


Review

# A Review for Artificial-Intelligence-Based Protein Subcellular Localization

Hanyu Xiao <sup>1</sup>, Yijin Zou <sup>2</sup>, Jieqiong Wang <sup>3</sup>  and Shibiao Wan <sup>1,\*</sup>

<sup>1</sup> Department of Genetics, Cell Biology and Anatomy, College of Medicine, University of Nebraska Medical Center, Omaha, NE 68198, USA; haxiao@unmc.edu

<sup>2</sup> College of Veterinary Medicine, China Agricultural University, Beijing 100193, China; 2020305010318@cau.edu.cn

<sup>3</sup> Department of Neurological Sciences, College of Medicine, University of Nebraska Medical Center, Omaha, NE 68198, USA; jiwang@unmc.edu

\* Correspondence: swan@unmc.edu; Tel.: +1-402-559-6560

**Abstract:** Proteins need to be located in appropriate spatiotemporal contexts to carry out their diverse biological functions. Mislocalized proteins may lead to a broad range of diseases, such as cancer and Alzheimer's disease. Knowing where a target protein resides within a cell will give insights into tailored drug design for a disease. As the gold validation standard, the conventional wet lab uses fluorescent microscopy imaging, immunoelectron microscopy, and fluorescent biomarker tags for protein subcellular location identification. However, the booming era of proteomics and high-throughput sequencing generates tons of newly discovered proteins, making protein subcellular localization by wet-lab experiments a mission impossible. To tackle this concern, in the past decades, artificial intelligence (AI) and machine learning (ML), especially deep learning methods, have made significant progress in this research area. In this article, we review the latest advances in AI-based method development in three typical types of approaches, including sequence-based, knowledge-based, and image-based methods. We also elaborately discuss existing challenges and future directions in AI-based method development in this research field.

**Keywords:** protein subcellular localization; machine learning; deep learning; artificial intelligence; gene ontology; sequence analysis



**Citation:** Xiao, H.; Zou, Y.; Wang, J.; Wan, S. A Review for Artificial-Intelligence-Based Protein Subcellular Localization. *Biomolecules* **2024**, *14*, 409. <https://doi.org/10.3390/biom14040409>

Academic Editor: Le Nguyen Quoc Khanh

Received: 29 February 2024

Revised: 21 March 2024

Accepted: 25 March 2024

Published: 27 March 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

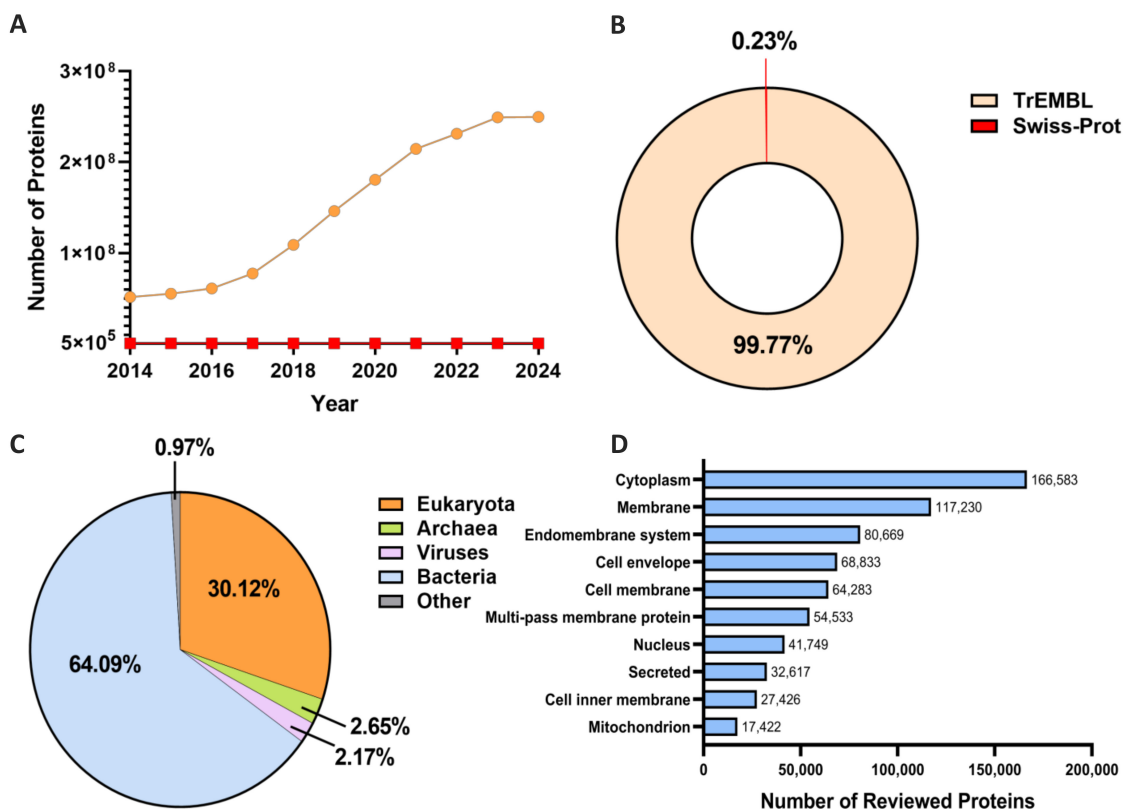
## 1. Introduction

Within a cell, mature proteins must reside in specific subcellular structures to properly perform their biological roles, as different cellular compartments provide distinct chemical environments (e.g., pH and redox conditions), potential interacting partners, or substrates for diverse functions [1,2]. Most cellular biological processes, such as the nucleocytoplasmic shuttling of transcription factors [3], the relocalization of mitochondrial proteins during apoptosis [4], and the endocytic uptake of cell-surface cargo receptors, all rely on precise protein localization. Conversely, mislocalization is often associated with cellular dysfunction and diseases, such as cancer [5,6], neurodegenerative diseases [7,8], and metabolic disorders [9,10].

Conventionally, identifying subcellular localization of proteins primarily relies on wet lab experimental methods. Fluorescence microscopy imaging, which applies fluorescent dyes or fluorescent protein tags to label target proteins, has commonly been used for observing their distribution within cells [11,12]. This method has become one of the preferred tools for studying protein subcellular localization due to its high resolution and real-time observation advantages [13]. By using labeled antibodies against target proteins, the immunoelectron microscopy technique is regarded as a gold standard to provide the high resolution of electron microscopy [14]. Another method involves the use of fluorescent biomarker tags [15] like the protein A-GFP tag, which fuses a fluorescent protein with the target protein, allowing it to emit a fluorescent signal among different cell

compartments [16]. These experimental methods yield high-resolution location of targeted proteins for researchers, enabling direct observation to uncover biological processes and metabolic mechanisms.

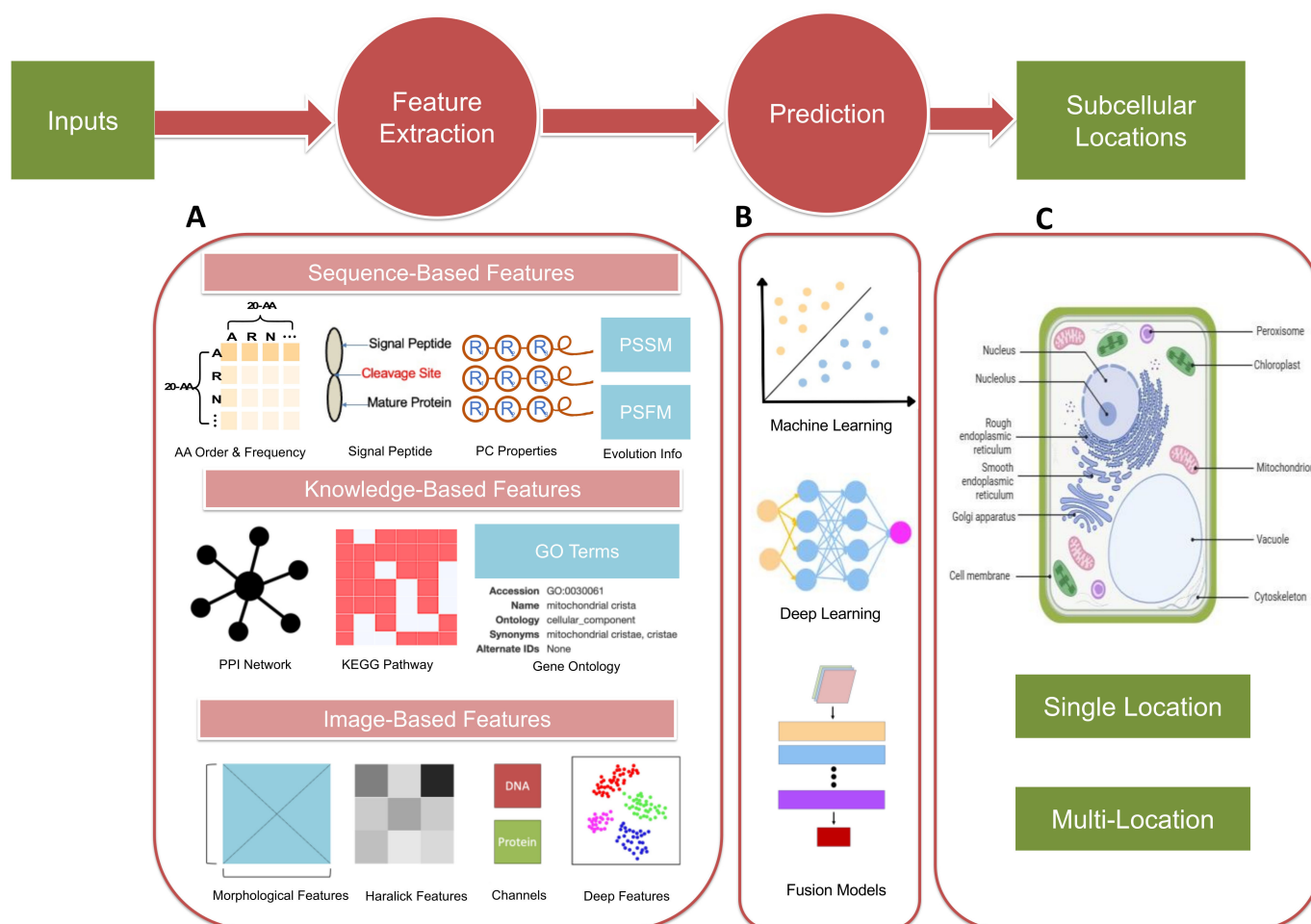
However, these wet lab experimental methods also have some significant drawbacks: they often require expensive equipment and time-consuming steps, making them costly for large-scale studies. These problems are exacerbated given that the number of newly discovered proteins has increased exponentially in the post-genomic era. Take the UniProt Database [17] as an example. The gap between the reviewed and unreviewed proteins has significantly expanded during the past decade (Figure 1A). Specifically, as shown in Figure 1B, in the latest 2024\_01. version of UniProt, a notable majority of data entries are unreviewed proteins in TrEMBL. In this case, implementing wet lab experiments alone for subcellular localization determination for remarkably large amounts of data from different species (Figure 1C) becomes an impossible mission. Moreover, the rich collection of accurately annotated protein data in databases (Figure 1D) can facilitate the development of robust prediction methods. It is noteworthy that, compared to TrEMBL, the smaller size of Swiss-Prot can be attributed to the rigorous manual curation of proteins. Conversely, TrEMBL comprises computationally analyzed records, leading to a plethora of protein sequences awaiting annotation before being entered into Swiss-Prot. Fortunately, the necessity of manual curation might be alleviated if transcript-translated sequences can be validated through proteomics detection. An example of such an approach can be observed in the Human Protein Atlas (HPA) [18,19], as we will elaborate in subsequent sections, where RNA-seq data were employed to corroborate immunofluorescence subcellular localization findings. In this context, leveraging computational models, particularly AI-assisted methodologies renowned for their adeptness in handling large-scale datasets, can offer substantial benefits.



**Figure 1.** Statistical analysis of UniProtKB [17] (2024\_01.version). (A) The trend of protein number growth in TrEMBL (unreviewed proteins) and Swiss-Prot (reviewed proteins). The number of newly discovered unannotated proteins far exceeds that of newly added experimentally validated proteins. (B) The proportion of newly added protein counts between the two databases in the 2024\_01.version. (C) Taxonomic distribution of protein sequences. (D) Number of proteins in the top 10 subcellular locations.

Recent decades have witnessed the booming of *in silico* methods for protein subcellular location prediction. Based on features used for computational modeling, most existing methods can be generally divided into three main categories: (1) sequence-based methods, which only use the amino acid sequence of the query protein as inputs; (2) knowledge-based methods, using protein annotations from multiple databases to correlate the information with their subcellular locations; and (3) image-based methods, extracting subcellular location features from bioimages and then identifying the likelihood of proteins being located in various subcellular compartments. The primary sequence for a protein is much easier to obtain with existing sequencing technologies. With remarkable advances in machine learning and deep learning, coupled with an increasing number of proteins with experimentally determined localization information as well as functional annotations and imaging records in publicly available databases, accurate and efficient computational frameworks provide a promising way for protein subcellular localization.

In this review, we will first present some remarkable progress in *in silico* models, including the three major types of models mentioned above. In Section 2, we will introduce common features and algorithms used in sequence-based methods and also for knowledge-based and image-based frameworks in Sections 3 and 4, respectively. The simplified flowchart for the prediction frameworks mentioned is illustrated in Figure 2. In Section 5, we will give an overview of protein subcellular localization models that are specially designed for different species. Lastly, we will explore the existing challenges and future trajectories of this research domain and propose our expectations.



**Figure 2.** The flowchart of three major types of AI-based prediction methods. The procedures include sequences or images as input, feature extraction, model prediction, and subcellular location output.

(A) Key features extracted from sequences, annotations, and image inputs. Different classifiers extract composition information, encompassing AA order and frequency, physicochemical properties, and identifying signal peptide cleavage sites from sequence inputs. In addition to straightforward data, evolutionary profiles are also considered through homology alignment with the Position-Specific Scoring Matrix (PSSM) and the Position-Specific Frequency Matrix (PSFM). Knowledge-based methods involve the establishment of Gene Ontology (GO) vectors, derived from GO terms collected from specific databases with protein sequences or accession numbers as keywords. Other functional annotations, such as protein–protein interaction (PPI) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway information, can also be fused as input features. Imaging features mainly consist of morphological, Haralick data and information from different channels, namely hand-crafted features, and deep features captured by deep learning algorithms. (B) Three types of algorithms used for prediction modules in computational models. (C) Major subcellular locations in a plant cell as an example of potential outputs for proteins with single or multiple locations.

