*Bioinformatics*, 2019, 1–7
doi: 10.1093/bioinformatics/btz909
Advance Access Publication Date: 5 December 2019
Original Paper



Bioimage informatics

# ImPLoc: a multi-instance deep learning model for the prediction of protein subcellular localization based on immunohistochemistry images

**Wei Long[1], Yang Yang [1,2,*] and Hong-Bin Shen [3]**

[1]Department of Computer Science and Engineering, Center for Brain-Like Computing and Machine Intelligence, Shanghai Jiao Tong University, Shanghai 200240, China, [2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai 200240, China and [3]Key Laboratory of System Control and Information Processing, Institute of Image Processing and Pattern Recognition, Ministry of Education of China, Shanghai Jiao Tong University, Shanghai 200240, China

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on May 6, 2019; revised on October 28, 2019; editorial decision on November 23, 2019; accepted on December 4, 2019

## Abstract

**Motivation:** The tissue atlas of the human protein atlas (HPA) houses immunohistochemistry (IHC) images visualizing the protein distribution from the tissue level down to the cell level, which provide an important resource to study human spatial proteome. Especially, the protein subcellular localization patterns revealed by these images are helpful for understanding protein functions, and the differential localization analysis across normal and cancer tissues lead to new cancer biomarkers. However, computational tools for processing images in this database are highly underdeveloped. The recognition of the localization patterns suffers from the variation in image quality and the difficulty in detecting microscopic targets.

**Results:** We propose a deep multi-instance multi-label model, ImPLoc, to predict the subcellular locations from IHC images. In this model, we employ a deep convolutional neural network-based feature extractor to represent image features, and design a multi-head self-attention encoder to aggregate multiple feature vectors for subsequent prediction. We construct a benchmark dataset of 1186 proteins including 7855 images from HPA and 6 subcellular locations. The experimental results show that ImPLoc achieves significant enhancement on the prediction accuracy compared with the current computational methods. We further apply ImPLoc to a test set of 889 proteins with images from both normal and cancer tissues, and obtain 8 differentially localized proteins with a significance level of 0.05.

**Availability and implementation:** https://github.com/yl2019lw/ImPloc.

**Contact:** yangyang@cs.sjtu.edu.cn

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Benefitting from the advances of high-throughput experiments and microscopic imaging techniques, image-based proteome has experienced a significant development for the past decade. The abundant spatial expression information captured in the microscopic images have greatly sped up the understanding of biological functions and mechanisms of proteins. The Human Protein Atlas (HPA) (www.proteinatlas.org) (Uhlen *et al.*, 2010) is an important public resource for the studies of the spatial human proteome, which houses both tissue-based and cell-based images, including the expression, metabolism and localization information from the tissue level to the subcellular level.

Especially, the tissue atlas of HPA collects over 13 million tissue-based immunohistochemistry (IHC) images, including 44 human tissues and organs. On the one hand, this data resource is commonly used in the quantitative analysis of expression variations in different tissues; on the other hand, since the images also capture protein distribution down to the subcellular level, they are useful for studying the variations in protein subcellular localization across tissues (Uhlen *et al.*, 2015).

As proteins should be located at the right compartments to function properly, the information of protein subcellular localization is very helpful in revealing protein functions (Emanuelsson *et al.*, 2000). Till now, a lot of computational tools for predicting protein subcellular localization have emerged (Briesemeister *et al.*, 2010; Chi and Nam, 2012; Park and Kanehisa, 2003; Pierleoni *et al.*, 2006; Zhou *et al.*, 2016), whereas most of them are based on amino acid sequences. The most common sequence features are residue-based statistical characteristics, such as the *k*-mer frequencies,

position specific scoring matrix (Xie *et al.*, 2005) and sorting signals, like mitochondrial-targeting signal and nuclear-localization signals (Nakai and Horton, 1999). Although a lot of efforts have been put on mining the sequence features, the sequence-based predictors can hardly achieve satisfying accuracy, as protein sequences carry limited information of protein subcellular localization. According to a recent study (Zhou *et al.*, 2016), the sequence features only obtained an $F_1$ of 0.22 and accuracy of 0.42 on the independent test set of human proteins.

In recent years, the microscopic images captured at the single-cell level have accumulated rapidly, which make the prediction of protein subcellular localization more accurate and allow tissue-specific studies. The microscopic images can visualize the complex localization patterns of proteins, and the image-based subcellular location prediction has attracted a lot of attention from both academia and industry. Especially, in the Kaggle competition for classifying subcellular protein patterns in human cells held in January 2019 (www.kaggle.com/c/human-protein-atlas-image-classification), over 2000 teams took part in the contest. The datasets of this competition and most of the other image-based studies were the immunofluorescence (IF) microscopic images from the cell atlas of the HPA database (Thul *et al.*, 2017), whereas the studies using IHC images have been very few (Kumar *et al.*, 2014; Newberg and Murphy, 2008; Xu *et al.*, 2013). Actually, as aforementioned, the tissue-based IHC images also contain subcellular localization information. Moreover, with samples from both cancer and normal tissues, this data repository can be very helpful in the differential analysis of protein subcellular localization between cancer and normal tissues. The results will lead to the identification of another kind of biomarker for diseases, i.e. the localization biomarker. Mislocalized proteins play roles in the progression of many human diseases (Hung and Link, 2011). Since their expression levels may have no obvious change in quantity, they are often missed in the biomarker screening based on differential expression analysis.

Protein subcellular localization based on the IHC images can be regarded as an image annotation task, where the annotation terms are the cellular compartments. However, this task is very different from common image annotation and much more challenging.

