Subcellular Protein Localisation in the Human Protein Atlas using Ensembles of Diverse Deep Architectures

Syed Sameed Husain1*, Eng-Jon Ong1, Dmitry Minskiy1, Mikel Bober-Irizar1,2, Amaia Irizar2 and Miroslaw Bober1

1CVSSP, University of Surrey, Guildford, GU27XH, Surrey, UK.
2ForecomAI, Leatherhead, Surrey, UK.

*Corresponding author(s). E-mail(s): s.husain@surrey.ac.uk; e.ong@surrey.ac.uk; d.minskiy@surrey.ac.uk; mikel@mxbi.net; a.irizar@forecom.ai; m.bober@surrey.ac.uk;

Abstract
Automated visual localisation of subcellular proteins can accelerate our understanding of cell function in health and disease. Despite recent advances in machine learning (ML), humans still attain superior accuracy by using diverse visual cues. We show how this gap can be narrowed by addressing three key aspects: (i) automated improvement of cell annotation quality, (ii) new Deep Neural Network (DNN) architectures supporting unbalanced and noisy data, and (iii) informed selection and fusion of multiple & diverse machine learning models. We introduce a new “AI-trains-AI” method for improving the quality of weak labels and propose novel DNN architectures exploiting wavelet filters and Weibull activations. We also explore key factors in the multi-DNN ensembling process by analysing correlations between image-level and cell-level predictions. Finally, in the context of the Human Protein Atlas, we demonstrate that our system achieves state-of-the-art performance in the multi-label single-cell classification of protein localisation patterns, while strengthening generalisation ability.

Proteins play a vital role in almost all cellular processes crucial for our survival. The specific biological tasks that proteins perform depend on their spatial distribution at the subcellular level. Therefore, the distribution of different proteins within a cell provides important clues about cell functions and state, leading to a better understanding of diseases and their treatments [1]. This can be further aided by visually detecting and localising a specific protein within a cell [2]. An invaluable resource for studying cell biology is the Human Protein Atlas (HPA) [3, 4]. The HPA aims to provide a repository of information on all proteins present in the human body, from cells to organs. Importantly, the HPA contains a large collection of images of specific protein localisations at a subcellular level, acquired using immunofluorescence staining followed by confocal microscopy imaging [5]. This resource is key for an in-depth understanding of the human cell and the complex molecular mechanisms enabling its function [6, 7], as antibody-based multiplexed protein imaging methods enable extended capabilities [8].

In this new era of quantitative cell imaging, Artificial Intelligence (AI) can enable new opportunities, such as the identification of disease biomarkers and drug discovery, by analysing and interpreting complex image data to gain biological insights [9–11]. The challenge of designing an intelligent algorithmic approach to HPA analysis gained significant interest from the Kaggle ML community, and two international competitions gathering leading ML researchers and practitioners were organised [12, 13]. The latest,
Human Protein Atlas - Single Cell Classification 2021, aimed to develop an AI system for precisely modelling the spatial organisation of human cells. Provided with an HPA dataset consisting of high-resolution confocal microscopy images, the task was to predict protein organelle localisation labels for each cell in the image.

This task brings several challenges. The primary difficulty from a ML perspective is the lack of accurate ground truth for training. Each HPA image comprises several cells jointly labelled with one set of labels (image-level labels) given by the union of individual cell labels. However, the task is to classify each cell individually, and image-level labels are often incorrect for some of the cells in an image; this phenomenon is referred to as weak labelling. The second difficulty arises due to dramatic variability in cells’ shapes and textures, further compounded by the use of different cell lines and poor cell image quality caused by staining or segmentation failures (Figure 1a). Another issue is the extremely imbalanced frequencies of the nineteen localisation classes (Figure 1b), along with the multi-label setting where a single cell can take on up to 6 of the 19 classes.

The approaches used in Kaggle competition [14] for classifying individual cells employed deep learning with two distinct models. The first is a Cell-level model (CLM), which takes normalised segmented cells as input and outputs the predictions. Examples of CLMs are the Bag-of-Cells [15] and the Data-centric [16] models. The disadvantage of CLM is inferior classification performance due to training using weak cell labels. To address this, Multi-head models (MHMs) were developed. The MHM comprises an image head that models the entire population of cells and a cell head that classifies patterns in each cell in the image. Examples of MHMs are the Fair Cell Activation Network (FCAN) [17], Duo-branch DNN [18] and modified Puzzle-CAM [19]. These models are trained using Focal loss [20] or weighted Binary Cross-Entropy (BCE) loss [21] to address the challenge of class imbalance. Other important HPA tasks include detecting and down-weighting poorly segmented cells. Two key methods are to scale the confidence score based on the area [22] and completeness [17] of a cell.

The results of the HPA competitions demonstrated that despite significant improvements in classification performance, a significant gap remained between automatic methods and human experts. Given the complex nature of the task and lacking performance of AI systems, we believe some fundamental questions need to be answered to ultimately bridge this gap: (1) How to learn effectively from weakly-labelled multi-cell images? (2) How to develop deep architectures capable of dealing with extreme data variability? (3) How to ensemble diverse models to highlight their strengths?