## 2. Sequence-Based Methods

Sequence-based methods directly use the amino acid sequence of a query protein as model inputs and attempt to find the correlations between protein sequences and their subcellular locations. With the advent of high-throughput sequencing technologies, large-scale genomic and proteomic data are easily obtained, allowing new big-data-based models to be constructed. In addition, as proteins consist of sequences of amino acids, they are fit for computational models that extract features for subcellular localization. However, protein sequences might not capture full information for protein subcellular localization, particularly in the cases of protein post-translational modifications or protein dynamics processes within cells once the protein is synthesized, which may influence where proteins reside.

### 2.1. Sequence-Based Features

In protein primary sequences, the 20 standard amino acids (AA) exert different biochemical properties such as hydrophobicity, hydrophilicity, side-chain characters, etc. Sequence-based methods intend to make predictions out of the correlations between protein subcellular locations and the information embedded in amino acid sequences. There are three major types of features used for model construction: AA composition information, sorting signal information, and evolutionary information.

The composition-based features, which include AA occurrences and order in the query sequence, were commonly used in the earliest subcellular prediction methods. Moreover, previous studies have confirmed a better performance of the model by combining AA original sequence, gapped amino acid composition (GapAA) [20], and amino-acid-pair composition (PairAA) [21]. Based on AA-composition features, Chou [22] proposed pseudo-amino-acid composition (PseAA) using the sequence-order correlation factor for greater biomedical property discovery when avoiding the high-dimensional vector formation. The simplicity of composition features helps the generalization and interpretation of the computational models since they capture the most basic trends in protein sequences associated with their locations. However, they may not provide sufficient resolution for a high accuracy rate, since there is a loss of information about important sequences or structural motifs highly related to proteins' subcellular location.

The sorting signal sequences or signal peptides, including transit peptides like mitochondrial transit peptides (mTPs) and chloroplast transit peptides (cTPs) [23], are short and cleavable segments of amino acid sequences added to newly synthesized proteins, determining their destination in the transportation process. These short peptides possess the directions mature proteins should be transported, reflecting the possible location event for one protein [24]. Available approaches with signal peptides for protein localization mainly refer to finding their cleavage sites [25]. As described in previous studies, sorting-signal sequences vary in length and composition but have similar structures: the N-terminal flanking region, also known as the n-region, the central hydrophobic region (h-region), and

the C-terminal flanking region (c-region) [26]. The hydrophobicity in the h-region and a large proportion of nonpolar residues in the c-region are used to label the cleavage sites by computational methods [27,28]. According to the location signal embedded in those short peptides, one can mimic the de facto information processing in cells and find the target spot of the test protein.

In addition, based on the fact that homologous sequences are likely to share the same subcellular location, the unknown protein can be assigned the same subcellular location as its homologs generated from PSI-BLAST [29]. Moreover, the evolutionary similarity profiles extracted from the position-specific scoring matrix (PSSM) and position-specific frequency matrix (PSFM) derived from multiple sequence alignment results can contribute as classification features providing valuable information such as conserved motifs or targeting signals among different protein families. This representation can also be extended by integrating pseudo-analysis [30]. Once aligned with known homologs in the database, this method can achieve high accuracy. However, as one amino acid change can directly influence the characters of one protein sequence, this method is more likely to be one of the sources of the feature basis of prediction models.

## 2.2. Sequences-Based AI Approaches

Most computational frameworks include three major steps: feature extraction, feature selection, and final classification. Considering common features discussed above, the complexity of the models developed also increases with the amount of data processed and the dimension of input features, from traditional machine learning classification to complex deep learning analytical models. Besides the development of computational frameworks, we will also introduce techniques that are used to improve the algorithms dealing with multi-location proteins in the following.

For conventional classification, the Support Vector Machine (SVM) [31], K-Nearest Neighbor (KNN) [32], and Random Forest (RF) [33,34] are widely chosen classifiers for training. Their simplicity makes them easy to use for prediction protocols with fast speed and low computational cost, suitable for limited data and low-dimensional inputs. Combined with highly efficient feature extraction methods, these frameworks will work well in most cases [35]. For instance, Du et al. [36] proposed two novel feature extraction methods that utilize evolutionary information via the transition matrix of the consensus sequence (CTM) and PSSM before adopting SVM, which, in the end, reached an overall accuracy of 99.7% in CL317 dataset. A feature-extraction-based hierarchical extreme learning machine (H-ELM) introduced by Zhang et al. [37] can handle high-dimension feature inputs directly without demanding dimension reduction for acceptable results. Alaa et al. [38] exploits an extended Markov chain to provide the latent feature vector, which records micro-similarities between the given sequence and their counterparts in reference models. These methods help extract more abundant features of query sequences and provide better performance.

However, these conventional models may not perform well in complex scenarios [1], especially multi-locational protein prediction [30]. Though many proteins only stay in one subcellular space, studies have discovered many multi-location proteins that have special functions or are involved in crucial biological steps [39]. Moreover, rather than staying in one place, proteins move from one subcellular compartment to another or simultaneously reside at two locations and participate in different cellular processes [40]. Recent studies have also shown the remarkable significance of multilocation proteins in cell growth and development [41]. For instance, phosphorylation-related multilocation proteins can function as a “needle and thread” via protein–protein interactions (PPI), thus playing an important role in organelle communication and regulating plant growth [42]. Under these circumstances, there are mainly two ways for predicting multi-location proteins based on conventional classifiers: algorithm adaption and problem transformation. The former method extends existing algorithms to deal with multi-label problems. Jiang et al. [43] considers weighted prior probabilities with a multi-label KNN algorithm to increase the model accuracy. Library of SVM (LIBSVM) toolbox [36,44], instead, uses a one-versus-one



(OVO) strategy to solve multi-class classification problems. Customization of well-known algorithms enhances their ability for specific requirements but there is a risk of overfitting and it may require significant computational resources. The problem transformation approach focuses on transforming the original problem into a different representation or formulation that is solvable with existing algorithms [45,46], such as converting a multi-location classification problem into multiple single-label classification problems [47]. Shen et al. [30] introduces multi-kernel SVM by training multiple independent SVM classifiers to solve single-label problems before combining their results, one classifier for each class. Following this idea, an algorithm can be easily extended to solve multi-label classification.

In summary, traditional machine learning algorithms can achieve fast training times and high accuracy in scenarios with well-organized feature spaces and clear decision boundaries; their performance may degrade quickly when faced with large-scale data inputs, even with tailored classifiers featuring more selected features. Dimension reduction [48] and parallel processing [49] can be applied to mitigate the challenges, allowing an improved computational method scalability.

As multi-layered structure provides better performance compared to traditional approaches [33], more methods based on deep networks, especially neural networks, have become increasingly popular in protein subcellular localization research [50,51]. Starting as effective feature extractors which automatically obtain deep features embedded in sequences [52], convolutional neural network (CNN) is widely implanted in multi-locus protein localization framework. Mining deeper, Kaleel et al. [53] ensemble Deep N-to-1 Convolutional Neural Networks that predict the location of the endomembrane system and secretory pathway versus all others and outperform many state-of-the-art web servers. Cong et al. [54] proposed a self-evolving deep convolutional neural network (DCNN) protocol to solve the difficulties in feature correlation between sites and avoid the impact of unknown data distribution while using the self-attention mechanism [55] and a customized loss function to ensure the model performance. In addition, a long short-term memory network (LSTM) which combines the previous states and current inputs is also commonly used [56,57], with Generative Adversarial Network (GAN) [58] and Synthetic Minority Over-sampling Technique (SMOTE) [59] used for synthesizing minority samples to deal with data imbalance. Developing data augmentation methods by deep learning algorithms has also made protein language model construction possible [60,61]. Through transfer learning [62], pretrained models can be fine-tuned on different downstream tasks, reducing the need for large amounts of labeled data for training. For example, Heinzinger et al. [63] proposed Sequence-to-Vector (SeqVec) that embeds biophysical properties of protein sequences as continuous vectors by using the natural language processing model ELMo on unlabeled big data. This represents a way to speed up the prediction process independent of the size of inputs. As protein sequences can also be tokenized and coded with a certain pattern as natural languages [64], some well-developed models (e.g., Universal Language Model Fine-tuning (ULMFiT)) [65] have also been repurposed to protein-related questions, like AlphaFold [66] for protein 3D structure prediction and ProteinBERT [67] for bidirectional sequence modeling and Gene Ontology annotation prediction. ESM2 [67], the pretrained deep language model inspired by Bidirectional Encoder Representations from Transformers (BERT) [68], allows fast training of localization predictors with limited labeled protein sequence data [69]. Transformer-based architectures leverage self-attention mechanisms, enabling the capture of key hidden vectors for subcellular localization from sequences. This intrinsic ability facilitates information exchange across all positions without necessitating pooling operations like CNN, greatly enhancing the ability of extracting deeper information. However, this will take more time and larger computational resources for training to gain higher model performance, since the results may be similar to simple classifiers when the prediction scale is small [60]. Details of the computational models mentioned above can be found in Table 1.

**Table 1.** A summary of state-of-the-art sequence-based protein subcellular localization frameworks. S: Single-Location; M: Multi-Location; Pub: Publication Cited; BERT: Bidirectional Encoder Representations from Transformers; XGBoost: Extreme Gradient Boosting; GAN: Generative Adversarial Network; CNN: Convolutional Neural Network; LM: Language Model; MLP: Multilayer Perceptron; SP: Signal Peptide; PC: Physicochemical Properties; PSSM: Protein-Specific Scoring Matrix; LSTM: Long Short-Term Memory; CTM: Consensus Sequence; AECA: Absolute Entropy Correlation Analysis; LDA: Linear Discriminant Analysis; SVM: Support Vector Machine; MAM: Multi-Attention Mechanism; PseAAC: Pseudo Amino Acid Composition; SAAC: Split Amino Acid Composition; KNN: K-Nearest Neighbor; AAF: Amino Acid Frequencies; GCF: Gene Co-expression Features; DNN: Deep Neural Network; AAC: Amino Acid Composition; Acc: Accuracy; Prec: Precision; Rec: Recall; F1: F1 Score; GM: Grand Mean; MicroF1: MicroF1 Score; MacroF1: MacroF1 Score; MCC: Matthews Correlation Coefficient; Jaccard: Jaccard Value; AUC: Area Under the Curve; Spec: Specificity; Sen: Sensitivity; FPR: The False Positive Rate; HL: Hamming Loss; RL: Ranking Loss; OE: One Error; CV: Coverage.

Method	Features	Algorithm	S/M-Location	Species	Performance Metrics <sup>1</sup>	Pub	Year
DaDL-SChlo	Deep- and Hand-crafted features	ProtBERT, XGBoost, GAN, CNN	M	Plants	Acc: 0.86~0.94 Prec: 0.88~0.95 Rec: 0.86~0.94 F1: 0.86~0.95 GM: 0.84~0.94	[61]	2023
DeepLoc—2.0	Masked-LM Objective	MLP, Protein LM	M	Eukaryotes	Acc: 0.39~0.73 MicroF1: 0.60~0.73 MacroF1: 0.46~0.66 MCC: 0.17~0.90 Jaccard: 0.53~0.69	[70]	2022
SignalP—6.0	SP	Transformer Protein LM	M	Archaea, Gram-positive Bacteria, Gram-negative Bacteria and Eukaryotes	MCC: 0.65~0.89 Prec: 0.53~0.94 Rec: 0.50~0.88	[28]	2022
MULocDeep <sup>2</sup>	PC, PSSM	LSTM	M	Viridiplantae, Metazoa, Fungi	AUC: 0.74~0.95	[71]	2021
SCLpred-EMS <sup>3</sup>	Sequence Motifs	Deep N-to-1 CNN	S	Eukaryotes	MCC: 0.75~0.86 Spec: 0.89~0.97 Sen: 0.75~0.89 FPR: 0.02~0.05	[53]	2020
CTM-AECA-PSSM-LDA	CTM, AECA-PSSM	LDA, SVM	S	Apoptosis Proteins on CL317 and ZW225 datasets	Acc: 0.95~0.99 MCC: 0.90~1.00 Spec: 0.94~1.00 Sen: 0.91~0.95	[36]	2020

Table 1. Cont.

Method	Features	Algorithm	S/M-Location	Species	Performance Metrics <sup>1</sup>	Pub	Year
TargetP—2.0	SP	LSTM, MAM	S	Plants and Non-plants	Prec: 0.75~0.98 Rec: 0.75~0.98 F1: 0.75~0.98 MCC: 0.75~0.97	[27]	2019
Javed and Hayat	PseAAC, SAAC	ML-KNN, Rank-SVM	M	Bacteria, Virus	Acc: 0.80~0.85 Prec: 0.88~0.90 HL: 0.07~0.09 RL: 0.07~0.08 OE: 0.17~0.20 CV: 0.26~0.51	[35]	2019
MU-LOC <sup>4</sup>	AAF, PSSM, GCF	DNN, SVM	S	Plants (Mitochondrian)	Acc: 0.74~0.94 Prec: 0.74~0.82 MCC: 0.50~0.67 Spec: 0.88~0.97 Sen: 0.60~0.70	[72]	2018
MultiP-SChlo	PseAAC	SVM	M	Plants (Subchloroplast)	Acc: 0.55~0.60 Prec: 0.64~0.65 Rec: 0.66~0.71 F1: 0.65~0.67	[73]	2015
SlocX	AAC, Gene Expression Profile	SVM	S	Plants	Prec: 0.83 MCC: 0.48 Sen: 0.33	[74]	2011

<sup>1</sup> The entries in this column are directly collected from the respective original publications. <sup>2</sup> Web server available at <http://mu-loc.org>. <sup>3</sup> Web server available at <http://distilldeep.ucd.ie/SCLpred2/>. <sup>4</sup> Available at <http://mu-loc.org>.



Deep learning will demonstrate exceptional outcomes dealing with high-dimensional inputs with deep feature extraction, eliminating the need for manual feature engineering and capturing intricate patterns in sequences. However, large, labeled, and high-quality datasets are still needed for original model training, which results in too many hyper-parameters and makes it hard to interpret the model itself [33].