First, unlike the conventional image annotation tasks, the location labels are actually assigned to proteins, while a protein may correspond to a set of images from different experimental batches in different conditions. Therefore, the labeling may be based on all localization patterns implied in the image features considering all images in the set. Different proteins have different numbers of images, i.e. the input is a set with a varying size, which falls into a special machine learning framework, the multi-instance learning (Foulds and Frank, 2010; Zhou, 2004). Moreover, proteins can function at multiple different subcellular locations, thus this is a multi-instance multi-label problem (Zhou *et al.*, 2012).

In multi-instance learning, each sample contains several instances, all of which make up a single bag. The label of the sample is owned by all instances within a bag. A positive sample is present when the bag contains at least one positive instance, otherwise it is a negative sample. Different samples can contain different numbers of instances. In multi-label learning, each sample is associated with a set of labels instead of one. Using the multi-instance multi-label learning approach, we predict multiple possible subcellular locations based on multiple IHC images of each protein. The problem is shown in Figure 1.

Second, the recognition of subcellular localization is performed at the very microscopic level. Due to the limitation in the resolution of the tissue-based IHC images, the detailed localization patterns are hard to be detected. This is also a major reason that most of the image-based prediction methods focus on the IF images in the cell atlas rather than the IHC images. For such a micro-target detection task, (i) extra efforts on the image feature representation are in need to extract useful information for discriminating the locations; (ii) some images may be useless because of noise or experimental contamination, making this task a special multi-instance learning problem.

In order to tackle this difficult task, a few computational methods have been developed (Kumar *et al.*, 2014; Newberg and Murphy, 2008; Xu *et al.*, 2013), but all of them simplify this task as a conventional image classification problem, under the assumption that each image in a bag is associated with all the subcellular locations assigned to that bag. Such simplification may introduce noise and lead to a high false positive rate. This motivates us to develop a multi-instance learning model for protein subcellular localization based on IHC images. Besides, deep neural networks (DNNs) have been widely adopted in image processing for the past decade, because they can capture high-level abstraction features. Most of the top-ranked teams in the Kaggle competition for protein localization used DNNs, while most of the existing studies on the IHC images in HPA have still adopted traditional classifiers.

In this study, we propose a new deep learning framework, ImPLoc, to address the protein subcellular localization using IHC images. We first utilize a transfer-learning strategy and extract image features from a pre-trained convolutional neural network (CNN). Then, in order to learn localization patterns at the protein-level, we design a model with stacked processing units, accepting multiple feature vectors as input. Moreover, the model is equipped with a self-attention mechanism to focus on the salient image features. For assessing the new learning framework, we collect a benchmark dataset covering six subcellular compartments. The experimental results show that the abstraction features extracted from DNNs have obvious advantages over traditional features, and ImPLoc achieves effective feature encoding and classification. It outperforms traditional feature aggregation methods and the existing multi-instance learning models by a large margin. Moreover, we use ImPLoc to identify proteins whose localizations change in cancer tissues, i.e. the potential localization biomarkers. Through a rigorous screening process, we obtain 8 candidates out of the 889 proteins. Among them, six candidates' roles in cancer development have been reported in previous studies. ImPLoc provides a computational method to speed up the understanding of the mechanism of protein mislocalization and the identification of cancer biomarkers.

## 2 Materials and methods

### 2.1 Dataset

We download IHC images from HPA Version 18 (released December 1, 2017). All the IHC images in HPA were stained with chemical dyes, the brown sections contain protein and the purple sections contain DNA. We only choose images whose staining intensity level is strong or moderate and quantity is higher than 75%. We use images of proteins from four organs: liver, bladder, breast and prostate.

In HPA, there are a total of 28 subcellular locations. However, the numbers of samples in many locations are very small, resulting in a severely imbalanced distribution across labels. Besides, due to the limited resolution, only the major organelles are distinguishable from IHC images. Therefore, we merge the labels into six categories (i.e. nucleus, mitochondria, vesicles, golgi apparatus, endoplasmic reticulum and cytoplasm) according to the hierarchical structure of organelles. We divide the dataset into training and test sets at protein-level, i.e. the images belonging to the same protein are either in the training set (including validation) or the test set. Detailed numbers of data samples are shown in Table 1. Note that there are some multi-label proteins, i.e. proteins with more than one subcellular locations. The ratio of the number of labels to the number of proteins in the whole dataset is about 1.15. Table 2 shows the label distribution. Following previous study (Xu *et al.*, 2013), we transfer the manually curated annotation from IF images in the cell atlas into our dataset, as the annotations in the cell atlas are more accurate and comprehensive, including four confidence levels, namely enhanced, supported, approved and uncertain. In order to ensure the label reliability in the training data, we only use proteins whose label confidence level is 'enhanced'.

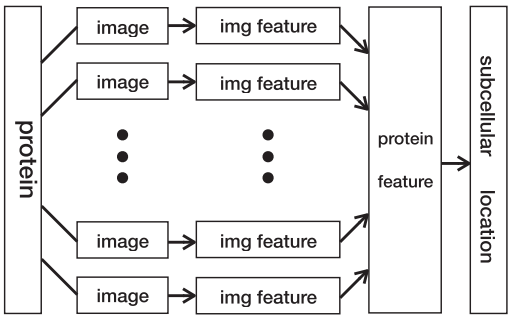


**Fig. 1.** Predicting multiple subcellular locations of a protein from multiple IHC images

**Table 1.** Data partition

| Split | Training[a] | Test | Total |
|---|---|---|---|
| Protein # | 1067 | 119 | 1186 |
| Image # | 7617 | 238 | 7855 |

[a]The training set includes validation set.