Results

This section details innovative ML solutions to answer the above questions. We first design a novel multi-task DNN architecture called Dual-stream Actnet (DSA). The DSA model learns to extract and fuse information from images and cells and mitigates the vast amounts of false-positive cell predictions caused by weak cell labelling. We then focus on capturing the biological variability and richness of patterns present in the HPA data by designing two low-complexity and high-performance models: Cell-level Hybrid (CLH) and Cell-level Actnet (CLA).

Next, we introduce an “AI-trains-AI” framework to improve the quality of the imprecise labels. It employs two robust techniques: (1) DNN-powered re-mining and (2) Automatic bad cell down-weighting. We also discuss the result of improved label quality on system performance.

Finally, we highlight the importance of diversity in expert opinions (provided by different DNNs) and describe how to achieve optimal ensembling of classifiers by exploiting the diversity between different architectures, giving rise to our novel system called Hybrid subCellular Protein Localiser (HCPL). We compare HCPL with the state-of-the-art to demonstrate the effectiveness of this approach. The overview of our system is presented in Figure 1c.

Notably, inference for all subsequent systems completes in under 9 hours on identical hardware with an NVIDIA P100 GPU [23]. All reported numerical results were obtained using the HPA private dataset, while the analysis was performed using a validation set partitioned from the HPA training dataset.

The importance of cell-level and image-level information fusion

To maximise classification performance and increase robustness, we exploit both local (cell-level) and global (image-level) visual cues. Our novel DSA architecture learns to fuse both types of information by controlling DNN activations representing distinctive
Fig. 1: Overview of the HPA challenge and the proposed solution. (a) The primary challenge is weak labelling: the image is labelled as mitotic spindle, but only a single cell (circled in red) represents this label. Another difficulty stems from challenging imaging conditions. Furthermore, the projection of 3D cell structures onto a 2D image often causes ambiguities: structures lying above but outside the nucleus (e.g. mitochondria or centrosome) appear within the nucleus on some images but not on others. The detection of poor-quality cells is also an important task. (b) Cell statistics in the HPA dataset. The class imbalance is apparent: 400K cells have the nucleoplasm label, but only 8K have the mitotic spindle label. (c) All images in the dataset are represented by four channels, the nucleus (blue), the antibody-stained protein of interest (green), microtubules (red) and the endoplasmic reticulum (yellow). The cells are extracted from RGBY images using HPA Cell Segmentator (Methods) and passed to our novel DNNs (DSA, CLH and CLA) for training. The quality of HPA cell labelling is improved using DNN-powered iterative data re-mining and filtering of ambiguous cells using our Bad Cell Detector system. We introduce a data-driven approach to model ensembling and present the HCPL system to generate multi-label predictions for 19 classes\(^a\). Finally, we conduct an in-depth study to understand the outcomes using CAM and UMAP plots.

\(^a\)The dataset and its 19 classes are visualised at https://www.kaggle.com/code/lnhtrang/single-cell-patterns/notebook.
visual features. The DSA incorporates a parametric activation block using a non-linear transformation, amplifying meaningful activations and optimising information fusion.

The structure of DSA is presented in Figure 2a. From each image, N cells are selected, resized and flattened as a batch (typically \( N = 20 \)). Let \( X \in \mathbb{R}^{A \times B \times 4} \) denote a cell image of resolution \( A \times B \). Each cell \( X \), is processed by a base DNN (e.g. EfficientNet [24]), which embeds an input into the space of compact deep features. The output tensor of the final convolutional layer, denoted as \( R \in \mathbb{R}^{W \times H \times D} \), is forwarded to a learnable activation layer [25], where \( W \) and \( H \) are the width and height of the feature map and \( D \) is the feature dimensionality. The activation layer maximises the Signal-to-Noise (SNR) ratio of the last convolutional feature map by applying the Weibull function and is crucial for training the DSA with weak labels. It transforms the last convolutional features map in a non-linear and robust way: dampening the weaker responses and equalising a learnable proportion of strong responses, thereby improving the discriminative power of the feature map (Methods). The output of the activation layer is forwarded to a Global Average-Pooling (GAP) layer and power-normalisation layer to generate global descriptors, which are then passed to the image stream and cell stream. The image stream outputs a prediction for the entire population of cells in an image, whereas the cell stream generates predictions for each cell in an image (Methods).

The DSA network is trained end-to-end using a weighted sum of Binary Cross-Entropy losses from cell and image streams. At inference, the final confidence score for each cell is computed as the correlation-driven product of image and cell stream predictions (Methods). Figure 3a shows that the image stream individually achieves 42.1% mAP. This baseline performance comes from mapping image-level labels to all cells in that image, resulting in vast overlabelling. The cell stream achieves a better 51.1% mAP, still relatively low due to the weak labels used in training. However, our DSA architecture achieves 55.2%, a gain of (+4.1%) stemming from the intelligent fusion of both streams.