### 3. Knowledge-Based Methods

There is a strong correspondence between annotations and subcellular locations of proteins. Knowledge-based methods for protein localization prediction mainly extract information from annotation databases and convert them into numeric features as model inputs. Since the annotations are generated based on biological processes, functions, or protein interactions within cells, models can provide more interpretable results for subcellular localization. But knowledge data are limited and only applicable to well-curated proteins, which limits the predictive power of this kind of method for novel or newly discovered proteins. In recent studies [75–77], different kinds of information are fused together for better model performance, given that computational methods excel with high dimensional data as inputs.

#### 3.1. Legitimacy of Using Gene Ontology (GO) Features

Knowledge-based methods tend to dig into the correlation between the annotation of one protein and its subcellular location to establish predictors. Compared to Swiss-Prot keywords [78,79] or PubMed abstracts [80,81], Gene Ontology (GO)-terms-based methods are more attractive for the following reasons.

GO terms describe reviewed knowledge of the biological domain in three aspects: (1) Molecular Function, representing activities that can be performed by individual or by assembled complexes of gene products at the molecular level; (2) Cellular Component, labeling locations relative to cellular compartments; and (3) Biological Process, describing the events achieved by one or more ordered assemblies of molecular functions. This well-organized information can be used for protein subcellular localization because, (1) instead of table-lookup, which is dependent on cellular component GO terms, they perform deeper mining into items to accumulate every related GO category to improve prediction results; (2) the methods outperform previous sequence-based methods without compromising either inputs or outputs [82]. Mining deeper, the GO term itself is structurally organized but loosely hierarchical, consisting of cellular components, biological processes, and molecular functions of gene products. The relationship between GO terms can be “part-of” (part and whole), which may embed some similarity information, and “is-a” (parent and child), which may result in more than one parent term. Starting from semantic similarity measurement, SS-Loc [83] incorporates a richer source of homologs and generates more features for prediction. Making use of the loosely hierarchical structure, relevance similarity (RS) considers the “distance” between the parent and child nodes. Take HybridGO-Loc [2], for example; it combines the frequency of occurrences of GO terms and semantic similarity between extracted GO terms to form a hybridized vector as input features, giving outstanding performance.

Mapping AA entries of a query protein or accession number (AC) of its homologs to the GO database [84] will result in a list of GO items representing the possible functions and biological metabolism process this protein is involved in. For further computational method implementation, reorganizing and transferring the list of data into numerical vectors is of high significance. Gneg-mPLoc [85], Euk-pLoc [86], and Hum-pLoc [87] consider GO terms as the basis of forming a Euclidean space, which only consists of 0 or 1 for co-ordinates. ProLoc-GO [43], on the other hand, represented the hit of annotated GO terms mined from Gene Ontology Annotation (GOA) with an n-dimensional binary feature vector. The constructed GO vectors are used for the following training.

### 3.2. Knowledge-Based AI Approaches

Originally, most machine learning methods used GO terms as the only input sources in a simple classification model [88,89]. Given the growing richness of comprehensive protein annotation like related metabolism pathways and structural information, the integration of various input sources, including annotations, interaction networks, and pathway enrichment knowledge, contributes to a multi-view foundation for model improvement [75,90,91]. Applying deep learning algorithms enables a more comprehensive understanding of these high-dimensional and complex features and furthers the combination of sequence and knowledge as input sources. According to the number of input sources, the methods can be roughly divided into GO terms only and fusion methods.

For a single input source, mGOASVM [92] introduces a new decision scheme in SVM multi-class classifiers to collect all the positive decisions, enabling both single- and multi-label localization. AD-SVM [93] enhances the binary relevance methods by integrating an adaptive decision scheme, thereby transforming the linear SVMs into piecewise linear SVMs, reducing the over-prediction instances. By using the frequency of the appearance of one protein in different places, Euk-mPLoc 2.0 [94] creates a virtual sample counting the appearance of protein to separate the total sequence input and the number of locations. However, a large number of proteins, especially newly discovered proteins, have not been functionally annotated yet and directly using homologs cannot guarantee the availability of enough GO terms to be found in the GOA Database. Moreover, the GO is not related to the representation of dynamics or pathway dependencies for protein, which will result in the risk of noise and overestimation of the novel proteins [95]. More details of the methods mentioned can be found in Table 2.

To improve the interpretability of the proposed model, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways is also considered as a functional annotation that can be incorporated in the computational approaches [96]. Since *in vivo* protein interaction is likely to reside within the same subcellular locations, it is possible to reveal protein subcellular localization with protein–protein interaction (PPI) networks [97–99], which is sensitive to mislocalization events [100]. The BioPlex network [101,102], which systematically explores the human interactome developed from affinity purification–mass spectrometry analyses, has also reflected protein function and localization information. As a multi-scale map (MuSIC 1.0) with 69 subcellular systems of human cells generated from BioPlex and Human Protein Atlas (HPA) data integration by Qin et al. [103], protein interactions can be observed from a spatial dimension, providing rich features for knowledge-based model development.

The fusion methods can basically be divided into two categories: feature-level fusion [77,104,105] and decision-level fusion [106]. Feature-level fusion is mostly based on average pooling, weighted combination [107], serial combination, or concatenation of selected values. Liu et al. [77] utilized the latent semantic index method to represent multi-label information, while Yu et al. [49] constructed a novel parallel framework of attribute fusion to avoid the impact of duplicated information. This fusion level enhances the information from multiple sources and allows flexibility in fusion techniques, such as early integration, intermediate integration, and late integration [108]. But low data quality and difficulty in feature selection will affect building one efficient computational model. At the decision level, basic classifiers are used for different data sources, first for selecting the suitable ones; then, the results of each chosen method are ensembled as part of the determination protocol [109], as for the decision voting process [106]. Though the integration strategy is simple, this method can help create various decision-making systems that lead to more robust and accurate predictors. For instance, a multi-view model like ML-FGAT [76] incorporates most of the feature types (e.g., sequence, evolutionary information, physicochemical property, etc.), which minimizes the perturbation of extraneous data in predictive tasks while concurrently enhancing the descriptive capability.

**Table 2.** A summary of state-of-the-art knowledge-based and fusion models for protein subcellular localization prediction. S: Single-Location; M: Multi-Location; Pub: Publication Cited; PsePSSM: Pseudo Position-Specific Scoring Matrix; PC: Physicochemical Properties; CT: Conjoint Triad; DE: Differential Evolution; wMLDAe: Weighted Linear Discriminant Analysis; F-GAN: Feature-Generative Adversarial Networks; GAT: Graph Attention Networks; KNN: K-Nearest Neighbor; CNN: Convolutional Neural Network; RF: Random Forest; CDD: Conserved Functional Domain; PseAAC: Pseudo Amino Acid Composition; PSSM: Position-Specific Scoring Matrix; NN: Nearest Neighbor; PPI: Protein-Protein Interaction Network; KEGG: KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway; mRMR: Minimum Redundancy Maximum Relevance; MCFS: Monte Carlo Feature Selection; LightGBM: Light Gradient Boosting Machine; IFS: Incremental Feature Selection; SVM: Support Vector Machine; SMOTE: Synthetic Minority Over-sampling Technique; EBGW: Encoding Based on Grouped Weight; RPT: Residue Probing Transformation; EDT: Evolutionary Distance Transformation; MCD: Multiscale Continuous and Discontinuous; MLSI: Multi-Label Information Latent Semantic Index; IRWLS: Newton-Weighted Least Squares Iterative Method; MLFE: Multi-Label Learning with Feature Induced Labeling Information Enrichment; DT: Decision Tree; DC: Dipeptide Composition; BR: Binary Relevance Method; CC: Classifier Chain; ECC: Ensemble Classifier Chain; SCF: Self-consistency Formulation; ML-KNN: Multi-Label K-Nearest Neighbor; FunD: Functional Domain; OET-KNN: Optimized Evidence-Theoretic K-Nearest Neighbor; SwissSCL: Swiss-Prot Subcellular Location Annotation; Acc: Accuracy; Prec: Precision; F1: F1 Score; HL: Hamming Loss; RL: Ranking Loss; OE: One Error; CV: Coverage; AT: Absolute Ture; AF: Absolute False; MCC: Matthews Correlation Coefficient; AUC: Area Under the Curve; OLA: Overall Location Accuracy; Rec: Recall.

Method	Features	Algorithm	S/M-Location	Species	Performance Metrics <sup>1</sup>	Pub	Year
ML-FGAT	GO, PsePSSM, PC, CT,	DE, wMLDAe, F-GAN, GAT, KNN, CNN	M	Human, Virus, Gram-negative Bacteria, Plants, SARS-CoV-2	Acc: 0.91~0.96 Prec: 0.92~0.99 F1: 0.94~0.98 HL: 0.01~0.04 RL: 0.02~0.06 OE: 0.04~0.07	[76]	2024
PMPSL-GRAKEL	GO	RF, Random k-label sets algorithm	M	Human, Bacteria, Animal	Acc: 0.89~0.97 CV: 0.92~0.98 AT: 0.82~0.95 AF: 0.01~0.02	[89]	2024
Wang et al.	GO, CDD, PseAAC, PSSM	NN	M	Human	Acc: 0.84 F1: 0.76	[75]	2023
Zhang et al.	PPI, KEGG, GO	mRMR, MCFS, LightGBM, IFS, RF, SVM, SMOTE	M	Human	Acc: 0.75~1.00 MCC: 0.80~0.85	[105]	2022

Table 2. Cont.

Method	Features	Algorithm	S/M-Location	Species	Performance Metrics <sup>1</sup>	Pub	Year
ML-locMLFE	GO, PseAAC, EBGW, RPT, EDT, MCD	MLSI, IRWLS, MLFE	M	Bacteria, Plants, Virus	Acc: 0.94~0.99 Prec: 0.99~1.00 AUC: 0.98~0.99 OLA: 0.99~1.00 HL: 0.00~0.01 CV: 0.07~0.08 RL: 0.00	[77]	2021
Chen et al.	GO, KEGG, PPI, PC	RF, mRMR, IFS, SVM, KNN, DT, SMOTE	S	Human	Acc: 0.56~0.80 MCC: 0.49~0.76	[96]	2021
Gpos-ECC-mPLoc	GO, DC	BR, CC, ECC, SVM	M	Gram-positive Bacteria	Acc: 0.90~0.93	[110]	2015
mGOASVM	GO	SVM	M	Virus, Plants	Acc: 0.87~0.89	[92]	2012
iLoc-Euk	GO, PseAAC, PSSM, SCF	ML-KNN	M	Eukaryotes	Acc: 0.79	[111]	2011
Gneg-mPLoc <sup>2</sup>	GO, FunD, PSSM	OET-KNN	M	Gram-negative Bacteria	Acc: 0.85~0.98	[85]	2010
PSORTb 3.0	SwissSCL	SVM	S	Eukaryotes, Prokaryotes	Acc: 0.97~0.98 Prec: 0.97~0.98 Rec: 0.93~0.94 MCC: 0.79~0.85	[112]	2010

<sup>1</sup> The entries in this column are directly collected from the respective original publications. <sup>2</sup> Web server available at <http://www.csbio.sjtu.edu.cn/bioinf/Gneg-multi/>.

## 4. Bioimage-Based Methods

Imaging data show direct visual evidence of protein localization within different cell components, allowing precise and accurate location determination. Through imaging processing, computational models can analyze the spatial distribution of proteins at the single cell level and quantify their localization patterns. The complexity of images offers different levels of features, which also requires multiple preprocessing steps, deep classification models, and a longer running time to deal with for better performance.

### 4.1. Bioimage-Based Features

Compared to amino acid sequences, representing proteins with 2D images is more interpretable and concise when determining the subcellular localization. With the rapid improvement in microscopic imaging technology, scientists have paid more attention to bioimage-based methods. Computer hardware improvement, especially in graphics processing units (GPUs), makes it possible to deal with more complex calculation problems. The development of neural network structure also accelerates deep learning algorithm architecture improvement for image analysis significantly. For high-quality data, with the mission of mapping all human proteins in cells, tissues, and organs, the Human Protein Atlas (HPA) program [113] was initialized in 2003 as an open-access database that consists of imaging data, mass-spectrometry-based proteomics data, transcriptomics data, etc. The subcellular section of HPA shows detailed expressions and spatial distribution conditions of proteins encoded by 13,147 genes. As it recently updated to version 23, it is one of the most powerful training data sources for computational method development [19,114]. According to most recent studies, immunofluorescence (IF) images and immunohistochemistry (IHC) images are commonly selected as benchmark training and testing data sources.

The subcellular location features (SLF) collected can be divided into two categories, namely, global features and local features [115]. Composed of DNA distribution information and global textures, the global features such as morphological features, local binary patterns (LBP) [116] and Zernike features [117] mainly describe the spatial structure of the whole image. The Haralick [118] texture feature, which obtains statistical features including contrast, correlation, and entropy from the gray-level co-occurrence matrix of input images, is one well-known global image descriptor in pattern recognition. Local features, instead, can describe the micro-patterns ignored in global features. Take scale-invariant feature transform (SIFT) [119] as an example. SIFT was originally used for salient point detection and is suitable for fluorescence object description, which guarantees good performance in fluorescence image studies, especially when combined with global features.

