>0.27</b>
<b>Nucleoli Fibrillar Center</b>	<b>0.09</b>	0.08	0.07	0.05
<b>Nucleoplasm</b>	0.78	0.74	0.74	<b>0.85</b>
<b>Plasma Membrane</b>	0.40	0.36	0.33	<b>0.58</b>
<b>Average</b>	0.16	0.18	0.14	<b>0.24</b>

### 4.3. Comparison With HPA Kaggle Challenge

The outcomes of the HPA Kaggle Challenge have been made accessible to the public through the HPA Kaggle challenge article (44). The challenge organisers have supplied both training and test data for participants. The metadata for these proteins is not included in the provided information. The available details are limited to IF images and SL information, with the assurance that all data originates from the HPA. In this study, although the challenge initially involved 28 subcellular localizations (SLs), only 22 SLs were considered. This reduction resulted from the exclusion of SLs with a protein count lower than 30 during the dataset preparation process. Hence, in this section, the performance is evaluated by comparing the results of 22 SLs. In the challenge's article, numerous performance results are provided with F1 score. To enhance the comprehensiveness of the comparison with the challenge results, the performance outcomes of the top 10 teams are considered, information can be found in Table 4.2. In this table 3 sets of information are merged which are the F1 score results of the top 10 teams in the challenge for each SL, the multi-location F1 scores and single-location macro F1 scores of HoliLoc, and the individual feature-based models displayed for each SL. The performance comparison is conducted with the best performer for each SL independently, and the best results

are shown in bold. This is a highly challenging comparison since HoliLoc is not compared with only one model; it is compared with the best performer for each location. For the half of the SLs (centrosome, cytokinetic bridge, cytoplasmic bodies, cytosol, ER, focal adhesion sites, mitochondria, mitotic spindle, MTOC, nucleoplasm, PM) HoliLoc single-location model had better macro F1 score compared to the best performers in each location class in the HPA challenge. The Wilcoxon signed-rank test conducted on the F1 score values resulted in a p-value of 0.61 shows that the observed differences in the F1 scores are not statistically significant. This implies the HoliLoc single-location model does not significantly differ from the best performers of the challenge and shows a similar performance. For the 3/22 of the SLs (cytosol, mitotic spindle, nucleoplasm) HoliLoc multi-location model had better F1 score compared to the best performers in each location class in the HPA challenge. The Wilcoxon signed-rank test, conducted on the F1 score, resulted in a p-value of  $9.06e-06$ , indicating a statistically significant difference between F1 scores of HoliLoc and challenge. While the HoliLoc multi-location prediction setting model exhibits a performance that is underperforming compared to the challenge results, it's crucial to recognize that this comparison is conducted against the best performer among the top ten teams for each location, rendering it a particularly challenging benchmark. For the cytosol SL both single and multi-location prediction models outperformed, while the best competitor in the challenge got 0.60 F1 score, all feature based multi-location models got average 0.64 F1 and HoliLoc 0.69. In the single-location setting both image and HoliLoc models have 0.72 macro F1 and sequence have 0.67. For the mitotic spindle SL single and multi-location models outperformed, while the best competitor in the challenge got 0.46 F1 score, multi-location sequence and HoliLoc models had the same 0.53, and in the single-location setting all models outperformed in which image 0.51, PPI and sequence 0.50, HoliLoc got 0.53 F1 score. For the nucleoplasm SL, while the best competitor in the challenge got 0.80 F1 score, both single and multi-location HoliLoc models got 0.83 F1 score, and the single-location setting image model got 0.84 F1 score. For the MTOC SL single-location HoliLoc outperformed the challenge with 0.56 macro F1 while the best competitor got 0.47 F1. However, the single-location PPI model gave a much better result with 0.64 macro F1.

