Genomics Proteomics Bioinformatics 20 (2022) 989–1001

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

METHOD

DeepNoise: Signal and Noise Disentanglement Based on Classifying Fluorescent Microscopy Images via Deep Learning

Sen Yang [1,#], Tao Shen [1,#], Yuqi Fang [2,#], Xiyue Wang [3], Jun Zhang [1,*], Wei Yang [1], Junzhou Huang [1], Xiao Han [1,*]

[1] *Tencent AI Lab, Shenzhen 518057, China*
[2] *Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong Special Administrative Region 999077, China*
[3] *College of Computer Science, Sichuan University, Chengdu 610065, China*

Received 16 October 2021; revised 26 November 2022; accepted 11 December 2022
Available online 3 January 2023

Handled by Kun Huang

KEYWORDS

Fluorescent microscopy image;
Biological signal;
Classification;
Deep learning;
Genetic perturbation

Abstract The high-content image-based assay is commonly leveraged for identifying the phenotypic impact of **genetic perturbations** in biology field. However, a persistent issue remains unsolved during experiments: the interferential technical noises caused by systematic errors (*e.g.*, temperature, reagent concentration, and well location) are always mixed up with the real **biological signals**, leading to misinterpretation of any conclusion drawn. Here, we reported a mean teacher-based **deep learning** model (DeepNoise) that can disentangle biological signals from the experimental noises. Specifically, we aimed to classify the phenotypic impact of 1108 different genetic perturbations screened from 125,510 **fluorescent microscopy images**, which were totally unrecognizable by the human eye. We validated our model by participating in the Recursion Cellular Image **Classification** Challenge, and DeepNoise achieved an extremely high classification score (accuracy: 99.596%), ranking the 2nd place among 866 participating groups. This promising result indicates the successful separation of biological and technical factors, which might help decrease the cost of treatment development and expedite the drug discovery process. The source code of DeepNoise is available at https://github.com/Scu-sen/Recursion-Cellular-Image-Classification-Challenge.

Introduction

Usually, it takes over 10 years and billions of dollars to find a new drug. Gene knockdown, a genetic technique which can make an organism's genes inoperative, is widely used as a screening tool in accelerating drug discovery and development.

* Corresponding authors.
E-mail: junejzhang@tencent.com (Zhang J), haroldhan@tencent.com (Han X).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2022.12.007

Gene knockdown based on RNA interference (RNAi) can be achieved by introducing a small interfering RNA (siRNA), which is designed to be fully complementary to a portion of an messenger RNA (mRNA), interfering with the gene and protein expression [1]. The siRNA transfection is to perturb the morphology, such as count of cells, creating a distinctive phenotype corresponding to each siRNA. To further investigate the results of siRNA transfection, high-content imaging techniques are often used to screen, visualize, and quantitatively analyze cellular feature representations [2,3], which have been widely adopted in many fields of biology, *e.g.*, genetics [4,5], and drug discovery and development [3,6–8]. In this study, we adopted Cell Painting [9], a high-content image-based assay for morphological profiling, to identify the phenotypic impact of genetic perturbations. Specifically, cells perturbed with different treatments were plated in multi-well plates, which were imaged on a high-throughput microscope using multiplexed fluorescent dyes. By analyzing these images, morphologically relevant similarities and differences among samples caused by genetic perturbations can be identified, and the valuable biological information about cellular states can be captured.

However, there exists a big challenge in identifying cellular phenotypes monitored in the readout assay, *i.e.*, technical noises (batch effects and plate effects) can significantly invalidate the biological conclusion drawn. For batch effects, even the experiments are carefully designed to control for systematic variables (*e.g.*, temperature and humidity), the biological measurements derived from the assay screens can mix up with the non-biological artifacts (**Figure 1**A). For plate effects, phenotypes are distinct across different plates even they are generated by the same siRNA targeting the same cell in the same batch (Figure 1B). Both batch effects and plate effects are inherent and unavoidable during the execution of biological experiments. Additionally, as demonstrated in Figure 1C and D, it is impossible to detect phenotypes generated by the same/distinct siRNAs by the human eye.

In order to capture real biological signals from measurements taken from high-throughput screens, effectively eliminating these noises is in highly demand [10–13]. RxRx1 (https://www.rxrx.ai/rxrx1) is the first publicly available biological dataset that is systematically created to study this noise removal problem. Our study adopted this dataset and tackled the task that classified 125,510 screened microscopy images of cells under one of 1108 different genetic perturbations. A high classification score suggests that the biological and technical factors can be effectively separated.

To achieve that, we in this work developed an intelligent deep learning-based model, *i.e.*, DeepNoise. Traditional methods of classification tasks mainly depend on the handcrafted features, *e.g.*, shapes [14–16] and textures [17,18]. Whereas the acquisition and quantification of these features highly depend on domain knowledge and manual design, the accuracy and robustness of traditional methods remain unsatisfying. For instance, images of phenotypes perturbed with two different siRNAs are visually similar, thus it becomes quite difficult for traditional approaches to extract the distinct features for each siRNA (Figure 1A). Recently, regarding the prosperity of deep learning in automatic feature mining, many studies have applied this kind of artificial intelligent technique for image classification [19–22]. Compared with the handcrafted ones, features learned by deep learning methods are with much diversity and may mine for the inherent difference in the phenotypic profiles induced by different siRNAs, enabling a better generalization ability of the models. The proposed deep learning model is presented in **Figure 2**, and it achieves a classification accuracy of 99.596% in this challenging 1108 classification task. The near-perfect result indicates the effectiveness of our proposed method on disentangling biological signals from the experimental noises, which may dramatically decrease the cost of treatment development and expedite the process of drug discovery.

Method

Datasets

In this study, we adopted the RxRx1 dataset (https://www.rxrx.ai/rxrx1) released by Recursion Pharmaceuticals (https://www.recursionpharma.com/). This dataset contains 125,510 fluorescent microscopy images representing 1108 classes. Each fluorescent microscopy image contains 6 channels with different wavelengths, which respectively represent 6 different cell organelles, *i.e.*, the nucleus, nucleolus, mitochondria, endoplasmic reticulum, actin cytoskeleton, and golgi apparatus [9].

The entire dataset consisted of 51 experiments, and each experiment was executed in one batch. Each batch held a single cell type: 24 in human umbilical vein endothelial cells (HUVECs), 11 in retinal pigment epithelial (RPE) cells, 11 in human hepatocellular carcinoma (HepG2), and 5 in human bone osteosarcoma epithelial (U2OS) cells. As shown in Figure 2A, microscopy images generated from four distinct cell types perturbed with different siRNAs showed much diverse phenotypes. And Figure 2B displays the diagram of full complementarity of an siRNA to an mRNA to knock down a particular target gene. In one batch, there were four plates, each with 384 (16×24) wells. One of 1108 different siRNAs was introduced into each well to create distinct genetic conditions. Images located in the outer rows and columns of the plate were not utilized since they were hugely affected by environmental effects, so for each plate there remained 308 wells. Each plate contained the same 30 control siRNAs, 277 different perturbed siRNAs, and one intact well. In each experiment, the locations of 1108 (277×4) perturbed siRNAs were randomized. Recursion Pharmaceuticals released this dataset and held a competition on Kaggle (https://www.kaggle.com/c/recursion-cellular-image-classification), attracting 866 teams from all over the world to participate in.