## 2.2 Overview of ImPLoc

We use pre-trained CNNs to extract features from raw images, then use a multi-head self-attention model to aggregate features from multiple instances into protein-level feature representations, and finally project the bag-level representations into the label space. The model architecture is shown in Figure 2.

## 2.3 Feature extraction

The existing studies extracted features from IHC images mainly based on patches. Commonly, the area containing proteins are first located and segmented out, then a traditional image feature extraction method is adopted to get subcellular location features (SLF), e.g. the SIFT descriptor (Lowe, 2004). As the rise of deep learning models, features learned using neural networks, such as AlexNet (Krizhevsky *et al.*, 2012), VGG (Simonyan and Zisserman, 2014), GoogLeNet (Szegedy *et al.*, 2015) and ResNet (He *et al.*, 2016), have been demonstrated to have much more powerful representation capacity than traditional image features, thus in this study we also explore the potential of using DNNs for learning features from IHC images.

Note that most of the IHC images are $3000 \times 3000$ pixels. It is very difficult to handle such large sizes using classic CNNs, due to heavy computational load and limited GPU memory. A common technique is to scale down the images before feeding them into CNNs. However, as mentioned in Section 1, this is a micro-target recognition task, i.e. the regions we focus on are very small in the original images. If we use the compressed images for training, we would lose important information. Thus, instead of using an end-to-end training scheme, we detach the feature representation from the training process. Specifically, we adopt the ResNet_18 model pre-trained on ImageNet (Krizhevsky *et al.*, 2012) to extract features from IHC images, as many studies have demonstrated that such a transfer learning strategy is effective for representing features of biomedical images (Yang *et al.*, 2019). The feature vectors are then fed into our deep learning framework.

According to previous studies on interpreting the feature map of CNNs (Zhang *et al.*, 2018), the low-level features describe detailed information, e.g. textures, colors and edges, while high-level features, which are more abstract, capture more position-independent semantic information. To investigate the impact of different levels of features, we extract feature from different blocks of ResNet_18. The corresponding feature dimensions are 64, 128, 256 and 512, respectively.

**Table 2.** Label distribution in the dataset[a]

| Label | Nuclear | Cytoplasm | Vesicles | Mito | Golgi | ER | Total |
|---|---|---|---|---|---|---|---|
| Training # | 587 | 345 | 77 | 120 | 52 | 47 | 1228 |
| Test # | 67 | 42 | 5 | 16 | 5 | 3 | 138 |

[a]Mito, Golgi and ER denote mitochondria, Golgi apparatus and endoplasmic reticulum, respectively. The training set includes validation set.

## 2.4 The core module

The core part of ImPLoc aggregates the feature vectors from each input set of images and learns a combined feature representation for subsequent classification. Unlike conventional CNNs or RNNs, the model needs to not only accept multiple input feature vectors but also consider all the feature vectors comprehensively and find useful information from them. To meet these needs, we design a DNN with attention mechanisms. Inspired by the self-attention and multi-head attention used in the Transformer model (Vaswani *et al.*, 2017), which has been widely used in machine translation tasks, we implement the similar attention layers and apply to the image features. The core module is shown in Supplementary Figure S1. Different from the Transformer model, ImPLoc only has the encoder part. Note that the input of ImPLoc is a set of feature vectors instead of a sequence, i.e. the model should be insensitive to the input order of image feature vectors, thus we remove the position-encoding section. The encoder consists of a stack of $N$ identical blocks. Each unit is comprised of two components, the first one is a multi-head self-attention mechanism and the second one is a fully connected feed forward network, where the output of the previous part is used as the input for the latter part. Each part adopts a residual connection and layer normalization. The feature aggregation process is formulated as follows.

Let $\mathbf{X}^{(1)}$ be the matrix representation for an input set of image features, defined in Equation (1),

$$\mathbf{X}^{(1)} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T, \tag{1}$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are the feature vectors (column vectors with dimensionality $d_m$) for the images, and $n$ is the number of images in the input set. $\mathbf{X}^{(1)}$ is inputted into the first block of the stack. The output of the $j$th block is the input of the $(j+1)$th block, where $j$ is the index of blocks, $j = \{1, 2, 3, \ldots, N\}$.

In the $j$th block, the attention component yields an output based on the query and key-value pairs generated by its input. As it is a multi-head attention component, we compute three matrices for the $i$th head,

$$Q^{(i,j)} = \mathbf{X}^{(j)} \times P_q^{(i)}, \tag{2}$$

$$K^{(i,j)} = \mathbf{X}^{(j)} \times P_k^{(i)}, \tag{3}$$

$$V^{(i,j)} = \mathbf{X}^{(j)} \times P_v^{(i)}, \tag{4}$$

where $\mathbf{X}^{(j)}$ is the input to the attention component in the $j$th block. $Q^{(i,j)}, K^{(i,j)}, V^{(i,j)}$ are the query, key, value matrices obtained from different projections of $\mathbf{X}^{(j)}$ for the $i$th head. $P_q$, $P_k$ and $P_v$ represent projection matrices corresponding to the query, key and value, whose sizes are $d_m \times d_q$, $d_m \times d_k$, $d_m \times d_v$, respectively, where $d_q$ is equal to $d_k$. The output of the $i$th single-head attention is a weighted sum of values, which is determined by the correlation of the query and the key, i.e.

$$\text{Attention}\ (Q^{(i,j)}, K^{(i,j)}, V^{(i,j)}) = W^{(i,j)} \times V^{(i,j)}, \tag{5}$$

$$W^{(i,j)} = \text{softmax}\left(\frac{Q^{(i,j)} \times K^{(i,j)T}}{\sqrt{d_k}}\right), \tag{6}$$

where $W^{(i)}$ (of size $n \times n$) represents attention weights.