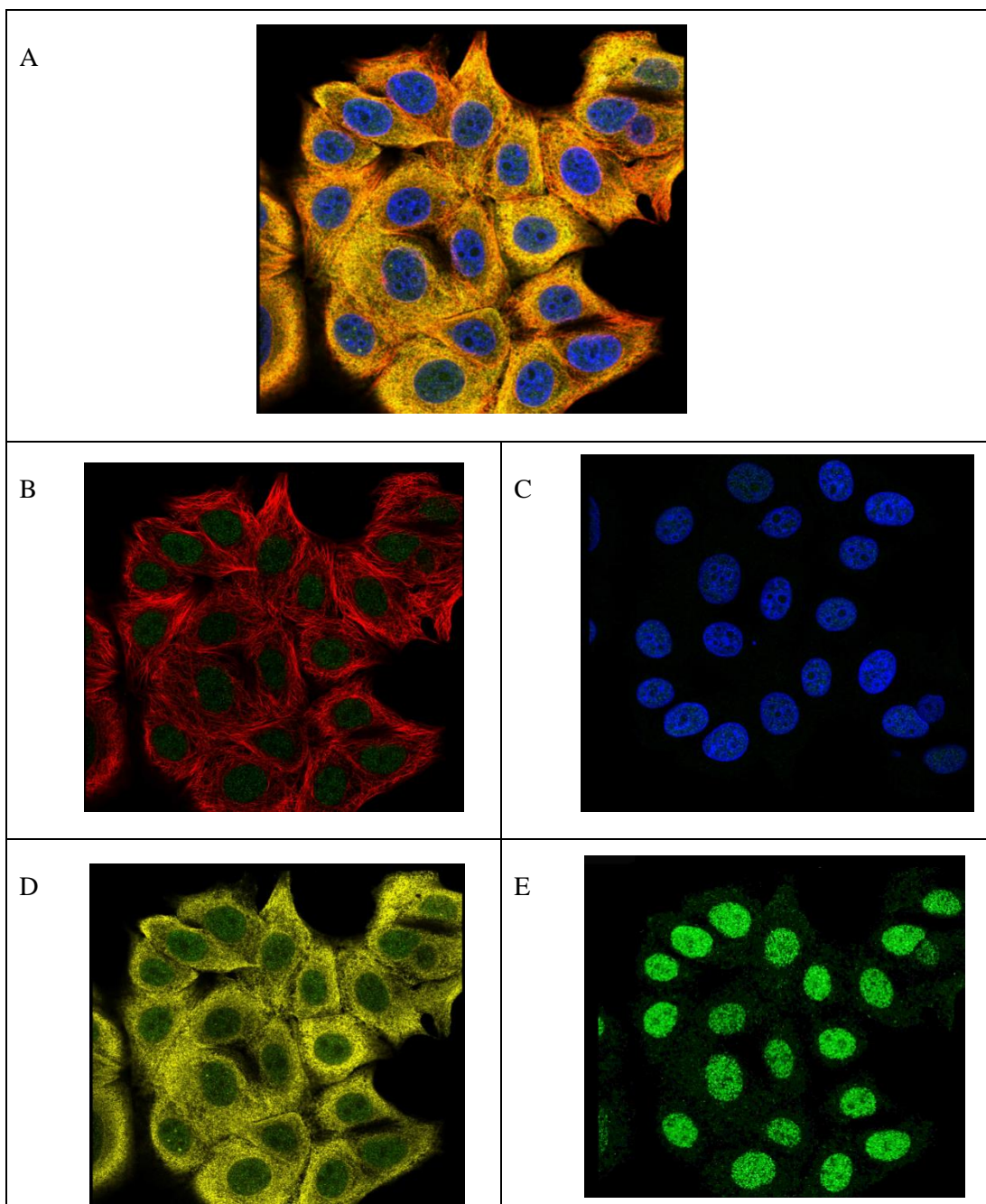
#### 4.4. Use Case

In our use case study, we decided to evaluate the HoliLoc multi-location model to predict the subcellular localizations of a protein that was neither included in our training nor test data sets. The protein that we selected for this analysis is “Histone H3.1” of the human (UniProt accession number: P68431). This protein is a core component of the nucleosome and plays a key role in regulating transcription, repairing DNA, replicating DNA, and maintaining chromosomal stability. Figure 4.4 displays the confocal microscopy images of “Histone H3.1”, obtained from the Human Protein Atlas database. The HoliLoc model predicts the subcellular location of Histone H3.1 as actin filaments, cell junctions, cytosol, plasma membrane (PM), and nucleoplasm. However, the HPA subcellular section only assigns nucleoplasm as the SL of Histone H3.1. Therefore, we decided to conduct a literature-based qualitative analysis to judge whether HoliLoc predictions could be undocumented/unknown true SLs of the protein or just false positives.

Newly synthesised histones must traverse the cytosol before being incorporated into the chromatin (47). Despite the absence of explicit mentions of cytosol in the HPA, UniProt, or Gene Ontology (GO) databases, the observation of the green-coloured target protein indicators surrounding the nucleus from confocal microscopy images in Figure 4.4 panel E supports the likelihood that this protein is present in the cytosol. Histones establish a strong binding with F-actin through robust electrostatic and hydrophobic interactions, influencing actin structure. Additionally, H1 histone has been observed to polymerize G-actin, while the H2A-H2B histone dimer has been found to bundle F-actin filaments (48). The presence of cell-surface proteins capable of binding to advanced glycation end products has been researched, and histone has been identified as a previously unrecognised binding protein for advanced glycation end products. Histone serves as a cell-surface receptor for advanced glycation end products (49), which could explain the PM prediction of HoliLoc. Moreover, the GO database supports this observation, as the information page for this protein includes the “GO:0016020; membrane” cellular component annotation

(<https://www.ebi.ac.uk/QuickGO/annotations?geneProductId=P68431&aspect=cellul>

ar\_component). Our investigation showed it is highly probable that HoliLoc SL predictions of cytosol, plasma membrane are true localizations that are not documented yet in HPA. Additional analysis is required to evaluate actin filaments and cell junctions predictions.



**Figure 4.4.** Confocal microscopy images of target protein P68431, obtained from Human Protein Atlas database in which green is target protein, blue is nucleus, red is microtubules and yellow is ER. A: all channels are visible, B: red and green channels are visible, C: blue and green channels are visible, D: yellow and green channels are visible, E: Only green channel is visible.



## 5. DISCUSSION

The subcellular localization of proteins holds great importance in explaining protein functions and roles within cellular processes. Understanding the subcellular distribution of proteins is crucial for unravelling complex mechanisms in systems biology, facilitating drug development, and advancing our understanding of protein-protein interactions. Despite the significant importance, the lack of experimentally verified annotations of subcellular localization is still a great challenge, which emphasises the necessity for advanced computational models that can effectively predict protein subcellular localization and facilitate protein subcellular localization studies by reducing the cost and time of lab experiments and imaging techniques. In response to this challenge, this thesis introduces HoliLoc, a holistic approach to protein subcellular localization prediction with deep learning utilising various information about proteins. This multi-modal learning strategy aims to exploit synergies among diverse features, providing a more comprehensive understanding of the factors influencing subcellular localization without using costly and resource-intensive AI models. HoliLoc's simple model architecture enables a clear and comprehensive examination of the effects of holistic data integration. HoliLoc is positioned as an option for problems related to subcellular localization prediction because of its simple model architecture and comprehensive data integration. Uniref50 clusters similar proteins according to their sequence similarity; hence, the inclusion of only one protein from each Uniref50 cluster in the HoliLoc dataset ensures the uniqueness of each protein, thereby preventing highly similar proteins from influencing the training and testing phases and mitigates the risk of the model overfitting, which could lead to falsely high-performance metrics. Moreover, with this diversity in the dataset, HoliLoc can generalise better on novel data. However, there is a notable class imbalance that presents a challenge to the task. Deep learning algorithms face a great challenge when dealing with the skew in the class distribution, which adds another level of complexity to the classification task.