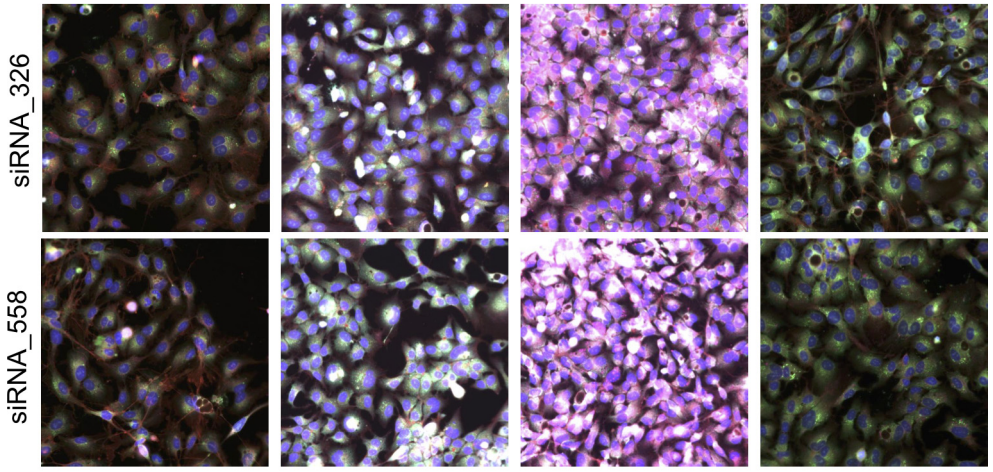
Evaluation metric

As the challenge organizers specified, multi-class accuracy was adopted to evaluate the model performance, which was simply the average number of observations with the correct label. The metric improved with the increased number of correctly classified images.

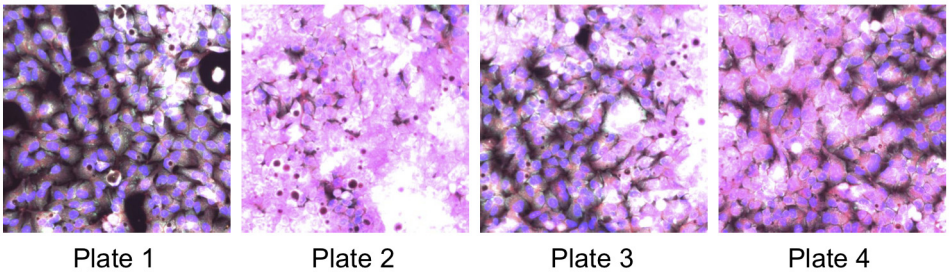
Training details

For model training, a mini-batch of size 64 was adopted and the Adam optimizer [23] with weight decay (2×10^{-4}) was used as the optimization method. The initial learning rate

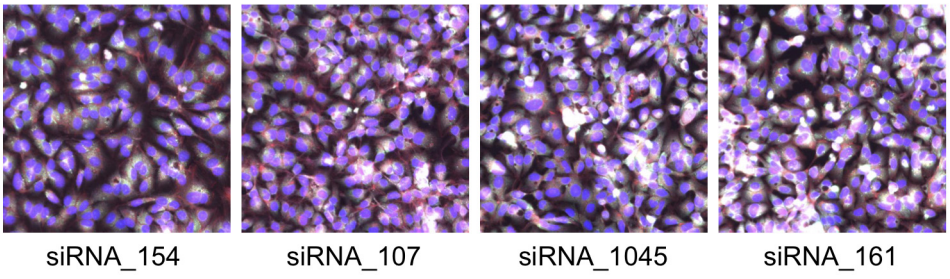
A Images of phenotypes generated by two different siRNAs in HepG2 cells across four batches



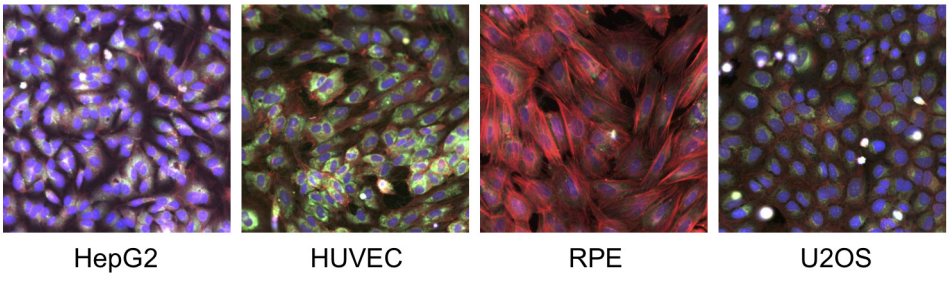
B Images of phenotypes generated by the same siRNA in HepG2 cells in four plates in Batch 2



C Images of phenotypes generated by different siRNAs in HepG2 cells (Batch 4, Plate 2)