**The importance of capturing the variability in HPA**

We develop two novel cell-level architectures to capture the variability and richness of patterns in the HPA. The first, Cell-level Hybrid (CLH), has an inductive hybrid architecture (Figure 2b). The network uses two complementary representations: hand-crafted features provided by a scattering transform [26] and features learnt from the data. The scattering transform, formed by cascading wavelets and modulus non-linearities, guarantees translation invariance and a linear response to deformations. When combined with learnt features, the model benefits from mathematically well-defined and data-adaptable filters. To ensure these properties are propagated into the main DNN flow, two Hybrid Fusion blocks are inductively integrated into an EfficientNet architecture [27]. The Hybrid-DNN module works on cell-level images and outputs convolutional features. The features are aggregated and passed to the classification module (fully connected and Softmax) to generate predictions.

The second architecture is the Cell-level Act-net (CLA), which employs learnable parametric activations to deliver adaptive and robust aggregation (Figure 2c). At its core, it uses a base DNN as a deep feature extractor. These features are fed to a non-linear activation layer [25] comprising the Weibull function. The Weibull function amplifies the responses corresponding to distinctive features of objects that are important for the classification tasks relative to the background. The transformed tensors are forwarded to the GAP layer, power-normalisation layer and classification module to compute predictions.

The training of both CLH and CLA is performed using a weighted Binary Cross-Entropy loss function. At inference time, the final confidence for each cell is computed as a correlation-based product of the cell-level architecture output and DSA image stream output. A class-wise evaluation demonstrates that the CLH better represents rare classes (e.g. 0.22% mAP gain over CLA on aggresome). The main gain comes from the wavelet filters at the base of the Hybrid Fusion Block, which are pre-defined and help the network to generalise with few training examples. Hand-crafted features also benefit the CLH by imposing translation invariance and geometric priors when describing blob-shaped organelles such as vesicles and Golgi apparatus. Conversely, the ability of CLA to select the most discriminative features helped it perform better on visually similar intermediate filaments (+0.13%) and actin filaments (+0.31%). We can observe in Figure 3a that cell-level architectures achieve better mAP than DSA. However, the performance of cell-level networks strongly depends on the strong cell labels generated by DSA.
Fig. 2: Key components of the HCPL system. (a) The DSA comprises an image stream that models the entire population of cells and a cell stream that classifies patterns in each cell in the image. (b) CLH is a single stream inductive hybrid architecture that uses a scattering transform as a complimentary source of information about each cell. (c) CLA is a single-stream architecture which employs Weibull activation pooling to aggregate deep features before classification.
**Fig. 3: Performance and functional details of selected elements of our system.** (a) The classification performance of different modules of HCPL including image stream, cell stream, DSA, CLH, CLA, Ensemble without BCD system and Ensemble with BCD system. (b) The figure on the left shows the histogram of confidence scores assigned to labels based on the ground-truth data. Approximately 60% of labels have scores below 0.1 indicating two issues: (1) DNNs are trained using weak labels, (2) training of DNNs is complex due to segmentation faults and cell variability. The CRA transforms the scores using a power-normalisation operation to generate a new set of probabilities $K$. The probabilities $K$ then replace their associated labels, producing a new set of improved continuous valued cell labels. The histogram of improved cell labels is presented on the right. (c) The BCD system consists of two modules: one that extracts geometric features from a cell and employs an XGBoost classifier, and a DNN to predict the visual integrity of the cell.
Improving labels quality using an “AI-trains-AI” approach

DNN-powered re-mining. Our approach leverages the generalisation ability of the DSA model to learn from noisy data and subsequently assigns confidence scores to ground-truth data. A confidence score represents the probability that a ground-truth label is correct; hence it takes values in [0, 1]. We use confidence scores to perform iterative training where subsequent models focus less on low confidence cell labels while emphasising those with high confidence.

This is achieved by our Cells Re-labelling Algorithm (CRA). As the first step, it computes the per-cell component probabilities using three DSAs trained on weak labels (bases EfficientNet-B4, NFNet and ResNest50d [28]). These three component probabilities are averaged across each label for each cell to obtain a new set of combined confidence factors.

In the second step, we re-evaluate the true-positive labels for each cell in the ground-truth data based on these combined confidence factors (Figure 3b left). The CRA transforms the factors using a power-normalisation operation \( c \mapsto c^\beta \), \( c \) is a probability) to generate a new set of probabilities \( K \), which replace the original labels, yielding improved continuous-valued cell labels (Figure 3b right). The CRA effectively re-evaluates cell labels originally inherited from the image level.

Next, we retrain the DSA model using relabelled cells. Experimental results show that the DSA retrained on improved labels achieved classification performance of 55.2% mAP compared to DSA trained on original weak labels (52.5% mAP).

The above process is employed twice to further improve cell labels’ quality. The cell labels obtained after round two are used to train the final CLA and CLH networks. As a result, the CLA and CLH trained on improved labels achieved an improvement exceeding +4% mAP over the models trained on the original labels.