The performance of HoliLoc is evaluated using the macro-F1 score at 22 subcellular locations, which demonstrates how well it handles the challenge of multi-class and multi-label prediction with an unbalanced data distribution. These

comparative analyses extend across both single-location prediction and multi-location prediction scenarios. In the single-location prediction setting, models are trained to provide prediction for specific locations individually. Meanwhile, in the multi-location prediction setting, these models exhibit their capability by simultaneously predicting for all 22 locations. Figure 4.3 illustrates a notable shift in the performance metrics of individual feature-based models. The average precision has improved, rising from 0.14 to 0.31. However, this improvement comes at the expense of a decrease in recall, dropping from 0.42 to 0.29. The distinctive feature of the HoliLoc model lies in its design for reducing false positives, emphasising its practical usage in assisting researchers to focus their efforts during wet lab experiments.

According to the comparative analysis results of the single-location setting, presented in Table 4.1, HoliLoc demonstrates the most significant performance improvement in PM, ER, mitochondria, and nucleoplasm. Comparing the average macro F1 scores of individual feature-based models to those achieved by HoliLoc reveals the following order: PM (0.54 to 0.79), ER (0.55 to 0.74), mitochondria (0.59 to 0.79), and nucleoplasm (0.65 to 0.83). Notably, despite the relatively small sample size, the enhancement in ER's performance with HoliLoc attracts attention, highlighting the effectiveness of HoliLoc in significantly improving predictive accuracy for these subcellular localizations in a single-location setting. It is observed that mitochondria demonstrate significant potential with a macro F1 score of 0.74 for the sequence model, exceeded by HoliLoc model, which achieved an even higher macro F1 score of 0.79. This accomplishment may be influenced by the presence of specific signal sequences in the regions of mitochondrial proteins (50). Despite their relatively large sample sizes within the dataset, nuclear speckles and nucleoli fail to demonstrate significant performance improvement with HoliLoc. Comparing the average macro F1 scores of individual feature-based models with HoliLoc, for nuclear speckles, the average macro F1 score increases from 0.52 to 0.55, for nuclear bodies from 0.53 to 0.55 and for nucleoli, from 0.52 to 0.54. These findings suggest that HoliLoc's impact on predictive accuracy for nuclear speckles, nuclear bodies and nucleoli in the single-location setting is limited, indicating potential complexities or challenges associated with these specific subcellular localizations such as; nucleoli

are dynamic structure can vary in size and number, and their appearance can be influenced by the cell's metabolic state (51), nuclear bodies can change in number and size. Their appearance can be influenced by the cell cycle and cellular stress (52), nuclear speckles are dynamic structures that can vary in size, number, and distribution (53). The notable superiority of HoliLoc over single feature-based models across various subcellular locations is observed. Specifically, when evaluating accuracy, HoliLoc exhibits enhanced performance in 9 out of 22 locations, including centrosome, cytokinetic bridge, cytoplasmic bodies, cytosol, ER, intermediate filaments, mitochondria, nucleoli fibrillar centre, and PM. Moreover, HoliLoc exceeds single feature-based models in terms of recall for 17 out of the 22 subcellular locations, indicating a heightened ability to correctly identify instances of interest. Furthermore, precision analysis corroborates the robustness of HoliLoc, showing its superior precision in 17 out of the 22 locations. These findings collectively underscore the effectiveness of HoliLoc in single-location prediction scenarios, positioning it as a promising and advanced model for subcellular localization tasks.

The optimization of deep learning models involves automatically distributing weights to PPI, image, and sequence modalities in an optimal manner. This optimization becomes particularly visible when assessing the performance of single-location prediction models, as demonstrated in Table 4.2. In this context, the evaluation is unaffected by the sample sizes of other localizations, offering a clearer perspective on individual model capabilities. Notably, HoliLoc demonstrates a performance pattern mirroring that of the best-performing model of individual feature based models in terms of macro F1 score. This alignment underscores the efficacy of HoliLoc in leveraging the strengths of each modality, affirming its competence in subcellular localization prediction. Particular examples such as mitochondria (Image: 0.53, Sequence: 0.74, PPI: 0.51, HoliLoc: 0.79) and microtubules (Image: 0.49, Sequence: 0.49, PPI: 0.53, HoliLoc: 0.58), further substantiates HoliLoc's ability to seamlessly integrate information from multiple sources.