D Images of phenotypes of four cell types generated by the same siRNA



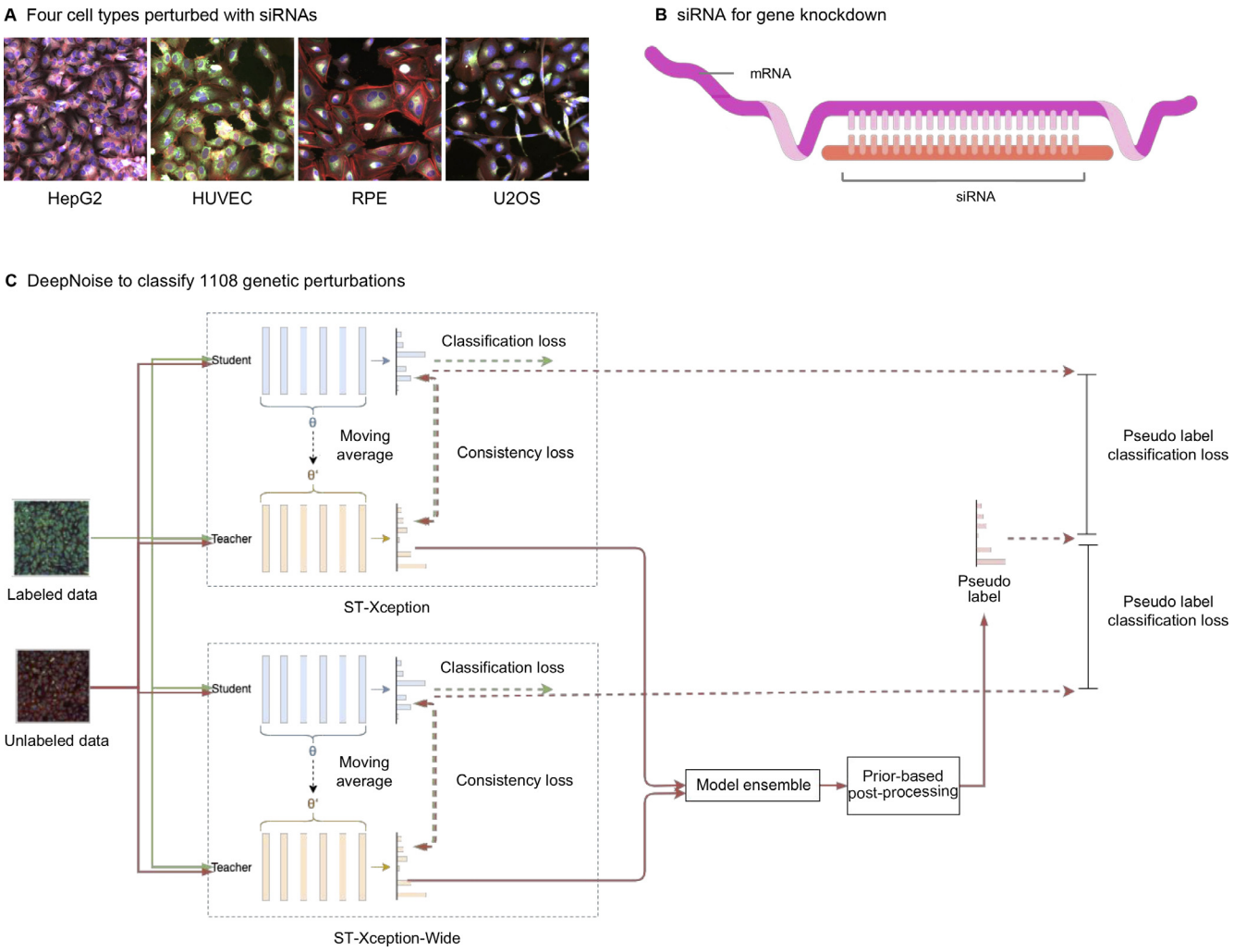


Figure 2 The full pipeline of DeepNoise

A. Fluorescent microscopy images generated from four distinct cell types perturbed with different siRNAs, showing much diverse phenotypes. **B.** A diagram showing the full complementarity of an siRNA to an mRNA to knock down a particular target gene. **C.** The architecture of the proposed DeepNoise network. The architecture is based on the mean teacher strategy [23], a semi-supervised teacher–student deep learning network that averages the model weights of the student network. The classification loss, *i.e.*, $L_{ArcFace}$, is applied on the student network, and a consistency loss, *i.e.*, $L_{Consist}$, is used to minimize the difference between the outputs of the student network and the teacher network. Two models (ST-Xception and ST-Xception-Wide) are integrated, and the final classification result is derived by averaging these two models' prediction outputs. The detailed description of the model can refer to the "Successful signal and noise disentanglement via deep learning based on 1108 genetic perturbations" section in Results. mRNA, messenger RNA.

Figure 1 The human has much difficulty in identifying the phenotypic impact of genetic perturbations

A. Microscopy images of phenotypes generated by two different siRNAs (rows) in HepG2 cells across four different batches (columns). Four morphological phenotypes in each row are all derived from the same siRNA targeting the same cell type but in different experimental batches. There exists large visual difference across four batches (*i.e.*, batch effects). Different siRNAs may generate visually similar expression by comparing morphological phenotypes in each column, which introduces large disturbance on real biological signal capture. **B.** Microscopy images of phenotypes generated by the same siRNA in four plates in one experimental batch. Four phenotypes are all derived from the same siRNA targeting the same cell type in one batch but in different plates. Visually different phenotypes are shown across four plates (*i.e.*, plate effects). **C.** Microscopy images of phenotypes generated by four distinct siRNAs in HepG2 cells in the same experiment and the same plate, which are totally unrecognizable by the human eye. **D.** Microscopy images of phenotypes generated by the same siRNA in four different cell types. siRNA, small interfering RNA; HUVEC, human umbilical vein endothelial cell; RPE, retinal pigment epithelial; HepG2, human hepatocellular carcinoma; U2OS, human bone osteosarcoma epithelial.

was set to 3×10^{-4}, which was reduced to 1×10^{-4} after 100 training epochs. After the 140th epoch, the learning rate decreased by a factor of 10, and there were totally 160 training epochs. All models were performed using the PyTorch package [24] and all experiments were implemented on a workstation equipped with four 24 GB memory NVIDIA Tesla P40 GPU cards.

Standard real-time data augmentation methods such as horizontal flip, vertical flip, 90° flip, random erasing, and random scale were performed to make the model learn features invariant to geometric perturbations.

Different normalization strategies in pre-processing step

Cell-based normalization

Cell-based normalization was to calculate the mean value and standard deviation (std) for microscopy images of each cell type, and then standardize these images based on their corresponding cell types.

$$x_{cell\_norm} = (x^i - x^i_{mean})/x^i_{std} \tag{1}$$

where i means HUVEC, RPE, HepG2, or U2OS; x^i represents the input image; x^i_{mean} and x^i_{std} denote the mean and the std of the corresponding cell type of the input image, respectively.

Batch-based normalization

There were totally 51 experimental batches in this study, and batch effects inevitably created factors of variation within the data that were irrelevant to the biological information, even when these batches were carefully designed to control for technical variables. Therefore, we considered conducting the batch-based normalization for these microscopy images.

$$x_{batch\_norm} = (x^j - x^j_{mean})/x^j_{std} \tag{2}$$

where j means one of the 51 experimental batches; x^j_{mean} and x^j_{std} denote the mean and the std of the corresponding batch of the input image, respectively.

Plate-based normalization

Since plate effects also interfere with the biological conclusion drawn, we applied the third strategy, *i.e.*, plate-based normalization, on these microscopy images.

$$x_{plate\_norm} = (x^k - x^k_{mean})/x^k_{std} \tag{3}$$

where k means one of the 204 (51×4) experimental plates; x^k_{mean} and x^k_{std} denote the mean and the std of the corresponding plate of the input image, respectively.

ArcFace loss

The ArcFace loss, $L_{ArcFace}$ [25], was leveraged to minimize the difference between prediction outputs of the student network and the one-hot classification label. Compared with other classification cost functions, *e.g.*, softmax loss and its variants [26–31], the ArcFace loss we adopted incorporated an additive angular margin in the loss function to enforce extra intra-class compactness and inter-class discrepancy simultaneously.

$$L_{ArcFace} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(cos(\theta_{y_i}+m))}}{e^{s(cos(\theta_{y_i}+m))}+\sum_{j=1,j\neq y_i}^{n}e^{s\cos\theta_j}} \tag{4}$$

where i-th sample belongs to the y_i-th class; N is the number of samples within one mini-batch; n is the number of classes. θ_{yi} is the angle between the weight W_j and the feature x_i, herein, $||W_{yi}||$ is fixed to 1, and the embedding feature $||x_i||$ is normalized and re-scaled to s ($s = ||x_i||$). m is an additive angular margin between x_i and W_{yi} to enforce intra-class compactness and inter-class discrepancy simultaneously. s and m are hyper-parameters. Experimentally, s is set to 30, and m is set to 0.1.

Consistency loss

The consistency loss, $L_{Consist}$, was designed to minimize the difference between the prediction of the student network (with weights θ) and that of the teacher network (with weights θ'), which was defined as follows:

$$L_{Consist} = \mathbb{D}_x[||f(x,\theta') - f(x,\theta)||^2] \tag{5}$$

where f represents mean squared error (MSE) in our study.

For training images with annotated labels, both $L_{ArcFace}$ and $L_{Consist}$ were leveraged to optimize DeepNoise network, while for those unlabeled examples, only $L_{Consist}$ was adopted.

Pseudo label softmax loss

The pseudo label softmax loss, $L_{pSoftmax}$, was introduced to minimize the difference between the averaged predictions derived from two teacher networks [in both student–teacher–Xception (ST-Xception) and ST-Xception-Wide] and the output of each student network, which can be regarded as another constraint consistency loss function.