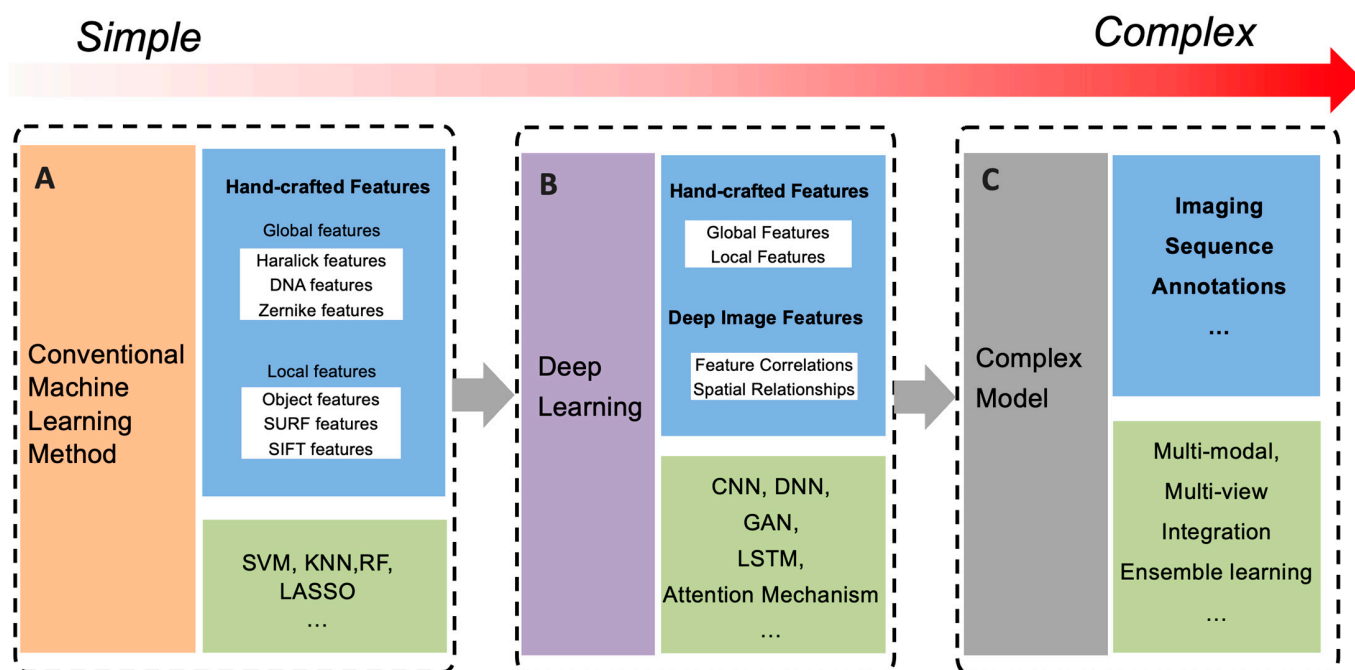
### 4.2. Bioimage-Based AI Methods

Image-related methods can be roughly organized into three phases based on the algorithms and the number of data types used, namely conventional or traditional machine learning methods, deep learning methods, and complex fusion methods, respectively. Figure 3 shows the development of these models from simple to complicated.

Traditional machine learning methods construct the prediction models with the aforementioned hand-crafted features for classification [120–122]. For instance, Li et al. [123] extended a logistic regression algorithm with structured latent variables for underlying components in different image regions for further classification. With two-layer deep-learned feature selection, Ulah et al. [124] established an SVM model based on both radial basis function and linear kernel for location prediction. However, these convolutional methods can be sensitive to noise and variability of imaging data collected, resulting in decreased model robustness. Spatial relationships embedded in images are rarely detected as well, due to manual feature engineering. As deep learning predictors are employed and have achieved high performance on various image-based tasks, recent advances in protein subcellular location rely more on deep learning methods [120].

Deep neural network implementation is the starting point, which increases the inner feature extraction power and the model's learning ability for large and complicated datasets.

In addition to selecting and integrating key features during the image preprocessing steps, most of the deep neural networks consider processed image segmentation as inputs for multi-layer convolutional neural networks (ML-CNN) [125]. Moreover, some predictors can integrate both low- and high-level features embedded in bioimages for a more in-depth view. For multi-label prediction, traditional CNN is extended with a criterion learning strategy to leverage label–attribute relevancy and label–label relevancy to determine the final location [126,127].



**Figure 3.** Three primary categories of computational methodologies for processing imaging data. The red arrow depicts the progressive complexity of prediction models, reflecting advancements toward more sophisticated computational frameworks. Blue rectangle: features used for model training; green rectangle: algorithms for location prediction. (A) Conventional Machine Learning Methods. Hand-crafted figures representing global and local information of images are extracted and trained for simple models. (B) Deep Learning Methods. Coupled with hand-crafted features, deep image features are obtained by deep neuro networks. (C) Complex Fusion Models. This method integrates multi-modality data like sequence, annotation texts, and imaging data as model inputs to gain a more comprehensive and interpretable model for protein subcellular localization. SURF: Speeded Up Robust Features. SIFT: Scale-Invariant Feature Transform. SVM: Support Vector Machine. KNN: K-Nearest Neighbor. RF: Random Forest. LASSO: Least Absolute Shrinkage and Selection Operator. CNN: Convolutional Neural Network. DNN: Deep Neural Network. GAN: Generative Adversarial Network. LSTM: Long Short-Term Memory.

Implementing attention mechanisms is another successful attempt for image classification tasks [128]. With a conventional neural network backbone, Long et al. [129] introduced self-attention and multi-head attention layers as encoders to aggregate multiple feature vectors to construct a combined representation of all immunohistochemistry images input for subsequent analysis. Wang and Wei [126] applied Vision Transformer (ViT) [128] to learn multi-scale feature representations and integrate them globally before entering into the fully connected network. Through different types of transformers (e.g., vision, graph, resolution, etc.), Zhao et al. [130] optimizes the full extent of information embedded in the imaging data. However, there is still a lack of protein subcellular localization studies from this perspective, partly because it lacks efficiency compared to convolutional architectures at large-scale analysis.



In addition, the diversity in input data types across various dimensions shapes the complexity of the entire model. To be more specific, from image datasets, DeepPSL automatically learns meaningful features and their correlations for prediction improvement [131]. Xue et al. [132] unmixed the IHC images into protein and DNA channels for representation construction while segmenting the images into patches for fine-tuning network training. Ding et al. [133] ensemble different classification models using different depths of feature vectors constructed from images as inputs to achieve high-accuracy outputs. By collecting different imaging types, Wei et al. [134] built another parallel integrative deep network for label-free cell optical images. More details about the models can be found in Table 3. Though further techniques can be applied during the pretraining step [129,135–138], image-only methods still lack generalization capability and external validation. When incorporating greater modality of data that are not directly observable from imaging alone but related to protein subcellular localization during model establishment, it will take more contextual information into consideration and overcome the limitations in model performance.

**Table 3.** A summary of state-of-the-art image-based methods for protein subcellular localization prediction. S: Single-Location; M: Multi-Location; Pub: Publication Cited; LBP: Local Binary Pattern; PSSM: Position-Specific Scoring Matrix; PseACC: Pseudo Amino Acid Composition; PC: Physicochemical Properties; LASSO: Least Absolute Shrinkage and Selection Operator; BR: Binary Relevance; SDA: Stepwise Discriminant Analysis; CNN: Convolutional Neural Network; MSA: Multihead Self-attention; Swin: Swin Transformer; CAFE: Cross Attention Feature Enhancement; DNN: Deep Neural Network; CLH: Cell-level Hybrid Model; CLA: Cell-level Actnet; VID: Visual Integrity Detector; ResNet: Residual Network; SE: Squeezeand-Excitation; DenseNet: Dense Convolutional Network; MIL: Multi-instance Learning; SRS: Stimulated Raman Scattering; MPFNet: Multiple parallel Fusion Network; MLP: Multi-Layer Perceptron; SLFs: Subcellular Location Features; CLBP: Completed Local Binary Pattern; LETRIST: Locally Encoded Transform Feature Histogram; RICLBP: Rotation Invariant Co-occurrence Among Adjacent Local Binary Patterns; GDA: Generalized Discriminant Analysis; DCF: Deep-cascade Forest; IF: Immunofluorescence Microscopic; GNT-Xent: The Gradient-Stabilized and Normalized Temperature-Scaled Cross-Entropy Loss; Acc: Accuracy; Prec: Precision; Rec: Recall; MAE: Mean Absolute Error; NRMSE: Normalized Root Mean Square Error; SSIM: Structural Similarity Index; PCC: Pearson’s Correlation Coefficient; R2: Coefficient Determination; F1: F1 Score; MicroF1: MicroF1 Score; MacroF1: MacroF1 Score; Dice: Dice Similarity Coefficient; mIOU: The Mean Intersection Over Union (IOU); MCC: Matthews Correlation Coefficient.

Method	Features	Algorithm	S/M-Location	Species	Performance Metrics <sup>1</sup>	Pub	Year
Zou et al.	Haralick, LBP, PSSM, PseAAC, PC	LASSO, BR, SDA, CNN	S	Human	Acc: 0.75~0.86 Prec: 0.80~0.85 Rec: 0.74~0.85	[122]	2023
ST-Net	Low- and High-Level features	MSA, Swin, CAFE, CNN,	S	Human	MAE: 0.15~0.23 NRMSE: 0.30~0.31 SSIM: 0.78~0.89 PCC: 0.94~0.95 R2: 0.87~0.88	[139]	2023
HCPL	Cell- and Image-Level Information	DNN, CLH, CLA, VID	M	Human	Prec: 0.55~0.57	[140]	2023
Ding et al.	Features Generated from ResNet	ResNet-34, SE, GAP-net, DNN	M	Yeast	Acc: 0.91 Prec: 0.89 Rec: 0.90 F1: 0.89	[133]	2023
Muti-task Learning Strategy	Features Generated from ResNet and DenseNet	ResNet, DenseNet, MIL, CNN	M	Human	MicroF1: 0.78 MacroF1: 0.71	[135]	2022
MPFnetwork	SRS and Fluorescence Signal	MPFNet, CNN, MSA, MLP	M	Human	NRMSE: 0.19~0.20 SSIM: 0.89~0.92 PCC: 0.90~0.91 Dice: 0.93~0.94 mIOU: 0.87~0.88	[134]	2022

Table 3. Cont.

Method	Features	Algorithm	S/M-Location	Species	Performance Metrics <sup>1</sup>	Pub	Year
PScL-DDCFPred	SLFs, LBP, CLBP, LETRIST, RICLBP	SDA-GDA, DNN-DCF	M	Human	Acc: 0.88 Rec: 0.88 Prec: 0.89 F1: 0.88 MCC: 0.86	[141]	2022
PLCNN	Image block structure	CNN	M	Human, Yeast	Acc: 0.91~1.00	[142]	2022
SIFLoc	IF images	GNT-Xent, RandAugment, ResNet18	M	Human	Acc: 0.67~0.73 Prec: 0.77~0.81 Rec: 0.69~0.74 F1: 0.73~0.77	[137]	2022
DeepYeast	Haralick, Gabor, Zernike Features	CNN, DNN	M	Yeast	Acc: 0.97~0.99 Prec: 0.70~0.95 Rec: 0.65~0.92	[125]	2017

<sup>1</sup> The entries in this column are directly collected from the respective original publications.

## 5. Protein Subcellular Localization in Different Species

Analyzing species separately allows a more accurate model generalization, since specific proteins and their subcellular localization patterns may differ in various cell organizations and organelle structures. Take bacteria as an example. As prokaryotes, they exhibit significant structural differences from eukaryotic organisms, like lacking common cellular organelles such as mitochondria, endoplasmic reticulum, and Golgi apparatus. However, within bacteria, a notable class of self-assembling microstructures, known as bacterial microcompartments (BMCs), consist of a protein shell encapsulating an enzymatic core [143,144], creating an internally enclosed space for protein to reside. Furthermore, bacteria possess special cell walls that can be classified as Gram-positive and Gram-negative bacteria [145], which are closely associated with different protein localization modes. For real-world application [146], the subcellular localization changes in host cells, like plants that need precise localization after viral infection, can give insights into the interactions of host cells and viruses, which helps in genetic resistance target identification [147].

Many models have been specially designed for distinct species (e.g., iLoc-Euk [111], iLoc-Virus [148], iLoc-Plant [149], and mPLR-Loc [150]). Gram-LocEN [151] is a predictor for large-scale datasets of both single- and multi-location proteins in bacteria. It created two databases called ProSeq and ProSeq-GO for query protein from Swiss-Prot and GOA databases [152], respectively, to guarantee the effectiveness and decrease storage complexity. After defining GO space and constructing GO vectors, the model demonstrated elastic net (EN) to enable automatic feature selection and further classification. DeepYeast [125], on the other hand, is a neural network trained specially for classifying fluorescent protein subcellular localization in yeast cells with images. As benchmark dataset construction is the foundation of building precise AI-based models, new methods tend to use datasets that have been collected and tested by previous models [28,92], like the Gram-positive and the Gram-negative bacteria dataset [153], the virus dataset [148,154], the plant dataset [149,155], the SARS-CoV-2 dataset [156], the animal cell lines [157], etc. Like Zou et al. [122], some models obtain the data directly by a manual literature search from UniProt and HPA database [70,112]. Multi-species database Compartments [158], fungal database FunSecKB2 [159], plant database PlantSecKB [160], and human and animal database MetazSecKB [161] mostly obtained and arranged from UniProt have also provided efficient searches for each organism and high-quality protein subcellular location annotation datasets across species.

## 6. Current Challenges and Future Directions

### 6.1. Challenges

Despite the significant advances, challenges still exist for AI-based method development in the protein subcellular localization field. The interpretability of the model will be one of the big concerns. Actually, we have developed a series of interpretable machine learning approaches [162–165] for protein subcellular localization and membrane protein function prediction. However, most of them are based on linear models. As deep learning algorithms have complicated training processes that generate high dimensional and nonlinear deep features for prediction, it is of great importance to interpret the decision-making procedures of the model for a better understanding of the essential factors that influence protein localization. SHAP [166], DeepExplainer [167] based on DeepLift [168], and other methodologies major in capturing the importance of features for overall prediction tasks have been implemented in recent studies for increasing model interpretability. Luo et al. [69] have also reduced the dimensionality of feature vectors by constructing autoencoders to obtain a better feature representation for downstream analysis. In ML-FGAT [76], the interpretability is strengthened by analyzing the attention weight parameters. Explainable and understandable frameworks will give more reliable predictions that benefit further studies from a biological perspective.

Moreover, protein subcellular location is influenced by multiple factors. AI-based methods mostly rely on original sequences or images as inputs, which lack the information

after protein biosynthesis. There is also a chance that the prediction model provides the same subcellular location when the mutant protein resides in a different place [147]. Post-translational modifications (PTMs), which refer to amino acid side-chain modification after the synthesis of some proteins, can contribute significant changes to their subcellular location [169]. There are many kinds of PTMs, such as phosphorylation, glycosylation, and acetylation, which dynamically regulate the protein within the cell simultaneously [170], resulting in sparse and incomplete experimental data for model training. As more post-translational positions are discovered [170], AI-based predictions that consider PTMs as key features can also be further investigated [170,171].

Establishing models to leverage both annotated and unannotated proteins for localization can also be a challenge, with a large proportion of unreviewed data reported each year (Figure 1A,B). Though data augmentation methods like SMOTE and GAN are widely used to handle data imbalance, semi-supervised learning can also be established to solve the problem [138,172]. To be more specific, EnTrans-Chlo [173] incorporates multi-modal features and converts them into sample-to-sample similarity features with assigned weights for feeding a highly efficient learning model. LNP-Chlo [174] extended the previous approach by adopting a quadratic programming algorithm to optimize the weights of nearest neighbors. These semi-supervised models remarkably outperformed state-of-the-art supervised methods when integrating different data modalities and dimensionalities with less of a requirement for sufficient labeled data.