According to the Table 4.2 multi-location setting section, HoliLoc demonstrates a notable improvement in macro F1 performance compared to

individual feature-based models. The average macro F1 score for individual feature-based models is 0.18, whereas HoliLoc achieves a higher score of 0.26. Furthermore, in the same setting, the individual feature-based models show a significant difference between average recall (0.42) and average precision (0.14). In contrast, HoliLoc shows a more balanced relationship between recall (0.29) and precision (0.31). This suggests that HoliLoc not only enhances overall performance but also achieves a more harmonised trade-off between recall and precision compared to the individual feature-based models. HoliLoc demonstrates its most significant impact on ER, mitochondria, and mitotic spindle SLs. The average F1 scores for individual feature-based models at these locations, along with HoliLoc, follow this order: ER (0.17, 0.46), mitochondria (0.23, 0.55), and mitotic spindle (0.18, 0.53). Despite their relatively small sample sizes, both ER and mitotic spindle SLs get attention which can be observed in Figure 3.5, emphasising the significance of HoliLoc's impact on cellular localization, especially in instances where the data may be limited in terms of sample size. Although HoliLoc couldn't manage to get better performance in multi-location prediction setting compared with challenge winners in multi-location prediction setting, despite its relatively small sample size, HoliLoc's performance improvement draws attention while comparing with single feature-based models. This success can be the result of ER protein's characteristic, often they contain specific subcellular localization signals or motifs that help target them to the ER (54). Detecting these signals can aid in differentiation, and IF images of HPA have colours such that target proteins are shown in green, nucleus in blue, microtubules in red and ER in yellow. This specific colour advantage can be fused with differentiable sequence signal information and boost performance in HoliLoc. Also, HoliLoc exhibits a relatively modest impact on the performance success of cytosol and nucleoplasm SLs, despite these being two of the most abundant SLs in the dataset (Figure 3.5). The average F1 scores for individual feature-based models are notably high, measuring 0.64 for cytosol and 0.75 for nucleoplasm. In comparison, HoliLoc achieves F1 scores of 0.69 for cytosol and 0.83 for nucleoplasm. These results suggest that while HoliLoc may not outperform individual feature-based models for cytosol and nucleoplasm SLs, it still maintains competitive performance, considering the high baseline set by the individual models. The inconsistency in performance

across different subcellular localizations underlines the complexity of the multilocation environment and the nuanced impact of HoliLoc on various cellular structures.

Some subcellular locations consistently yielded 0 F1 performance across the individual models: aggresome, cytokinetic bridge, cytoplasmic bodies, intermediate filaments, mitotic spindle, MTOC scored 0 F1 from the image model. Additionally, aggresome, cytoplasmic bodies, MTOC, and nuclear membrane obtained 0 from the sequence model. The PPI model registered 0 F1 for aggresome, cytoplasmic bodies, focal adhesion sites, intermediate filaments, and mitotic spindle. In the HoliLoc model, aggresome, cytoplasmic bodies, nuclear membrane, and nucleoli fibrillar centre showed 0 F1 performance. Locations with 0 performance in the HoliLoc model also exhibited 0 or very close to 0 F1 scores in their individual image, sequence, and PPI models. This suggests that the challenge lies in the inherent difficulty of predicting these particular locations. Considering the imbalanced distribution of localizations, zero performers are closely biologically related to more abundant sample size localizations which demonstrate notably better performance. For instance, cytoplasmic bodies, receiving 0 F1 for all models, and cytokinetic bridge, obtaining 0 from the image model, contrast with the high-performing cytosol ( $F1=0.69$  for HoliLoc), showcasing the impact of ample sample size on prediction. Similarly, while the nuclear membrane and nucleoli fibrillar centre received 0 from the HoliLoc model, nucleoplasm emerged as the top performer among all locations ( $F1=0.83$ ). Following the identification of consistent 0 F1 performances across diverse subcellular locations and models, a comprehensive evaluation through 10-fold analysis was conducted. According to the comparison between 10-fold analysis results on Table 4.3 and Table 4.2 multi location prediction part, some 0 F1 score performed SLs (cytokinetic bridge, focal adhesion sites, intermediate filaments, mitotic spindle, MTOC, nuclear membrane and nucleoli fibrillar center) obtained better performance different than 0. This pattern suggests a correlation between these 0 performers and the characteristics of the data. The observed drop in 0 performers across subcellular locations, highlights the impact of data characteristics on model performance and emphasises that 0 performances are more likely a consequence of chance in the data split process rather than deficiencies in the models' themselves.

The comparison with the HPA Kaggle Challenge in Table 4.2 was conducted among the top 10 scoring teams in the competition. The results reveal superior performance for our approach compared to the maximum score achieved by any of the top 10 teams, which is highlighted in bold for emphasis. HoliLoc demonstrates superior performance in comparison to half of the single-location models and 3/22 of SLs (nucleoplasm, mitotic spindle, cytosol) in the multi-location setting. In the context of simpler model designs and limited computational resources, achieving superior performance in half of the single-location models and 3 out of the 22 SLs in the multi-location model indicates the effectiveness of the holistic approach. These outcomes underline HoliLoc's capability to capture complicated relationships among diverse data types, demonstrating its effectiveness in challenging scenarios with unbalanced data. This highlights the robustness and adaptability of HoliLoc in addressing complex tasks within resource constraints. There could be several reasons why nucleoplasm, mitotic spindle and cytosol show better performance compared with challenge in multi-location prediction setting. Firstly, cytosol and nucleoplasm are the most abundant SLs in the dataset in which HoliLoc's potential could be observed fully. In addition, these SLs demonstrate comparable performance in single-feature-based models. Specifically, the average F1 score for cytosol in feature-based models is 0.64, while HoliLoc achieves 0.69. Similarly, for nucleoplasm, the scores are 0.75 for feature-based models and 0.84 for HoliLoc. Despite its relatively small sample size, the mitotic spindle exhibits high performance in both the sequence-based model and HoliLoc, with identical F1 scores of 0.53. This could be because of the localization of proteins related to mitosis on the mitotic spindle is likely an evolutionarily conserved mechanism, ensuring timely mitotic events. In sea urchin embryos and mammalian cells, RNA transcripts encoding mitosis-related proteins were identified at the spindle. Disruption of microtubule processes or motor proteins led to loss of spindle localization. Notably, the cytoplasmic polyadenylation element within Aurora B's 3'UTR, a cytoplasmic polyadenylation element binding protein recognition site, is crucial for RNA localization to the mitotic spindle, highlighting the significance of a specific sequence in proteins' spatial distribution (55). In the context of single-location prediction, nucleoplasm image model exhibits comparable performance to HoliLoc with F1 scores of 0.84 and 0.83, respectively.

