nature biotechnology

Deep learning is combined with massive-scale citizen science to improve large-scale image classification

Devin P Sullivan[1,7], Casper F Winsnes[1,7], Lovisa Åkesson[1], Martin Hjelmare[1], Mikaela Wiking[1], Rutger Schutten[1], Linzi Campbell[2], Hjalti Leifsson[2], Scott Rhodes[2], Andie Nordgren[2], Kevin Smith[3], Bernard Revaz[4], Bergur Finnbogason[2], Attila Szantner[4] & Emma Lundberg[1,5,6]

Pattern recognition and classification of images are key challenges throughout the life sciences. We combined two approaches for large-scale classification of fluorescence microscopy images. First, using the publicly available data set from the Cell Atlas of the Human Protein Atlas (HPA), we integrated an image-classification task into a mainstream video game (EVE Online) as a mini-game, named Project Discovery. Participation by 322,006 gamers over 1 year provided nearly 33 million classifications of subcellular localization patterns, including patterns that were not previously annotated by the HPA. Second, we used deep learning to build an automated Localization Cellular Annotation Tool (Loc-CAT). This tool classifies proteins into 29 subcellular localization patterns and can deal efficiently with multi-localization proteins, performing robustly across different cell types. Combining the annotations of gamers and deep learning, we applied transfer learning to create a boosted learner that can characterize subcellular protein distribution with F1 score of 0.72. We found that engaging players of commercial computer games provided data that augmented deep learning and enabled scalable and readily improved image classification.

Analysis of large data sets is an increasingly important challenge[1]. Although machine learning, artificial intelligence and citizen science offer potential solutions to coping with this explosion of data[2–6], the large amounts of data that are generated as automated fluorescence microscopy systems become ever more widely used in quantitative biology create new challenges for automated image analysis.

The Human Protein Atlas (HPA) is an open-access database using antibody labeling and microscopy to systematically build an image-based map that details the spatial distribution of proteins in human cells and tissues (http://www.proteinatlas.org)[7]. Subcellular compartmentalization is fundamental to eukaryotic cells enabling multiple cellular processes to occur in parallel. The Cell Atlas in the HPA is building a proteome scale map of protein subcellular localization via hundreds of thousands of high-resolution confocal immunofluorescent images[8]. This map aids researchers in understanding protein function, interactions, cellular biology and, ultimately, disease. Given the magnitude of images continuously collected by the HPA Cell Atlas, a detailed analysis of the data requires a very large number of accurate image classifications.

Previous efforts to automate the classification of protein subcellular distribution from images have included methods such as k-NN classifiers, support vector machines, artificial neural networks and decision trees. Most studies have used hand-crafted feature sets[9–14], whereas others have used inference and multi-resolution techniques[15]. Recently, deep convolutional neural networks (CNNs) have been successful in classifying protein localization of single localizing proteins in budding yeast[16,17] and human cells[18]. These approaches, however, have focused on a limited number of single patterns (9–18 patterns), most often in a single cell type. This number of labels only provides a coarse description of biology as cellular architecture is more refined with specialized sub-organelle compartments and dynamic structures. However, the severe class imbalance introduced when considering rare cellular structures makes it harder to create a classifier that is capable of accurately predicting all localizations[19]. An even greater limitation to previous methods is that they only consider proteins in a single subcellular location, making them unsuitable for the ~50% of the human proteome that are multi-localizing[8]. Multi-localizing proteins are likely to be important for the inter-connectedness and adaptivity of cellular processes; thus, correctly localizing these proteins is key to our understanding of cell biology. Although methods of 'unmixing' a pair of known individual patterns have been put forward[20,21], to the best of our knowledge no global image-based subcellular protein classification method that handles multi-localizing proteins has been presented until now[22].

Crowd-sourced citizen science offers an alternative for large-scale image classification[6]. Projects such as FoldIt[23,24], Galaxy Zoo[25–27], EyeWire, EteRNA[28] and Quantum moves[29] represent implementations of citizen science in which large numbers of non-expert volunteers have contributed valuable scientific information. The major drawback of this approach is that implementing an engaging citizen

[1]Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH - Royal Institute of Technology, Stockholm, Sweden. [2]CCP hf, Reyjkavik, Iceland. [3]Science for Life Laboratory, School of Computer Science and Communication, KTH - Royal Institute of Technology, Stockholm, Sweden. [4]MMOS Sàrl, Monthey, Switzerland. [5]Visiting appointment: Department of Genetics, Stanford University, Stanford, California, USA. [6]Visiting appointment: Chan Zuckerberg Biohub, San Francisco, San Francisco, California, USA. [7]These authors contributed equally to this work. Correspondence should be addressed to E.L. (emma.lundberg@scilifelab.se).

Received 24 December 2017; accepted 19 July 2018; published online 20 August 2018; doi:10.1038/nbt.4225

science project requires resources, knowledge and time that most laboratories lack. Furthermore, creating and maintaining an engaged user base is difficult in one-off citizen science projects. One method of dealing with this is paying for citizen science efforts, as in Amazon's mechanical turk (mturk)[30]; however, this method is prone to exploitation and low data quality[31].

Here we demonstrate two complementary and successful approaches for large-scale classification of protein localization patterns in microscopy images from the HPA Cell Atlas. The first utilizes the power of massive multiplayer online (MMO) games to create a new approach to citizen science and was a collaborative effort between the HPA, Massive Multiplayer Online Science (MMOS) and the video game developer CCP Games. This partnership substantially reduced the effort to the lab by allowing CCP Games to develop the interface and MMOS to handle data management and serving. The result was the scientific project of image classification seamlessly integrated into the EVE Online universe, an MMO science fiction game with ~500,000 active players each month. The resulting mini-game, Project Discovery (PD), was successful in terms of participation, player retention, number of images classified and accuracy. In the second approach, we present Loc-CAT, a model for automated image classification of subcellular protein distribution patterns using deep neural networks (DNNs). To the best of our knowledge, this method represents the first tool for classifying protein distribution in human cells in microscope images capable of predicting robustly across cell types for proteins with an unknown number of locations. Furthermore, we compared the performance of the respective approaches and found that the gamer output could be used to improve deep learning models. Altogether, both approaches provide a refinement of the biological details in the HPA Cell Atlas. We believe that integration of scientific tasks into established computer games can be a valuable approach in the future with the power of rapidly leveraging the output of large-scale science efforts.

RESULTS

Subcellular distribution of proteins in microscopy images

Each sample in the HPA Cell Atlas consists of human cells that are immunofluorescently labeled for one protein of interest and three reference markers: DAPI for the nucleus and antibody-based labeling of microtubules and the endoplasmic reticulum. High-resolution images were acquired using confocal microscopy (**Fig. 1a**). The resulting images were annotated to determine the localization(s) of the protein of interest with the help of the three cellular reference markers. This study was performed using the Cell Atlas of the Human Protein Atlas version 14.0 (HPA Cell Atlas v14; **Supplementary Data Set 1**) in which protein distributions were classified into 20 organelles and subcellular structures (**Fig. 1b**). To refine the biological details of the HPA Cell Atlas, players in PD were asked to re-classify these images and classify the protein distribution into an additional ten patterns (**Fig. 1c,d**), for a total of 29 patterns in 17 human cell lines (**Supplementary Table 1**).

Some protein localization patterns, such as the centrosome, were small and easily overlooked, whereas others, such as the cytokinetic bridge, occurred in only a small fraction of cells. Adding to the complexity, some compartments, such as actin filaments, the Golgi apparatus and mitochondria, displayed highly heterogeneous morphologies across cell lines, making them more difficult to recognize (**Fig. 1e**). Class frequency (that is, protein localization) varied widely in human cells, from 0.016–24.3% in the HPA Cell Atlas v14. Furthermore, ~50% of proteins were localized to multiple cellular compartments[8] (**Fig. 1f**). Cell-to-cell variability could create further confusion (**Fig. 1g**). Together, these findings demonstrate that location classification is hardly a trivial task.

Image classification by citizen science task in EVE Online

In PD, players in EVE Online performed the aforementioned protein image classification. This project represents the first time a scientific task has been directly and seamlessly integrated into a mainstream video game narrative (**Fig. 2a**). The resulting mini-game was accessible from anywhere and at any time in the virtual universe of EVE Online. Participants were trained using a small set of preselected images gradually increasing in difficulty. Classification options were initially restricted to ease players into the complexity of the task. Participants in PD were motivated with leveled badges and in-game currency with which they could purchase exclusive items. This approach was able to easily gather and maintain participants, something other citizen science projects have struggled with, as measured by 'project appeal'[32] (**Fig. 2b** and Online Methods). This participation drive can particularly be seen when EVE became free to play, causing an increase in PD participation (day 250; **Fig. 2c**).

Participation peaked on day 3, with 5,507 players contributing 292,374 classifications (**Fig. 2c**). In total, 322,006 players of EVE Online played PD and contributed ~33 million image classifications. Of these, 59,901 players passed the training and tutorial phases and had above threshold performance, leading to 23.7 million high-quality image classifications. From this set of 59,901 players, on average 6,846 unique players contributed each month with a 30-d monthly retention rate of 32% (68% churn), and a rolling retention of 53% (47% churn) over the first 6 months, which was very good compared with other in-game features over the same period and vastly improves on previous citizen science efforts (**Supplementary Fig. 1**)[33].

Measuring player performance

To assess data quality, we used the F1 score, a measure of accuracy suitable for multi-label data, with the HPA Cell Atlas v14 image labels as ground truth. Initially, players received an additional reward for agreeing with the eventual community consensus. This reward was quickly exploited by gamers converging to a single annotation (cytoplasm, day 0–20; **Fig. 2d**) and was therefore removed. This resulted in a rapid improvement of accuracy (**Fig. 2d**). On the basis of player feedback, we created and implemented a larger set of more difficult control images including multi-localizing proteins and image artifacts. This led to a significant increase in data quality (day 50, $P < 4 \times 10^{-70}$, day 0–50 versus 50+, two-tailed t test; **Fig. 2d**).

To guard against erroneous annotations, we required a minimum of 12 votes per image before evaluating each task for a consensus using a hypergeometric test. Consensus was considered to be reached only if the number of votes for at least one class was significantly greater than would be expected at random ($P < 0.01$) and no other classes were near the decision threshold ($P < 0.1$). If consensus was not reached, the task was kept open and more votes were acquired. On average, each task required 15 player votes (median = 13) to reach a consensus. Given the speed of players, the data set was annotated six times, resulting in a median of 78 annotations per image. This statistical approach, together with the high number of annotations per image, allowed us to tolerate annotations from players performing worse than naively guessing the single most common class, accounting for ~10% of the annotations (**Fig. 3a**). The overall F1 score was 0.55 with a mean per-class F1 score of 0.50. In general, gamers performed better for common categories, presumably because they are more accustomed to these. Microtubules were a notable exception, as the gamers had a reference channel, allowing them to easily recognize this pattern (F1 = 0.78).