Automatic bad cell detection. Poor imaging or segmentation failures may affect the quality of cell images. Hence, an important feature of our system is a Bad Cell Detector (BCD) that explicitly detects such events and adjusts classification confidence accordingly, thereby increasing the reliability of classification labels. Figure 3c demonstrates the internal operation of the BCD system. It consists of two modules; the first extracts the eight most representative geometric features from each cell and uses an XGBoost classifier [29] to compute the probability of a cell being bad. Then, the output cell confidences are obtained by multiplying the probability of a cell being ‘good’ with the previously obtained per class confidences.

The competition organisers specified that cells with a significant part (over 50%) truncated would not be considered for classification by annotators. Therefore, the second module of the BCD is a DNN with base EfficientNet-B2 [24], which predicts the visual integrity of each cell. If the cell is predicted to be less than 50% visible, the confidence score of that cell is down-weighted (Methods).

The results in Figure 3a show that the inclusion of the BCD module improves the overall system performance by 0.8% mAP.

The result of intelligent information fusion

Correlation of image-level and cell-level predictions. To maximise the system’s classification performance, we exploit correlation-based fusion of cell and image-level information. Figure 4a demonstrates nineteen correlation matrices plotted by contrasting image and cell-level predictions for cells from the corresponding images. We can observe that most classes have high correlation coefficients except mitotic spindle, centrosome and aggresome. The reason for low correlation values is that only a small subset of cells in an image contains the class. Image-level predictions fail to reflect this since the image-level classifier must label all cells present in an image, causing over-labelling in cases where only a small subset of cells may contain a particular label (exacerbated on rare classes).

For classes where the correlation is greater than a certain threshold \( th \), the confidence score of a cell is obtained as a product of image-level and cell-level predictions. We assign the cell-level prediction to the confidence score for classes where the correlation coefficient is less than \( th \) since the cell-level network was trained using improved cell-level labels.

Capturing HPA variability with ensemble diversity. Classification accuracy can be improved by combining the outputs from diverse classifiers with comparable performance. A set of classifiers is considered diverse if they perform well on different dataset examples. We aim to achieve optimal ensembling of classifiers by exploiting the diversity between different architectures (DSA, CLH, CLA) employing various base DNNs.
Fig. 4: Correlation based ensembling of diverse DNNs. (a) Correlation between image-level and cell-level predictions for each class. Classes mitotic spindle, centrosome and aggresome have very low correlation coefficients $r$. The values of $r$ are displayed with the class names at the top of the correlation plot. The image stream consistently labels all cells as having aggresome, visible in the bright horizontal line at the top of the correlation matrix. However, we know that only around 40%-60% of cells in an image have aggresome highlighted on the green channel. The same phenomenon can be seen for the class mitotic spindle, where the number of cells showing the mitotic spindle is even fewer (1-3 cells per image). We again see that the image stream consistently outputs high values for many cells, with little correlation to the cell-level outputs. (b) Correlation between confidence scores generated by different DNNs. (c) HCPL inference framework employed to generate submissions for Kaggle.
The diversity inherent in the DNNs can be visualised with a correlation matrix of confidence scores produced by each network (Figure 4b). We notice strong correlations between DNNs of similar architectures, for example, EfficientNet-B3 and EfficientNet-B4, or ResNet101 and ResNet50d. Furthermore, the SwinT model shows a low correlation to all other convolutional architectures. The Hybrid-DNN has a marginal amount of correlation with EfficientNet only. The selection of the final network set (Figure 4c) is based on a greedy algorithm using Pearson’s correlation between pairs of individual DNN outputs [30].

The result of diversity based ensembling of DNNs. Figure 4c demonstrates our HCPL system that applies the principles discussed in this work. The first phase is to extract individual cells from each image using HPA Cell Segmentator. The cells are forwarded to the DSA, CLH and CLA models to compute the predictions. The predictions from these diverse networks are averaged to compute the confidence scores. Simultaneously, the cells are passed to the BCD to compute the bad cell weighting. The confidence scores are multiplied by the bad cell weights to generate final score.

This diversity-driven multi-DNN ensembling brings a considerable improvement to the final classification score. Figure 5a demonstrates that none of the individual DNNs (DSA, CLH or CLA) exceeds 55.8% mAP. The mAP can be improved to 56.6% by ensembling five Dual-stream architectures with bases ResNet50d [28] (55.2%), EfficientNet-B4 [24] (55.0%), NFNNet-ECA [31] (54.9%), Regnety-32 [32] (54.7%) and DenseNet161 [33] (54.6%). We can achieve a mAP of 56.7% by ensembling the predictions from CLAs with bases EfficientNet-B4 (55.8%) and ResNet50D (55.5%), CLH (55.6%) and SwinT [34] (55.4%). Importantly, the ensembling of all DSA and CL networks obtains a state-of-the-art classification score of 57.1%.