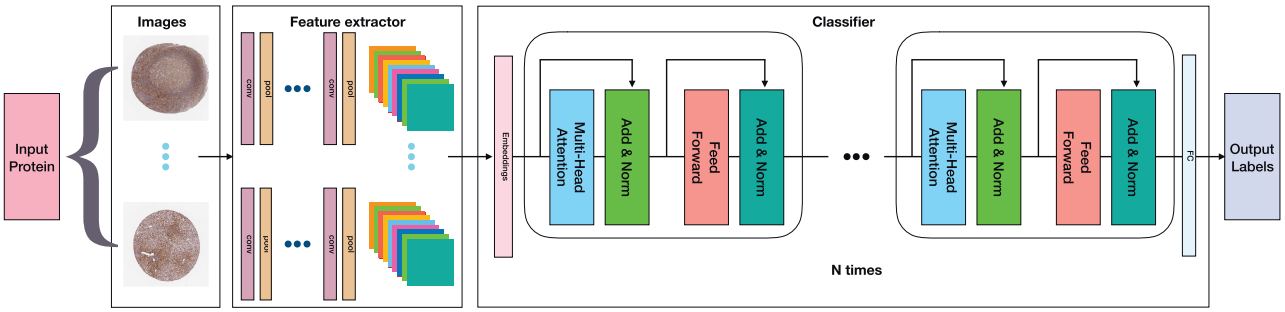


**Fig. 2.** Model architecture. An input protein has a series of images. First, a feature vector is extracted using a CNN from each image, then the features are aggregated in the core module using attention mechanism and passed to fully connected layers and finally the probabilities of labels are output

In the attention component, the above operations [Equations (2)–(6)] are conducted, respectively, in multiple heads, then the multiple single-headed outputs are concatenated together and projected again. In this way, the model can utilize information from different representation subspaces at different positions comprehensively.

$$\text{head}_{i,j} = \text{Attention}\ (Q^{(i,j)}, K^{(i,j)}, V^{(i,j)}) \tag{7}$$

$$\text{MultiHead}_j = \text{Concat}\ (\text{head}_{1,j}, \ldots, \text{head}_{h,j}) \times P_o, \tag{8}$$

where $P_o$ represents the output projection matrix, $h$ is the number of heads, $\text{MultiHead}_j$ is the output of the attention component of the $j$th block.

## 3 Experimental results

### 3.1 Experimental setup

There are four blocks in the ImPLoc encoder and six heads in each block. We partition the training data into 10 equal-size folds, using 9 of them for training and the remaining one for validation. We repeat this procedure for 10 times, thus get 10 models. The models are trained with Adam optimizer (Kingma and Ba, 2014) using adaptive learning rate. In order to avoid overfitting, both dropout and weight decay techniques are used, with rates 0.1 and 0.001, respectively. The 10 models are used to predict on the same test set, respectively. The averaged performance and max/min values of various metrics are shown in Figures 3–5.

### 3.2 Evaluation metrics

In order to assess the performance of ImPLoc, we adopt both the criteria common in supervised learning problems (label-based metrics), including accuracy, precision, recall, $F_1$ and the criteria specific for the multi-label learning (example-based metrics), including subset accuracy, example-based accuracy, precision, recall and $F_1$. The difference is that the former type of metrics are evaluated over the labels, while the latter type is computed per sample and then averaged over all samples in the dataset. The formal definitions of these measures can be found in the Supplementary Materials.

### 3.3 Investigation on feature representation

In this study, we adopt a transfer learning strategy to obtain original feature representation from each image, i.e. the images are passed to a pre-trained ResNet_18 model and converted into feature vectors. As it is known that different conv-layers of CNNs capture different levels of abstraction features, here we compare the prediction performance using features extracted from different layers of ResNet_18. Specifically, we append global average pooling to the end of 4 different blocks of ResNet_18 to get the feature vectors of 64, 128, 256 and 512 dimensions, respectively.

Figure 3 shows the comparison results of the four types of feature vectors extracted from different blocks of Resnet (for details see Supplementary Table S1). Res18-128D performs the best at most of the metrics (7 of 9). As it is known that, natural image processing tasks usually adopt the features from the last convolutional layer, because the features extracted from the last block are high-level abstract features, representing more semantic information. However, microscopic images are very different from natural images. Since the protein subcellular localization is a micro-target detection task, the low-level features that describe detailed information may be more important.

### 3.4 Investigation on feature aggregation

As described in Section 2.4, for a certain protein, we aggregate feature vectors extracted from all its corresponding images into a fixed-length feature representation. This operation is performed automatically by the new learning model. Actually, there are some alternative methods for the feature aggregation.