Model ensemble

Model ensemble is an effective and commonly used approach in machine learning. Usually, it can efficaciously improve the overall performance of the network by aggregating predictions of each base model. Similar to seeking an answer about something from many people, we can often get a better solution from their combined answers than asking one person. The model ensemble strategy can reduce the variance and bias caused by a single contributing model. Hence, the final aggregated prediction can be less noisy, helping reduce the prediction errors. Moreover, since predicted outputs are aggregated from different and diverse well-trained models, we can get more robust prediction, and also the ensemble model can generalize well on other new scenarios. It has been reported that model ensemble achieves satisfying performance on the segmentation of natural images [32,33], magnetic resonance images [34–36], pathological images [37–39], fundus images [40], aerial images [41], *etc*. In this work, we integrated two deep learning-based models, *i.e.*, ST-Xception and ST-Xception-Wide to derive the final classification results by averaging their respective predictions. Xception-Wide was developed based on Xception [42] but with wider convolutional channels. Specifically, the numbers of the output convolutional channels of ST-Xception-Wide and ST-Xception were 64 and 32, respectively. Usually, as channel number increases, bigger the capacity the network will hold for capturing different patterns. That is, we can regard each convolution as a feature extractor over the input data. Networks with more channel numbers often perform better in more complex datasets.

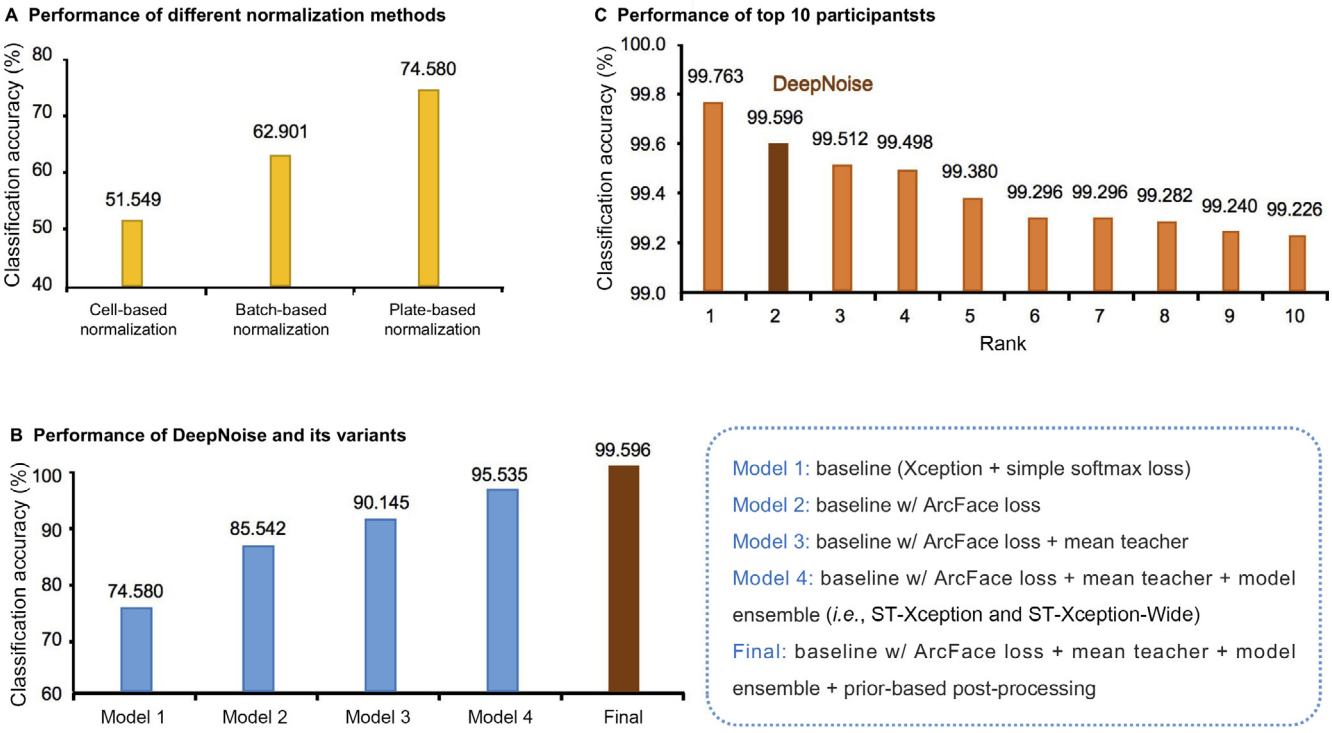


Figure 3 Experimental results of this cellular image classification task

A. Prediction results under different normalization strategies in pre-processing step. **B.** Ablation studies to demonstrate each component introduced into DeepNoise. The detailed model description can refer to Figure 4. **C.** Leaderboard of CellSignal competition. Our proposed DeepNoise ranked 2nd place among 866 participants with a classification accuracy of 99.596%.

However, if the channel number is too large, the model will suffer from performance degradation, which may be due to the overfitting problem. Here, we used a relatively large and small number (*i.e.*, 64 and 32) to capture diverse and complementary features of the cellular images for final classification.

Prior-based post-processing

In RxRx1 dataset, same siRNA does not occur twice in an experimental plate. Based on this prior information, we added a post-processing step to adjust the initial prediction of DeepNoise. The idea of the algorithm was as follows: for the popular classes (appear a few times more than minor classes), we decreased their predicted probabilities iteratively by a small value (an initial value is 0.001, and it is halved iteratively), until all classes were equally presented and the prediction score no longer improved. This balancing was not easy to enforce, but as a soft constraint in the loss, it worked quite well.

Results

Successful signal and noise disentanglement via deep learning based on 1108 genetic perturbations

In this study, we proposed a semi-supervised deep learning network (DeepNoise) to identify the phenotypic impact of 1108 genetic perturbations. The network architecture was comprised of two base models (Figure 2C), *i.e.*, ST-Xception and ST-Xception-Wide, which were both based on the mean teacher strategy [43]. Taking ST-Xception (Figure 2C, upper) for example, the student network adopted Xception module [42], which replaced Inception modules with depthwise separable convolutions, enriching the diversity of learned feature representations. The teacher network sharing the same architecture with the student network was initialized based on pretrained weights of ImageNet [44]. During model training, the model weights of student network were first averaged and passed to the teacher network, and then the teacher network was updated by combining the updated student network with the historical information of teacher network using an exponential moving average strategy. The student network learned from the teacher network by minimizing the classification loss computed from the input annotated data (*i.e.*, $L_{ArcFace}$), and the consistency loss between the teacher and student networks computed from the unannotated data (*i.e.*, $L_{Consist}$).

For ST-Xception-Wide (Figure 2C, bottom), both student and teacher networks used Xception-Wide, which had wider convolutional channels and larger model capability compared with Xception [42]. To achieve the final prediction, the results of both ST-Xception and ST-Xception-Wide were averaged, and went through a prior-based post-processing step to derive the pseudo labels. The pseudo labels were updated using the highest validation score generated by teacher networks during the training procedure. Eventually, we introduced $L_{pSoftmax}$ to

minimize the difference between the averaged prediction derived from two teacher networks and the output of each student network, which can be regarded as another consistency loss function.

Plate-based normalization achieves best performance

The adopted dataset was comprised of 125,510 fluorescence microscopy images of human cells of four different types, *i.e.*, HUVEC, RPE, HepG2, and U2OS, and it was collected from 51 batches with 4 plates in each batch. Since phenotype profiles produced visually differ from each other, it becomes necessary to apply normalization on these microscopy images to mitigate this variation to some extent. In this work, we explored three normalization strategies before model training, *i.e.*, cell-based normalization, batch-based normalization, and plate-based normalization. Except the difference of normalization strategies, all training models were the same, *i.e.*, Xception [42] optimized with the simple softmax loss function.