Figure 5e demonstrates the performance of our system in each of the nineteen classes. The performance depends on several variables such as the difficulty of recognising different localisation patterns, number of training samples and extreme visual variations. Despite a lower number of training samples, the mAP is higher for aggresome (63.9%) than plasma membrane (50.3%) because aggresome has visually distinguishable features, whereas plasma membrane is often confused with cytosol. The endoplasmic reticulum has the lowest mAP of 36.3%; it is also confused with cytosol. Centrosome also has a low mAP of 42.1% due to its proximity to the nucleus, causing it to often appear within the nucleus; a consequence of 2D projection of a 3D structure. This means it can be confused with nucleoli or nuclear bodies. Despite the extreme rarity of the mitotic spindle in training samples, our system achieves a good score of 64.3%.

Comparison with the state-of-the-art. In this section, we first compare the performances of single models without ensembling. From Figure 5b, we observe that the proposed Dual-stream Actnet (DSA) outperforms state-of-the-art MHMs [14]. Compared to the FCAN [17] and Duo-branch [18] models, DSA provides an improvement of +1.1% and +1.4% in mAP respectively. Among cell-level architecture [14], Figure 5c demonstrates that our CLA achieves the best classification performance of 55.8%.

When ensemble-based architectures are considered (Figure 5d), our HCPL system achieves the highest classification accuracy of 57.1% mAP. The second-best result (56.7%, Bestfitting, [17]) was trained with antibody information; this was discouraged by the competition organisers as it is known to generalise poorly on new data. A previous study [12] reported that gains achieved using this information may be due to the exploitation of data leakage between the training and testing examples via the antibody identifiers and cell-line information specific to the competition data. The next best approach [18] that did not utilise antibody data achieved significantly lower performance (55.3%).

The ability of DNNs to interpret biological information

Comprehending whether output from a DNN for a given HPA image is based on biologically meaningful information is important; DNNs that fail to focus on this information generally have lower cell classification performance. We use the Grad-CAM method [35] to visualise the spatial attention of DNNs. Given a target label, gradients with respect to a loss function from each DNN convolutional layer are combined into a class activation map (CAM), highlighting regions whose visual features are used by the DNN to predict the target label. In Figure 6a-c, we use CAMs to show the effect of improving the quality of weak labels using our Cells Re-labelling Algorithm (CRA). We can observe that the staining patterns of the respective classes overlap well with the CAMs of the high-scoring model trained using improved cell labels.
Fig. 5: Evaluation of the HCPL system and comparison with the state-of-the-art. (a) The impact of diversity-based ensembling of DNNs. (b) Comparison of our DSA with MHMs. (c) Comparison of our CLA with existing CLMs. (d) Comparison of our final system with state-of-the-art ensembled systems. (e) The performance of our system for different label classes. Performance of the two rarest classes is relatively good: mitotic spindle 64.3% and aggresome 63.9%. The nuclear bodies class achieves only 43.0% as it is often mistakenly classified as nucleoplasm. Furthermore, there is a significant amount of confusion between cytosol, endoplasmic reticulum and plasma membrane classes resulting in relatively poor performance.
Deep features visualisation using UMAP. When performing multi-label classification on cells, each of the DNNs extracts a global descriptor in the form of a high-dimensional vector representing the salient features of a cell image. We can visualise these global descriptors using uniform manifold approximation and projection for dimension reduction (UMAP) [36]. The visualisation results for two different DNNs (DSA and CLH) are presented in Figure 6d-e. Here, each point on the plot represents a single cell. Only cells with single labels are chosen, coloured by their respective label. Firstly, we observe that the DNNs cluster the majority of cells from the same class together, which shows their underlying discriminative power (since UMAP produces visualisations in an unsupervised manner). Secondly, we can notice the presence of two larger clusters at a global level: one for nuclear sub-compartments (nucleoplasm, nuclear membrane, nucleoli, nucleoli fibrillar center, nuclear speckles and nuclear bodies) and another for organelles outside the nucleus (e.g. cytosol, mitochondria). Thirdly, instances of some classes are located within clusters belonging to other classes, which links to the relative decrease in performance (Figure 5e), for example: cytosol partially overlaps with endoplasmic reticulum and plasma membrane, nuclear bodies with nucleoplasm and vesicles with Golgi apparatus. Lastly, the different amount of overlap observed in the clusters generated by DSA and CLH confirms the diversity and complementarity in the deep features originating from different DNN architectures.

Discussion

The HPA competition aimed to develop AI solutions to precisely model the spatial subcellular protein organisation on a dataset with weak labels, high-class imbalance, and significant visual variations. Despite significant efforts, existing systems suffer from low performance, hallucination and poor generalisation [37].

Our core contributions include a novel DSA architecture for improved image and cell information fusion and two novel Cell-level architectures (CLA & CLH) to capture the wide variability between cells. Further, we introduce an “AI-trains-AI” approach with DNN-powered re-mining & automatic bad cell down-weighting and propose an effective strategy to optimally ensemble diverse DNNs. The experimental evaluation shows that HCPL outperforms existing approaches and is expected to generalise well to unseen cell-lines and proteins. Further improvements to the state-of-the-art could be achieved by obtaining more examples of rare localisation classes, accurate cell labels, and feedback from human experts. It is recognised that the annotation of unknown proteins needs to be greatly extended to fully anchor the opportunities that the sequencing of the human genome affords [38]. HCPL reduces the performance gap between AI methods and human experts. This will help accelerate the characterisation of unknown proteins and our understanding of cellular function and biology, in order to advance our knowledge of disease-related phenotypes and drug discovery.