- Traditional methods: traditionally, as a part of feature engineering, the feature aggregation is an independent step before classification. There are some classic feature fusion methods, e.g. the bag of visual words (BOV) (Yang *et al.*, 2007), Fisher vector (Perronnin *et al.*, 2010) and vector of locally aggregated descriptors (VLAD) (Arandjelovic and Zisserman, 2013). By treating image features as words, BOV is a vector of occurrence counts for a vocabulary of local image features. The Fisher vector stores the mixing coefficients of the Gaussian mixture model as well as the mean and covariance deviation vectors of the individual components. The VLAD method computes the distance of each feature point to the cluster center closest to it. We experiment these three methods and adopt the traditional machine learning model, support vector machines (SVMs), as the classifier. For a fair comparison, we try different kernels and perform a grid search to obtain the optimal performance of SVMs.
- Other deep learning methods: till now, deep models for multi-instance learning have been very few, whereas some related models have emerged very recently. For instance, Feng and Zhou (2017) proposed a DNN model, DeepMIML, to address the object detection task in images, where an image is regarded as a bag of objects, thus the authors treated the task as a multi-instance multi-label learning problem. Another example is AnnoFly (Yang *et al.*, 2019), which was designed to annotate gene expression patterns from *Drosophila* embryogenesis images. The latter task is more similar to this study, and AnnoFly also allows multiple images as input while DeepMIML was designed for single-image input. AnnoFly adopts RNN to treat the bag of images as a sequence of instances. In order to compare with the RNN-based multi-instance learning, we use the gated recurrent units as the implementation.

In order to assess the performance of the new model, we compare it with both the traditional methods, i.e. BOV, Fisher vector and VLAD working with SVMs and the RNN-based deep learning

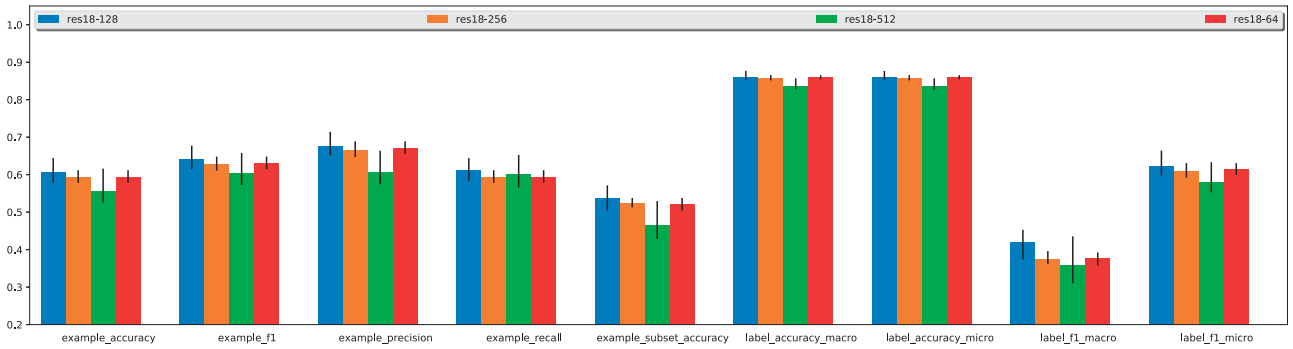


**Fig. 3.** Comparison of features extracted from different blocks of ResNet_18 with the SLFs. The bars show average performance of 10 repeated experiments, and small line segments on top of the bars show the max/min values

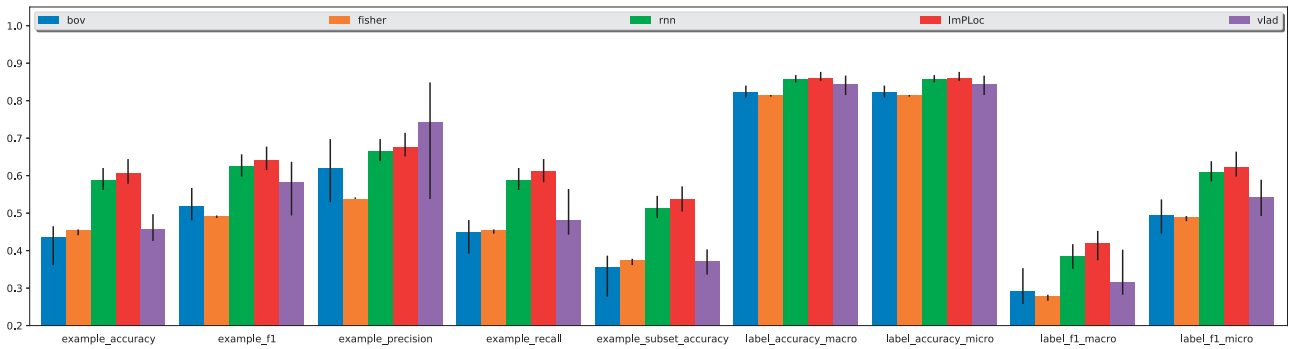


**Fig. 4.** Comparison of feature aggregation methods working with Res18-128D feature vectors

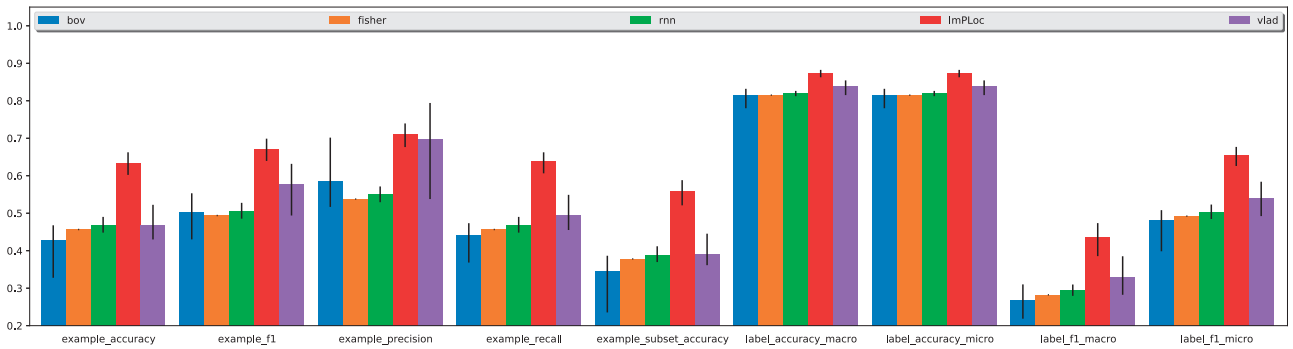


**Fig. 5.** Comparison of feature aggregation methods working with SLFs

model. As BOV, Fisher vector and VLAD are traditional feature aggregation methods, we also investigate their performance working together with traditional features, i.e. SLF. The comparison results are depicted in Figures 4 and 5. For both types of feature vectors, our model achieves the best performance on most of the evaluation metrics. The detailed discussions are listed below.