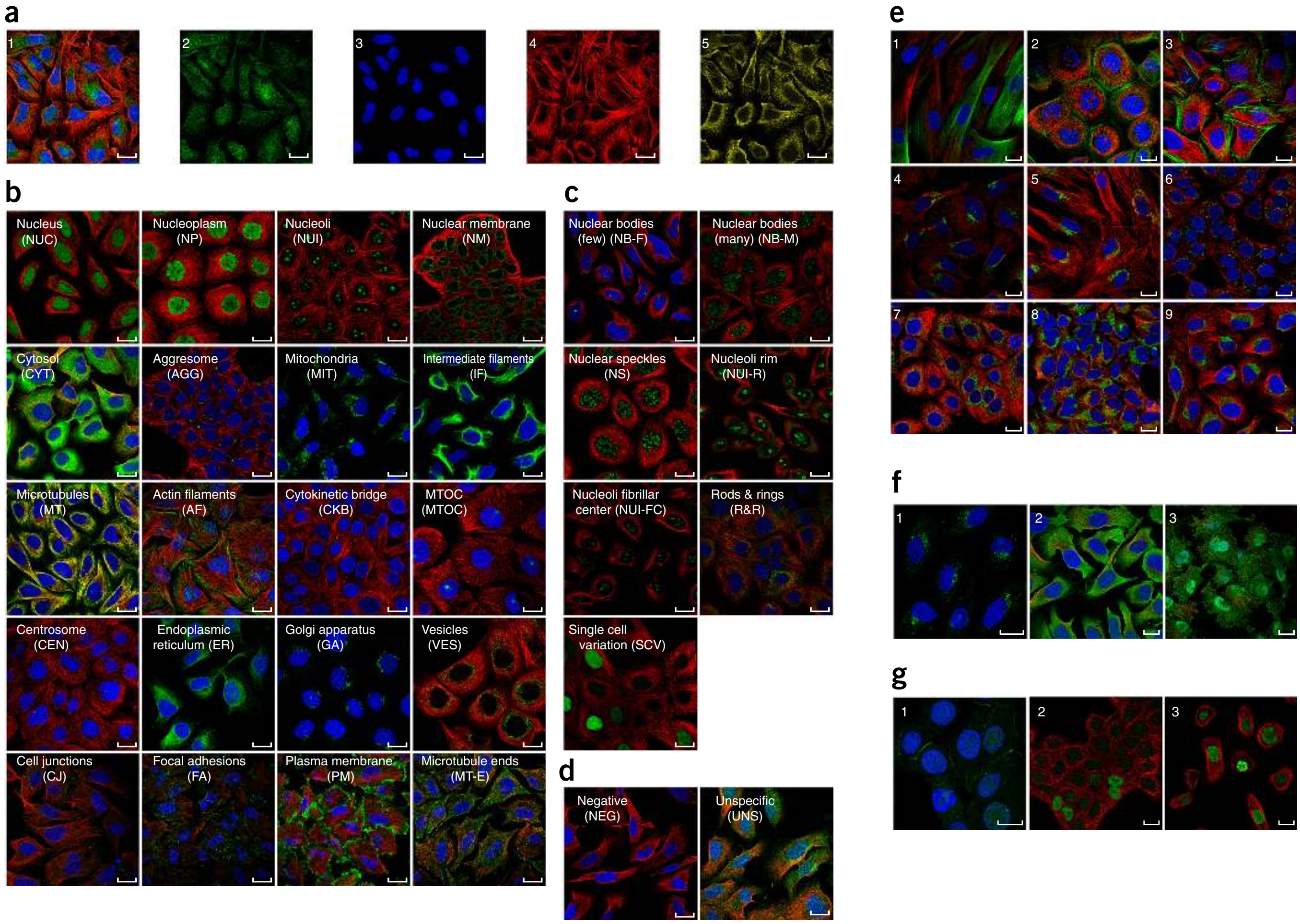


Figure 1 Illustrative data from the HPA Cell Atlas. The HPA Cell Atlas contains four-channel confocal images for the majority of all human proteins. Scale bars represent 20 μm (inner length). The experimental procedure was described previously[8]. (**a**) Example composite image (1) consisting of false-colored channels representing the protein of interest (green, 2), with DAPI labeling the nucleus (blue, 3), an antibody labeling the microtubules (red, 4). Each assay also contains an antibody labeling the endoplasmic reticulum (yellow, 5). (**b**) The images in HPA Cell Atlas v14 were classified into 20 organelle patterns of major organelles. (**c**) In PD, players classified seven additional patterns. (**d**) Two additional patterns were classified to filter negative and unspecific experiments. (**e**) Organelles displayed morphological differences across cell lines such as actin filaments (1–3), Golgi (4–6) and mitochondria (7–9). (**f**) Multi-localizing proteins offer another challenge for accurate pattern classification. Often patterns were hard to distinguish when occurring together, such as Golgi and vesicles (1), or cytoplasm and plasma membrane (2). Some localizations were not visible or easily distinguishable in every field of view, particularly when they occurred in variable focal planes. For example, although focal adhesions, nucleoplasm, and cytosol were in focus, a Golgi apparatus localization in another focus plane could not be seen (3). (**g**) Cell-to-cell variations can be challenging and viewing images with various channel combinations can aid annotations of patterns such as cell junctions (1) or nucleoplasm (2 and 3).

An alternative would be expectation maximization for jointly estimating player bias and protein localization[34]. This approach was not feasible during gameplay given the computation time, and did not perform as well as the aforementioned consensus approach in *post hoc* analysis (**Supplementary Table 1**). As a result of the large number of annotations per image, it is also unlikely that additional annotations would improve the accuracy of PD without a shift in player behavior (retraining) or task (sub-set annotation)[35]. This is supported by the lack of correlation between the number of images analyzed, time-on-task per player and performance (**Supplementary Fig. 2**).

The frequency at which players co-annotated classes was compared with the co-annotation frequencies in the HPA Cell Atlas v14 using independent Bonferroni corrected binomial tests to estimate multi-label confusion for each class (**Fig. 3b** and **Supplementary Fig. 3a**). Patterns of structurally similar organelles appeared to be frequently confused, such as centrosomes and microtubule organizing centers (**Fig. 3c**). Although much rarer, confusion across the nuclear and cytoplasmic spaces could also be observed, for example, when gamers annotated vesicles instead of nuclear bodies, both of which are dot-like structures.

Players were also able to report unusual findings in images. A review of all images with more than 20 such reports mainly revealed rare cellular morphologies such as blebs and membrane protrusions, or staining artifacts. However, this demonstrates that the players are capable of finding patterns that deviate from the common patterns, and identified several interesting and previously unannotated patterns such as vesicle fronts and condensed chromosomes.

Adjusting for PD player bias leads to improved data quality

Of the 29 patterns classified in PD, 20 were previously annotated in the HPA Cell Atlas v14 and 3 additional patterns were annotated

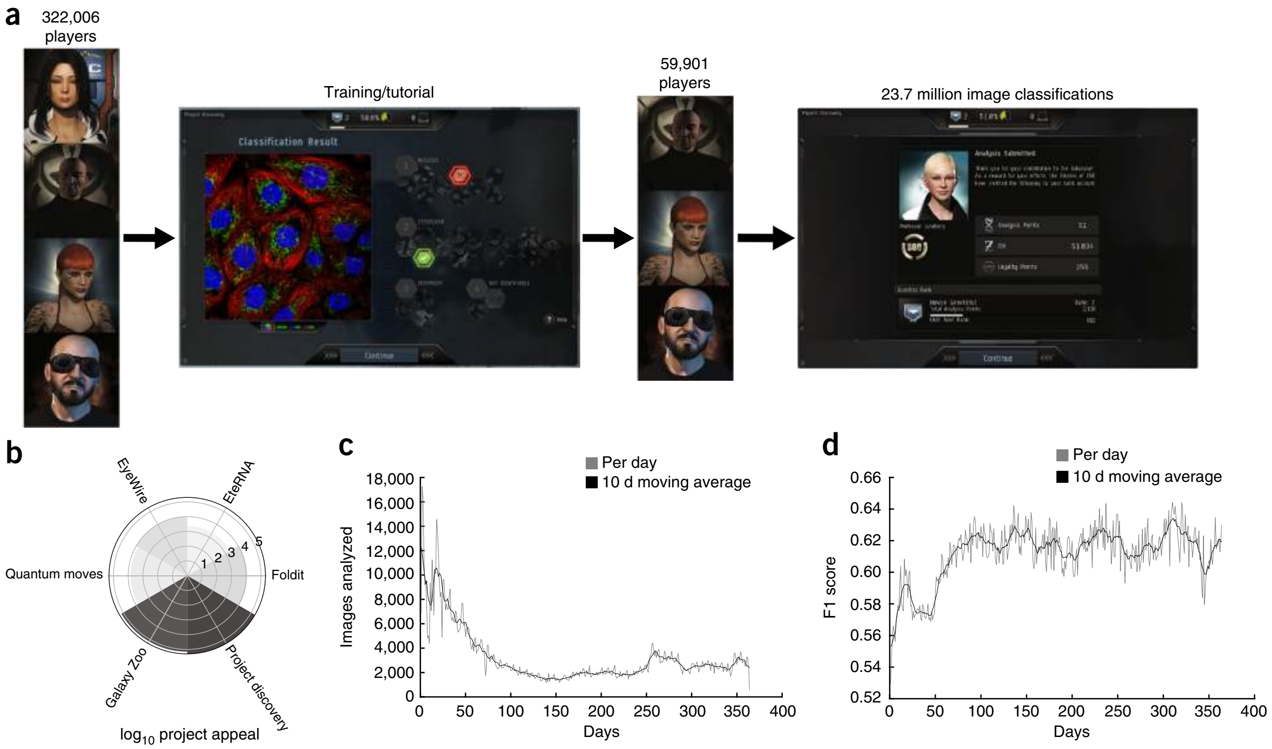


Figure 2 PD workflow. (**a**) Nearly 350,000 players in EVE Online played PD. Before contributing classifications, players were trained via an interactive tutorial with an increasing number of classes available and increasingly difficult samples. Once they passed the training phase, close to 60,000 players went on to contribute over 23.7 million image classifications. (**b**) The $\log_{10}$ project appeal, defined as (number of volunteers)/(project time^2), demonstrates the way that PD participation dwarfed previous citizen science efforts (gray indicates project). (**c**) Participation peaked on day 3, with over 17,000 images analyzed (minimum 12 players per image); however, throughput decreased over the first 150 d of the project (per day, gray; 10-d moving average, black). Participation substantially increased around day 250 of the project, when EVE Online went free to play, demonstrating that in-game mechanics can be directly used to influence participation. (**d**) Data quality also took time to stabilize, and removing rewards for community consensus agreement, together with new control samples integrated around day 50, significantly improved accuracy ($P < 4 \times 10^{-70}$, day 0–50 versus 50+, two-tailed t test).

internally while PD was active (nucleoli fibrillar center, nuclear speckles and nuclear bodies). This allowed us to examine the annotation trends of the players. After initial assessment, it was clear that some classes such as cytoplasm, nucleus and vesicles were over-annotated (**Fig. 4a**). To correct for this bias, we used the class distribution in the reference HPA Cell Atlas v14 data set to create per-class cutoffs