Methods

HPA Dataset

Our work uses the HPA dataset provided by the “Human Protein Atlas - Single Cell Classification” Kaggle challenge [13]. This consists of images from the freely accessible Human Protein Atlas project. In particular, images from the subcellular section of the HPA were used. A total of 97K images (internal and external dataset) were made available for training purposes whilst an additional 2270 unseen images were retained by the challenge organisers for testing purposes, further split into 559 images for the public leaderboard and 1711 images for the private leaderboard. Each of the provided images contains multiple cells and consists of four channels, displayed in red, green, blue, and yellow. The task considered in this paper is to localise the protein of interest (green channel) in 18 possible subcellular organelles in each cell in an image. An additional “negative” class is added for negative staining and unspecific patterns. The 19 labels and their names are shown in Figure 1b. Our DNNs are trained on approximately 1.2 Million cell images segmented from HPA images using the HPA Cell Segmentator.

Evaluation metric

To ensure a fair comparison of results, all experiments were evaluated by computing mean Average Precision (mAP) [39]. The mean value was calculated over the 19 segmentable classes (C) of the challenge with a mask-to-mask IoU > 0.6 as described below:

\[
mAP = \frac{1}{C} \sum_{i}^{C} prec_i
\]
Fig. 6: Visual interpretations of learned deep features. (a-c) CAMs of a low-scoring CLA trained on weak labels and a high-scoring CLA trained on improved labels. (a) The low-scoring DNN uses visual features belonging to the aggresome class to assign a high probability for the Nucleoplasm Class. As a result, the cells indicated by green circles receive a strong output for Nucleoplasm, which is biologically incorrect as these cells do not have their nucleoplasm highlighted on the green channel. The high-scoring DNN focuses on biologically relevant regions of the cell. (b) The CAMs for a Golgi reassembly stacking protein, localised to the Golgi apparatus demonstrate attention of accurate size for the high-scoring model only. (c) Green circles show the areas where visual features of the microtubules are not captured by the low-scoring DNN, yielding low confidence for the Microtubules class. However, the high-scoring DNN focuses on relevant visual features. (d-e) UMAP visualisation of the deep features for DSA and CLH DNNs. UMAP is a dimensionality reduction method capable of projecting high-dimensional data onto a 2D plane such that local structures in the original space are captured, whilst simultaneously retaining the global structure of the data. The deep features from both DNNs distinguish the different subcellular locations.
where $Prec_i$ is Precision for class $i$ which is calculated according to the two-stage method described in article [40]. All mAP scores are reported based on the Kaggle private leaderboard.

**Improving the HPA Cell Segmentator (HCS)**

The HCS [41] segments input images into individual cell instances for multi-label classification. Since the HCS software provided by organisers procedure occupied 60% of the total permitted processing time (9 hours for the entire system’s inference), improving the efficiency of the segmentation algorithm is important. The algorithm consists of three main stages: i) prediction of the nuclei; ii) general cell prediction and iii) post processing procedure. However, cell segmentation training data was not publicly available, and so gains could only be obtained by modifying the post-processing procedure rather than training a new model. In the segmentation architecture, the nuclei and cell maps are first obtained via the corresponding predictor U-Net [42]. The next step is the post-processing of the outputs. To improve the efficiency of HCS, we introduce down-scaling and up-scaling blocks at the start and the end of the process respectively. Since post-processing is largely based on various morphological operations, its complexity is proportional to the product of the processed image dimensions. Therefore, reducing the spatial dimensions by 50% resulted in a speedup of a factor of at least 2. Another effect of the reduced dimensionality was a significantly lowered amount of high-frequency noise. This allowed simplification of the pipeline by removing extra processing in two fine-tuning blocks and in the "Segmentation and gradual object removal" block. These changes resulted in a 2x speed-up. However, such speed improvement resulted in a deterioration in system accuracy of around 0.2% mAP.

**Configurations for DNNs training**

The DSA consists of an image stream and a cell stream. The image stream applies Global Max-Pooling to a bag of $N$ cell descriptors originating from a single image to generate a unified image representation $V$, which is then passed to a fully connected layer $FC_1$ and Softmax to generate an image-level prediction. The cell stream takes $N$ cell descriptors as an input and outputs the predictions for each cell using a fully-connected layer $FC_2$ and Softmax. The predictions from the image stream are passed to classification loss layer. The loss layer computes the weighted Binary Cross-Entropy loss ($L_f$) between the image label and bag-prediction. Similarly, the cell stream weighted Binary Cross-Entropy loss $L_2$ is calculated between cell predictions and cell labels. The final loss ($L_f$) is the weighted sum of cell stream loss and image stream loss $L_f = W_1 \times L_1 + W_2 \times L_2$. For the cell stream, the labels are weak and we therefore intuitively assign a much lower weight to cell stream loss ($W_2 = 0.2 \times W_1$). The DSA is trained using an Adam optimiser and cosine annealing learning rate scheduler. We applied data augmentation in the form of random cropping, flipping, shifting, rotation, scaling and cutout.