Similarly, cytosol demonstrates similar performance to HoliLoc with F1 scores of 0.72 for both. These SLs either show better or equivalent performance compared to HoliLoc when integrated with image feature-based models. The colour scheme used by HPA, where target proteins are represented in green, the nucleus in blue, and microtubules in red in the nucleoplasm, contributes to the anatomical association of nucleoplasm with the nucleus and cytosol with microtubules. This strong anatomical correlation could provide an advantage to image feature-based models. MTOC is the only SL, in single-location prediction setting, with the best performing interactome feature-based model with a significant difference model (F1: 0.64) while image, sequence feature-based models and HoliLoc underperform (F1 scores: 0.52, 0.51, 0.56) and best performer of the challenge gets 0.47. MTOC plays a significant role in mediating protein-protein interactions and is involved in various cellular processes such as cell division, intracellular transport, and cell shape maintenance. Microtubule anchoring factors serve to anchor or bind microtubules to MTOC. Hence, PPI information could have very distinctive features for MTOC located proteins (56).

In the development of the PPI model, a critical consideration was the generation of protein-protein interaction (PPI) embeddings. To prevent leakage between the training and testing phases, we adopted a cautious approach, obtaining node2vec embeddings separately for the training and test datasets. However, it is crucial to note that in real-life scenarios, proteins interact, and their embeddings should ideally be influenced by the broader context of the entire human protein interactome. Hence, it should be noted that potential limitation in our approach, wherein the model is trained and tested on embeddings that may not fully capture the intricate relationships within the complete protein-protein interaction network. To enhance the usability and robustness of our inference system which is available at GitHub (<https://github.com/huBioDataLab/HoliLoc>), we have taken a proactive step by providing users with a more comprehensive set of embeddings. These embeddings, derived from the entire human protein interactome, offer a broader representation of protein relationships. In the upcoming versions of HoliLoc trains and tests will be conducted by using this more comprehensive PPI embeddings covering the entire human protein interactome.

Researchers can employ the HoliLoc tool to analyse images acquired under optimal conditions in which cells should be fixed in 4% formaldehyde and permeabilized with Triton X-100. Co-application of the antibody targeting the protein of interest with markers for microtubules (gamma tubulin) and the endoplasmic reticulum (calreticulin) is essential. Nuclei should be counterstained using 4',6-diamidino-2-phenylindole (DAPI). Primary antibody detection involves species-specific secondary antibodies labelled with distinct fluorophores (Alexa Fluor 488 for the protein of interest, Alexa Fluor 555 for microtubules, and Alexa Fluor 647 for the ER). Imaging should be performed with a laser scanning confocal microscope (63X objective). Images, in PNG or JPG format of any size, can be integrated into the system. In multicolour images, fluorophores are represented as different channels, with the protein of interest in green, the nucleus in blue, microtubules in red, and the ER in yellow. By comparing the results with existing literature on subcellular localization, researchers can identify potential novel subcellular localizations of proteins. This capability not only enhances the understanding of cellular dynamics but also serves as a catalyst for generating new research topics. Researchers can leverage the system to uncover previously unexplored aspects, thereby creating opportunities for groundbreaking investigations and contributing to the expansion of protein subcellular localization knowledge.



## 6. CONCLUSION

Protein subcellular localization is an important data source in biology and bioinformatics studies, playing a key role in various areas such as, protein function and mechanism, drug discovery and development, disease mechanism, cell signalling, and structural biology. This study introduces HoliLoc, a novel approach for predicting protein subcellular localization through multi-modal deep learning. Our method incorporates 2D confocal microscopy images, amino acid sequences, and protein-protein interactions. The first purpose of this study is to explore the impact of employing various types of data on the prediction of protein subcellular localization and secondly, to develop a high-performance multi-modal deep learning model for accurate predictions in this task. The source code of this thesis and HoliLoc's dataset are publicly available at GitHub (<https://github.com/huBioDataLab/HoliLoc>) where it can be used as a programmatic tool for reproducibility in version 0.1.0.

HoliLoc stands as a significant study in predicting protein subcellular localization, making a valuable contribution to the field. Its unique strength lies in a holistic approach that integrates data from various sources, including image, amino acid sequences, and protein-protein interactions. HoliLoc's observed improvements in predictive performance, especially in difficult scenarios with unbalanced data, demonstrate its practical usage. These outcomes not only confirm the model's efficacy but also demonstrate how flexible it is in managing complex situations in the real world. The success of HoliLoc inspires an important change in the way we address complex biological problems, like subcellular protein location prediction.

Several limitations should be acknowledged in the context of this thesis. The most important one arises from the inherent nature of the protein subcellular localization data, leading to potential imbalances in the dataset. The distribution of proteins across different cellular locations is not uniform because of their various functionalities, which introduces bias into the generalizability of the models. Also, limitations in computational resources made it difficult to construct large models or conduct a more comprehensive hyperparameter search. Furthermore, the lack of protein IDs in the train and test splits of the HPA Kaggle challenge dataset made it

impossible to directly compare HoliLoc and the participating methods on the challenge dataset. Finally, issues regarding the public availability of confocal microscopy image data for proteins limit the usability of our methods for new proteins.