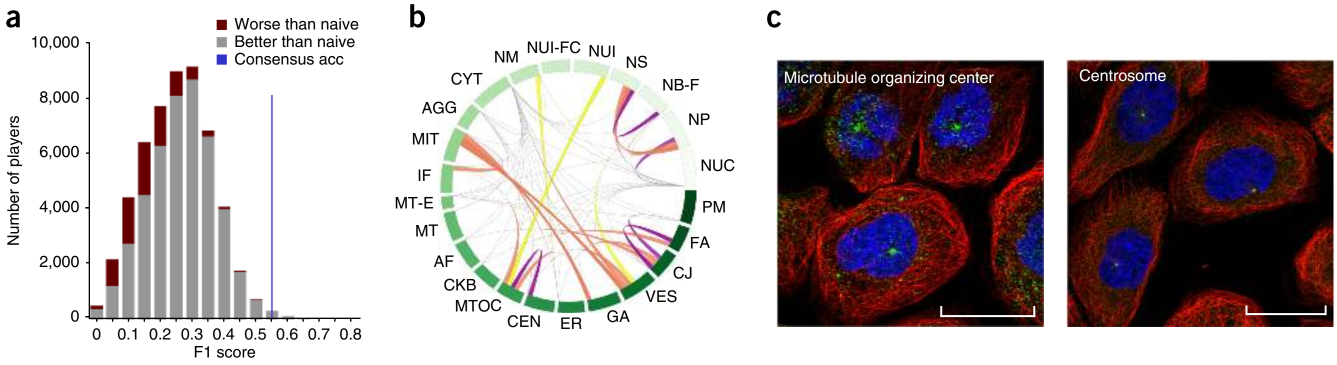


Figure 3 PD performance and confusion. (**a**) F1 score distribution of players with a minimum of ten analyzed images (n = 59,901). Despite nearly 10% of players performing worse than naively guessing the most common class (nucleoplasm, maroon bar sections), the consensus accuracy (blue line, 65,596 hypergeometric test consensuses, $P < 0.01$, $n \geq 12$ per test) remained markedly higher than the player average. (**b**) Circular plot showing player over-represented co-annotations with solution classes from the HPA Cell Atlas v14 ($P < 10^{-3}$, Bonferroni corrected one-tailed Binomial test, per-class sample sizes are found in **Supplementary Fig. 3**) serving as a proxy for multi-label confusion. Bars for each class are scaled to $\log_2$ class size and color (green) is used for visual clarity. Classes containing more than five over-represented co-annotations are considered 'uniformly over-annotated' and are shown in gray (NUC, CYT, MT-E and AGG). Colors represent the distance between classes (nodes) in the hierarchical structure, with purple, salmon and yellow representing d = 2, 4 or 6, respectively. Over-annotations were directional, with the thick end of the ribbon indicating which class was over-annotated by players (confused with) relative to the HPA Cell Atlas v14. Ribbons with two thick ends indicate a bi-directional over-representation of the co-annotation. (**c**) Centrosome (n = 1,498 images) and MTOC (n = 424 images) is an example of a bi-directional over-representation co-annotation. Scale bars represent 20 μm (inner length).

(**Supplementary Table 2**). This correction is not possible for data where the approximate proportions of classes are unknown. This approach led to a large improvement in per-class F1 score for over-annotated classes such as cytosol and nucleus, resulting in an average per-class F1 score of 0.53 (**Fig. 4b**) and an overall mean F1-score of 0.68. This includes novel classes for which we chose the most permissive cutoff to maximize discovery (recall).

Refined classifications of images in the Cell Atlas

A major contribution of the participants in PD was to refine the classifications in the Cell Atlas. Although work is still underway to incorporate all this data, version 18 of the HPA Cell Atlas includes five new categories annotated by the gamers, in total refining localization information for 2,902 proteins. Gamer annotations of classifications such as 'nucleoli fibrillar center' or 'nucleoli rim' were nearly entirely contained in their previously annotated parent class 'nucleoli' (99%), indicating that many of these annotations are indeed refining annotations in v14 (**Supplementary Fig. 3b**). Cytoophidium, or Rods and Rings (R&R), are an excellent example of a fairly uncharacterized transient cellular structure, with only three previously known protein members[36,37]. In addition to the known component protein encoded by IMPDH1, the gamers identified ten additional R&R proteins that were confirmed by colocalization analysis to localize to R&R after induction with ribavirin (**Supplementary Table 3**), such as UPF0769 protein C21orf59 (**Fig. 4c**). By expanding the set of known R&R proteins, players in PD have shed new light on this structure that may help in understanding its biological function.

Image classification with Loc-CAT using deep learning

Another approach for classification of image patterns is machine learning. Toward this end, we used a deep neural network to create Loc-CAT. Inputs for this network were previously optimized subcellular localization features (SLFs; **Supplementary Data Set 2**) calculated on segmented single cells[9,20]. As with PD, the ground truth was the labels for images in the HPA Cell Atlas v14. Predictions by Loc-CAT were made per-cell. The mean per-cell predictions were then used to create a per-image classifier on which class-specific decision boundaries were adjusted in a parameter tuning step (**Supplementary Table 4**). Along these lines, a major challenge with this approach is recognizing patterns that may occur in only a few cells in the image and discerning them from a false-positive prediction, such as the cytokinetic bridge, aggresome, centrosome and microtubule organizing center (**Fig. 5a**). Another challenge when classifying segmented cells is to recognize patterns at the cell periphery, such as plasma membrane, focal adhesions and cell junctions. Representative samples for classes with poor performance demonstrate that Loc-CAT struggles to recognize cell-to-cell variable patterns and patterns at the cell periphery (**Fig. 5b**).

Most previous methods[12–14,17,38] have controlled for biological variance by restricting classification to a single cell line. To test the robustness of Loc-CAT, we trained models on all 17 of the cell lines present in the HPA Cell Atlas v14 individually and applied each model to each cell line in turn (**Fig. 5c**). On the basis of this comparison, we can conclude that the performance of Loc-CAT was significantly higher when training on cell lines with more data ($P < 10^{-10}$, two tailed *t* test). The generalized model trained on all of the cell lines was capable of predicting subcellular localization across variations in morphology with high accuracy and performed best on nearly all of the cell lines tested.

Previous methods have also been limited to single label predictions. To test the performance of Loc-CAT in classifying single labels, multiple labels and mixed labels (data set containing both single and multi-label images), we trained models on these groups separately and applied them to each group in turn (**Fig. 5d**). Loc-CAT significantly improved localization accuracy when predicting on multi-label or mixed single and multi-label data relative to a comparable single-label-based approach ($P < 10^{-4}$, two-tailed students *t* test), indicating that this method is more generally applicable for images where the number of localizations is not known a priori.

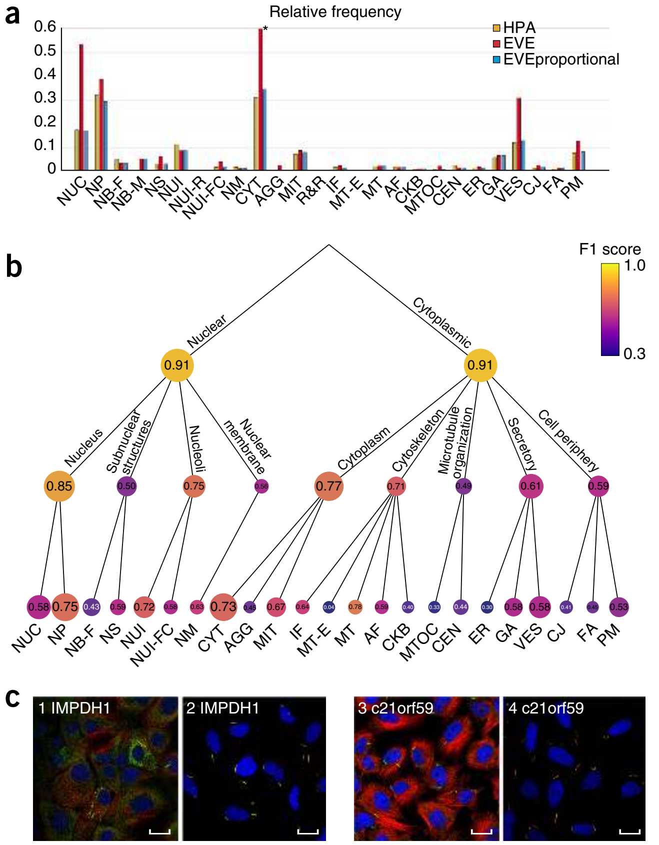
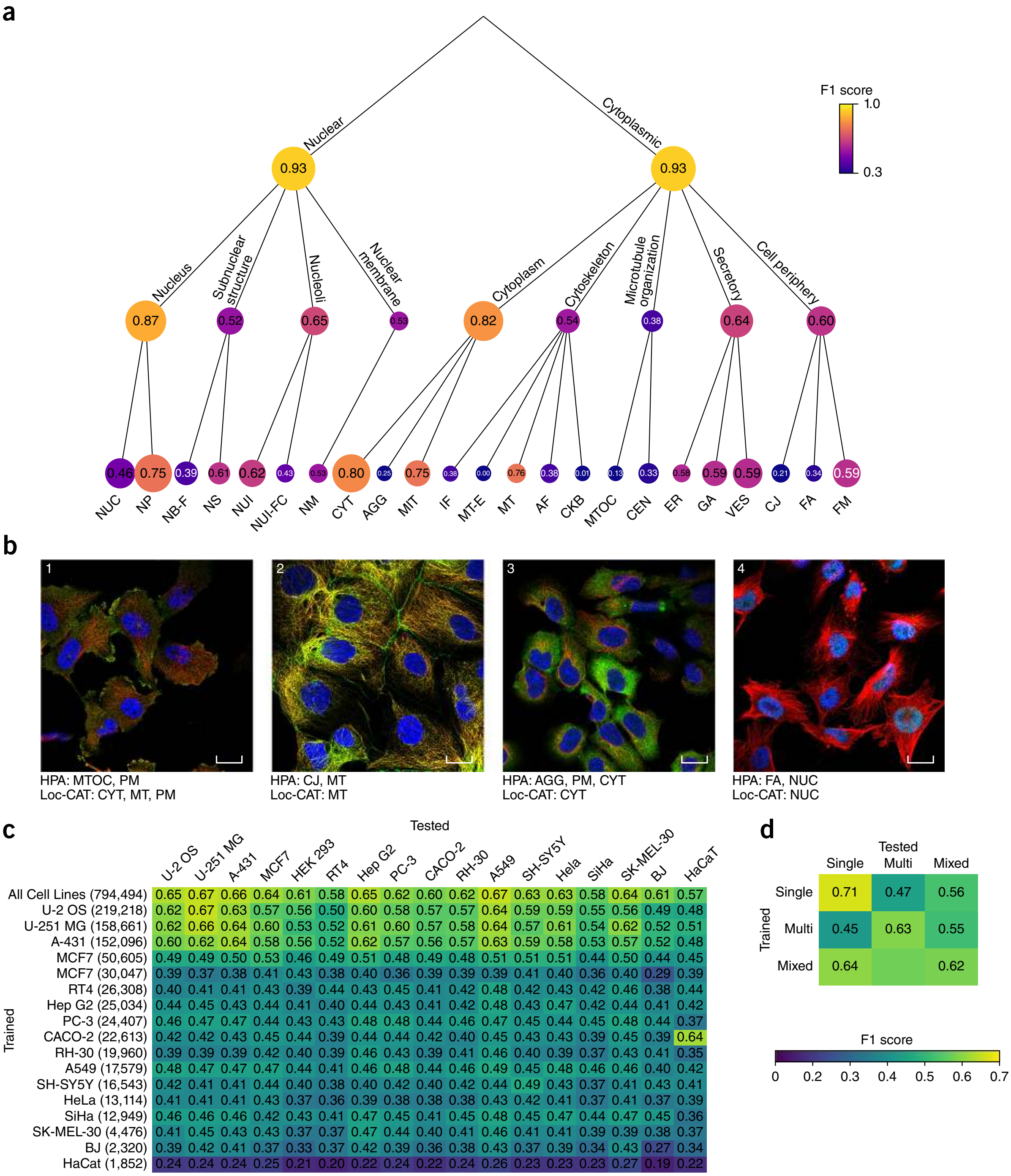


Figure 4 Correcting player bias. To improve data quality, the proportions of each class in the HPA Cell Atlas v14 were used to tune the significance levels for each class. (**a**) Bar plot showing relative proportion of data with an annotation in each class before (red) and after (blue) tuning as compared with the HPA Cell Atlas v14 (yellow). The bar for cytosol extends off the graph to a value of 0.73 before tuning (red bars). (**b**) Evaluation of classes in a tree-based hierarchy allowed evaluation of confusion between functionally and spatially similar patterns (ball size = $\log_{10}$ class size). As the depth in the tree decreased, cellular compartments were merged into meta-compartments. Votes from leaves were pooled before calculating consensus for each meta-compartment, showing player performance at different granularities. Vote pooling results in a stricter cutoff for each meta-class and, for this reason, nuclear membrane accuracy dropped in the second layer of the tree despite there being no branch. (**c**) Participants in PD shed new light on the little-known R&R structure, including correct identification of the IMPDH1 encoded protein known to localize to R&R and the novel discovery of the localization to R&R of the UPF0769 protein C21orf59. For each protein (green), localizations were verified by an independent colocalization study with a marker for R&R (red, 2, 4) after induction of R&R formation with ribavirin. Colocalization assays (2, 4) are shown beside the standard HPA Cell Atlas image (1, 3). High colocalization (yellow) confirmed that these proteins (green) were localized R&Rs (red). These experiments were performed in triplicate.