## 6.2. Future Directions

Currently, cutting-edge research directions in subcellular localization mainly lie in spatial proteomics [9] and RNA subcellular localization.

With the blooming of single-cell research, it is possible to gain a full understanding of disease from cell and tissue heterogeneity. Since the exact location of proteins at the subcellular, cellular, or tissue levels directly links to their functions, it is essential for protein localization with a single-cell and spatial resolution [18]. Zhu et al. [175] have created cell-based methods with a pseudo-label assignment to discover protein subcellular localization results across distinct cells with heterogeneity among single cells. Husain et al. [140] presents the Hybrid subCellular Protein Localiser (HCPL) that robustly localizes single-cell subcellular protein patterns. Wang's work with mass spectrometry (MS)-based spatial proteomics [176] shows the possibility of larger dimensional feature maps and higher learning ability of computational models.

System-wide studies of RNA subcellular localization (e.g., mRNA [177]) have also paved the way for a more comprehensive analysis of the cellular dynamics [178,179], as proteins are usually transcribed by RNA molecules. Moreover, except for RNA transcripts for protein, other RNAs, like long noncoding RNAs (lncRNAs), may also be involved in many biological functions [180]. Predicting their subcellular locations with AI-based methods [180] can significantly reduce costs and time expenditure, enabling the investigation of their functionalities with limited data [178]. In addition, common [181] and rare cellular-compartment-specific prediction models can be further explored [182]. As for the data imbalance issue, most of the prediction models mainly focus on some of the subcellular components, since they have more manually adjusted records for model construction.

In addition, other promising future directions in this field include web server or tool/software development for protein subcellular localization. Though accurate and efficient models are continuously published, only a few are freely available to the public. Moreover, since models are getting complex with multiple processing blocks, developing methods and/or algorithms into a web-based platform [53] or software service [183] would greatly facilitate experimental research and interdisciplinary collaboration. For downstream applications, Wang et al. [100] have detected mislocated proteins under drug treatments with established models. Xue et al. [132] developed a machine-learning model and validated its ability by identifying biomarker proteins related to colon cancer. Pang et al. [52] proposed the CNN-XGBoost model for Alzheimer's Disease and achieved

competitive performance among general methods. As model accuracy and consistency have been greatly increased, it will be more beneficial to apply complete models for different biomedical scenarios.

## 7. Conclusions

In this review, we have reviewed three types of computational methods using machine learning or deep learning models to construct predictors for protein subcellular localization. For different kinds of inputs, such as protein sequence, GO terms, or IHC images, the predictors will first convey the biological data to numerical or mathematical representations of essential features embedded in the source and apply widely used classifiers for single or multi-class tasks. Traditional machine learning methods can combine various features and manage the high-dimensional data by dimensionality reduction techniques like random projection [184] to avoid the curse of dimensionality and achieve interpretable outcomes under large data scales. Alternatively, they can combine the results of different classifiers, which run the calculation parallelly, to improve the overall performance. Deep learning methods that are mostly based on neural networks will learn and extract high-level features and their correlations from the inputs before the classification. When dealing with large-scale datasets, prediction with a language model is also available with deep learning. For future direction, in addition to faster and more effective algorithm development, we also assume that the localization prediction will incorporate more biochemical interactions like protein–protein interaction networks (PPI), metabolic networks, gene co-expression interaction, etc., into consideration, since proteins intricately engage in complex physiological reactions within the cellular space. Above all, we are confident that the computational methods will raise more and more attention for (1) systematic research like proteomics and metabolomics, (2) provide dynamic insights into cells and reveal what the influence will be when the target protein is muted; and (3) assist the experimental side with data analysis, experimental design, and so on. In the long run, this research area will benefit clinical drug development and contribute to disease detection, diagnosis, prognosis, and treatment.

**Author Contributions:** H.X., Y.Z., J.W. and S.W. wrote the manuscript. S.W. supervised the manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** The research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number P30CA036727. This work was supported by the American Cancer Society under award number IRG-22-146-07-IRG and by the Buffett Cancer Center, which is supported by the National Cancer Institute under award number CA036727. This work was supported by the Buffet Cancer Center, which is supported by the National Cancer Institute under award number CA036727, in collaboration with the UNMC/Children’s Hospital & Medical Center Child Health Research Institute Pediatric Cancer Research Group. This study was supported, in part, by the National Institute on Alcohol Abuse and Alcoholism (P50AA030407-5126, Pilot Core grant). This study was also supported by the Nebraska EPSCoR FIRST Award (OIA-2044049). This work was also partially supported by the National Institute of General Medical Sciences under Award Number P20GM130447. The content is solely the responsibility of the authors and does not necessarily represent the official views from the funding organizations.

**Acknowledgments:** Some figures in the manuscript are generated by BioRender (Available at <https://www.biorender.com>).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Xu, Q.; Hu, D.H.; Xue, H.; Yu, W.; Yang, Q. Semi-Supervised Protein Subcellular Localization. *BMC Bioinform.* **2009**, *10*, S47. [CrossRef]
2. Wan, S.; Mak, M.-W.; Kung, S.-Y. HybridGO-Loc: Mining Hybrid Features on Gene Ontology for Predicting Subcellular Localization of Multi-Location Proteins. *PLoS ONE* **2014**, *9*, e89545. [CrossRef] [PubMed]
3. Stewart, M. Molecular Mechanism of the Nuclear Protein Import Cycle. *Nat. Rev. Mol. Cell Biol.* **2007**, *8*, 195–208. [CrossRef] [PubMed]



4. Mayor, S.; Pagano, R.E. Pathways of Clathrin-Independent Endocytosis. *Nat. Rev. Mol. Cell Biol.* **2007**, *8*, 603–612. [[CrossRef](#)] [[PubMed](#)]
5. Lee, K.; Byun, K.; Hong, W.; Chuang, H.-Y.; Pack, C.-G.; Bayarsaikhan, E.; Paek, S.H.; Kim, H.; Shin, H.Y.; Ideker, T.; et al. Proteome-Wide Discovery of Mislocated Proteins in Cancer. *Genome Res.* **2013**, *23*, 1283–1294. [[CrossRef](#)] [[PubMed](#)]
6. Wang, X.; Li, S. Protein Mislocalization: Mechanisms, Functions and Clinical Applications in Cancer. *Biochim. Biophys. Acta—Rev. Cancer* **2014**, *1846*, 13–25. [[CrossRef](#)] [[PubMed](#)]
7. Barmada, S.J.; Skibinski, G.; Korb, E.; Rao, E.J.; Wu, J.Y.; Finkbeiner, S. Cytoplasmic Mislocalization of TDP-43 Is Toxic to Neurons and Enhanced by a Mutation Associated with Familial Amyotrophic Lateral Sclerosis. *J. Neurosci.* **2010**, *30*, 639–649. [[CrossRef](#)] [[PubMed](#)]
8. Ziff, O.J.; Harley, J.; Wang, Y.; Neeves, J.; Tyzack, G.; Ibrahim, F.; Skehel, M.; Chakrabarti, A.M.; Kelly, G.; Patani, R. Nucleocytoplasmic mRNA Redistribution Accompanies RNA Binding Protein Mislocalization in ALS Motor Neurons and Is Restored by VCP ATPase Inhibition. *Neuron* **2023**, *111*, 3011–3027.e7. [[CrossRef](#)]
9. Lundberg, E.; Börner, G.H.H. Spatial Proteomics: A Powerful Discovery Tool for Cell Biology. *Nat. Rev. Mol. Cell Biol.* **2019**, *20*, 285–302. [[CrossRef](#)]
10. Xiang, L.; Yang, Q.-L.; Xie, B.-T.; Zeng, H.-Y.; Ding, L.-J.; Rao, F.-Q.; Yan, T.; Lu, F.; Chen, Q.; Huang, X.-F. Dysregulated Arginine Metabolism Is Linked to Retinal Degeneration in *Cep250* Knockout Mice. *Investig. Ophthalmol. Vis. Sci.* **2023**, *64*, 2. [[CrossRef](#)]
11. Kohnhorst, C.L.; Schmitt, D.L.; Sundaram, A.; An, S. Subcellular Functions of Proteins under Fluorescence Single-Cell Microscopy. *Biochim. Biophys. Acta (BBA)—Proteins Proteom.* **2016**, *1864*, 77–84. [[CrossRef](#)]
12. Feng, S.; Sekine, S.; Pessino, V.; Li, H.; Leonetti, M.D.; Huang, B. Improved Split Fluorescent Proteins for Endogenous Protein Labeling. *Nat. Commun.* **2017**, *8*, 370. [[CrossRef](#)] [[PubMed](#)]
13. Brzozowski, R.S.; White, M.L.; Eswara, P.J. Live-Cell Fluorescence Microscopy to Investigate Subcellular Protein Localization and Cell Morphology Changes in Bacteria. *J. Vis. Exp.* **2020**, *153*, e59905.
14. Liang, F.-X.; Sall, J.; Petzold, C.; Van Opbergen, C.J.M.; Liang, X.; Delmar, M. Nanogold Based Protein Localization Enables Subcellular Visualization of Cell Junction Protein by SBF-SEM. In *Methods in Cell Biology*; Elsevier: Amsterdam, The Netherlands, 2023; Volume 177, pp. 55–81. ISBN 978-0-323-91607-3.
15. Schornack, S.; Fuchs, R.; Huitema, E.; Rothbauer, U.; Lipka, V.; Kamoun, S. Protein Mislocalization in Plant Cells Using a GFP-binding Chromobody. *Plant J.* **2009**, *60*, 744–754. [[CrossRef](#)] [[PubMed](#)]
16. Orbán, T.I.; Seres, L.; Özvegy-Laczka, C.; Elkind, N.B.; Sarkadi, B.; Homolya, L. Combined Localization and Real-Time Functional Studies Using a GFP-Tagged ABCG2 Multidrug Transporter. *Biochem. Biophys. Res. Commun.* **2008**, *367*, 667–673. [[CrossRef](#)]
17. The UniProt Consortium; Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E.H.; Britto, R.; Bye-A.-Jee, H.; et al. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51*, D523–D531. [[CrossRef](#)]
18. Digre, A.; Lindskog, C. The Human Protein Atlas—Spatial Localization of the Human Proteome in Health and Disease. *Protein Sci.* **2021**, *30*, 218–233. [[CrossRef](#)]
19. Thul, P.J.; Lindskog, C. The Human Protein Atlas: A Spatial Map of the Human Proteome. *Protein Sci.* **2018**, *27*, 233–244. [[CrossRef](#)] [[PubMed](#)]
20. Park, K.-J.; Kanehisa, M. Prediction of Protein Subcellular Locations by Support Vector Machines Using Compositions of Amino Acids and Amino Acid Pairs. *Bioinformatics* **2003**, *19*, 1656–1663. [[CrossRef](#)]
21. Chou, K.-C. Using Pair-Coupled Amino Acid Composition to Predict Protein Secondary Structure Content. *J. Protein Chem.* **1999**, *18*, 473–480. [[CrossRef](#)]
22. Chou, K. Prediction of Protein Cellular Attributes Using Pseudo-amino Acid Composition. *Proteins* **2001**, *43*, 246–255. [[CrossRef](#)]
23. Von Heijne, G.; Steppuhn, J.; Herrmann, R.G. Domain Structure of Mitochondrial and Chloroplast Targeting Peptides. *Eur. J. Biochem.* **1989**, *180*, 535–545. [[CrossRef](#)] [[PubMed](#)]
24. Wan, S.; Mak, M.-W. *Machine Learning for Protein Subcellular Localization Prediction*; De Gruyter: Berlin, Germany, 2015; ISBN 978-1-5015-1048-9.
25. Xue, S.; Liu, X.; Pan, Y.; Xiao, C.; Feng, Y.; Zheng, L.; Zhao, M.; Huang, M. Comprehensive Analysis of Signal Peptides in *Saccharomyces Cerevisiae* Reveals Features for Efficient Secretion. *Adv. Sci.* **2023**, *10*, 2203433. [[CrossRef](#)] [[PubMed](#)]
26. Martoglio, B.; Dobberstein, B. Signal Sequences: More than Just Greasy Peptides. *Trends Cell Biol.* **1998**, *8*, 410–415. [[CrossRef](#)]
27. Almagro Armenteros, J.J.; Salvatore, M.; Emanuelsson, O.; Winther, O.; von Heijne, G.; Elofsson, A.; Nielsen, H. Detecting Sequence Signals in Targeting Peptides Using Deep Learning. *Life Sci. Alliance* **2019**, *2*, e201900429. [[CrossRef](#)]
28. Teufel, F.; Almagro Armenteros, J.J.; Johansen, A.R.; Gíslason, M.H.; Pihl, S.I.; Tsirigos, K.D.; Winther, O.; Brunak, S.; Von Heijne, G.; Nielsen, H. SignalP 6.0 Predicts All Five Types of Signal Peptides Using Protein Language Models. *Nat. Biotechnol.* **2022**, *40*, 1023–1025. [[CrossRef](#)]
29. Hirano, Y.; Ohno, Y.; Kubota, Y.; Fukagawa, T.; Kihara, A.; Haraguchi, T.; Hiraoka, Y. Ceramide Synthase Homolog Tlc4 Maintains Nuclear Envelope Integrity via Its Golgi Translocation. *J. Cell Sci.* **2023**, *136*, jcs260923. [[CrossRef](#)]
30. Shen, Y.; Tang, J.; Guo, F. Identification of Protein Subcellular Localization via Integrating Evolutionary and Physicochemical Information into Chou’s General PseAAC. *J. Theor. Biol.* **2019**, *462*, 230–239. [[CrossRef](#)] [[PubMed](#)]
31. Mak, M.-W.; Guo, J.; Kung, S.-Y. PairProSVM: Protein Subcellular Localization Based on Local Pairwise Profile Alignment and SVM. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2008**, *5*, 416–422. [[CrossRef](#)]