(i) Using the Res18-128D features, although RNN-based model is superior to traditional feature aggregation methods, ImPLoc outperforms RNN on almost all the metrics.
(ii) Using SLF features, RNN performs worse than VLAD, whereas ImPLoc still has an apparent advantage over VLAD.
(iii) Among the three traditional feature aggregation methods, VLAD performs the best. Fisher vector performs well on label-based metrics, but very bad on example-based ones. For the multi-label classification problems, the example-based metrics could be more important for performance assessment.
(iv) Both RNN and ImPLoc have relatively small variability on different metrics, whereas BOV and VLAD have larger variability, especially on the example-based metrics.

These results demonstrate that the new feature aggregation method not only achieves the highest accuracies but also has a stable performance no matter what kind of features are used.

### 3.5 Comparison with the existing predictor

To further assess the performance of the whole model, we compare it with *i*Locator, which is the state-of-the-art method for image-based protein subcellular localization (Xu *et al.*, 2013). Note that all the existing methods including *i*Locator perform prediction at the image-level rather on protein-level, i.e. their inputs are single images. To evaluate the performance at protein-level, we integrate the prediction results within each image set (corresponding to a protein). The comparison results are shown in Table 3. Besides *i*Locator and ImPLoc, we also list the accuracies of a randomized classifier in

**Table 3.** Performance comparison with the existing image-based predictor and randomized predictor[a]

| Method / Metric | Subset accuracy | Example accuracy | Example precision | Example recall | Example $F_1$ | Label accuracy | Label precision | Label recall | Label $F_1$ |
|---|---|---|---|---|---|---|---|---|---|
| *i*Locator | 0.303 | 0.354 | 0.408 | 0.356 | 0.380 | 0.772 | 0.311 | 0.249 | 0.277 |
| ImPLoc | 0.538 | 0.608 | 0.677 | 0.611 | 0.642 | 0.861 | 0.819 | 0.283 | 0.420 |
| Randomized | 0.077 | 0.083 | 0.092 | 0.083 | 0.087 | 0.671 | 0.696 | 0.183 | 0.290 |

[a]The metric values for ImPLoc are averaged over 10 repeated experiments.

Table 3 to show the significance of the achieved results, as the subcellular localization patterns present in IHC images are not as obvious as those in IF images. The randomized predictor is designed by shuffling rows of the label matrix (a binary matrix of the size $n \times m$, where $n$ is the number of samples and $m$ is the number of labels), i.e. the mapping relation between proteins and locations are permutated, while the numbers and distribution of locations in the dataset remain the same.

As can be seen in Table 3, ImPLoc outperforms *i*Locator on all metrics, suggesting the great advantages of the new deep learning-based method over traditional classifiers. As for the randomized predictor which uses the exact same feature vectors and model as ImPLoc, we find that the training process hardly converges and the prediction performance is quite bad. The randomized classifier obtains a higher value only on label precision, because it classifies all samples to the major classes (negative classes). Thus, it actually has no discriminant ability. This experiment demonstrates the effectiveness of the proposed model trained on the collected data.

### 3.6 Identification of cancer biomarkers

An important goal of IHC image-based analysis of protein subcellular localization is to identify mislocalized proteins, which may serve as biomarkers for cancers. In this section, we explore the potential of ImPLoc in the identification of such localization biomarkers.

As described in Section 2.1, the benchmark data used for training and test was collected from normal tissues. Here, we construct a new dataset, including both normal and cancer samples from three organs, i.e. liver, breast and prostate. This dataset contains a total of 889 proteins, which have IHC images from both normal and cancer tissues. The screening process of biomarker candidates is as follows.

- Step 1: For each protein, we divide its images into a normal set and a cancer set and use ImPLoc to predict the label vector. If the two vectors differ in at least one element, this protein will be kept for subsequent screening.
- Step 2: We use ImPLoc to predict the label probabilities (6-D vector) for all images one by one (each image can be regarded as a single-member set and passed into the model for test). In this way, for each protein, we obtain two groups of probability vectors.
- Step 3: Given the protein set returned by Step 1, for each subcellular location, we perform a $t$-test on the two groups of probability scores computed in Step 2 for each protein. If the $P$-values are less than 0.05 for all locations, we regard the protein as a candidate biomarker.