Evaluation of Loc-CAT and citizen science performance

Despite the high performance of Loc-CAT, players in PD (average per-class F1 = 0.53) outperformed Loc-CAT (average per-class

F1 = 0.47), particularly in many of the less common classes, for example, microtubule ends, which has only 32 images. Loc-CAT outperformed PD in most other classes, particularly on classes with large amounts of training data and endoplasmic reticulum (ER) where Loc-CAT has access to an additional reference channel players in PD did not (**Fig. 5a**). This makes the two methods closer in performance when comparing



Trained \ Tested	U-2 OS	U-251 MG	A-431	MCF7	HEK 293	RT4	Hep G2	PC-3	CACO-2	RH-30	A549	SH-SY5Y	Hela	SiHa	SK-MEL-30	BJ	HaCaT
All Cell Lines (794,494)	0.65	0.67	0.66	0.64	0.61	0.58	0.65	0.62	0.60	0.62	0.67	0.63	0.63	0.58	0.64	0.61	0.57
U-2 OS (219,218)	0.62	0.67	0.63	0.57	0.56	0.50	0.60	0.58	0.57	0.57	0.64	0.59	0.59	0.55	0.56	0.49	0.48
U-251 MG (158,661)	0.62	0.66	0.64	0.60	0.53	0.52	0.61	0.60	0.57	0.58	0.64	0.57	0.61	0.54	0.62	0.52	0.51
A-431 (152,096)	0.60	0.62	0.64	0.58	0.56	0.52	0.62	0.57	0.56	0.57	0.63	0.59	0.58	0.53	0.57	0.52	0.48
MCF7 (50,605)	0.49	0.49	0.50	0.53	0.46	0.49	0.51	0.48	0.49	0.48	0.51	0.51	0.51	0.44	0.50	0.44	0.45
MCF7 (30,047)	0.39	0.37	0.38	0.41	0.43	0.38	0.40	0.36	0.39	0.39	0.39	0.41	0.40	0.36	0.40	0.29	0.39
RT4 (26,308)	0.40	0.41	0.41	0.43	0.39	0.44	0.43	0.45	0.41	0.42	0.48	0.42	0.43	0.42	0.46	0.38	0.44
Hep G2 (25,034)	0.44	0.45	0.43	0.44	0.41	0.40	0.44	0.43	0.41	0.42	0.48	0.43	0.47	0.42	0.44	0.41	0.42
PC-3 (24,407)	0.46	0.47	0.47	0.44	0.43	0.43	0.48	0.48	0.44	0.46	0.47	0.45	0.44	0.45	0.48	0.44	0.37
CACO-2 (22,613)	0.42	0.42	0.43	0.45	0.44	0.39	0.44	0.44	0.42	0.40	0.45	0.43	0.43	0.39	0.45	0.39	0.64
RH-30 (19,960)	0.39	0.39	0.39	0.42	0.40	0.39	0.46	0.43	0.39	0.41	0.46	0.40	0.39	0.37	0.43	0.41	0.35
A549 (17,579)	0.48	0.47	0.47	0.47	0.44	0.41	0.46	0.48	0.44	0.46	0.49	0.45	0.48	0.46	0.46	0.40	0.42
SH-SY5Y (16,543)	0.42	0.41	0.41	0.44	0.40	0.38	0.40	0.42	0.40	0.42	0.44	0.49	0.43	0.37	0.41	0.43	0.41
HeLa (13,114)	0.41	0.41	0.41	0.43	0.37	0.36	0.39	0.38	0.38	0.38	0.43	0.42	0.41	0.37	0.41	0.37	0.39
SiHa (12,949)	0.46	0.46	0.46	0.42	0.43	0.41	0.47	0.45	0.41	0.45	0.48	0.45	0.47	0.44	0.47	0.45	0.36
SK-MEL-30 (4,476)	0.41	0.45	0.43	0.43	0.37	0.37	0.47	0.44	0.40	0.41	0.46	0.41	0.41	0.39	0.39	0.38	0.37
BJ (2,320)	0.39	0.42	0.41	0.37	0.33	0.37	0.42	0.39	0.36	0.38	0.43	0.37	0.39	0.34	0.43	0.27	0.34
HaCat (1,852)	0.24	0.24	0.24	0.25	0.21	0.20	0.22	0.24	0.22	0.24	0.26	0.23	0.23	0.23	0.27	0.19	0.22

Trained \ Tested	Single	Multi	Mixed
Single	0.71	0.47	0.56
Multi	0.45	0.63	0.55
Mixed	0.64		0.62

Figure 5 Loc-CAT DNN performance. (**a**) Loc-CAT, a DNN trained on HPA Cell Atlas v14 data, performed similarly to the participants in PD when viewing per-class F1 scores in a hierarchical tree format. Here, each level of the tree represents an independently trained version of the Loc-CAT DNN (ball size = $\log_{10}$ class size). (**b**) Worst-case images for classes in which Loc-CAT performed poorly are shown with the HPA annotation and Loc-CAT DNN annotation (images selected from false-negative rank list, lowest confidence in class of interest). Scale bars represent 20 µm (inner length). (**c**) Cross cell-line performance (average F1, fivefold cross validation) of Loc-CAT for the 17 cell types present in the HPA Cell Atlas v14. Number of cells used in training are shown in parentheses for each data set. (**d**) Loc-CAT networks trained (*y* axis) versus tested (*x* axis) on single-label (*n* = 55,083 cells, 5,234 images), multi-label (*n* = 58,133 cells, 5,560 images) and a mixture of single and multi-label data (*n* = 57,320 cells, 5,416 images, 51% multi-label) using U-2 OS cells from the HPA Cell Atlas v14 demonstrate that a model trained on a mixture of single and multi-localizing proteins generalizes best to novel data in which the number of locations a protein is present in are not known a priori.

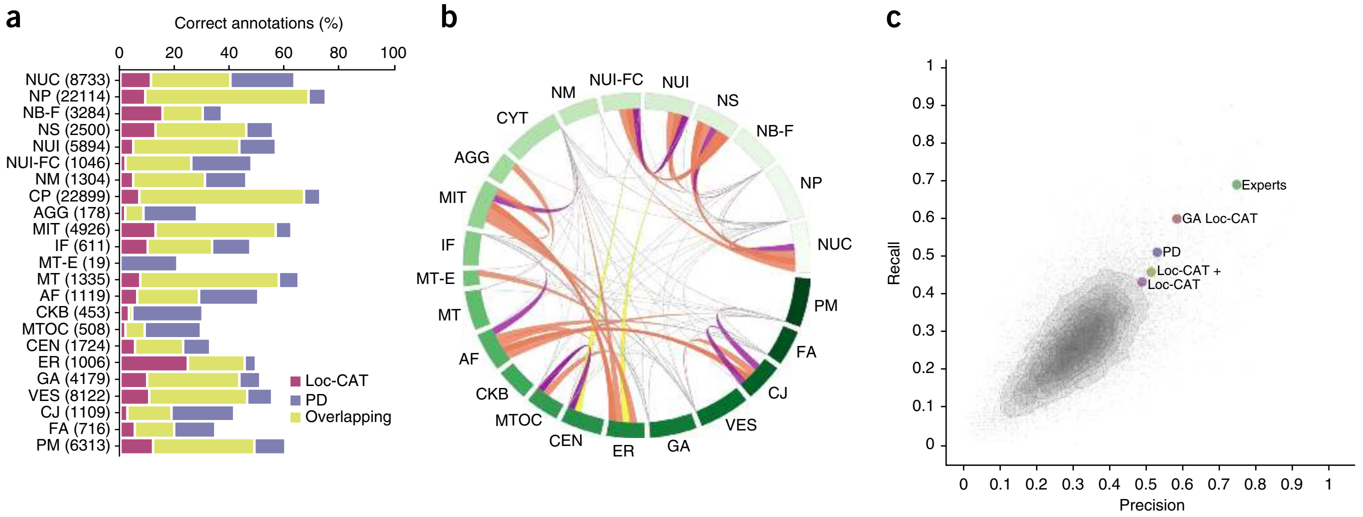


Figure 6 Transfer learning boosts Loc-CAT DNN performance. True positive percentages (recall) for each class identified by Loc-CAT (pink), PD (purple) and both (overlapping, yellow). (**b**) Loc-CAT over-represented co-annotations with solution classes from the HPA Cell Atlas v14 ($P < 10^{-3}$, Bonferroni corrected one-tailed Binomial test, per-class sample sizes are found in **Supplementary Fig. 5**) serves as a proxy for multi-label confusion. Bars for each class are scaled to $\log_2$ class size and color (green) is used for visual clarity. Classes containing more than five over-represented co-annotations were considered to be 'uniformly over-annotated' and are shown in gray. Given that over-annotations are directional, tapering was used to indicate directional confusion. Colors represent the distance between classes (nodes) in the hierarchical structure with purple, salmon and yellow representing $d = 2$, 4 or 6 respectively. Over-annotations were directional, with the thick end of the ribbon indicating which class was over-annotated by Loc-CAT relative to the HPA Cell Atlas v14. Two thick ends indicate a bi-directional over-representation. (**c**) Average per-class precision versus recall plot on HPA Cell Atlas v14 (all cell lines). PD bias corrected consensus (purple) compared with players in PD (gray points, contours). Loc-CAT (pink) performance was dragged down by low-frequency classes; however, transfer learning using both PD player input and computational features (GA Loc-CAT, red) outperformed both Loc-CAT and the PD results. Generation of 'pseudo-gamer' data for use in the transfer learner when player data was not available improved Loc-CAT (Loc-CAT+, gold); however, experts in the HPA Cell Atlas (experts, green) still vastly outperformed all of the other methods.

overall F1 score (Loc-CAT = 0.65, PD = 0.68). PD continued to outperform Loc-CAT when examining the middle layer of resolution in the organelle hierarchy (**Figs. 4b** and **5a**). Notably, gamers appeared to be more accurate at identifying nucleoli-related patterns and continued to outperform Loc-CAT in the cytoskeleton and microtubule organization meta-classes.

Loc-CAT and PD were also evaluated relative to previous methods for classification of localization patterns in images (**Supplementary Table 1**). A direct comparison was made by testing the proposed methods on the lower-complexity single-label data set used in other studies. Despite being trained on over twice as many classes, Loc-CAT was able to predict protein localization in these images with nearly equivalent per-class precision and recall as previous methods trained on this data set. PD was substantially better than all of the methods in per-class recall and overall precision, but struggled with some classes, lowering the per-class precision. In addition, a convolutional architecture based on SimpleNet was introduced to Loc-CAT instead of using traditional image features[39]. Although other convolutional architectures may perform well, this approach did not outperform the SLFs used in this work (**Supplementary Table 1**).