32. Tahir, M.; Khan, F.; Hayat, M.; Alshehri, M.D. An Effective Machine Learning-Based Model for the Prediction of Protein–Protein Interaction Sites in Health Systems. *Neural Comput. Appl.* **2024**, *36*, 65–75. [[CrossRef](#)]
33. Wang, J.; Nabil, M.; Zhang, J. Deep Forest-Based Prediction of Protein Subcellular Localization. *Curr. Gene Ther.* **2018**, *18*, 268–274. [[CrossRef](#)]
34. Sui, J.; Chen, Y.; Cao, Y.; Zhao, Y. Accurate Identification of Submitochondrial Protein Location Based on Deep Representation Learning Feature Fusion. In Proceedings of the ICIC 2023: Advanced Intelligent Computing Technology and Applications, Zhengzhou, China, 10–13 August 2023; Huang, D.-S., Premaratne, P., Jin, B., Qu, B., Jo, K.-H., Hussain, A., Eds.; Springer Nature: Singapore, 2023; pp. 587–596.
35. Javed, F.; Hayat, M. Predicting Subcellular Localization of Multi-Label Proteins by Incorporating the Sequence Features into Chou’s PseAAC. *Genomics* **2019**, *111*, 1325–1332. [[CrossRef](#)] [[PubMed](#)]
36. Du, L.; Meng, Q.; Chen, Y.; Wu, P. Subcellular Location Prediction of Apoptosis Proteins Using Two Novel Feature Extraction Methods Based on Evolutionary Information and LDA. *BMC Bioinform.* **2020**, *21*, 212. [[CrossRef](#)] [[PubMed](#)]
37. Zhang, S.; Zhang, T.; Liu, C. Prediction of Apoptosis Protein Subcellular Localization via Heterogeneous Features and Hierarchical Extreme Learning Machine. *SAR QSAR Environ. Res.* **2019**, *30*, 209–228. [[CrossRef](#)] [[PubMed](#)]
38. Alaa, A.; Eldeib, A.M.; Metwally, A.A. Protein Subcellular Localization Prediction Based on Internal Micro-Similarities of Markov Chains. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 1355–1358.
39. Murphy, R.F. Communicating Subcellular Distributions. *Cytom. Pt. A* **2010**, *77A*, 686–692. [[CrossRef](#)]
40. Cohen, S.E.; Erb, M.L.; Selimkhanov, J.; Dong, G.; Hasty, J.; Pogliano, J.; Golden, S.S. Dynamic Localization of the Cyanobacterial Circadian Clock Proteins. *Curr. Biol.* **2014**, *24*, 1836–1844. [[CrossRef](#)]
41. Kim, W.; Jeon, T.J. Dynamic Subcellular Localization of DydA in Dictyostelium Cells. *Biochem. Biophys. Res. Commun.* **2023**, *663*, 186–191. [[CrossRef](#)]
42. Xiong, E.; Cao, D.; Qu, C.; Zhao, P.; Wu, Z.; Yin, D.; Zhao, Q.; Gong, F. Multilocation Proteins in Organelle Communication: Based on Protein–Protein Interactions. *Plant Direct* **2022**, *6*, e386. [[CrossRef](#)]
43. Jiang, Z.; Wang, D.; Wu, P.; Chen, Y.; Shang, H.; Wang, L.; Xie, H. Predicting Subcellular Localization of Multisite Proteins Using Differently Weighted Multi-Label k-Nearest Neighbors Sets. *THC* **2019**, *27*, 185–193. [[CrossRef](#)]
44. Chang, C.-C.; Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27. [[CrossRef](#)]
45. Ding, Y.; Tang, J.; Guo, F. Human Protein Subcellular Localization Identification via Fuzzy Model on Kernelized Neighborhood Representation. *Appl. Soft Comput.* **2020**, *96*, 106596. [[CrossRef](#)]
46. Hasan, M.A.M.; Ahmad, S.; Molla, M.K.I. Protein Subcellular Localization Prediction Using Multiple Kernel Learning Based Support Vector Machine. *Mol. BioSyst.* **2017**, *13*, 785–795. [[CrossRef](#)] [[PubMed](#)]
47. Boutell, M.R.; Luo, J.; Shen, X.; Brown, C.M. Learning Multi-Label Scene Classification. *Pattern Recognit.* **2004**, *37*, 1757–1771. [[CrossRef](#)]
48. Wang, S.; Liu, S. Protein Sub-Nuclear Localization Based on Effective Fusion Representations and Dimension Reduction Algorithm LDA. *Int. J. Mol. Sci.* **2015**, *16*, 30343–30361. [[CrossRef](#)] [[PubMed](#)]
49. Yu, D.; Wu, X.; Shen, H.; Yang, J.; Tang, Z.; Qi, Y.; Yang, J. Enhancing Membrane Protein Subcellular Localization Prediction by Parallel Fusion of Multi-View Features. *IEEE Trans. Nanobiosci.* **2012**, *11*, 375–385. [[CrossRef](#)] [[PubMed](#)]
50. Jing, R.; Li, Y.; Xue, L.; Liu, F.; Li, M.; Luo, J. autoBioSeqpy: A Deep Learning Tool for the Classification of Biological Sequences. *J. Chem. Inf. Model.* **2020**, *60*, 3755–3764. [[CrossRef](#)] [[PubMed](#)]
51. Semwal, R.; Varadwaj, P.K. HumDLoc: Human Protein Subcellular Localization Prediction Using Deep Neural Network. *Curr. Genom.* **2020**, *21*, 546–557. [[CrossRef](#)]
52. Pang, L.; Wang, J.; Zhao, L.; Wang, C.; Zhan, H. A Novel Protein Subcellular Localization Method With CNN-XGBoost Model for Alzheimer’s Disease. *Front. Genet.* **2019**, *9*, 751. [[CrossRef](#)]
53. Kaleel, M.; Zheng, Y.; Chen, J.; Feng, X.; Simpson, J.C.; Pollastri, G.; Mooney, C. SCLpred-EMS: Subcellular Localization Prediction of Endomembrane System and Secretory Pathway Proteins by Deep N-to-1 Convolutional Neural Networks. *Bioinformatics* **2020**, *36*, 3343–3349. [[CrossRef](#)]
54. Cong, H.; Liu, H.; Chen, Y.; Cao, Y. Self-Evolving Framework of Deep Convolutional Neural Network for Multilocus Protein Subcellular Localization. *Med. Biol. Eng. Comput.* **2020**, *58*, 3017–3038. [[CrossRef](#)]
55. Cong, H.; Liu, H.; Cao, Y.; Chen, Y.; Liang, C. Multiple Protein Subcellular Locations Prediction Based on Deep Convolutional Neural Networks with Self-Attention Mechanism. *Interdiscip. Sci. Comput. Life Sci.* **2022**, *14*, 421–438. [[CrossRef](#)] [[PubMed](#)]
56. Liao, Z.; Pan, G.; Sun, C.; Tang, J. Predicting Subcellular Location of Protein with Evolution Information and Sequence-Based Deep Learning. *BMC Bioinform.* **2021**, *22*, 515. [[CrossRef](#)] [[PubMed](#)]
57. Jiang, Y.; Wang, D.; Yao, Y.; Eubel, H.; Künzler, P.; Møller, I.M.; Xu, D. MULocDeep: A Deep-Learning Framework for Protein Subcellular and Suborganellar Localization Prediction with Residue-Level Interpretation. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4825–4839. [[CrossRef](#)] [[PubMed](#)]
58. Wu, L.; Gao, S.; Yao, S.; Wu, F.; Li, J.; Dong, Y.; Zhang, Y. Gm-PLoc: A Subcellular Localization Model of Multi-Label Protein Based on GAN and DeepFM. *Front. Genet.* **2022**, *13*, 912614. [[CrossRef](#)] [[PubMed](#)]
59. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority over-Sampling Technique. *J. Artif. Int. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]

60. Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7112–7127. [[CrossRef](#)] [[PubMed](#)]
61. Wang, X.; Han, L.; Wang, R.; Chen, H. DaDL-SChlo: Protein Subchloroplast Localization Prediction Based on Generative Adversarial Networks and Pre-Trained Protein Language Model. *Brief. Bioinform.* **2023**, *24*, bbad083. [[CrossRef](#)]
62. Hosna, A.; Merry, E.; Gyalmo, J.; Alom, Z.; Aung, Z.; Azim, M.A. Transfer Learning: A Friendly Introduction. *J. Big Data* **2022**, *9*, 102. [[CrossRef](#)] [[PubMed](#)]
63. Heinzinger, M.; Elnaggar, A.; Wang, Y.; Dallago, C.; Nechaev, D.; Matthes, F.; Rost, B. Modeling Aspects of the Language of Life through Transfer-Learning Protein Sequences. *BMC Bioinform.* **2019**, *20*, 723. [[CrossRef](#)]
64. Ofer, D. The Language of Proteins: NLP, Machine Learning & Protein Sequences. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 1750–1758.
65. Howard, J.; Ruder, S. Universal Language Model Fine-Tuning for Text Classification. *arXiv* **2018**, arXiv:1801.06146.
66. Jumper, J. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [[CrossRef](#)] [[PubMed](#)]
67. Brandes, N.; Ofer, D.; Peleg, Y.; Rappoport, N.; Linial, M. ProteinBERT: A Universal Deep-Learning Model of Protein Sequence and Function. *Bioinformatics* **2022**, *38*, 2102–2110. [[CrossRef](#)] [[PubMed](#)]
68. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
69. Luo, Z.; Wang, R.; Sun, Y.; Liu, J.; Chen, Z.; Zhang, Y.-J. Interpretable Feature Extraction and Dimensionality Reduction in ESM2 for Protein Localization Prediction. *Brief. Bioinform.* **2024**, *25*, bbad534. [[CrossRef](#)] [[PubMed](#)]
70. Thumulari, V.; Almagro Armenteros, J.J.; Johansen, A.R.; Nielsen, H.; Winther, O. DeepLoc 2.0: Multi-Label Subcellular Localization Prediction Using Protein Language Models. *Nucleic Acids Res.* **2022**, *50*, W228–W234. [[CrossRef](#)]
71. Jiang, Y.; Jiang, L.; Akhil, C.S.; Wang, D.; Zhang, Z.; Zhang, W.; Xu, D. MULocDeep Web Service for Protein Localization Prediction and Visualization at Subcellular and Suborganellar Levels. *Nucleic Acids Res.* **2023**, *51*, W343–W349. [[CrossRef](#)] [[PubMed](#)]
72. Zhang, N.; Rao, R.S.P.; Salvato, F.; Havelund, J.F.; Møller, I.M.; Thelen, J.J.; Xu, D. MU-LOC: A Machine-Learning Method for Predicting Mitochondrially Localized Proteins in Plants. *Front. Plant Sci.* **2018**, *9*, 634. [[CrossRef](#)] [[PubMed](#)]
73. Wang, X.; Zhang, W.; Zhang, Q.; Li, G.-Z. MultiP-SChlo: Multi-Label Protein Subchloroplast Localization Prediction with Chou’s Pseudo Amino Acid Composition and a Novel Multi-Label Classifier. *Bioinformatics* **2015**, *31*, 2639–2645. [[CrossRef](#)]
74. Ryngajllo, M.; Childs, L.; Lohse, M.; Giorgi, F.M.; Lude, A.; Selbig, J.; Usadel, B. SLocX: Predicting Subcellular Localization of Arabidopsis Proteins Leveraging Gene Expression Data. *Front. Plant Sci.* **2011**, *2*, 43. [[CrossRef](#)]
75. Wang, S.; Zou, K.; Wang, Z.; Zhu, S.; Yang, F. A Novel Multi-Label Human Protein Subcellular Localization Model Based on Gene Ontology and Functional Domain. In Proceedings of the 2023 3rd International Conference on Bioinformatics and Intelligent Computing, Sanya, China, 10–12 February 2023; Association for Computing Machinery: New York, NY, USA, 2023; pp. 376–380.
76. Wang, C.; Wang, Y.; Ding, P.; Li, S.; Yu, X.; Yu, B. ML-FGAT: Identification of Multi-Label Protein Subcellular Localization by Interpretable Graph Attention Networks and Feature-Generative Adversarial Networks. *Comput. Biol. Med.* **2024**, *170*, 107944. [[CrossRef](#)] [[PubMed](#)]
77. Liu, Y.; Jin, S.; Gao, H.; Wang, X.; Wang, C.; Zhou, W.; Yu, B. Predicting the Multi-Label Protein Subcellular Localization through Multi-Information Fusion and MLSI Dimensionality Reduction Based on MLFE Classifier. *Bioinformatics* **2022**, *38*, 1223–1230. [[CrossRef](#)] [[PubMed](#)]
78. Lu, Z.; Szafron, D.; Greiner, R.; Lu, P.; Wishart, D.S.; Poulin, B.; Anvik, J.; Macdonell, C.; Eisner, R. Predicting Subcellular Localization of Proteins Using Machine-Learned Classifiers. *Bioinformatics* **2004**, *20*, 547–556. [[CrossRef](#)] [[PubMed](#)]
79. Nair, R.; Rost, B. Sequence Conserved for Subcellular Localization. *Protein Sci.* **2002**, *11*, 2836–2847. [[CrossRef](#)] [[PubMed](#)]
80. Fyshe, A.; Liu, Y.; Szafron, D.; Greiner, R.; Lu, P. Improving Subcellular Localization Prediction Using Text Classification and the Gene Ontology. *Bioinformatics* **2008**, *24*, 2512–2517. [[CrossRef](#)] [[PubMed](#)]
81. Brady, S.; Shatkay, H. EpiLoc: A (working) text-based system for predicting protein subcellular location. In *Biocomputing 2008*; WORLD SCIENTIFIC: Kohala Coast, HI, USA, 2007; pp. 604–615.
82. Huang, W.-L.; Tung, C.-W.; Ho, S.-W.; Hwang, S.-F.; Ho, S.-Y. ProLoc-GO: Utilizing Informative Gene Ontology Terms for Sequence-Based Prediction of Protein Subcellular Localization. *BMC Bioinform.* **2008**, *9*, 80. [[CrossRef](#)] [[PubMed](#)]
83. He, L.; Liu, X. The Development and Progress in Machine Learning for Protein Subcellular Localization Prediction. *Open Bioinform. J.* **2022**, *15*, e187503622208110. [[CrossRef](#)]
84. Wan, S.; Mak, M.-W.; Kung, S.-Y. GOASVM: A Subcellular Location Predictor by Incorporating Term-Frequency Gene Ontology into the General Form of Chou’s Pseudo-Amino Acid Composition. *J. Theor. Biol.* **2013**, *323*, 40–48. [[CrossRef](#)]
85. Shen, H.-B.; Chou, K.-C. Gneg-mPLoc: A Top-down Strategy to Enhance the Quality of Predicting Subcellular Localization of Gram-Negative Bacterial Proteins. *J. Theor. Biol.* **2010**, *264*, 326–333. [[CrossRef](#)] [[PubMed](#)]
86. Shen, H.-B.; Yang, J.; Chou, K.-C. Euk-PLoc: An Ensemble Classifier for Large-Scale Eukaryotic Protein Subcellular Location Prediction. *Amino Acids* **2007**, *33*, 57–67. [[CrossRef](#)]
87. Chou, K.-C.; Shen, H.-B. Hum-PLoc: A Novel Ensemble Classifier for Predicting Human Protein Subcellular Localization. *Biochem. Biophys. Res. Commun.* **2006**, *347*, 150–157. [[CrossRef](#)] [[PubMed](#)]
88. Mei, S.; Fei, W.; Zhou, S. Gene Ontology Based Transfer Learning for Protein Subcellular Localization. *BMC Bioinform.* **2011**, *12*, 44. [[CrossRef](#)] [[PubMed](#)]