The prediction results of 1108 genetic perturbations based on these three normalization methods are displayed in **Figure 3**A. We observed that only 51.549% classification accuracy was achieved under the cell-based normalization. This result was expected because neither patch effects nor plate effects were removed, thus biological information and interferential technical noises in experiments were mixed up. Batch-based normalization was developed for tackling batch effects, using which the classification accuracy reached 62.901%. Although there existed a dramatic improvement of the prediction accuracy compared with the cell-based normalization method, this batch-based normalization strategy may not be optimal because four plates in one batch still affected each other. To tackle this problem, we finally applied plate-based normalization on this dataset, and the prediction accuracy of 1108 genetic perturbations reached up to 74.580%.

Ablation studies

We reported each component introduced into DeepNoise network that improved the prediction accuracy of genetic perturbation classification. The following ablation studies were all based on the plate-based normalization strategy. The baseline model (Model 1 in **Figure 4**) was Xception [42] optimized with a simple softmax loss function, which achieved a classification accuracy of 74.580%. For Model 2, we replaced the simple softmax loss in Model 1 with the ArcFace loss [25]. Since the number of classes we identified was quite large (*i.e.*, 1108 classes), how to enhance the intra-class compactness and inter-class discrepancy becomes important. The ArcFace loss [25] added with an angular margin was really good at separating inter-class distance, thus it was leveraged to optimize the neural network. Compared with Model 1 (Figure 3B), we could see that an accuracy of 10.962% was improved for Model 2 when using ArcFace loss. For both Model 1 and Model 2, any information of the unannotated (test) fluorescent microscopy images was not taken in account, that is, both models were fully supervised.

The following three models were the semi-supervised approaches, which considered utilizing the unannotated (test) microscopy images to assist the model training. Model 3 (Figure 4) was designed based on the mean teacher strategy [43]. Both student and teacher networks adopted Xception [42] as the training network. For annotated data (training microscopy images), a classification loss and a consistency loss were jointly used to train the network, while for unannotated data (testing microscopy images), only the consistency loss was adopted to optimize the network. Comparing the classification accuracy derived from Model 2 and Model 3 (Figure 3B), we could see that feature representations derived from unannotated images indeed enhance the prediction accuracy. To further improve the classification performance, we adopted the model ensemble strategy by aggregating predictions of two base models, *i.e.*, ST-Xception and ST-Xception-Wide (Model 4 in Figure 4). By model ensemble, the strengths of both models were taken advantage of, and the features generated from two teacher networks could be more representative. By aggregating the complementary information of two models, the classification accuracy could be significantly improved (from 90.145% to 95.535%) (Figure 3B). Eventually, we applied a prior-based post-processing before pseudo label generation (Figure 2C), which took into consideration the prior information that same siRNA does not occur twice in an experimental plate. As expected, the classification accuracy increased after all classes were equally balanced (from 95.535% to 99.596%) (Figure 3B).

The training and validation loss curves during the training process are displayed in **Figure 5**. At the beginning, all microscopy images no matter what the cell type they belonged to were used for training, and the training loss and validation loss are shown in Figure 5A. With the aim of extracting the cell-specific feature representations, we fine-tuned DeepNoise on four cell types, respectively.The training loss and validation loss in terms of four cell-specific models are shown in Figure 5B. During the inference procedure, for the microscopy images under the specific cell types, we applied the corresponding trained model and generated the final prediction results.

Comparison with top 10 teams participating in the challenge

The organizer of RxRx1 dataset sponsored a challenge named CellSignal to encourage researchers to explore methods of disentangling biological signals from technical noises. Totally 866 teams participated in this competition, and the scores of top 10 teams of the leaderboard are shown in Figure 3C. Our proposed DeepNoise ranked 2nd place, achieving a classification accuracy of 99.596%. According to the solution of the competition champion released on the official website (https://www.kaggle.com/c/recursion-cellular-image-classification/discussion/110543), we found that they adopted an ensemble of 11 deep learning models (6× DenseNet-161 and 5× DenseNet-201 [45]), while our proposed solution only integrated two models (ST-Xception and ST-Xception-Wide) to derive the final classification results. It is likely that the increasing number of ensemble models contributes to their championship.

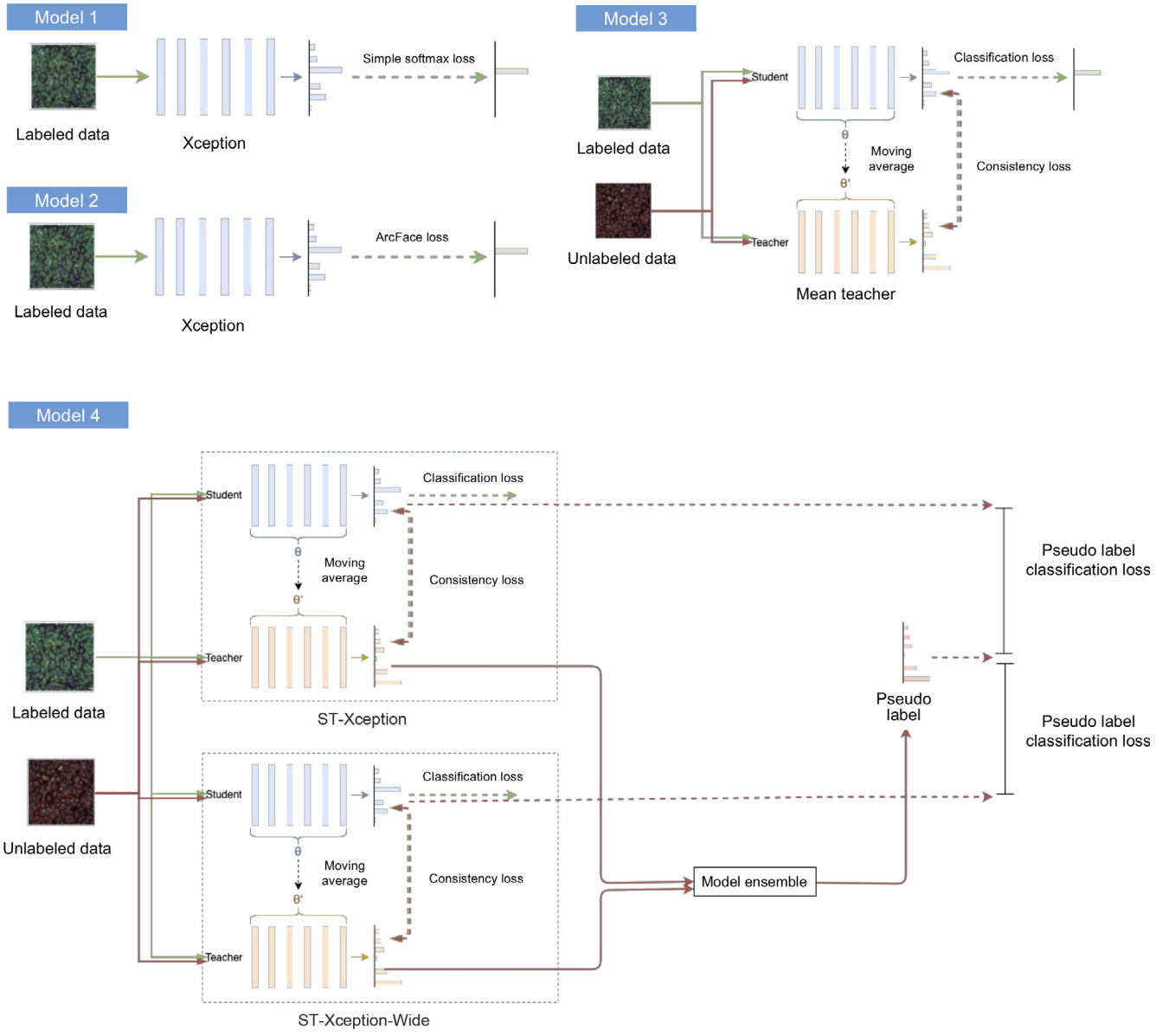


Figure 4 Four ablation studies to demonstrate each component introduced into DeepNoise