Gamer augmented transfer learning improves Loc-CAT accuracy

Although the overall accuracies of PD and Loc-CAT are relatively similar (**Figs. 4b** and **5a**), per-class true-positive overlap revealed that correctly annotated images varied widely (**Fig. 6a**). This suggests that labels generated in PD represent a substantial amount of per-image information in addition to the five novel classes. To leverage this information, we applied a transfer-learning approach in which we fed gamer annotations as a set of additional input features to Loc-CAT, resulting in increased performance (GA Loc-CAT; **Fig. 6c**). Because we will not have gamer input for all future tasks, we extended this approach by training a shallow 'pseudo-gamer' network (**Supplementary Fig. 4**). The resulting pseudo-gamer predictions were then fed into the Loc-CAT DNN as additional input features. This combined network, henceforth referred to as Loc-CAT+ (**Fig. 6c** and **Supplementary Fig. 4**), displays many of the same overrepresented co-annotations (**Fig. 6b** and **Supplementary Fig. 5**) as players in PD (**Fig. 3c**). Notably, however, overrepresented co-annotations between major compartments (**Figs. 3c** and **6b**) differed between the two approaches. For example, Loc-CAT+ annotates endoplasmic reticulum together with nucleoli more frequently than expected, a behavior that was not seen by the gamers. Nevertheless, the Loc-CAT+ model allowed us to incorporate some of the insights of the gamers, improving the performance of Loc-CAT by raising the average per-class F1 score from 0.44 to 0.47. However, experts in the HPA Cell Atlas (**Fig. 6c**) still outperformed all of the methods in a randomized blind annotation test (per-class F1: 0.71, overall F1: 0.76; **Supplementary Data Set 3**), suggesting that there is room for further improvement in computational image classification.

DISCUSSION

This work presents two complementary approaches to high-throughput classification of subcellular localizations in fluorescent microscope images from the HPA Cell Atlas. Multi-localizing proteins, large class imbalance, cell line variations and rare patterns that may not be present in all of the cells in an image make annotation of this dataset challenging.

The first approach uses the power of MMO games through the PD mini-game in EVE Online to perform large-scale image classification. This is the first implementation of a scientific task into an existing mainstream video game. This approach reduces development costs to labs for citizen science and demonstrates that players in MMO games can produce high-quality data despite potentially being motivated by alternative in-game dynamics or fun, rather than connection to a cause. An equivalent annotation using mechanical turk and a reward of \$0.01–0.05 per task would result in costs of \$0.33–1.65 million to obtain an equal number of annotations in addition to requiring the same effort in preparation, data management and analysis. This approach also solves issues surrounding the creation and maintenance of a user base in citizen science, as in-game rewards can be used to drive participation.

Training of players proved to be important for obtaining good results. The initial training images were too simple relative to the general population, and player performance improved significantly when more challenging training images were introduced ($P < 4 \times 10^{-70}$, day 0–50 versus 50+, two-tailed *t* test). Vote aggregation and statistics allowed us to tolerate noise in player annotations, and basic knowledge of the background distribution of classes allowed us to mitigate the effects of player bias. In future efforts, simplifying the task (for example, binary classification for the presence of a single class) may improve accuracy in a cost-effective manner, as throughput is not a large concern for this gamification paradigm. Through PD, players assisted in the refinement of annotations for thousands of samples, including several members of the largely uncharacterized R&R structure.

Participation in PD on behalf of the gaming company (CCP games) is voluntary based on their desire to promote scientific research and foster good will in their player base. This approach was highly rewarding and is promising for other massive analysis problems, with a major caveat being that the data set needs to be large, as the players were very fast. In addition to providing-high throughput image analysis, scientific outreach was a huge benefit of this method, reaching a broad community that is not necessarily invested in science. Future projects can further benefit from the development of the PD citizen science platform, even across disciplines, as exemplified by the recently launched Project Discovery Exoplanets in EVE Online.

Although PD represents one of the most successful citizen science efforts to date, it relies on continuous manual efforts of many participants administered by a third party and is therefore not sustainable for long-term generalized future use, as the gaming company may decide to take down the game. For this purpose, our DNN-based approach, Loc-CAT+, provides a promising method for annotating protein localizations in future work as it is fully automated. Loc-CAT+ represents major improvements on previous efforts, as a result of its ability to accurately classify a large number of patterns and mixtures thereof, as well as generalize across cell types with different morphologies. Thanks to this generalizability and ability to classify multi-localizing proteins, it is, to the best of our knowledge, the first automated image-based protein localization method capable of accurately classifying images where no information about the protein is known a priori. Furthermore, by augmenting the quantitative image features used in Loc-CAT with PD gamer annotations we improved Loc-CAT to nearly human performance. One major challenge in machine learning remains the recognition of rare and novel classes in which there is little or no training data. In our study, humans still clearly outperformed the algorithmic approaches. The refinement of the annotations in the HPA Cell Atlas made through PD and Loc-CAT, with the novel classification of seven additional subcellular localizations, present an exciting new resource for understanding cell biology. Although preliminary tests of convolutional neural networks in this work did not improve results over the quantitative image features used, different model architectures and hyperparameters may provide the improvements needed to reach expert performance.

To summarize, we demonstrated two alternative approaches for large-scale classification of protein distribution patterns in microscopy images. Furthermore, we showed how gamers and DNNs excel at different types of classifications and that gamer output can be used to augment and improve deep learning models. Finally, we speculate that the integration of scientific tasks into established computer games will be a commonly used approach in the future to harness the brain processing power of humans and that intricate designs of citizen science games feeding directly into machine learning models through techniques such as reinforcement learning have the power of rapidly leveraging the output of large-scale science efforts.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We acknowledge the staff of the Human Protein Atlas program for valuable contributions. We acknowledge the EVE Development team, the University of Reykjavik and the University of Iceland for assistance with the game implementation. We acknowledge MMOS Sarl for serving images and managing response collection and CCP hf and MMOS Sarl for financially supporting the image storage and serving throughout Project Discovery. Funding to E.L. was provided by the Knut and Alice Wallenberg Foundation. Finally we acknowledge all EVE Online players that participated in Project Discovery!

AUTHOR CONTRIBUTIONS

A.S., B.R., B.F., A.N. and E.L. conceived the study. M.H., A.S., B.F., E.L., D.P.S. and C.F.W. developed the methodology for the study. A.S. and B.R. developed the citizen science engine. L.C., H.L., S.R. and B.F. developed the game narrative and implementation. Project Discovery was played by thousands of players of EVE Online. D.P.S., L.Å., M.W., R.S. and E.L. provided game support. C.F.W., K.S. and D.P.S. developed the machine learning. D.P.S., C.F.W. and E.L. carried out data analysis and investigation. D.P.S., C.F.W. and E.L. wrote the manuscript. D.P.S. and C.F.W. created the figures. E.L. supervised and administered the project and acquired funding.

COMPETING INTEREST

A.S. and B.R. are founders of MMOS Sarl.

1. Bouwer, J. *et al.* Petabyte data management and automated data workflow in neuroscience: delivering data from the instruments to the researcher's fingertips. *Microsc. Microanal.* **17**, 276–277 (2011).
2. Ferrucci, D. *et al.* Building Watson: an overview of the DeepQA project. *AI Magazine* **31**, 59–79 (2010).
3. Larrañaga, P. *et al.* Machine learning in bioinformatics. *Brief. Bioinform.* **7**, 86–112 (2006).
4. Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
5. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
6. Cohn, J.P. Citizen science: can volunteers do real research? *Bioscience* **58**, 192–197 (2008).
7. Uhlen, M. *et al.* Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* **28**, 1248–1250 (2010).
8. Thul, P.J. *et al.* A subcellular map of the human proteome. *Science* **356**, eaai3321 (2017).
9. Boland, M.V. & Murphy, R.F. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics* **17**, 1213–1223 (2001).
10. Huang, K. & Murphy, R.F. Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *BMC Bioinformatics* **5**, 78 (2004).
11. Newberg, J.Y. *et al.* Automated analysis of Human Protein Atlas immunofluorescence images. *Proc. IEEE Int. Symp. Biomed. Imaging* **5193229**, 1023–1026 (2009).
12. Li, J., Newberg, J.Y., Uhlén, M., Lundberg, E. & Murphy, R.F. Automated analysis and reannotation of subcellular locations in confocal images from the Human Protein Atlas. *PLoS One* **7**, e50514 (2012).

13. Li, J., Xiong, L., Schneider, J. & Murphy, R.F. Protein subcellular location pattern classification in cellular images using latent discriminative models. *Bioinformatics* **28**, i32–i39 (2012).
14. Coelho, L.P. *et al.* Determining the subcellular location of new proteins from microscope images using local features. *Bioinformatics* **29**, 2343–2349 (2013).
15. Chebira, A. *et al.* A multiresolution approach to automated classification of protein subcellular location images. *BMC Bioinformatics* **8**, 210 (2007).
16. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
17. Pärnamaa, T. & Parts, L. Accurate classification of protein subcellular localization from high-throughput microscopy images using deep learning. *G3 (Bethesda)* **7**, 1385–1392 (2017).
18. Kraus, O.Z., Ba, J.L. & Frey, B.J. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics* **32**, i52–i59 (2016).
19. Nathalie Japkowicz, S.S. The class imbalance problem: A systematic study. *Intell. Data Anal.* **6**, 429–449 (2002).
20. Coelho, L.P., Peng, T. & Murphy, R.F. Quantifying the distribution of probes between subcellular locations using unsupervised pattern unmixing. *Bioinformatics* **26**, i7–i12 (2010).
21. Zhao, T., Velliste, M., Boland, M.V. & Murphy, R.F. Object type recognition for automated analysis of protein subcellular location. *IEEE Trans. Image Process.* **14**, 1351–1359 (2005).
22. Shen, Y.-Y.X.L.-X.Y.H.-B. Bioimage-based protein subcellular location prediction: a comprehensive review. *Front. Comput. Sci.* **12**, 26–39 (2018).
23. Khatib, F. *et al.* Algorithm discovery by protein folding game players. *Proc. Natl. Acad. Sci. USA* **108**, 18949–18953 (2011).
24. Khatib, F. *et al.* Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nat. Struct. Mol. Biol.* **18**, 1175–1177 (2011).
25. Chris, J. *et al.* Galaxy Zoo: 'Hanny's Voorwerp', a quasar light echo? *Mon. Not. R. Astron. Soc.* **399**, 129–140 (2009).
26. Clery, D. Galaxy evolution. Galaxy zoo volunteers share pain and glory of research. *Science* **333**, 173–175 (2011).
27. Raddick, M.J. *et al.* Galaxy Zoo: exploring the motivations of citizen science volunteers. *Astron. Educ. Rev.* **9**, 18 (2010).
28. Lee, J. *et al.* RNA design rules from a massive open laboratory. *Proc. Natl. Acad. Sci. USA* **111**, 2122–2127 (2014).
29. Sørensen, J.J. *et al.* Exploring the quantum speed limit with computer games. *Nature* **532**, 210–213 (2016).
30. Hughes, A. *et al.* Quantius: Generic, high-fidelity human annotation of scientific images at 10^5-clicks-per-hour. Preprint at https://www.biorxiv.org/content/early/2017/07/15/164087 (2017).
31. Danielle, N., Shapiro, J.C. & Mueller, P.A. Using mechanical turk to study clinical populations. *Clin. Pyschol. Sci.* **1**, 213–220 (2013).
32. Cox, J. *et al.* How is success defined and measured in online citizen science? A case study of Zooniverse projects. *Comput. Sci. Eng.* **17**, 28–41 (2015).
33. Feng, W., Brandt, D. & Shah, D. A long-term study of a popular MMORPG. *Proceedings of the 6th ACM SIGCOMM Workshop on Network and System Support for Games* 19–24 (2007).
34. Warfield, S.K., Zou, K.H. & Wells, W.M. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans. Med. Imaging* **23**, 903–921 (2004).
35. Snow, R., O'Connor, B., Jurafsky, D. & Ng, A. Cheap and fast, but is it good? Evaluating non-expert annotations for natural language tasks. *Conference on Empirical Methods in Natural Language Processing* 254–263 (2008).
36. Calise, S.J. *et al.* Glutamine deprivation initiates reversible assembly of mammalian rods and rings. *Cell. Mol. Life Sci.* **71**, 2963–2973 (2014).
37. Carcamo, W.C. *et al.* Induction of cytoplasmic rods and rings structures by inhibition of the CTP and GTP synthetic pathway in mammalian cells. *PLoS One* **6**, e29690 (2011).
38. Handfield, L.F., Chong, Y.T., Simmons, J., Andrews, B.J. & Moses, A.M. Unsupervised clustering of subcellular protein expression patterns in high-throughput microscopy images reveals protein complexes and functional relationships between proteins. *PLOS Comput. Biol.* **9**, e1003085 (2013).
39. Hasanpour, S., Rouhani, M., Fayyaz, M. & Sabokrou, M. Lets keep it simple, Using simple architectures to outperform deeper and more complex architectures. Preprint at https://arxiv.org/abs/1608.06037 (2016).