89. Chen, L.; Qu, R.; Liu, X. Improved Multi-Label Classifiers for Predicting Protein Subcellular Localization. *Math. Biosci. Eng.* **2024**, *21*, 214–236. [[CrossRef](#)] [[PubMed](#)]
90. Zhang, Q.; Li, S.; Yu, B.; Zhang, Q.; Han, Y.; Zhang, Y.; Ma, Q. DMLDA-LocLIFT: Identification of Multi-Label Protein Subcellular Localization Using DMLDA Dimensionality Reduction and LIFT Classifier. *Chemom. Intell. Lab. Syst.* **2020**, *206*, 104148. [[CrossRef](#)]
91. Chen, C.; Chen, L.-X.; Zou, X.-Y.; Cai, P.-X. Predicting Protein Structural Class Based on Multi-Features Fusion. *J. Theor. Biol.* **2008**, *253*, 388–392. [[CrossRef](#)] [[PubMed](#)]
92. Wan, S.; Mak, M.-W.; Kung, S.-Y. mGOASVM: Multi-Label Protein Subcellular Localization Based on Gene Ontology and Support Vector Machines. *BMC Bioinform.* **2012**, *13*, 290. [[CrossRef](#)] [[PubMed](#)]
93. Wan, S.; Mak, M.-W. Predicting Subcellular Localization of Multi-Location Proteins by Improving Support Vector Machines with an Adaptive-Decision Scheme. *Int. J. Mach. Learn. Cyber.* **2018**, *9*, 399–411. [[CrossRef](#)]
94. Chou, K.-C.; Shen, H.-B. A New Method for Predicting the Subcellular Localization of Eukaryotic Proteins with Both Single and Multiple Sites: Euk-mPLoc 2.0. *PLoS ONE* **2010**, *5*, e9931. [[CrossRef](#)]
95. Mei, S. Multi-Label Multi-Kernel Transfer Learning for Human Protein Subcellular Localization. *PLoS ONE* **2012**, *7*, e37716. [[CrossRef](#)]
96. Chen, L.; Li, Z.; Zeng, T.; Zhang, Y.-H.; Zhang, S.; Huang, T.; Cai, Y.-D. Predicting Human Protein Subcellular Locations by Using a Combination of Network and Function Features. *Front. Genet.* **2021**, *12*, 783128. [[CrossRef](#)]
97. Garapati, H.S.; Male, G.; Mishra, K. Predicting Subcellular Localization of Proteins Using Protein-Protein Interaction Data. *Genomics* **2020**, *112*, 2361–2368. [[CrossRef](#)] [[PubMed](#)]
98. Jiang, J.Q.; Wu, M. Predicting Multiplex Subcellular Localization of Proteins Using Protein-Protein Interaction Network: A Comparative Study. *BMC Bioinform.* **2012**, *13*, S20. [[CrossRef](#)] [[PubMed](#)]
99. Li, M.; Li, W.; Wu, F.-X.; Pan, Y.; Wang, J. Identifying Essential Proteins Based on Sub-Network Partition and Prioritization by Integrating Subcellular Localization Information. *J. Theor. Biol.* **2018**, *447*, 65–73. [[CrossRef](#)] [[PubMed](#)]
100. Wang, R.-H.; Luo, T.; Zhang, H.-L.; Du, P.-F. PLA-GNN: Computational Inference of Protein Subcellular Location Alterations under Drug Treatments with Deep Graph Neural Networks. *Comput. Biol. Med.* **2023**, *157*, 106775. [[CrossRef](#)] [[PubMed](#)]
101. Huttlin, E.L. Architecture of the Human Interactome Defines Protein Communities and Disease Networks. *Nature* **2017**, *545*, 505–509. [[CrossRef](#)]
102. Huttlin, E.L. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **2015**, *162*, 425–440. [[CrossRef](#)] [[PubMed](#)]
103. Qin, Y. A Multi-Scale Map of Cell Structure Fusing Protein Images and Interactions. *Nature* **2021**, *600*, 536–542. [[CrossRef](#)] [[PubMed](#)]
104. Li, B.; Cai, L.; Liao, B.; Fu, X.; Bing, P.; Yang, J. Prediction of Protein Subcellular Localization Based on Fusion of Multi-View Features. *Molecules* **2019**, *24*, 919. [[CrossRef](#)] [[PubMed](#)]
105. Zhang, Y.-H.; Ding, S.; Chen, L.; Huang, T.; Cai, Y.-D. Subcellular Localization Prediction of Human Proteins Using Multifeature Selection Methods. *BioMed Res. Int.* **2022**, *2022*, 3288527. [[CrossRef](#)]
106. Wang, G.; Zhai, Y.-J.; Xue, Z.-Z.; Xu, Y.-Y. Improving Protein Subcellular Location Classification by Incorporating Three-Dimensional Structure Information. *Biomolecules* **2021**, *11*, 1607. [[CrossRef](#)]
107. Zhang, Q.; Zhang, Y.; Li, S.; Han, Y.; Jin, S.; Gu, H.; Yu, B. Accurate Prediction of Multi-Label Protein Subcellular Localization through Multi-View Feature Learning with RBRL Classifier. *Brief. Bioinform.* **2021**, *22*, bbab012. [[CrossRef](#)] [[PubMed](#)]
108. Picard, M.; Scott-Boyer, M.-P.; Bodein, A.; Périn, O.; Droit, A. Integration Strategies of Multi-Omics Data for Machine Learning Analysis. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 3735–3746. [[CrossRef](#)]
109. Guo, X.; Liu, F.; Ju, Y.; Wang, Z.; Wang, C. Human Protein Subcellular Localization with Integrated Source and Multi-Label Ensemble Classifier. *Sci. Rep.* **2016**, *6*, 28087. [[CrossRef](#)]
110. Wang, X.; Zhang, J.; Li, G.-Z. Multi-Location Gram-Positive and Gram-Negative Bacterial Protein Subcellular Localization Using Gene Ontology and Multi-Label Classifier Ensemble. *BMC Bioinform.* **2015**, *16*, S1. [[CrossRef](#)]
111. Chou, K.-C.; Wu, Z.-C.; Xiao, X. iLoc-Euk: A Multi-Label Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Eukaryotic Proteins. *PLoS ONE* **2011**, *6*, e18258. [[CrossRef](#)] [[PubMed](#)]
112. Yu, N.Y.; Wagner, J.R.; Laird, M.R.; Melli, G.; Rey, S.; Lo, R.; Dao, P.; Sahinalp, S.C.; Ester, M.; Foster, L.J.; et al. PSORTb 3.0: Improved Protein Subcellular Localization Prediction with Refined Localization Subcategories and Predictive Capabilities for All Prokaryotes. *Bioinformatics* **2010**, *26*, 1608–1615. [[CrossRef](#)] [[PubMed](#)]
113. Uhlén, M.; Fagerberg, L.; Hallström, B.M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, Å.; Kampf, C.; Sjöstedt, E.; Asplund, A.; et al. Tissue-Based Map of the Human Proteome. *Science* **2015**, *347*, 1260419. [[CrossRef](#)] [[PubMed](#)]
114. Ouyang, W.; Winsnes, C.F.; Hjelmare, M.; Cesnik, A.J.; Åkesson, L.; Xu, H.; Sullivan, D.P.; Dai, S.; Lan, J.; Jinmo, P.; et al. Analysis of the Human Protein Atlas Image Classification Competition. *Nat. Methods* **2019**, *16*, 1254–1261. [[CrossRef](#)] [[PubMed](#)]
115. Xu, Y.-Y.; Yao, L.-X.; Shen, H.-B. Bioimage-Based Protein Subcellular Location Prediction: A Comprehensive Review. *Front. Comput. Sci.* **2018**, *12*, 26–39. [[CrossRef](#)]
116. Nanni, L.; Lumini, A.; Brahnam, S. Survey on LBP Based Texture Descriptors for Image Classification. *Expert. Syst. Appl.* **2012**, *39*, 3634–3641. [[CrossRef](#)]
117. Tahir, M.; Khan, A.; Majid, A. Protein Subcellular Localization of Fluorescence Imagery Using Spatial and Transform Domain Features. *Bioinformatics* **2012**, *28*, 91–97. [[CrossRef](#)] [[PubMed](#)]

118. Haralick, R.M.; Shanmugam, K.; Dinstein, I. Textural Features for Image Classification. *IEEE Trans. Syst. Man Cybern.* **1973**, SMC-3, 610–621. [[CrossRef](#)]
119. Godil, A.; Lian, Z.; Wagan, A. Exploring Local Features and the Bag-of-Visual-Words Approach for Bioimage Classification. In Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics, Washington, DC, USA, 22 September 2013; pp. 694–695.
120. Liu, G.-H.; Zhang, B.-W.; Qian, G.; Wang, B.; Mao, B.; Bichindaritz, I. Bioimage-Based Prediction of Protein Subcellular Location in Human Tissue with Ensemble Features and Deep Networks. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2020**, *17*, 1966–1980. [[CrossRef](#)] [[PubMed](#)]
121. Newberg, J.; Murphy, R.F. A Framework for the Automated Analysis of Subcellular Patterns in Human Protein Atlas Images. *J. Proteome Res.* **2008**, *7*, 2300–2308. [[CrossRef](#)] [[PubMed](#)]
122. Zou, K.; Wang, S.; Wang, Z.; Zou, H.; Yang, F. Dual-Signal Feature Spaces Map Protein Subcellular Locations Based on Immunohistochemistry Image and Protein Sequence. *Sensors* **2023**, *23*, 9014. [[CrossRef](#)] [[PubMed](#)]
123. Li, J.; Xiong, L.; Schneider, J.; Murphy, R.F. Protein Subcellular Location Pattern Classification in Cellular Images Using Latent Discriminative Models. *Bioinformatics* **2012**, *28*, i32–i39. [[CrossRef](#)] [[PubMed](#)]
124. Ullah, M.; Han, K.; Hadi, F.; Xu, J.; Song, J.; Yu, D.-J. PScL-HDdeep: Image-Based Prediction of Protein Subcellular Location in Human Tissue Using Ensemble Learning of Handcrafted and Deep Learned Features with Two-Layer Feature Selection. *Brief. Bioinform.* **2021**, *22*, bbab278. [[CrossRef](#)] [[PubMed](#)]
125. Pärnamaa, T.; Parts, L. Accurate Classification of Protein Subcellular Localization from High-Throughput Microscopy Images Using Deep Learning. *G3 Genes Genomes Genet.* **2017**, *7*, 1385–1392. [[CrossRef](#)]
126. Wang, F.; Wei, L. Multi-Scale Deep Learning for the Imbalanced Multi-Label Protein Subcellular Localization Prediction Based on Immunohistochemistry Images. *Bioinformatics* **2022**, *38*, 2602–2611. [[CrossRef](#)]
127. Su, R.; He, L.; Liu, T.; Liu, X.; Wei, L. Protein Subcellular Localization Based on Deep Image Features and Criterion Learning Strategy. *Brief. Bioinform.* **2021**, *22*, bbaa313. [[CrossRef](#)]
128. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2021**, arXiv:2010.11929.
129. Long, W.; Yang, Y.; Shen, H.-B. ImPLoc: A Multi-Instance Deep Learning Model for the Prediction of Protein Subcellular Localization Based on Immunohistochemistry Images. *Bioinformatics* **2020**, *36*, 2244–2250. [[CrossRef](#)] [[PubMed](#)]
130. Zhao, C.; Xu, Z.; Wang, X.; Tao, S.; MacDonald, W.A.; He, K.; Poholek, A.C.; Chen, K.; Huang, H.; Chen, W. Innovative Super-Resolution in Spatial Transcriptomics: A Transformer Model Exploiting Histology Images and Spatial Gene Expression. *Brief. Bioinform.* **2024**, *25*, bbae052. [[CrossRef](#)] [[PubMed](#)]
131. Wei, L.; Ding, Y.; Su, R.; Tang, J.; Zou, Q. Prediction of Human Protein Subcellular Localization Using Deep Learning. *J. Parallel Distrib. Comput.* **2018**, *117*, 212–217. [[CrossRef](#)]
132. Xue, Z.-Z.; Wu, Y.; Gao, Q.-Z.; Zhao, L.; Xu, Y.-Y. Automated Classification of Protein Subcellular Localization in Immunohistochemistry Images to Reveal Biomarkers in Colon Cancer. *BMC Bioinform.* **2020**, *21*, 398. [[CrossRef](#)] [[PubMed](#)]
133. Ding, J.; Xu, J.; Wei, J.; Tang, J.; Guo, F. A Multi-Scale Multi-Model Deep Neural Network via Ensemble Strategy on High-Throughput Microscopy Image for Protein Subcellular Localization. *Expert. Syst. Appl.* **2023**, *212*, 118744. [[CrossRef](#)]
134. Wei, Z.; Liu, W.; Yu, W.; Liu, X.; Yan, R.; Liu, Q.; Guo, Q. Multiple Parallel Fusion Network for Predicting Protein Subcellular Localization from Stimulated Raman Scattering (SRS) Microscopy Images in Living Cells. *Int. J. Mol. Sci.* **2022**, *23*, 10827. [[CrossRef](#)]
135. Zhang, P.; Zhang, M.; Liu, H.; Yang, Y. Prediction of Protein Subcellular Localization Based on Microscopic Images via Multi-Task Multi-Instance Learning. *Chin. J. Electron.* **2022**, *31*, 888–896. [[CrossRef](#)]
136. Hu, J.; Yang, Y.; Xu, Y.; Shen, H. Incorporating Label Correlations into Deep Neural Networks to Classify Protein Subcellular Location Patterns in Immunohistochemistry Images. *Proteins* **2022**, *90*, 493–503. [[CrossRef](#)]
137. Tu, Y.; Lei, H.; Shen, H.-B.; Yang, Y. SIFLoc: A Self-Supervised Pre-Training Method for Enhancing the Recognition of Protein Subcellular Localization in Immunofluorescence Microscopic Images. *Brief. Bioinform.* **2022**, *23*, bbab605. [[CrossRef](#)]
138. Xu, Y.-Y.; Yang, F.; Zhang, Y.; Shen, H.-B. Bioimaging-Based Detection of Mislocalized Proteins in Human Cancers by Semi-Supervised Learning. *Bioinformatics* **2015**, *31*, 1111–1119. [[CrossRef](#)] [[PubMed](#)]
139. Liu, S.; Huang, M.; Liu, X.; Han, K.; Wang, Z.; Sun, G.; Guo, Q. Swin Transformer Based Neural Network for Organelles Prediction from Quantitative Label-Free Imaging with Phase and Polarization (Qlipp) in Unlabeled Live Cells and Tissue Slices; SSRN. 2023. Available online: <https://ssrn.com/abstract=4604469> (accessed on 28 February 2024). [[CrossRef](#)]
140. Husain, S.S.; Ong, E.-J.; Minskiy, D.; Bober-Irizar, M.; Irizar, A.; Bober, M. Single-Cell Subcellular Protein Localisation Using Novel Ensembles of Diverse Deep Architectures. *Commun. Biol.* **2023**, *6*, 489. [[CrossRef](#)] [[PubMed](#)]
141. Ullah, M.; Hadi, F.; Song, J.; Yu, D.-J. PScL-DDCFPred: An Ensemble Deep Learning-Based Approach for Characterizing Multiclass Subcellular Localization of Human Proteins from Bioimage Data. *Bioinformatics* **2022**, *38*, 4019–4026. [[CrossRef](#)]
142. Tahir, M.; Anwar, S.; Mian, A.; Muzaffar, A.W. Deep Localization of Subcellular Protein Structures from Fluorescence Microscopy Images. *Neural Comput. Appl.* **2022**, *34*, 5701–5714. [[CrossRef](#)]
143. Kerfeld, C.A.; Aussignargues, C.; Zarzycki, J.; Cai, F.; Sutter, M. Bacterial Microcompartments. *Nat. Rev. Microbiol.* **2018**, *16*, 277–290. [[CrossRef](#)]