Using the above screening process, we finally obtain eight candidates (listed in Supplementary Table S2). We search the literature and find that six of them have direct evidence about their correlation with cancer. Especially, multiple studies revealed the role of LIF in the development of breast cancer (Estrov *et al.*, 1995; Jung Eun *et al.*, 2011; Ravandi and Estrov, 2001). In our prediction results, it has an obvious tendency to change location from cytoplasm to nuclear from normal to cancer tissue. Two proteins have the opposite changing pattern, i.e. from nuclear to cytoplasm. One is STRAP, which is an inhibitor of TGF-$\beta$ and plays important roles in the pathways related to cancer development (Kim *et al.*, 2007); and the other is CACYBP, human calcyclin-binding protein, which was reported to be an important regulator in gastric cancer. Besides, RSRC2 was reported to be a tumor suppressor (Kurehara *et al.*, 2007), whose location changes from mitochondria to nuclear. Goncalves *et al.* (2017) studied the effects of copy number variation (CNV) on the proteome of tumors. The large-scale study on tumor cells showed that the CNV events frequently happen in protein complex members, and identified GTF2E2 as a representative case. In Cheng *et al.* (2018), the authors reported PSMD4 as a therapeutic target in chemoresistant colorectal cancer. Interestingly, they also found that PSMD4 gives rise to the localization change of Nrf2 and further promotes cancer development (Lin *et al.*, 2016). The changing patterns of subcellular localization provide new hints for further exploring protein function in the pathogenesis of cancer. As for the remaining two proteins, ADPRHL2 and PDCL3, the former one lacks functional annotation, while the latter one was reported to play a crucial role in angiogenesis (Srinivasan *et al.*, 2015), which is also closely related to tumor progression.

## 4 Discussion

In this study, we prepare a dataset including a total of 1186 proteins and 7855 images. As far as we know, this is the largest dataset among current studies on IHC image-based protein subcellular localization. Previous studies conducted conventional single-instance learning (Newberg and Murphy, 2008; Xu *et al.*, 2013), where the training and test data was split at image-level; while our dataset enables multi-instance learning, where the training and prediction are both performed at protein-level. Moreover, the large data scale ensures the sufficient training of DNNs.

In addition to the benchmark set used in the experiments, we also constructed two larger datasets. As mentioned in Section 2.1, the annotations of subcellular localization in the HPA database have four confidence levels, namely enhanced, supported, approved, uncertain and we only use the proteins annotated by the most reliable labels (i.e. enhanced). The results reported in previous sections are all conducted on the 'enhanced' dataset. Considering that a larger dataset may lead to better prediction accuracy, we experiment on two other datasets with different annotation qualities. Specifically, we name the three datasets as $D_0$, $D_1$ and $D_2$, corresponding to the confidence level of enhanced, supported, and approved, which have 1186, 3441 and 5710 proteins, respectively. $D_0$ is included in $D_1$, and $D_1$ is included in $D_2$. Apparently, the larger the size, the less reliable of the data labels. The results are shown in Supplementary Table S1. For each dataset, we experiment four kinds of feature vectors with different dimensionality, i.e. 64, 128, 256 and 512.

Unfortunately, the prediction performance does not benefit from such data augmentation. $D_1$ has close performance with $D_0$, and $D_2$ has a significant decrease on nearly all metrics. A possible reason is that there is a lot of noise in these two levels of annotations, although marked as supported and approved.

Besides the data scale, we also notice that the input IHC images have large sizes. Since the subcellular localization is only related to the local area of the image, there may be potential performance gain if the model could focus more on the salient areas. However, the organelles in the IHC image are very small, and the localization patterns are difficult to recognize even with naked eyes, thus it is impossible to segment out the salient regions due to the lack of annotations. We can only attempt to execute unsupervised segmentation based on some prior information such as color distribution. A simple

implementation is to use sliding window technique to obtain patches by heuristic approaches, but the area, aspect ratio of the patch and sample density need to be carefully designed. Meanwhile, since this is a multi-instance learning problem, we need to summarize all results for patches coming from the same image and we also need to aggregate results from multiple images belonging to the same protein, thus resulting in a large number of instances in a bag, which will lead to another challenge. We will explore the segmentation methods to improve our model performance in our future work.

## 5 Conclusions

In this study, we propose a predictor of protein subcellular localization based on IHC images, called ImPLoc. The core parts include a feature aggregator using the DNN with a multi-head self-attention mechanism. The attention mechanism not only provides an effective way to encode the multiple input feature vector into a unified representation, but also extracts valuable information for subsequent classification, as the input usually consists of tens of images while many of them have poor quality. In order to assess the performance of ImPLoc, we construct a benchmark dataset from the tissue atlas of the HPA database. Compared with other feature extraction and aggregation methods, ImPLoc shows significant advantages of prediction accuracy. Besides, we perform a screening process for candidate biomarkers. In this process, we conduct hypothesis test on each location and identify potention locational biomarkers by observing the significance level of difference between normal and cancer tissues for each of the six cell compartments. One of our future research direction is to improve the screening process and find a proper quantitative indicator to describe the variation in multiple subcellular locations. This study can help the identification of cancer biomarkers that are hard to find in differential expression analysis, and also provide a general learning framework for tissue/cell-based microscopic image processing.

## Funding

This work was supported by the National Key Research and Development Program of China (No. 2018YFC0910500), the National Natural Science Foundation of China (Nos. 61972251, 61725302, 61671288) and the Science and Technology Commission of Shanghai Municipality (No. 17JC1403500).

*Conflict of Interest*: none declared.

## References

Arandjelovic,R. and Zisserman,A. (2013) All about VLAD. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1578–1585.

Briesemeister,S. *et al.* (2010) Yloc: an interpretable web server for predicting subcellular localization. *Nucleic Acids Res.*, **38**, W497–W502.

Cheng,Y. *et al.* (2018) Psmd4 is a novel therapeutic target in chemoresistant colorectal cancer activated by cytoplasmic localization of nrf2. *Oncotarget*, **9**, 26342–26352.