ONLINE METHODS

Images from the HPA Cell Atlas. In this article, we are classifying protein distribution patterns from the publically available Human Protein Atlas (HPA), Cell Atlas database (https://www.proteinatlas.org). The HPA Cell Atlas project aims to characterize the subcellular distribution patterns of the entire human proteome using an antibody-based approach and confocal microscopy. Here, we have used the images and annotations from v14 of the HPA Cell Atlas, where proteins were classified into one or more of 20 organelles and cellular structures (in total 226,732 images of which 65,596 were public in v14).

Proteins are cataloged serially using in-house generated antibodies and immunostaining in a gene-centric manner as described in detail previously[7]. Briefly, the spatial distribution of each protein is studied in three cell lines out of a panel of 17; U-2 OS and two additional selected to have the highest RNA expression level of the corresponding gene. Each antibody-cell line 'sample' is then imaged to produce a minimum of two images per sample (average 2.93 images per sample). Each 'image' in the HPA Cell Atlas consists of four channels acquired sequentially with a Leica SP5 confocal microscope (DM6000CS) equipped with a 63× HCX PL APO 1.40 oil CS objective (Leica Microsystems). The settings for each image were as followed: Pinhole 1 Airy unit, 16bit acquisition and a pixel size of 0.08 μm. The detector gain measuring the signal of each antibody was adjusted to a maximum of 800 V to avoid strong background noise. A small part of the plates was imaged automatically using the MatrixScreener M3 in LAS AF software (Leica Microsystem). Here, z-stacks at six FOVs were acquired. False-colored channels represent the protein of interest (green), DAPI labeling of the nucleus (blue), microtubules (red), and the endoplasmic reticulum (yellow). Each channel is stored as a separate 2,048 × 2,048 16-bit ome-tiff.

Additional information on the experimental materials and reproducibility can be found in the Life Sciences Reporting Summary.

Tree structured annotations. Classes in the HPA Cell Atlas can be viewed as a tree structure, where depth in the tree increases, annotations become more specific. At its base, the cell is divided into the nuclear and cytoplasmic spaces. These two super-classes can be further divided into meta-classes. The nuclear super-class into; nucleus, subnuclear structures, nucleoli, and nuclear membrane. The cytoplasmic super-class into; cytoplasm, cytoskeleton, MTOC, secretory, and cell periphery. Lastly, these meta classes can be divided into the leaf node classes used in this publication. When discussing this structure in terms of PD, votes are first pooled and a hypergeometric test is performed to calculate consensus as described below at each level of the tree. As there are fewer options to choose from, nodes near the root of the tree require more evidence to be considered significant (hypergeometric test $P < 0.01$). When discussing this structure in terms of the DNN approach using the localization cellular annotation tool (Loc-CAT), each level of the tree represents a separately trained model.

Immunostaining after induction of R&R formation. U-2 OS cells were cultivated in McCoy's 5A modified medium (Sigma Aldrich) with 10% FBS and 1% L-glutamine (Sigma Aldrich), at 37 °C in a 5% CO_2 humidified environment. The cells were harvested at 60–70% confluency and seeded onto a glass bottom plate (Greiner Sensoplate Plus, Cat# 655892, Greiner Bio-One) coated with fibronectin (Sigma-Aldrich). 6 h before fixation Ribavirin was added to the growth medium to a final concentration of 0.15 mM. PBS-washed cells were fixed in 4% paraformaldehyde (PFA) in growth media supplemented with 10% FBS for 15 min, followed by permeabilization with 0.1% Triton X-100 in PBS for 3 × 5 min. After a washing step with PBS, cells were incubated with the primary antibody overnight at 4 °C. Rabbit polyclonal HPA antibodies were diluted to 2–4 μg/ml in blocking buffer (PBS with 4% FBS) containing the R&R marker (Abnova Corporation Cat#H00055466-M01, RRID:AB_426011) diluted to 1 μg/ml in blocking buffer. The next day, cells were washed 4 × 10 min with PBS followed by 90-min incubation at 20–22 °C with the following secondary antibodies (all from ThermoFisher Scientific) diluted to 1 μg/ml in blocking buffer: goat anti-rabbit AlexaFluor 488 (A11034, RRID: AB_2576217), goat anti-mouse AlexaFluor 555 (A21424, RRID:AB_2535845). Cells were finally counterstained with DAPI for 10 min, before being mounted in PBS containing 78% glycerol.

Ground truth for evaluation. The training labels for evaluating methods presented in this work are based on three rounds of manual curation performed for the HPA Cell Atlas v14 (**Supplementary Data Set 1**). Images were first annotated manually by a trained expert and labels were given based on all images in a sample. This was followed by a review in which stainings from multiple cell types were compared and consistency of staining was assessed. Lastly, a thorough literature review was performed. Annotations were corrected as needed throughout this process.

Expert reannotation. To assess consistency of labels in the HPA Cell Atlas, internal experts were presented with a random subset of 660 samples for reannotation from samples that were publicly available in the HPA Cell Atlas v14 (**Supplementary Data Set 3**). All reannotations and statistics measuring the accuracy of these reannotations were calculated at the per-sample (group of images) level. This gives an advantage to experts over the other methods in this work as some images in a sample do not contain the annotated label. For historical reasons, this reannotation did not include a distinction between Nucleus and Nucleoplasm, so the expert score is inflated slightly. Assuming the expert performance on the Nucleus-Nucleoplasm split was comparable to Loc-CAT and gamers, the expert per-class precision and recall would drop from 0.74 and 0.69 to ~0.70 and ~0.66 respectively, still well above all other methods. Microtubule ends were not present in the reannotation set as it is very rare. As a proxy, average performance was assumed for this class. This may be generous considering both gamers and Loc-CAT struggled with this class. In the worst case, were experts to entirely miss this class, per-class precision and recall would drop to ~0.69 (~0.67 with nucleoplasm) and 0.63 (~0.62 with nucleoplasm).

Statistics and reproducibility. In this work we used several metrics to measure the performance of both PD participants and the Loc-CAT DNNs, hence forth referred to as 'predictors'.

Assessing performance. To assess the agreement between generated predictor annotations and HPA Cell Atlas v14 annotations, we assessed precision and recall as defined in equations (1) and (2) below.

$$Precision = \frac{TP}{(TP+FP)} \tag{1}$$

$$Recall = \frac{TP}{(TP+FN)} \tag{2}$$

Here, true positives (TP) are annotations for which the predicted label matches the prior HPA label, false positives (FP) are predicted labels that the HPA has not identified, and false negatives (FN) are labels that the HPA had annotated which the predictor did not predict. Again, note that in cases of labels which are novel to the HPA, such as nucleoli (rim), the FP = 1 and FN = 0 by definition as the HPA has never previously annotated this localization.

In the case of multi-label data, accuracy cannot directly be assessed, as label confusion cannot be readily defined. To measure per-class performance we used F1 score which is the harmonic mean between precision and recall.

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{3}$$

Measuring cross-localization confusion. Due to the multi-label nature of the problem, it is impossible to construct a confusion matrix indicating what labels predictors select in comparison to those annotated in the HPA Cell Atlas v14. In an attempt to understand confusion, we compute a matrix indicating the probability that the frequency of specific multi-localizations occurs based on HPA Cell Atlas v14 colocalization probabilities. In doing this, we compare the probability of observing location B_{HPA} given location A_{HPA} as defined by the HPA Cell Atlas v14 annotations ($P(B_{\text{HPA}}|A_{\text{HPA}})$) with the frequency of prediction for the localization $\hat{B}$ given A_{HPA} using a one-tailed binomial test as indicated in equation (4) below. Note that this test only measures over co-annotation.

$$P_{\text{binomial}}\left(F(\hat{B}\,|\,A_{HPA})\middle|F(\hat{B}), P(B_{HPA}\,|\,A_{HPA})\right) \tag{4}$$

Where $F(\hat{B}\,|\,A_{HPA})$ is the frequency of predicted location $\hat{B}$ given an observed label A_{HPA}, and $F(\hat{B})$ is the number of all predicted localizations $\hat{B}$ independent of corresponding HPA annotation. The test is used to measure over co-annotation in the multilabel case, and will be significant if the predictor annotates one category significantly more frequently than we expect given

doi:10.1038/nbt.4225

the co-annotation probabilities by the HPA Cell Atlas v14. This can occur either via confusion, where one label is incorrectly identified as another with regularity, or general over-annotation where a predictor is biased to more frequent annotation of a given label. Note that the test is never significant on the diagonal where $P(B_{HPA}|A_{HPA}) = 1$. The resulting p-values were then subjected to a Bonferroni multiple hypothesis correction per-class (n = 29). These results are presented as a circular plot serving as a proxy for multi-label confusion (**Figs. 3** and **6**). As over-annotations are directional, tapering is used to indicate directional confusion, with the thick end of the ribbon indicating which class is over-annotated by the predictor (confused with) together with the HPA Cell Atlas v14. Ribbons with two thick ends indicate a bi-directional over-representation of the co-annotation. This can also be viewed in tabular form (**Supplementary Figs. 3** and **5**).

Reproducibility. All images in the HPA Cell Atlas v14 (n = 226,732 images of which 65,596 were public in v14) consist of 4 false-colored channels as shown in **Figure 1a**. The number of images containing each of the patterns in the **Figure 1b–d** can be found in **Supplementary Figure 3**. The authors note that this is not the number of images containing only this pattern as multi-localization of proteins causes more than one pattern per image. All 10,003 protein coding genes publicly available in HPA Cell Atlas v14 are assayed in three cell types; U-2 OS and two selected based on maximal RNA expression (FPKM/TPM) creating >10,003 such morphological variability replicates in HPA Cell Atlas v14 such as the examples seen in **Figure 1e**. The numbers for each specific cell type can be seen in **Figure 5c**. Of the proteins analyzed in this study, 44% (n = 101,903) of images (not proteins) contain multi-localizing proteins such as those shown in **Figure 1f**. Cell-to-cell variability (**Fig. 1g**) was a new category in v14 and therefore contained no true-positive images. In the updated Cell Atlas v16 containing this cell-to-cell variability analysis 1,896 protein coding genes (most with 6+ images per protein coding gene) are annotated as having variable patterns.