144. Yeates, T.O.; Crowley, C.S.; Tanaka, S. Bacterial Microcompartment Organelles: Protein Shell Structure and Evolution. *Annu. Rev. Biophys.* **2010**, *39*, 185–205. [[CrossRef](#)] [[PubMed](#)]
145. Forster, B.M.; Marquis, H. Protein Transport across the Cell Wall of Monoderm Gram-positive Bacteria. *Mol. Microbiol.* **2012**, *84*, 405–413. [[CrossRef](#)] [[PubMed](#)]
146. Yadav, A.K.; Singla, D. VacPred: Sequence-Based Prediction of Plant Vacuole Proteins Using Machine-Learning Techniques. *J. Biosci.* **2020**, *45*, 106. [[CrossRef](#)]
147. Rodriguez-Peña, R.; Mounadi, K.E.; Garcia-Ruiz, H. Changes in Subcellular Localization of Host Proteins Induced by Plant Viruses. *Viruses* **2021**, *13*, 677. [[CrossRef](#)]
148. Xiao, X.; Wu, Z.-C.; Chou, K.-C. iLoc-Virus: A Multi-Label Learning Classifier for Identifying the Subcellular Localization of Virus Proteins with Both Single and Multiple Sites. *J. Theor. Biol.* **2011**, *284*, 42–51. [[CrossRef](#)]
149. Wu, Z.C.; Xiao, X.; Chou, K.C. iLoc-Plant: A Multi-Label Classifier for Predicting the Subcellular Localization of Plant Proteins with Both Single and Multiple Sites. *Mol. BioSystems* **2011**, *7*, 3287–3297. [[CrossRef](#)] [[PubMed](#)]
150. Wan, S.; Mak, M.-W.; Kung, S.-Y. mPLR-Loc: An Adaptive Decision Multi-Label Classifier Based on Penalized Logistic Regression for Protein Subcellular Localization Prediction. *Anal. Biochem.* **2015**, *473*, 14–27. [[CrossRef](#)] [[PubMed](#)]
151. Wan, S.; Mak, M.-W.; Kung, S.-Y. Gram-LocEN: Interpretable Prediction of Subcellular Multi-Localization of Gram-Positive and Gram-Negative Bacterial Proteins. *Chemom. Intell. Lab. Syst.* **2017**, *162*, 1–9. [[CrossRef](#)]
152. Camon, E. The Gene Ontology Annotation (GOA) Database: Sharing Knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.* **2004**, *32*, 262D–266D. [[CrossRef](#)] [[PubMed](#)]
153. Dehzangi, A. Gram-Positive and Gram-Negative Protein Subcellular Localization by Incorporating Evolutionary-Based Descriptors into Chou's General PseAAC. *J. Theor. Biol.* **2015**, *364*, 284–294. [[CrossRef](#)] [[PubMed](#)]
154. Shen, H.-B.; Chou, K.-C. Virus-mPLoc: A Fusion Classifier for Viral Protein Subcellular Location Prediction by Incorporating Multiple Sites. *J. Biomol. Struct. Dyn.* **2010**, *28*, 175–186. [[CrossRef](#)] [[PubMed](#)]
155. Chou, K.-C.; Shen, H.-B. Plant-mPLoc: A Top-Down Strategy to Augment the Power for Predicting Plant Protein Subcellular Localization. *PLoS ONE* **2010**, *5*, e11335. [[CrossRef](#)] [[PubMed](#)]
156. Zhang, J.; Cruz-Cosme, R.; Zhuang, M.-W.; Liu, D.; Liu, Y.; Teng, S.; Wang, P.-H.; Tang, Q. A Systemic and Molecular Study of Subcellular Localization of SARS-CoV-2 Proteins. *Sig. Transduct. Target. Ther.* **2020**, *5*, 269. [[CrossRef](#)]
157. Boland, M.V.; Murphy, R.F. A Neural Network Classifier Capable of Recognizing the Patterns of All Major Subcellular Structures in Fluorescence Microscope Images of HeLa Cells. *Bioinformatics* **2001**, *17*, 1213–1223. [[CrossRef](#)]
158. Binder, J.X.; Pletscher-Frankild, S.; Tsafou, K.; Stolte, C.; O'Donoghue, S.I.; Schneider, R.; Jensen, L.J. COMPARTMENTS: Unification and Visualization of Protein Subcellular Localization Evidence. *Database* **2014**, *2014*, bau012. [[CrossRef](#)]
159. Meinken, J.; Asch, D.K.; Neizer-Ashun, K.A.; Chang, G.-H.; Cooper, C.R., Jr.; Min, X.J. FunSecKB2: A Fungal Protein Subcellular Location Knowledgebase. *Comput. Mol. Biol.* **2014**, *4*, 1–17. [[CrossRef](#)]
160. Lum, G.; Meinken, J.; Orr, J.; Frazier, S.; Min, X. PlantSecKB: The Plant Secretome and Subcellular Proteome KnowledgeBase. *Comput. Mol. Biol.* **2014**, *4*, 1–17. [[CrossRef](#)]
161. Meinken, J.; Walker, G.; Cooper, C.R.; Min, X.J. MetazSecKB: The Human and Animal Secretome and Subcellular Proteome Knowledgebase. *Database* **2015**, *2015*, bav077. [[CrossRef](#)] [[PubMed](#)]
162. Wan, S.; Mak, M.-W.; Kung, S.-Y. Mem-mEN: Predicting Multi-Functional Types of Membrane Proteins by Interpretable Elastic Nets. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2016**, *13*, 706–718. [[CrossRef](#)]
163. Wan, S.; Mak, M.-W.; Kung, S.-Y. FUEL-mLoc: Feature-Unified Prediction and Explanation of Multi-Localization of Cellular Proteins in Multiple Organisms. *Bioinformatics* **2017**, *33*, 749–750. [[CrossRef](#)] [[PubMed](#)]
164. Wan, S.; Mak, M.-W.; Kung, S.-Y. Sparse Regressions for Predicting and Interpreting Subcellular Localization of Multi-Label Proteins. *BMC Bioinform.* **2016**, *17*, 97. [[CrossRef](#)] [[PubMed](#)]
165. Wan, S.; Mak, M.-W.; Kung, S.-Y. mLASSO-Hum: A LASSO-Based Interpretable Human-Protein Subcellular Localization Predictor. *J. Theor. Biol.* **2015**, *382*, 223–234. [[CrossRef](#)]
166. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [[CrossRef](#)]
167. Ancona, M.; Ceolini, E.; Öztireli, C.; Gross, M. Towards Better Understanding of Gradient-Based Attribution Methods for Deep Neural Networks. *arXiv* **2018**. [[CrossRef](#)]
168. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning Important Features through Propagating Activation Differences. In Proceedings of the International Conference on Machine Learning, 2017; Available online: <http://proceedings.mlr.press/v70/shrikumar17a> (accessed on 28 February 2024).
169. Zecha, J.; Gabriel, W.; Spallek, R.; Chang, Y.-C.; Mergner, J.; Wilhelm, M.; Bassermann, F.; Kuster, B. Linking Post-Translational Modifications and Protein Turnover by Site-Resolved Protein Turnover Profiling. *Nat. Commun.* **2022**, *13*, 165. [[CrossRef](#)]
170. Ramazi, S.; Zahiri, J. Post-Translational Modifications in Proteins: Resources, Tools and Prediction Methods. *Database* **2021**, *2021*, baab012. [[CrossRef](#)]
171. Nickchi, P.; Jafari, M.; Kalantari, S. PEIMAN 1.0: Post-Translational Modification Enrichment, Integration and Matching ANALYSIS. *Database* **2015**, *2015*, bav037. [[CrossRef](#)] [[PubMed](#)]
172. Pacharawongsakda, E.; Theeramunkong, T. Predict Subcellular Locations of Singleplex and Multiplex Proteins by Semi-Supervised Learning and Dimension-Reducing General Mode of Chou's PseAAC. *IEEE Trans. Nanobiosci.* **2013**, *12*, 311–320. [[CrossRef](#)]



173. Wan, S.; Mak, M.-W.; Kung, S.-Y. Transductive Learning for Multi-Label Protein Subchloroplast Localization Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2017**, *14*, 212–224. [[CrossRef](#)] [[PubMed](#)]
174. Wan, S.; Mak, M.-W.; Kung, S.-Y. Ensemble Linear Neighborhood Propagation for Predicting Subchloroplast Localization of Multi-Location Proteins. *J. Proteome Res.* **2016**, *15*, 4755–4762. [[CrossRef](#)] [[PubMed](#)]
175. Zhu, X.-L.; Bao, L.-X.; Xue, M.-Q.; Xu, Y.-Y. Automatic Recognition of Protein Subcellular Location Patterns in Single Cells from Immunofluorescence Images Based on Deep Learning. *Brief. Bioinform.* **2023**, *24*, bbac609. [[CrossRef](#)]
176. Wang, B.; Zhang, X.; Xu, C.; Han, X.; Wang, Y.; Situ, C.; Li, Y.; Guo, X. DeepSP: A Deep Learning Framework for Spatial Proteomics. *J. Proteome Res.* **2023**, *22*, 2186–2198. [[CrossRef](#)] [[PubMed](#)]
177. Bi, Y.; Li, F.; Guo, X.; Wang, Z.; Pan, T.; Guo, Y.; Webb, G.I.; Yao, J.; Jia, C.; Song, J. Clarion Is a Multi-Label Problem Transformation Method for Identifying mRNA Subcellular Localizations. *Brief. Bioinform.* **2022**, *23*, bbac467. [[CrossRef](#)]
178. Li, J.; Zou, Q.; Yuan, L. A Review from Biological Mapping to Computation-Based Subcellular Localization. *Mol. Ther. Nucleic Acids* **2023**, *32*, 507–521. [[CrossRef](#)]
179. Villanueva, E.; Smith, T.; Pizzinga, M.; Elzek, M.; Queiroz, R.M.L.; Harvey, R.F.; Breckels, L.M.; Crook, O.M.; Monti, M.; Dezi, V.; et al. System-Wide Analysis of RNA and Protein Subcellular Localization Dynamics. *Nat. Methods* **2024**, *21*, 60–71. [[CrossRef](#)]
180. Cai, J.; Wang, T.; Deng, X.; Tang, L.; Liu, L. GM-IncLoc: LncRNAs Subcellular Localization Prediction Based on Graph Neural Network with Meta-Learning. *BMC Genom.* **2023**, *24*, 52. [[CrossRef](#)] [[PubMed](#)]
181. Zhang, B.; He, L.; Wang, Q.; Wang, Z.; Bao, W.; Cheng, H. Mit Protein Transformer: Identification Mitochondrial Proteins with Transformer Model. In Proceedings of the ICIC 2023: Advanced Intelligent Computing Technology and Applications, Zhengzhou, China, 10–13 August 2023; Huang, D.-S., Premaratne, P., Jin, B., Qu, B., Jo, K.-H., Hussain, A., Eds.; Springer Nature: Singapore, 2023; pp. 607–616.
182. Liu, M.-L.; Su, W.; Guan, Z.-X.; Zhang, D.; Chen, W.; Liu, L.; Ding, H. An Overview on Predicting Protein Subchloroplast Localization by Using Machine Learning Methods. *Curr. Protein Pept. Sci.* **2020**, *21*, 1229–1241. [[CrossRef](#)] [[PubMed](#)]
183. King, B.R.; Vural, S.; Pandey, S.; Barteau, A.; Guda, C. ngLOC: Software and Web Server for Predicting Protein Subcellular Localization in Prokaryotes and Eukaryotes. *BMC Res. Notes* **2012**, *5*, 351. [[CrossRef](#)]
184. Wan, S.; Kim, J.; Won, K.J. SHARP: Hyperfast and Accurate Processing of Single-Cell RNA-Seq Data via Ensemble Random Projection. *Genome Res.* **2020**, *30*, 205–213. [[CrossRef](#)] [[PubMed](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.