The Cell-level Actnet and Cell-level Hybrid are trained using Focal loss, Adam optimiser, and a cosine annealing scheduler with initial learning rate $2e^{-4}$.

**Bad Cell Detector system**

For training of first BCD module, we first compute the eight most representative features from each cell in the training dataset: bounding box height, width, aspect ratio, area, mask area, mask perimeter, the value of the largest dimension and a binary feature that is based on the pixel intensity and the ratio of blue and green to the total number of pixels. The training dataset contains 10K cells hand-labelled as either 'good', i.e. most of a cell is clearly visible or 'bad', i.e. a cell is damaged and not suitable for further processing. Note, this dataset with an extended set of properties and reference segmented cells are made publicly available [43]. We then train the XGBoost classifier on cell features using a five fold cross-validation strategy.

The second module consists of a base EfficientNet-B2 with a fully connected layer to output predictions for four classes. The data to train the DNN is generated by randomly cropping out some area on the border of the cell. If the cropped area is less than 30% of the original cell, that image belongs to class 1. Similarly, if the cropped area is between 30% to 50%, 50% to 80% and 80% to 100% then that image is assigned to classes 2, 3 and 4 respectively. The network takes cropped cell tiles as an input and outputs probabilities of the four classes. The training is performed using the cross-entropy loss function. At inference time, each cell is forwarded to a trained EfficientNet-B2, and the probability for each class is obtained.
The Weibull activation layer

The Weibull activation layer is aimed at maximising the Signal-to-Noise ratio (SNR) of the last convolutional feature map by applying the Weibull function to the tensor $\mathbf{R}$, the output tensor of the final convolutional layer. Each element of the tensor $\mathbf{R}$ is transformed by the Weibull function resulting in the output tensor $\mathbf{T} \in \mathbb{R}^{W \times H \times D}$ (where $0 \leq i < W \times H \times D$):

$$T_i = \left( \frac{R_i}{\lambda} \right)^{\xi - 1} \exp\left( - \left( \frac{R_i}{\gamma} \right)^\eta \right).$$ (2)

The learnable parameters of the activation layer are $\lambda$, $\xi$, $\gamma$, and $\eta$. The output of the activation layer is fed to the Global Average Pooling (GAP) layer, denoted as $P(\mathbf{T})$, to compute the global vector $\mathbf{S}$:

$$\mathbf{S} = \left( \frac{1}{WH} \sum_{i}^{W} \sum_{j}^{H} T_{ijk} \right)^D_{k=1}$$ (3)

Each element ($s \in \mathbf{S}$) of the tensor $\mathbf{S}$ is power-normalised to balance the non-linear scaling of the Weibull function. The power normalisation function is represented as $\delta : \mathbb{R}^D \rightarrow \mathbb{R}^D$, with the rule:

$$\delta(s) = \alpha_1^s, \alpha_2^s, ..., \alpha_D^s$$ (4)

where $\alpha$, $\beta$ are learnable scaling parameters.

References

[1] Paul I, White C, Turcinovic I, Emili A. Imaging the future: the emerging era of single-cell spatial proteomics. The FEBS Journal. 2021;288(24):6990–7001.

[2] Gnann C, Cesnik AJ, Lundberg E. Illuminating Non-genetic Cellular Heterogeneity with Imaging-Based Spatial Proteomics. Trends in Cancer. 2021;7(4):278–282.

[3] Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015;347.

[4] Sullivan D, Winsnes C, Åkesson L, Hjelmare M, Wiking M, Schutten R, et al. Deep learning is combined with massive-scale citizen science to improve large-scale image classification. Nature Biotechnology. 2018.

[5] Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Blal HA, et al. A subcellular map of the human proteome. Science. 2017;356(6340).

[6] Lundberg E, Borner G. Spatial proteomics: a powerful discovery tool for cell biology. Nature Reviews Molecular Cell Biology. 2019 01;20:285–302.

[7] Christopher J, Stadler C, Martin C, Morgenstern M, Pan Y, Betsinger C, et al. Subcellular proteomics. Nature Reviews Methods Primers. 2021 12;.

[8] Hickey J, Neumann E, Radvke A, Camarillo J, Beuschel R, Albanese A, et al. Spatial mapping of protein composition and tissue organization: a primer for multiplexed antibody-based imaging. Nature Methods. 2022 Mar;19:284–295.

[9] Chandrasekaran SN, Ceulemans H, Boyd J, Carpenter AE. Image-based profiling for drug discovery: due for a machine-learning upgrade? Nature Reviews Drug Discovery. 2020;20:145–159.