Players in PD contributed ~33 million image annotations. Annotations for each image are pooled into a consensus using a hypergeometric test (minimum 12 votes, see Online Methods). Results of these consensus annotations each day are compared with gold standard HPA Cell Atlas v14 to obtain an F1 score per day which demonstrates a stable behavior after 100 d (**Fig. 2d**).

Comparing the overall F1 score of players (n = 59,901) who have analyzed a minimum of ten images, with a consensus built on hypergeometric tests based on the cumulative consensus (pooling individual votes from all rounds, median 78 votes per image) demonstrates the power of pooling multiple votes (**Fig. 3a**). Over-represented co-annotations are measured using a set of pair-wise binomial tests where the null hypothesis is the expected co-localization probability in the HPA Cell Atlas (**Fig. 3b**). Replicate numbers of images containing proteins annotated to each class under both the HPA Cell Atlas, and PD can be found in **Supplementary Figure 3**. There were 1,498 images annotated centrosome and 424 images annotated MTOC in the HPA Cell Atlas (**Fig. 3c**).

Histograms showing the counts of images annotated for each class are shown in **Figure 4a**. These numbers are either directly counted from the HPA Cell Atlas v14 data set (gold), or based on the 65,596 public images in the HPA Cell Atls v14, where a hypergeometric test is performed on the pooled annotations for each image (median n = 78 annotations per image). Performances (F1 score) in the tree based hierarchy (**Fig. 4b**) are based on hypergeometric consensus for each image. The number of class instances in HPA Cell Atlas v14 can be found in **Supplementary Figure 3**, rows labels. Example images of Rods & Rings proteins are shown based on the ten proteins discovered by players of PD. Independent colocalization experiments under ribavirin induction with a marker for R&R for each protein were performed in triplicate (see Online Methods).

Performances in the tree based hierarchy (**Fig. 5a**) are based on the number of class instances in HPA Cell Atlas v14 (**Supplementary Fig. 5**). Example images (**Fig. 5b**) are the worst-case picked from a rank-list of each of the lowest performing classes from the hierarchical tree (**Fig. 5a**). These images are meant for illustrative purposes of the types of mistakes Loc-CAT makes in the worst case. Performance on each cell line (**Fig. 5c**) and compared across single and multi-label data (**Fig. 5d**) is based on the average of fivefold cross validation.

Overlap in Loc-CAT predictions and PD predictions (**Fig. 6a**) are based on the number of class instances in HPA Cell Atlas v14 (**Supplementary Fig. 5**). Over-represented co-annotations are measured using a set of pair-wise binomial tests where the null hypothesis is the expected co-localization probability in the HPA Cell Atlas (**Fig. 6b**). Replicate numbers of images containing proteins annotated to each class under both the HPA Cell Atlas, and Project Discovery can be found in **Supplementary Figure 5**. Individual player performance (**Fig. 6c**), compared to consensus performance of Project Discovery, Loc-CAT performance using various training architectures (Loc-CAT, Loc-CAT+, GA Loc-CAT), and expert annotations are based on the 65,596 images in the HPA Cell Atlas v14. Scores are computed per-class, where true-positive instances of each class can be seen in **Supplementary Figure 5**.

Experimental reproducibility. Additional information on the experimental materials and reproducibility can be found in the Life Sciences Reporting Summary.

PD: MMO science. This work presented a new approach to citizen science and gamification. Termed Project Discovery, this method is the first to utilize main-stream MMO games to perform real scientific research. This effort was a collaboration with CCP Games (EVE Online) and MMOS.

Image preparation. Images were converted from 2,048 × 2,048 16-bit greyscale tiff images to RGB false color 1,200 × 1,200 jpeg images with 89% compression. The resulting images were then given randomized names and uploaded to MMOS Amazon Web Servers. This configuration was chosen to limit server load as each color channel could be directly dropped on the EVE client side after the image was served. For this reason, the players did not receive a color channel for the endoplasmic reticulum (ER). This also limited the ability of colorblind players to participate, though we suggested that such players use a 'shader' to shift the screen into a colorblind-friendly palate.

For each batch of images, a tab separated plain text metadata file was generated and uploaded to an MMOS Amazon Web Server. Each row of a metadata file represented one image in the batch and the columns of each metadata file were used to provide information about the image. In addition, a json formatted 'control' file was generated for each batch specifying information about the batch including the number of images, and version number.

Game play. The mini-game within the EVE Online universe was accessible from anywhere in-game allowing maximum access for players. The game design was created by CCP games and students at Reykjavik University.

In the game, players were presented a false-color confocal microscopy image and are tasked with classifying the green pattern into up to 5 of the 29 pre-defined categories. Players could use the blue (nucleus) and red (microtubules) channels to assist them and could toggle these color channels on and off as well as zoom in on the image by hovering. Players could compare the patterns seen in each image with five reference images of each pattern visible upon hovering over each tool-tip in-game. These images were carefully selected to represent the diversity of the respective staining patterns across the multitude of cell lines.

After submitting a classification, players received an in-game reward in the form of in-game currencies that could be used to purchase in-game items exclusive to PD as well as level-badges. Players received one small reward per-sample analyzed, plus a larger reward for each time they leveled up. Initially players also received a bonus reward based on their agreement with the eventual community consensus, however this was quickly exploited with players converging on a single common class (Cytoplasm) and this reward was therefore discarded. Players were also provided with a 'pass' option after expressing that some images were too challenging and they would rather pass than make a bad guess.

In an additional attempt to control accuracy, control samples in which the solutions were known a priori were provided at random intervals. If player performance drops too low, the player is returned to the tutorial phase.

To view a tutorial of game play, please visit our youtube channel (https://www.youtube.com/channel/UCfUAILRafjldAom5lzSQD7A/videos?view_as=subscriber).

Tutorial and training. To control data quality, players were required to complete a tutorial and training phase before contributing to classifications of unknown samples in PD. Players were entered into the tutorial and first asked to classify images of easy, single-localizing protein to a restricted set of localizations to familiarize themselves with the user interface. Once past the tutorial, players entered training, where players were presented with increasingly difficult samples and were required to correctly annotate these before passing the training phase and being allowed to contribute annotations for unknown samples to the project. Player accuracy was measured with random control samples which were identical to test samples but had been pre-annotated by experts from the HPA Cell Atlas. If player performance dropped below a threshold, players were returned to the training phase of the game until their performance improved to a level that they were allowed to contribute to the consensus again.

Consensus calculation. Tasks were presented to gamers in a randomized order. To control for erroneous annotations, we asked multiple gamers to annotate each image. A minimum of twelve gamers assessed each image before it could be evaluated for consensus. We measured 'consensus' on a task using a hypergeometric test as described in equation (5) below assuming that each player chose the maximum of five classes per task. This test assumes each class is independent and as such it does not account for mutual exclusivity of tasks (for example, Nucleoplasm with Nucleoli). The test is given by

$$P=1-\sum_{i=1}^{m}\frac{\binom{p}{i}\binom{N*n-n}{m*n-i}}{\binom{N*n}{m*n}}=1-\sum_{i=1}^{5}\frac{\binom{n}{i}\binom{28*n}{5*n-i}}{\binom{29*n}{5*n}} \qquad (5)$$

where, n is the number of players that have voted on the task, N is the number of classes available (29), m is the maximum number of allowed classes per sample (5). This equation gives an estimate of the probability that each category had been selected m times given n tries (gamers). Once 12 votes were acquired, this CDF was evaluated after each subsequent vote. If the likelihood of at least one category was statistically significant ($P < 0.01$), and no other categories were near the significance boundary ($0.01< P < 0.1$) we considered a consensus reached and the task was closed. The hypergeometric test measures the probability of obtaining k 'hits' in n random draws from a set without replacement. This test is also extremely efficient to compute, making it feasible for the real-time computation with high server loads experienced of over 800 submissions per minute.

After six rounds of annotations, votes from each round were aggregated and the consensus recalculated (average 97 votes per image). As statistical significance was increased due to the increased number of samples, this created a more sensitive test for rare and under annotated classes, however it also exacerbated the over-annotation problem for common classes. To correct for this effect p-value cutoffs were tuned per class on a held out 10% of the data of the based on the expected class distribution from the previously annotated HPA data. Novel classes were set to the highest allowable p-value cutoff (0.01) for discovery.

When constructing consensuses for meta classes, votes were merged into super categories, and then re-evaluated using the hypergeometric test and the aforementioned procedure given the presence of fewer classes.

Expectation maximization. Jointly estimating individual player bias together with the true label can be done via expectation maximization (EM). In this work, we implemented a binary EM for each class based on the STAPLE method[39], as multi-label data makes direct multi-class evaluation impossible. Due to the computational time required, we ran the algorithm for 10–30% of the data set (n = 6,558–19,534) respectively. We observed no improvement when increasing the percentage of the data set evaluated and report the best accuracy of all runs (**Supplementary Table 1**). Unfortunately, as the number of single labels is very high (29), the frequency of most labels is very low (<0.1%), and the number of images analyzed per player is relatively low (mean = 44), this method did not improve results and the previously discussed consensus calculation was used instead.

Project appeal. Project appeal (**Fig. 2b**) was calculated as defined in[32] and given in equation (6) below.

$$Project\ Appeal=\frac{Number\ of\ volunteers}{(Project\ active\ period)^2} \qquad (6)$$

Loc-CAT: DNN protein localization. This work presented a feature based multi-label DNN model for predicting subcellular protein localization. This network outputs a real-valued confidence vector with a score for each possible class.

Image feature extraction. Quantitative image features were calculated using MATLAB 2016a. Image processing was performed at a per-cell level on each image consisting of four fluorescent microscopy images, one for each acquisition channel. The DAPI images were first treated with a low pass filter followed by active contour segmentation. Cells were then segmented using a combination of the microtubule and ER channels and seeded watershed. Cells with nuclei touching the image edge are removed from classification, though the cytoplasm of the cell can contact the edge of the image.

After segmentation, a set of 2,233 quantitative SLF image features were extracted based on work by the Murphy Lab[9,20]. Of these, the 719 features describing the green fluorescent channel in relation to the other channels, were passed to Loc-CAT (https://github.com/CellProfiling/Loc-CAT). These features describe the intensity, texture, and spatial relationships between the protein of interest (green) and remaining fluorescent channels of the image. The remaining unused features describe the relationship of reference channels to themselves and are used internally in other applications. Details on each feature can be found at (http://murphylab.web.cmu.edu/services/SLF/features.html, **Supplementary Data Set 2**).

Data partitioning. Images are shuffled at the sample level, meaning though images of the same antibody in another cell line may be present in the test set (for cross-cell line classification), all images of an antibody in a specific cell type will be in a single fold. Images are then partitioned into five folds by sample. Each training set contains 80% and each testing set contains 20% of the available data. Training sets are then split again by sample into training (90% of training set) and validation (10% of training set) sets. The resulting training set was then shuffled per-cell to avoid bias in the training. Because folds were shuffled created per-sample, it is possible that each fold contains variable number of images and cells. After data partitioning, input features to Loc-CAT are Z-normalized on the training set (excluding the validation subset).

Neural network architecture. In Loc-CAT, CPython 3.5.2 with CUDA gpu-accelerated TensorFlow (v1.3.0) was used to train a feed-forward deep artificial neural network containing three hidden ReLU6 layers with 800 neurons per layer and a sigmoid output function. Dropout was applied to reduce the risk of overfitting and better generalize the network, 20% on the input layer and 40% on each hidden layer. The network was optimized using the ADAM optimizer and a binary-cross entropy loss function. In the development of Loc-CAT, several network architectures were tested using a multidimensional parameter sweep (trained and tested on U-2 OS images from HPA Cell Atlas v.14).