As future work, we plan to provide HoliLoc as a stand-alone, online tool for predicting human proteins' subcellular localization. Researchers will be able to utilise confocal microscopy images of their proteins, receiving predictions regarding the proteins' subcellular localizations. Active learning techniques will drive the model's continual enhancement by securing permissions from collaborating researchers to share the images they submit. Hence, augmenting the existing small image dataset continuously, ensuring a more robust and comprehensive training environment. Also, different strategies adopted by the winning teams in the HPA Kaggle challenge such as more advanced loss functions, augmenting data and using pretrained networks will be applied. The performance of HoliLoc will be enhanced using these strategies, while considering cost, resource-intensity, and usability. In this thesis, we did not utilise any of these approaches not to overshadow the transparent observation of the impact of holistic data integration. Certain subcellular locations, including aggresome, cytoplasmic bodies, mitotic spindle, nucleoli fibrillar centre, nuclear membrane, MTOC, and cytokinetic bridge, achieved a 0 F1 score in the multi-localization prediction setting. These locations are notably part of the smallest sample-sized group within the HoliLoc data, with the largest among them, nucleoli fibrillar centre, having a sample size of 238. In contrast, the largest sample-sized location in the entire HoliLoc dataset, nucleoplasm, boasts 4138 samples. The evident discrepancy in sample sizes highlights an opportunity for future exploration, particularly zero-shot and few-shot learning techniques, which involve leveraging unseen classes during training and learning from a minimal set of examples. By addressing the data shortage issue, through these innovative learning approaches, there exists potential to enhance the predictive performance of the model for subcellular locations with limited training samples. This strategic consideration underscores the importance of ongoing efforts to mitigate data imbalances and improve the model's accuracy, particularly for locations represented by smaller sample sizes. Finally, adding new modalities using language models with text data

from scientific articles as well as GO annotations and enzyme commission numbers appears as promising options for future research.

## 7. REFERENCES

1. Hung M, Link W. Protein localization in disease and therapy. *J Cell Sci* [Internet]. 2011 Oct;15;124(20):3381–3392. Available from: <https://doi.org/10.1242/jcs.089110>
2. The human subcellular proteome - The Human Protein Atlas [Internet]. Available from: <https://www.proteinatlas.org/humanproteome/subcellular>
3. Jiang Y, Wang D, Yao Y, Eubel H, Künzler P, Møller IM, et al. MULocDeep: A deep-learning framework for protein subcellular and suborganellar localization prediction with residue-level interpretation. *Comput Struct Biotechnol J*. 2021;19:4825–39.
4. Jadot M, Boonen M, Thirion J, Wang N, Xing J, Zhao C, et al. Accounting for protein subcellular localization: a compartmental map of the rat liver proteome. *Mol Cell Proteomics*. 2017 Feb;1;16(2):194-212.
5. Allaire A, Picard-Jean F, Bisailon M. Immunofluorescence to monitor the cellular uptake of human lactoferrin and its associated antiviral activity against the hepatitis C virus. *JoVE J Vis Exp*. 2015 Oct;1(104).
6. Im K, Mareninov S, Díaz MF, Yong WH. An introduction to performing immunofluorescence staining. *Methods Mol Biol*. 2018;299–311.
7. Stadler C, Rexhepaj E, Singan V, Murphy RF, Pepperkok R, Uhlén M, et al. Immunofluorescence and fluorescent-protein tagging show high correlation for protein localization in mammalian cells [Internet]. *Nature Methods*. 2013. Available from: <https://doi.org/10.1038/nmeth.2377>
8. Assays and annotation - The Human Protein Atlas [Internet]. Available from: <https://www.proteinatlas.org/about/assays+annotation>
9. Al-Hamdani YS, Tkatchenko A. Understanding non-covalent interactions in larger molecular complexes from first principles. *J Chem Phys*. 2019 Jan 3;150(1):010901.
10. Plewcyński D, Ginalski K. The interactome: Predicting the protein-protein interactions in cells. *Cell Mol Biol Lett*. 2009;14(1):1–22.
11. Scott MS, Thomas DY, Hallett M. Predicting subcellular localization via protein motif Co-Occurrence [Internet]. *Genome Research*. 2004. Available from: <https://doi.org/10.1101/gr.2650004>
12. Shin CJ, Wong S, Davis MJ, Ragan MA. Protein-protein interaction as a predictor of subcellular location. *BMC Syst Biol* [Internet]. 2009;3(1). Available from: <https://doi.org/10.1186/1752-0509-3-28>
13. Bajpai AK, Davuluri S, Tiwary K, Narayanan S, Oguru S, Basavaraju K, et al. Systematic comparison of the protein-protein interaction databases from a user's perspective. *J Biomed Inform* [Internet]. 2020;103(103380). Available from: <https://doi.org/10.1016/j.jbi.2020.103380>
14. Embl-Ebi. IntACT Portal [Internet]. Available from: <https://www.ebi.ac.uk/intact/home>