Model 1: Xception optimized with a simple softmax loss function, which is regarded as a baseline. Model 2: replacing the simple softmax loss in Model 1 with the ArcFace loss. Both Model 1 and Model 2 are fully-supervised models. Model 3: a semi-supervised mean teacher strategy is utilized, in which both the student and teacher networks adopt Xception. For annotated data (training microscopy images), a classification loss, *i.e.*, $L_{ArcFace}$, and a consistency loss, *i.e.*, $L_{Consist}$, are jointly to train the network, whereas for unannotated data (testing microscopy images), only the consistency loss, *i.e.*, $L_{Consist}$, is adopted to optimize the network. Model 4: a model ensemble strategy is leveraged to aggregate predictions of two base models, *i.e.*, ST-Xception and ST-Xception-Wide. The pseudo labels update using the highest validation score generated by two teacher networks during the training procedure. The pseudo label classification loss function is leveraged to minimize the difference between the averaged predictions derived from two teacher networks and the output of each student network.

Moreover, among top 10 teams, the participants ranking 5th and 8th also made their method publicly available. The former one used a backbone of DenseNet-201 equipped with ResNeXt101-32x8d, HRNet-W18, and HRNet-W30. In their method, an exemplar memory technique [46] took into account the intra-domain variations in the target domain. Compared with our proposed solution, no pseudo label and no test-time augmentations were used in their method. The latter one utilized a backbone of DenseNet-121 embedded with EfficientNet-B1, SE-Restnet101, and DenseNet169. Moreover,

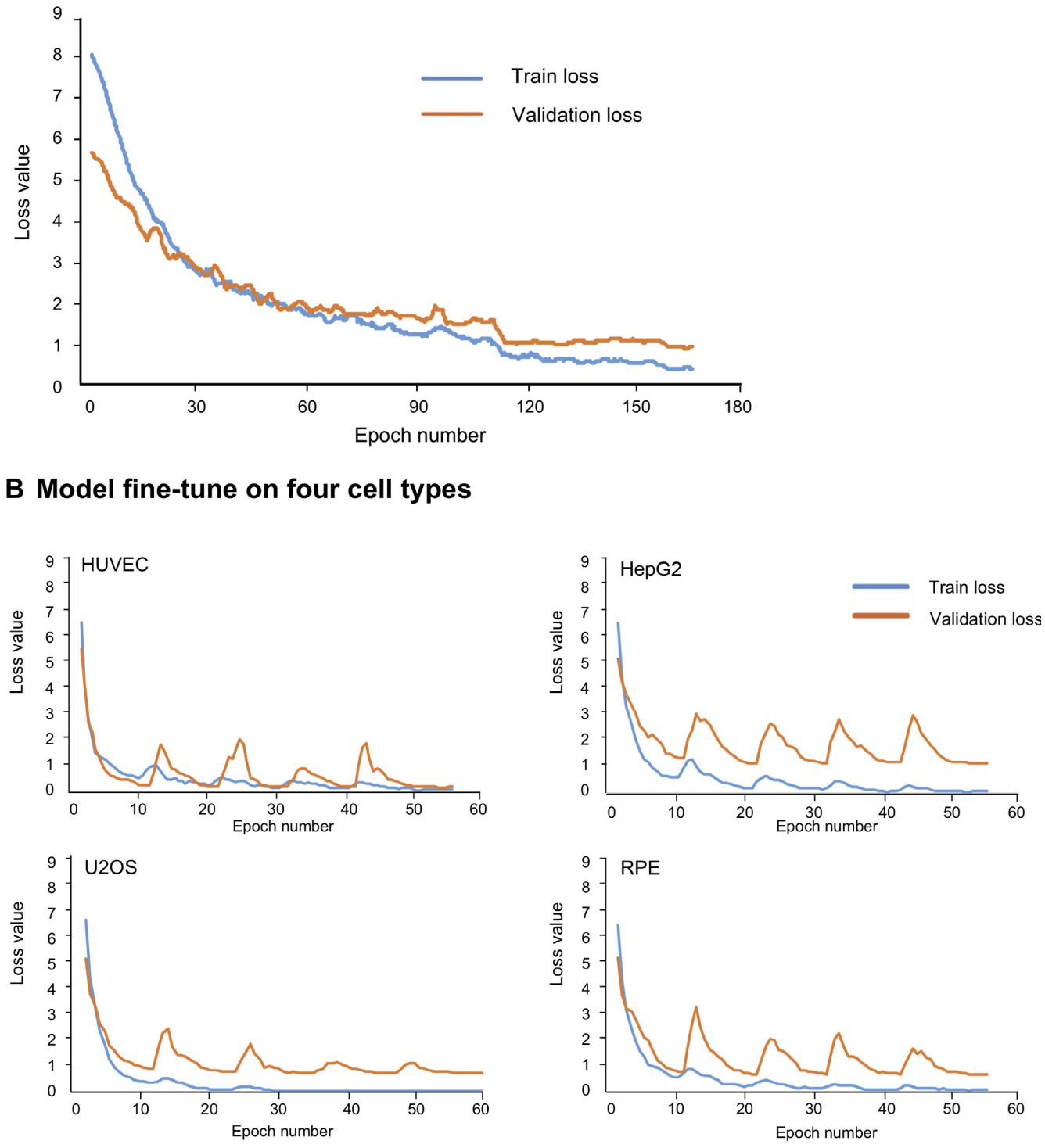


Figure 5 The training and validation loss curves during the training process
A. All microscopy images no matter what the cell type they belong to are used for training. **B.** The training and validation loss curves fine-tuned on four cell types, respectively. The fine-tuning is conducted after 160 training epochs under (A).

mixup regularization was performed to regularize the network to favor simple linear behavior in-between training examples. For the post-processing step, test-time augmentations were used to improve learning performance and reduce generalization error.

Discussion

In this work, we developed a deep learning based model (DeepNoise) that could disentangle the real biological signals from

the interferential technical noises by identifying phenotypic impact of 1108 different genetic perturbations screened from 125,510 fluorescent microscopy images. The proposed method achieved an extremely high classification score, with a multi-class accuracy of 99.596%. Moreover, among 866 participating groups that adopted the same database, our method competed favorably and won the second prize. This promising result verified the successful separation of biological and technical factors, helping the biology researchers derive conclusions from real biological information. Moreover, the persistent issue that biological experiments interfered with batch effects were hardly reproducible was also tackled to some extent.

Our work is motivated by that during the execution biological experiments, the experimental/technical noises (batch effects and plate effects) are always mixed up with the real biological signals, leading to misinterpretation of any conclusion drawn. This unavoidable problem is a persistent issue remaining unsolved for a long time, which dramatically hinders drug discovery process of researchers/companies. Our study is the first one to investigate this noise removal problem by classifying 125,510 screened microscopy images of cells under one of 1108 different genetic perturbations. Surprisingly, we can achieve an extremely high classification score (accuracy: 99.596%) in this difficult task, indicating that the biological and technical factors can be effectively separated. Our success is believed to benefit the society by expediting the process of drug discovery and decreasing the cost of treatment development. Note that our source code is made publicly available to benchmark the future studies in this research field.