Chi,S.-M. and Nam,D. (2012) Wegoloc: accurate prediction of protein subcellular localization using weighted gene ontology terms. *Bioinformatics*, **28**, 1028–1030.

Emanuelsson,O. *et al.* (2000) Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.

Estrov,Z. *et al.* (1995) Leukemia inhibitory factor binds to human breast cancer cells and stimulates their proliferation. *J. Interferon Cytokine Res.*, **15**, 905–913.

Feng,J. and Zhou,Z.-H. (2017) Deepmiml network. *In: Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, CA*, pp. 1884–1890.

Foulds,J. and Frank,E. (2010) A review of multi-instance learning assumptions. *Knowledge Eng. Rev.*, **25**, 1–25.

Goncalves,E. *et al.* (2017) Widespread post-transcriptional attenuation of genomic copy-number variation in cancer. *Cell Syst.*, **5**, 386–398.

He,K. *et al.* (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

Hung,M.C. and Link,W. (2011) Protein localization in disease and therapy. *J. Cell Sci.*, **124**, 3381–3392.

Jung Eun,S. *et al.* (2011) Epigenetic up-regulation of leukemia inhibitory factor (lif) gene during the progression to breast cancer. *Mol. Cells*, **31**, 181–189.

Kim,C. *et al.* (2007) Overexpression of serine-threonine receptor kinase-associated protein in colorectal cancers. *Pathol. Int.*, **57**, 178–182.

Kingma,D.P. and Ba,J. (2014) Adam: a method for stochastic optimization. *eprint arXiv: 1412.6980*.

Krizhevsky,A. *et al.* (2012) Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105.

Kumar,A. *et al.* (2014) Automated analysis of immunohistochemistry images identifies candidate location biomarkers for cancers. *Proc. Natl. Acad. Sci.*, **111**, 18249–18254.

Kurehara,H. *et al.* (2007) A novel gene, rsrc2, inhibits cell proliferation and affects survival in esophageal cancer patients. *Int. J. Oncol.*, **30**, 421.

Lin,P.-L. *et al.* (2016) Cytoplasmic localization of nrf2 promotes colorectal cancer with more aggressive tumors via upregulation of psmd4. *Free Radical Biol. Med.*, **95**, 121–132.

Lowe,D.G. (2004) Distinctive image features from scale-invariant keypoints. *Int. J. Comp. Vis.*, **60**, 91–110.

Nakai,K. and Horton,P.R.T. (1999) Psort: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, **24**, 34–35.

Newberg,J. and Murphy,R.F. (2008) A framework for the automated analysis of subcellular patterns in human protein atlas images. *J. Proteome Res.*, **7**, 2300–2308.

Park,K.-J. and Kanehisa,M. (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, **19**, 1656–1663.

Perronnin,F. *et al.* (2010) Improving the fisher kernel for large-scale image classification. In: *European Conference on Computer Vision*, Springer, pp. 143–156.

Pierleoni,A. *et al.* (2006) Bacello: a balanced subcellular localization predictor. *Bioinformatics*, **22**, e408–e416.

Ravandi,F. and Estrov,Z. (2001) *The Role of Leukemia Inhibitory Factor in Cancer and Cancer Metastasis*. Kluwer Academic Publishers, Springer, Dordrecht.

Simonyan,K. and Zisserman,A. (2014) Very deep convolutional networks for large-scale image recognition. *arXiv Preprint arXiv*, **1409**, 1556.

Srinivasan,S. *et al.* (2015) Hypoxia-induced expression of phosducin-like 3 regulates expression of vegfr-2 and promotes angiogenesis. *Angiogenesis*, **18**, 449–462.

Szegedy,C. *et al.* (2015) Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9.

Thul,P.J. *et al.* (2017) A subcellular map of the human proteome. *Science*, **356**, eaal3321.

Uhlen,M. *et al.* (2010) Towards a knowledge-based human protein atlas. *Nat. Biotechnol.*, **28**, 1248.

Uhlen,M. *et al.* (2015) Tissue-based map of the human proteome. *Science*, **347**, 1260419.

Vaswani,A. *et al.* (2017) Attention is all you need. In: *Advances in Neural Information Processing Systems*, pp. 5998–6008.

Xie,D. *et al.* (2005) Locsvmpsi: a web server for subcellular localization of eukaryotic proteins using SVM and profile of psi-blast. *Nucleic Acids Res.*, **33**, W105–W110.

Xu,Y.-Y. *et al.* (2013) An image-based multi-label human protein subcellular localization predictor (i locator) reveals protein mislocalizations in cancer tissues. *Bioinformatics*, **29**, 2032–2040.

Yang,J. *et al.* (2007) Evaluating bag-of-visual-words representations in scene classification. In: *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval*, ACM, pp. 197–206.

Yang,Y. *et al.* (2019) Annofly: annotating drosophila embryonic images based on an attention-enhanced RNN model. *Bioinformatics*, **35**, 2834–2842.

Zhang,Q. *et al.* (2018) Interpreting CNN knowledge via an explanatory graph. In: Thirty-Second AAAI Conference on Artificial Intelligence, pp. 4454–4463.

Zhou,Z.-H. (2004) Multi-instance learning: a survey. In: *Department of Computer Science and Technology. Nanjing University, Tech. Rep.*

Zhou,Z.H. *et al.* (2012) Multi-instance multi-label learning. *Artif. Intel.*, **176**, 2291–2320.

Zhou,H. *et al.* (2016) Hum-mploc 3.0: prediction enhancement of human protein subcellular localization through modeling the hidden correlations of gene ontology and functional domain features. *Bioinformatics*, **33**, 843–853.