15. Biochemistry, Essential Amino Acids - StatPearls - NCBI Bookshelf [Internet]. [cited 2023 Dec 10]. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK557845/>
16. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Protein Function. In: Molecular Biology of the Cell 4th edition [Internet]. Garland Science; 2002 [cited 2023 Dec 11]. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK26911/>
17. Consortium U. The universal protein resource (UniProt. Nucleic Acids Res. 2007 Nov;27;36(suppl\_1):D190-5.
18. Hadzic M, Wongthongtham P, Dillon T, Chang E. Introduction to Ontology. In: Ontology-Based Multi-Agent Systems Studies in Computational Intelligence [Internet]. Berlin, Heidelberg: Springer; 2009. Available from: [https://doi.org/10.1007/978-3-642-01904-3\\_3](https://doi.org/10.1007/978-3-642-01904-3_3)
19. Thomas PD. The Gene Ontology and the Meaning of Biological Function. In: Dessimoz C, Škunca N, editors. The Gene Ontology Handbook Methods in Molecular Biology, vol 1446 [Internet]. New York, NY: Humana Press; 2017. Available from: [https://doi.org/10.1007/978-1-4939-3743-1\\_2](https://doi.org/10.1007/978-1-4939-3743-1_2)
20. GO A. Gene Ontology Resource [Internet]. Available from: <https://geneontology.org/docs/introduction-to-go>
21. Russell SJ, Norvig P. Artificial intelligence: A modern approach. 4th ed. Pearson; 2021.
22. Sharma N, Sharma R, Jindal N. Machine Learning and Deep Learning Applications- A Vision. Glob Transit Proc. 2021;2(1):24–8.
23. Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. Electron Mark. 2021;31(3):685–95.
24. Bansal M, Goyal A, Choudhary A. A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. Decis Anal J [Internet]. 2022;3(100071). Available from: <https://doi.org/10.1016/j.dajour.2022.100071>
25. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature [Internet]. 2015 May;28;521(7553):436–444. Available from: <https://doi.org/10.1038/nature14539>.
26. Svozil D, Kvasnička V, Pospíchal J. Introduction to multi-layer feed-forward neural networks. Chemom Intell Lab Syst. 1997;39(1):43–62.
27. Chauhan R, Ghanshala KK, Joshi RC. Convolutional Neural Network (CNN) for Image Detection and Recognition. In: 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC [Internet]. p. 2018 278-282. Available from: <https://doi.org/10.1109/ICSCCC.2018.8703316>
28. Hao W, Yizhou W, Yaqin L, Zhili S. The Role of Activation Function in CNN. In: 2020 2nd International Conference on Information Technology and Computer Application (ITCA [Internet]. p. 2020 429-432. Available from: <https://doi.org/10.1109/ITCA52113.2020.00096>

29. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili AQ, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* [Internet]. 2021;8(1). Available from: <https://doi.org/10.1186/s40537-021-00444-8>
30. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15, 1:1929–58.
31. Chen F, Wang Y, Wang B, Kuo CJ. Graph representation learning: a survey. *APSIPA Trans Signal Inf Process* [Internet]. 2020;9(1). Available from: <https://doi.org/10.1017/atsip.2020.13>
32. Grover A, node2vec LJ. Scalable Feature Learning for Networks [Internet]. 2016. Available from: <https://doi.org/10.48550/arxiv.1607.00653>
33. Dallago C, Schütze K, Heinzinger M, Olenyi T, Littmann M, Lu AX, et al. Learned Embeddings from Deep Learning to Visualize and Predict Protein Sets. *Curr Protoc* [Internet]. 2021;1(5). Available from: <https://doi.org/10.1002/cpz1.113>
34. Raffel C, N S. Exploring the limits of transfer learning with a unified text-to-text transformer. 2019.
35. Clark K, M-T L. Electra: Pre-training text encoders as discriminators rather than generators. 2020.
36. Chang DMW. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
37. Lan Z, M C. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations.
38. Dai Z, Z Y. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. 2019.
39. Yang Z, Z D. XLNet: Generalized Autoregressive Pretraining for Language Understanding.
40. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: Towards Cracking the language of life’s Code through Self-Supervised Learning. *bioRxiv* (Cold Spring Harbor Laboratory). 2020 [Internet]. Available from: <https://doi.org/10.1101/2020.07.12.199554>
41. JJ AA, CK S, SK S, H N, O W. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* [Internet]. 2017 Jul;7;33(21):3387–3395. Available from: <https://doi.org/10.1093/bioinformatics/btx431>
42. Qin Y, Huttlin EL, Winsnes CF, Gosztyla ML, Wacheul L, Kelly MR, et al. A multi-scale map of cell structure fusing protein images and interactions. *Nature* [Internet]. 2021 Nov;24;600(7889):536–542. Available from: <https://doi.org/10.1038/s41586-021-04115-9>