[10] Bagheri N, Carpenter AE, Lundberg E, Plant AL, Horwitz R. The new era of quantitative cell imaging—challenges and opportunities. Molecular Cell. 2022;82(2):241–247.

[11] Mund A, Coscia F, Kriston A, Hollandi R, Kovács F, Brunner AD, et al. Deep Visual Proteomics defines single-cell identity and heterogeneity. Nature Biotechnology. 2022 May;.

[12] Ouyang W, Winsnes C, Hjelmare M, Cesnik A, Åkesson L, Xu H, et al. Analysis of the Human Protein Atlas Image Classification competition. Nature Methods. 2019;p. 1254–1261.

[13] HPA.: Human Protein Atlas - Single Cell Classification. https://www.kaggle.com/competitions/hpa-single-cell-image-classification/.

[14] HPA-Discussion.: Human Protein Atlas - Single Cell Classification Discussion. https://www.kaggle.com/competitions/hpa-single-cell-image-classification/discussion.

[15] Henkel C.: 3rd place Dieter part. https://www.kaggle.com/competitions/
hpa-single-cell-image-classification/discussion/238898.

[16] Banić N.: 4th Place Solution. https://www.kaggle.com/competitions/hpa-single-cell-image-classification/discussion/239071.

[17] Dai S.: Fair Cell Activation Network and Swin Transformer, the 1st place solution. https://www.kaggle.com/competitions/hpa-single-cell-image-classification/discussion/239001.

[18] Makarov I.: HPA 2nd Place Solution. https://www.kaggle.com/competitions/hpa-single-cell-image-classification/discussion/238645.

[19] Jo S, Yu IJ. Puzzle-CAM: Improved Localization Via Matching Partial And Full Features. In: 2021 IEEE International Conference on Image Processing (ICIP); 2021.

[20] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017. p. 2999–3007.

[21] Jadon S. A survey of loss functions for semantic segmentation. In: 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB); 2020. p. 1–7.

[22] Chubachi K.: 7th place solution. https://www.kaggle.com/competitions/hpa-single-cell-image-classification/discussion/238507.

[23] HPA-Leaderboard.: Human Protein Atlas - Single Cell Classification Private Leaderboard. https://www.kaggle.com/competitions/hpa-single-cell-image-classification/leaderboard.

[24] Tan M, Le Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: Proceedings of the 36th International Conference on Machine Learning; 2019. p. 6105–6114.

[25] Husain SS, Ong E, Bober M. ACTNET: end-to-end learning of feature activations and multistream aggregation for effective instance image retrieval. IJCV. 2021.

[26] Mallat S. Group invariant scattering. Communications on Pure and Applied Mathematics. 2012;65(10):1331–1398.

[27] Minskiy D, Bober M. Efficient Hybrid Network: Inducting Scattering Features. CoRR. 2022;abs/2203.15392. https://doi.org/10.48550/arXiv.2203.15392. https://arxiv.org/abs/2203.15392.

[28] Zhang H, Wu C, Zhang Z, Zhu Y, Zhang Z, Lin H, et al. ResNeSt: Split-Attention Networks. CoRR. 2020.

[29] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p. 785–794.

[30] Ong EJ, Husain SS, Bober-Irizar M, Bober M. Deep Architectures and Ensembles for Semantic Video Classification. IEEE Transactions on Circuits and Systems for Video Technology. 2019. p. 3568–3582.

[31] Brock A, De S, Smith SL, Simonyan K. High-Performance Large-Scale Image Recognition Without Normalization. In: Proceedings of the 38th International Conference on Machine Learning; 2021. p. 1059–1071.

[32] Xu J, Pan Y, Pan X, Hoi S, Yi Z, Xu Z. RegNet: Self-Regulated Network for Image Classification. IEEE Transactions on Neural Networks and Learning Systems. 2022. p. 1–6.

[33] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. p. 2261–2269.

[34] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: 2021
IEEE/CVF International Conference on Computer Vision (ICCV); 2021. p. 9992–10002.

[35] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017. p. 618–626.

[36] McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv e-prints. 2018.;

[37] Belthangady C, Royer L. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. Nature Methods. 2019 07;16:1215–1225.

[38] Kustatscher G, Collins T, Gingras AC, Guo T, Hermjakob H, Ideker T, et al. Understudied proteins: opportunities and challenges for functional proteomics. Nature Methods. 2022 May;.

[39] LIU L, ÖZSU MT. In: Mean Average Precision; 2009. p. 1703–1703.

[40] Everingham M, Gool LV, Williams CKI, Winn J, Zisserman A. The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision. 2009;88:303–308.

[41] Atlas HP.: HPA Cell Segmentator. Available from: https://github.com/CellProfiling/HPA-Cell-Segmentation.

[42] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015; 2015. p. 234–241.

[43] Ong EJ, Minskiy D, Bober-Irizar M, Husain SS, Bober M.: HPA Bad Cell Detection. Kaggle. Available from: https://www.kaggle.com/dsv/3349136.