Our proposed DeepNoise is a deep learning based model. Compared with conventional methods that depend on handcrafted features for classification (shape, color, and texture), deep learning methods can automatically discover intricate hierarchical feature representations for a large quantity of input data. Deep learning technique has been adopted in many biological fields, including gene expression modeling [47,48], protein structure prediction [49], DNA methylation [50], and protein localization [51]. Our proposed model is based on the semi-supervised mean teacher strategy [43], which has achieved the state-of-the-art performance on both CIFAR-10 and ImageNet 2012 challenges. Thanks to the averaged weight from student to teacher network, mean teacher strategy can effectively exploit the unlabeled data, which exactly suits our needs. In future work, we would like to explore other semi-supervised [52,53] or self-supervised learning models [54,55] to tackle this genetic perturbation problem.

The mean teacher model has demonstrated the state-of-the-art results in the computer vision domain, and been widely used in semi-supervised scenarios. Several other semi-supervised approaches are considered to be investigated, but finally do not meet our requirements, which are elaborated as follows. 1) Shared-weight approaches [56] let the model weights of the teacher network be identical to those of the student network, which may reduce model robustness. 2) Label-averaging approaches [56] adopt the label-averaging scheme, where the learned information can only be aggregated after each epoch, causing a slow feedback loop. Also, intermediate representations are lacking so that the unlabeled data cannot be fully exploited. 3) Generative adversarial network-based approaches [53] are usually hard to train and converge, and the unbalance between the generator and discriminator is easy to cause overfitting problem. 4) Single model approaches often derive noisy pseudo labels with one trained model, and no consistency regularization is used to generate more accurate pseudo labels. Compared with these approaches, our mean teacher strategy constructed a target-generating teacher model by averaging the student model's weights, which effectively improved the label quality of the unlabeled data. The algorithm describing the whole pipeline can be found in Algorithm 1.

Algorithm 1 Introduction of the whole pipeline of the proposed model

Definition
Labeled data, I_L; unlabeled data, I_U; ST-Xception model, M;
ST-Xception-wide, MW; pseudo labels from M, PL_M;
pseudo labels from MW, PL_{MW}.
Pre-processing
Plate-based normalization for I_L and I_U
Data augmentation
Model training
while $i \leq$ Epoch **do**
Train M and MW based on I_L via $L_{ArcFace} + L_{Consist}$ [Equations (4) and (5)]
Train M and MW based on I_U via $L_{Consist}$ [Equation (5)]
Generate pseudo labels PL via averaging PL_M and PL_{MW}
Train whole network based on PL via $L_{pSoftmax}$
end while
Prior-based post-processing
All classes are forced equally presented

The confirmation bias remains a problematic issue in the semi-supervised tasks, which stems from using incorrect predictions on unlabeled data for training in subsequent epochs, and thereby the model will constantly make wrong predictions. Our mean teacher strategy was exactly designed for tackling this problem. Specifically, we first trained the mean teacher model with labeled data only via $L_{ArcFace}$ loss using the student model, preventing the model getting stuck into uncertainty. Then, we added the consistency loss $L_{Consist}$ between the student and teacher outputs to learn knowledge from unlabeled data. Note that we gradually ramped up the coefficient of $L_{Consist}$ in the beginning over the first few epochs until the teacher model started giving good predictions. Moreover, during training we dedicated a portion of each mini-batch for labeled examples (a quarter or a half) and the rest samples were unlabeled, which aimed to stabilize the training process, preventing the network dominating by wrongly classified unlabeled data. Moreover, we added noise/perturbation to the student and teacher models, respectively, and applied a consistency cost between two models' predictions. Such noise regularization technique helped the network learn invariant features of different domains, and improved the quality of the generated pseudo labels, thus alleviating confirmation bias mainly caused by inaccurate pseudo labels.

The proposed method was specifically designed for a semi-supervised scenario where a large portion of data were unlabeled. Here, we utilized the mean teacher strategy, and its basic idea was to utilize abundant unlabeled data and construct a target-generating teacher model by averaging the student model's weights, which aims to improve the label quality of the unlabeled data. In this way, rich and diverse feature patterns of the unlabeled data can be leveraged, which can improve the generalizability and robustness of the constructed model. In real world, there exist tons of unannotated data in

biological and other medical image field due to labeling burden, and our method can be used as a benchmark in the semi-supervised settings.

The limitations of this study are four aspects. Firstly, since the labels of test dataset are not accessible, extensive comparisons between our proposed DeepNoise and other classification models cannot be conducted. The more detailed investigations also cannot be explored, including the predictions in each cell type, and other measurement metrics (*e.g.*, specificity, sensitivity, and precision) that further validate the proposed method. Secondly, our submitted solution does not consider any information of 30 control genes in each experimental plate, which are totally the same across different plates. This information is believed to be valuable. Our future work will seek to intelligently combine features derived from these reference images and current learned features to further improve the classification accuracy. Thirdly, current work simply averages the predictions of ST-Xception and ST-Xception-Wide, neglecting the contribution of each specific model. In the future, we will automatically learn weighted importance of both models for performance improvement. Moreover, we only investigate fluorescent microscopy images generated from two siRNAs in HepG2 cells in this study. More cell types can be explored to further validate the effectiveness of our proposed method.

Code availability

The source code of DeepNoise is available for research purposes at BioCode: https://ngdc.cncb.ac.cn/biocode/tools/BT007332 and GitHub: https://github.com/Scu-sen/Recursion-Cellular-Image-Classification-Challenge.

CRediT author statement

Sen Yang: Conceptualization, Methodology, Writing - original draft. **Tao Shen:** Methodology. **Yuqi Fang:** Conceptualization, Writing - original draft. **Xiyue Wang:** Investigation, Writing - review & editing. **Jun Zhang:** Visualization. **Wei Yang:** Validation. **Junzhou Huang:** Supervision. **Xiao Han:** Supervision, Writing - review & editing. All authors have read and approved the final manuscript.

Competing interests

Sen Yang, Jun Zhang, Wei Yang, Junzhou Huang, and Xiao Han are current employees of Tencent Technology (Shenzhen) Co., Ltd. Tao Shen is a former employee of Tencent Technology (Shenzhen) Co., Ltd. All the other authors have declared no competing interests.

Acknowledgments

We specifically express our gratitude to the Recursion organizers for their released dataset.