43. Özşarı G, Rifaioglu AS, Atakan A, Doğan T, Martin MJ, Çetin Atalay R, et al. SLPRED: A multi-view subcellular localization prediction tool for multi-location human proteins. *Bioinformatics*. 2022;38(17):4226–9.
44. Ouyang W, Winsnes CF, Hjelmare M, Cesnik AJ, Åkesson L, Xu H, et al. Analysis of the human protein atlas image classification competition. *Nat Methods*. 2019;Dec;16(12):1254–61.
45. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digit Med [Internet]*. 2020 Oct;16;3(1):136. Available from: <https://doi.org/10.1038/s41746-020-00339-7>
46. Protein Embeddings | UniProt help | UniProt [Internet]. [cited 2023 Dec 11]. Available from: <https://www.uniprot.org/help/embeddings>
47. Smith MJ, Bowman AJ. Observation of histone nuclear import in living cells: implications in the processing of newly synthesised H3.1 & H4 [Internet]. *bioRxiv*; 2017 [cited 2023 Dec 20]. p. 111096. Available from: <https://www.biorxiv.org/content/10.1101/111096v1>
48. Blotnick E, Sol A, Muhlrad A. Histones bundle F-actin filaments and affect actin structure. *PLoS ONE*. 2017 Aug 28;12(8):e0183760.
49. Itakura M, Yamaguchi K, Kitazawa R, Lim SY, Anan Y, Yoshitake J, et al. Histone functions as a cell-surface receptor for AGEs. *Nat Commun*. 2022 May 27;13(1):2974.
50. Murakami H, Blobel G, Pain D. Signal sequence region of mitochondrial precursor proteins binds to mitochondrial import receptor. *Proc Natl Acad Sci U S A*. 1993 Apr 15;90(8):3358–62.
51. Farley KI, Surovtseva Y, Merkel J, Baserga SJ. Determinants of mammalian nucleolar architecture. *Chromosoma*. 2015 Sep;124(3):323–31.
52. Dundr M, Misteli T. Biogenesis of Nuclear Bodies. *Cold Spring Harb Perspect Biol*. 2010 Dec;2(12):a000711.
53. Spector DL, Lamond AI. Nuclear speckles. *Cold Spring Harb Perspect Biol*. 2011 Feb 1;3(2):a000646.
54. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. The Endoplasmic Reticulum. In: *Molecular Biology of the Cell 4th edition [Internet]*. Garland Science; 2002 [cited 2023 Dec 11]. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK26841/>
55. Remsburg CM, Konrad KD, Song JL. RNA localization to the mitotic spindle regulated by kinesin-1 and dynein is essential for early development of the sea urchin embryo [Internet]. *bioRxiv*; 2022 [cited 2023 Dec 19]. p. 2022.08.16.504170. Available from: <https://www.biorxiv.org/content/10.1101/2022.08.16.504170v1>
56. Sanchez AD, Feldman JL. Microtubule-organizing centers: from the centrosome to non-centrosomal sites. *Curr Opin Cell Biol*. 2017 Feb;44:93–101.

## 8. APPENDIX

### EK-1: Tez Çalışması ile İlgili Etik Kurul İzinleri

Sayı : 16969557-203

Konu :

18.01.2022

**Doç. Dr. Tunca DOĞAN**  
Mühendislik Fakültesi  
Bilgisayar Mühendisliği Bölümü  
Yapay Zeka Mühendisliği Anabilim Dalı  
Öğretim Üyesi

Sayın Doç. Dr. DOĞAN,

Kurulumuza değerlendirilmek üzere sunduğunuz GO 22/116 kayıt numaralı ve "*Proteinlerin Subselüler Yerleşimlerinin Görüntü, Sekans ve İnteraktom Verisi Tabanlı Tahmini*" başlıklı proje Kurulumuzun 18.01.2022 tarihli toplantısında değerlendirilmiş olup, çalışmanın erişime açık veri tabanlarından veri toplanması yolu ile yapılacağı görülmüştür. Gönüllü insanlar üzerinde gerçekleştirilecek nitelikte olmayan bu tip çalışmalar Etik Kurulların kapsamı dışında kalmaktadır.

Bu yazı ilgili protokolün bilimsel ve etik açıdan incelendiğini belirtmek için Etik Kurul kararı yerine geçmek üzere hazırlanmıştır.

Prof. ~~Dr.~~ G. Burça AYDIN  
Başkan

EK \_\_\_\_\_ :  
Toplantı Katılım Tutanağı.



**EK-2: Tez Çalışması Orijinallik Raporu****Digital Receipt**

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author: Ecem Kuşçuoğlu  
Assignment title: EcemKuscuoglu\_MSc\_Thesis\_FirstDraft  
Submission title: EcemKuscuoglu\_MSc\_Thesis\_FinalDraft  
File name: HoliLoc\_CV.pdf  
File size: 3.68M  
Page count: 103  
Word count: 20,522  
Character count: 112,547  
Submission date: 29-Jan-2024 12:01AM (UTC+0300)  
Submission ID: 2249974212



## EcemKuscuoglu\_MSc\_Thesis\_FinalDraft

### ORJİNALLIK RAPORU

% <b>10</b>	% <b>6</b>	% <b>7</b>	% <b>3</b>
BENZERLİK ENDEKSİ	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCİ ÖDEVLERİ

### BİRİNCİL KAYNAKLAR

<b>1</b>	<a href="http://www.ncbi.nlm.nih.gov">www.ncbi.nlm.nih.gov</a> İnternet Kaynağı	<% <b>1</b>
<b>2</b>	<a href="http://www.proteinatlas.org">www.proteinatlas.org</a> İnternet Kaynağı	<% <b>1</b>
<b>3</b>	<a href="http://www.mdpi.com">www.mdpi.com</a> İnternet Kaynağı	<% <b>1</b>
<b>4</b>	Submitted to KTH - The Royal Institute of Technology Öğrenci Ödevi	<% <b>1</b>
<b>5</b>	<a href="http://link.springer.com">link.springer.com</a> İnternet Kaynağı	<% <b>1</b>
<b>6</b>	Yuexu Jiang, Duolin Wang, Yifu Yao, Holger Eubel, Patrick Künzler, Ian Max Møller, Dong Xu. "MULocDeep: A deep-learning framework for protein subcellular and suborganellar localization prediction with residue-level interpretation", Computational and Structural Biotechnology Journal, 2021 Yayın	<% <b>1</b>
<b>7</b>	<a href="http://openaccess.hacettepe.edu.tr">openaccess.hacettepe.edu.tr</a>	

## **9. CURRICULUM VITAE**