Stopping rule: during training, if the cost on the held-out validation set did not decrease for ten epochs, training was halted. The network weights were then reset to the epoch before those ten epochs.

Prediction aggregation. As the quantitative image features are extracted per cell, the classifier predicts localization for individual cells rather than images. When training the network, binary cross entropy was applied to these per-cell annotations, using the HPA annotation for the image the cells came from as the true label. Location predictions were aggregated for all cells from the same image by taking the mean predicted value for each class. The cutoffs for each class are then tuned at the IMAGE level for the first fold of the testing set to optimize the per-class performance. Average performance for each of the remaining four test-set accuracies are reported using these cutoffs.

All single-cell line classifiers reported the average statistics of fivefold cross validation. Cross-cell line classification statistics are based on predictions for all samples in the testing cell type and therefore cross validation does not apply. In the hierarchical tree (**Fig. 5**), each level of the tree was trained separately and tested using fivefold cross validation.

Gamer-augmented transfer learning. The gamer transfer learning network was trained using the same network structure as before with the P values calculated from the gamers' consensus added concatenated to the input features (**Supplementary Fig. 4**).

For the pseudo-gamer transfer learning network (Loc-CAT+), a secondary network was trained to predict the gamer consensus P values (**Supplementary Fig. 4**). The secondary network was trained for 100 epochs on the same SLF input features with two hidden ReLU6 layers containing 200 and 100 neurons, respectively. Dropout was applied to the secondary network as well, 20% on the input layer and 40% on each of the hidden layers. The predicted P values are then concatenated to the standard SLF features as input to the standard Loc-CAT network.

CNN. We evaluated a convolutional neural network using the SimpleNet architecture with dropout[39]. The network was trained using versions of HPA images scaled to 128 × 128 pixels as input. Each input image contained all four available image channels. The network was trained for 600 epochs with no substantial validation loss change seen for the last 200 of those epochs. The performance, although inferior to the other methods presented in this paper,

doi:10.1038/nbt.4225

showed great promise for a convolutional network properly trained and tuned for protein localization.

Protein of interest only classifier. Loc-CAT architecture was trained using only features from the protein of interest; however, performance of this classifier was substantially inferior to that of the model trained with the three cellular reference channels (data not shown). This suggests that the contextualization of the protein in a cell using such reference markers is crucial for accurate protein localization from images.

Life Sciences Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability statement. The images included in this study are available in the HPA Cell Atlas (https://www.proteinatlas.org), specifically the HPA Cell Atlas v14 can be found at (https://v14.proteinatlas.org). The data from Project Discovery is available upon request.

Code availability statement. Code for extracting features from images in the HPA Cell Atlas is available at: https://github.com/CellProfiling/FeatureExtraction. Code for the analysis of data from Project Discovery presented in this work is available at: https://github.com/CellProfiling/ProjectDiscovery. Code for the Loc-CAT presented in this publication is available at: https://github.com/CellProfiling/Loc-CAT.

nature research

Corresponding author(s): Emma Lundberg

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
□	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
□	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	□	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Project Discovery was run inside of EVE Online. Data was served by Massive Multi-player Online Science (MMOS). Analysis of the resulting data was performed in MATLAB 2016a. Quantitative image features were calculated using MATLAB 2016a. CPython 3.5.2 with CUDA gpu-accelerated TensorFlow (v1.3.0) was used to train the feed-forward deep artificial neural networks used in this work. All analysis and modeling software is available through the CellProfiling GitHub at the addresses noted in the manuscript.
Data analysis	Analysis of the resulting data was performed in MATLAB 2016a. Quantitative image features were calculated using MATLAB 2016a. CPython 3.5.2 with CUDA gpu-accelerated TensorFlow (v1.3.0) was used to train the feed-forward deep artificial neural networks used in this work. All analysis and modeling software is available through the CellProfiling GitHub at the addresses noted in the manuscript.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The images included in this study are available in the HPA Cell Atlas (www.proteinatlas.org)

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Samples in this paper were taken from version 14 of the Cell Atlas, a part of the Human Protein Atlas. The entire dataset was included.
Data exclusions	Two types of data were excluded in the analysis. First, internal data from experiments not passing the quality controls of the Human Protein Atlas, or having a "Negative" or "Unspecific" label were excluded from the analysis due to quality concerns. Second, proteins identified by the Human Protein Atlas as containing cell-to-cell variation were excluded as no ground-truth was available for which cells displayed a given pattern. Proteins identified as cell-to-cell variable by gamers within Project Discovery were not excluded if they were not previously annotated as such by the Human Protein Atlas. These exclusions were pre-determined.
Replication	All attempts at replication of results showed consistency via cross validation.
Randomization	In all generalized models and Project Discovery, sample allocation was randomized from Human Protein Atlas version 14 data using PRNG. Below are the specific non-random cases from the Loc-CAT machine learning approach. 1. Per-cell line models: only samples from a given cell line were used when training/testing the models respectively. All other aspects of the sampling were randomized. 2. "Single" localization model (Fig 5d): Only images containing a single annotation in the Human Protein Atlas v14 were included in this model. The number of samples was restricted to 10,000. All other aspects of the sampling were randomized. 3. "Multi" localization model (Fig 5d): Only images containing multiple annotations in the Human Protein Atlas v14 were included in this model. The number of samples was restricted to 10,000. All other aspects of the sampling were randomized. 4. "Mixed" localization model (Fig 5d): The number of samples was restricted to 10,000. 50% of the samples are taken from the "Single" dataset, and 50% from the "Multi" dataset. All other aspects of the sampling were randomized.
Blinding	For Project Discovery, players were blind to what images they were presented, what annotations the HPA Cell Atlas had made previously if any, and what other players had input until after submission of their analysis. For the machine learning approach, the learner was blind in the held out testing sets to the labels given to the image by the HPA Cell Atlas, and any labels derived from the gamer input excluding the Gamer-Augmented model, which used gamer input to improve accuracy (Fig 6b). For internal expert reannotation, experts were blind to the previous annotation(s) for each image.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study	
☒	□	Unique biological materials
□	☒	Antibodies
□	☒	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
☒	□	Human research participants

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Antibodies

Antibodies used

The entire collection of antibodies publically available in the Human Protein Atlas were included in this study. A list of them, validation and RRID for tracking can be found at www.proteinatlas.org.

The listed antibodies do not contain lot numbers as they are produced in-house and then sold by Atlas Antibodies. Therefore the lot number is identical to the antibody ID. We also provide RRID numbers as a means of identification. There is no CloneID as they are polyclonal antibodies.

The antibodies specifically investigated in this study were those used to characterize the Rods & Rings structure. Specifically:
Target Gene, Catalog number, Supplier, Clone, Name, Dilution, Original Concentration,RRID
ENSG00000204310, HPA048478, Atlas Antibodies, polyclonal, 1:56, 0.1128 mg/ml,AB_2680411
ENSG00000159079, sc-83559 (CAB034170), Santa Cruz-Biotechnology, polyclonal Y-18, 1:87, Unavailable, AB_1564311
ENSG00000168237, HPA006913, Atlas Antibodies, polyclonal, 1:20, 0.0425 mg/ml, AB_1078993
ENSG00000106348, HPA001400, Atlas Antibodies, polyclonal, 1:29, 0.0575 mg/ml, AB_1079139
ENSG00000188647, HPA021248, Atlas Antibodies, polyclonal, 1:40, 0.08 mg/ml, AB_1855872
ENSG00000133872, HPA040400, Atlas Antibodies, polyclonal, 1:94, 0.188 mg/ml, AB_2676962
ENSG00000180900, HPA064312, Atlas Antibodies, polyclonal, 1:200, 0.6489 mg/ml, AB_2685242
ENSG00000104375, HPA007120, Atlas Antibodies, polyclonal, 1:30, 0.06 mg/ml, AB_10601664
ENSG00000213186, HPA017750, Atlas Antibodies, polyclonal, 1:12, 0.0240 mg/ml, AB_1858312
ENSG00000118420, HPA027231, Atlas Antibodies, polyclonal, 1:91, 0.1825 mg/ml, AB_2256689
ENSG00000121417, HPA049967, Atlas Antibodies, polyclonal, 1:43, 0.0865 mg/ml, AB_2680973

Validation

As described in Thul et al. 2017 - "All antibodies generated and validated within the HPA project were rabbit polyclonal antibodies. They were designed to bind specifically to as many isoforms of the target protein as possible. The antigens consisted of recombinant protein epitope signature tags (PrEST) with a typical length between 50 and 100 amino acids. The resulting antibodies were affinity purified using the antigen as affinity ligand. All antibodies used were first approved for sensitivity and lack of cross-reactivity to other proteins, on arrays consisting of glass slides with spotted PrEST fragments. Commercial antibodies were provided by the suppliers and used according to the supplier's recommendations."

All antibodies produced internally within the Human Protein Atlas project (HPA antibodies) must pass steps 1-3 in the list below in order to be used for immunohistochemistry and immunocytochemistry/IF. Steps 4-6 provide the basis for evaluating and scoring the antibody reliability. All antibodies that provide a reasonable pattern of immunoreactivity are added to the Human Protein Atlas portal. Feedback from the research community is appreciated and needed for continuous curation of data.
Quality assurance steps for antibodies generated within the Human Protein Atlas project:

1. Plasmid inserts are sequenced to assure that the correct protein epitope signature tag (PrEST) sequence is cloned.
2. Size of the resulting recombinant protein (including the specific PrEST) is analyzed using mass spectrometry to assure that the correct antigen has been produced and purified.
3. To control for cross-reactivity, affinity purified antibodies are tested for sensitivity and specificity on protein arrays consisting of glass slides with spotted PrEST fragments.
4. Antibody specificity is analyzed using Western blot in a standardized setup. Total protein lysates from a limited number of tissues (liver and tonsil), cell lines (RT4 and U-251 MG), and human plasma are used to evaluate the antibody target binding in a Western blot setting. Antibodies with an uncertain standard Western blot are reanalyzed using an over-expression lysate as a positive control.
5. Immunohistochemical staining of normal and cancer tissue is examined and annotated by specially educated personnel, and the staining patterns are compared with available gene/RNA/protein characterization data.
6. High resolution confocal microscopy images of human cell lines stained by indirect immunofluorescence are annotated for subcellular localizations by trained cell biologists, and the subcellular localization patterns are compared with the immunohistochemical staining and available experimental protein characterization data.

For antibodies supplied through commercial or other academic sources (CAB antibodies), immunocytochemistry and immunohistochemistry have been performed and validated in a similar manner as for HPA antibodies. These antibodies have also been tested on Western blot in a standardized setup. For each commercially available antibody, a link to the antibody provider is given on the "Antibody validation" page. For further validation we refer to quality controls provided by the respective company.

detailed information on antibody sources can be found in the Protein Atlas database (https://www.proteinatlas.org/about/antibody+validation)

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	ATCC
Authentication	All cell lines used to generate the data in the Cell Atlas have been authenticated and their transcriptome sequenced as previously described (Thul 2017, Uhlen 2015). This includes the HeLa cell line sourced by DSMZ.
Mycoplasma contamination	All cell lines were tested mycoplasma negative

Commonly misidentified lines (See ICLAC register)	All cell lines used to generate the data in the Cell Atlas have been authenticated and their transcriptome sequenced as previously described (Thul 2017, Uhlen 2015) The commonly misidentified cell lines used in the Cell Atlas are: MCF-7: source - DSMZ, authenticated by transcriptome sequencing RT4: source - DSMZ/ECACC, authenticated by transcriptome sequencing