ORCID

ORCID 0000-0002-0639-4122 (Sen Yang)
ORCID 0000-0002-6129-3092 (Tao Shen)
ORCID 0000-0002-8769-496X (Yuqi Fang)
ORCID 0000-0002-3597-9090 (Xiyue Wang)
ORCID 0000-0001-5579-7094 (Jun Zhang)
ORCID 0000-0002-6488-2546 (Wei Yang)
ORCID 0000-0002-9548-1227 (Junzhou Huang)
ORCID 0000-0002-5151-6547 (Xiao Han)

References

[1] Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K, Tuschl T. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature 2001;411:494–8.
[2] Conrad C, Gerlich DW. Automated microscopy for high-content RNAi screening. J Cell Biol 2010;188:453–61.
[3] Boutros M, Heigwer F, Laufer C. Microscopy-based high-content screening. Cell 2015;163:1314–25.
[4] Zhou Y, Zhu S, Cai C, Yuan P, Li C, Huang Y, et al. High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. Nature 2014;509:487–91.
[5] Echeverri CJ, Perrimon N. High-throughput RNAi screening in cultured cells: a user's guide. Nat Rev Genet 2006;7:373–84.
[6] Swinney DC, Anthony J. How were new medicines discovered? Nat Rev Drug Discov 2011;10:507–19.
[7] Broach JR, Thorner J. High-throughput screening for drug discovery. Nature 1996;384:14–6.
[8] Macarron R, Banks MN, Bojanic D, Burns DJ, Cirovic DA, Garyantes T, et al. Impact of high-throughput screening in biomedical research. Nat Rev Drug Discov 2011;10:188–95.
[9] Bray MA, Singh S, Han H, Davis CT, Borgeson B, Hartland C, et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat Protoc 2016;11:1757–74.
[10] Soneson C, Gerster S, Delorenzi M. Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. PLoS One 2014;9:e100335.
[11] Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 2010;11:733–9.
[12] Nygaard V, Rødland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics 2016;17:29–39.
[13] Parker HS, Leek JT. The practical effect of batch on genomic prediction. Stat Appl Genet Mol Biol 2012;11:10.
[14] Kothari S, Phan JH, Young AN, Wang MD. Histological image classification using biologically interpretable shape-based features. BMC Med Imaging 2013;13:9.
[15] Zhang D, Lu G. Review of shape representation and description techniques. Pattern Recogn 2004;37:1–19.
[16] Krishnan MMR, Pal M, Paul RR, Chakraborty C, Chatterjee J, Ray AK. Computer vision approach to morphometric feature analysis of basal cell nuclei for evaluating malignant potentiality of oral submucous fibrosis. J Med Syst 2012;36:1745–56.
[17] Nanni L, Lumini A, Brahnam S. Survey on LBP based texture descriptors for image classification. Expert Syst Appl 2012;39:3634–41.

[18] Désir C, Petitjean C, Heutte L, Thiberville L, Salaün M. An SVM-based distal lung image classification using texture descriptors. Comput Med Imaging Graph 2012;36:264–70.
[19] Bayramoglu N, Kannala J, Heikkilä J. Deep learning for magnification independent breast cancer histopathology image classification. Proceeding of the 23rd International Conference on Pattern Recognition 2016:2440–5.
[20] Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 2018;24:1559–67.
[21] Korbar B, Olofson AM, Miraflor AP, Nicka CM, Suriawinata MA, Torresani L, et al. Deep learning for classification of colorectal polyps on whole-slide images. J Pathol Inform 2017;8:30.
[22] Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH. Patch-based convolutional neural network for whole slide tissue image classification. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2016:2424–33.
[23] Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv 2014;1412.6980.
[24] Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, et al. Automatic differentiation in PyTorch. Proceedings of the 31st International Conference on Neural Information Processing Systems 2017:1–4.
[25] Deng J, Guo J, Xue N, Zafeiriou S. ArcFace: additive angular margin loss for deep face recognition. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019:4690–9.
[26] Wen Y, Zhang K, Li Z, Qiao Y. A discriminative feature learning approach for deep face recognition. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer Vision — ECCV 2016, Cham: Springer; 2016, p.499–515.
[27] Deng J, Zhou Y, Zafeiriou S. Marginal loss for deep face recognition. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2017:60–8.
[28] Zhang X, Fang Z, Wen Y, Li Z, Qiao Y. Range loss for deep face recognition with long-tailed training data. Proc IEEE Int Conf Comput Vis 2017:5409–18.
[29] Wang F, Cheng J, Liu W, Liu H. Additive margin softmax for face verification. IEEE Signal Process Lett 2018;25:926–30.
[30] Liu W, Wen Y, Yu Z, Li M, Raj B, Song L. SphereFace: deep hypersphere embedding for face recognition. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2017:212–20.
[31] Wang H, Wang Y, Zhou Z, Ji X, Gong D, Zhou J, et al. CosFace: large margin cosine loss for deep face recognition. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2018:5265–74.
[32] Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2018:8759–68.
[33] Zhao W, Zheng B, Lin Q, Lu H. Enhancing diversity of defocus blur detectors via cross-ensemble network. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019:8905–13.
[34] Kamnitsas K, Bai W, Ferrante E, McDonagh S, Sinclair M, Pawlowski N, et al. Ensembles of multiple models and architectures for robust brain tumour segmentation. In: Crimi A, Bakas S, Kuijf H, Menze B, Reyes M, editors. Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. Cham: Springer; 2016, p.450–62.
[35] Li H, Jiang G, Zhang J, Wang R, Wang Z, Zheng WS, et al. Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. Neuroimage 2018;183:650–65.
[36] Chen Y, Shi B, Wang Z, Zhang P, Smith CD, Liu J. Hippocampus segmentation through multi-view ensemble convnets. Proceeding of the 14th IEEE International Symposium on Biomedical Imaging 2017:192–6.
[37] Pimkin A, Makarchuk G, Kondratenko V, Pisov M, Krivov E, Belyaev M. Ensembling neural networks for digital pathology images classification and segmentation. In: Campilho A, Karray F, ter Haar RB, editors. Image Analysis and Recognition. Cham: Springer; 2018, p.877–86.
[38] Qaiser T, Tsang YW, Epstein D, Rajpoot N. Tumor segmentation in whole slide images using persistent homology and deep convolutional features. In: Valdés Hernández M, González-Castro V, editors. Medical Image Understanding and Analysis. Cham: Springer; 2017, p.320–9.
[39] Zhao J, Li Q, Li X, Li H, Zhang L. Automated segmentation of cervical nuclei in pap smear images using deformable multi-path ensemble model. Proceeding of the 16th IEEE International Symposium on Biomedical Imaging 2019:1514–8.
[40] Tang P, Liang Q, Yan X, Zhang D, Coppola G, Sun W. Multi-proportion channel ensemble model for retinal vessel segmentation. Comput Biol Med 2019;111:103352.
[41] Marmanis D, Wegner JD, Galliani S, Schindler K, Datcu M, Stilla U. Semantic segmentation of aerial images with an ensemble of CNNs. ISPRS Ann Photogramm Remote Sens Spat Inf Sci 2016;3:473–80.
[42] Chollet F. Xception: deep learning with depthwise separable convolutions. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2017:1251–8.
[43] Tarvainen A, Valpola H. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. Proceedings of the 31st International Conference on Neural Information Processing Systems 2017:1195–204.
[44] Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: a large-scale hierarchical image database. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2009:248–55.
[45] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2017:4700–8.
[46] Zhong Z, Zheng L, Luo Z, Li S, Yang Y. Invariance matters: exemplar memory for domain adaptive person re-identification. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019:598–607.
[47] Chen L, Cai C, Chen V, Lu X. Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. BMC Bioinformatiocs 2016;17:9.
[48] Chen Y, Li Y, Narayan R, Subramanian A, Xie X. Gene expression inference with deep learning. Bioinformatics 2016;32:1832–9.
[49] Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 2015;12:931–4.
[50] Wang Y, Liu T, Xu D, Shi H, Zhang C, Mo YY, et al. Predicting DNA methylation state of CpG dinucleotide using genome topological features and deep networks. Sci Rep 2016;6:19598.
[51] Pärnamaa T, Parts L. Accurate classification of protein subcellular localization from high-throughput microscopy images using deep learning. G3 (Bethesda) 2017;7:1385–92.
[52] Berthelot D, Carlini N, Goodfellow I, Papernot N, Oliver A, Raffel C. MixMatch: a holistic approach to semi-supervised learning. Proceedings of the 33rd International Conference on Neural Information Processing Systems 2019:1–11.
[53] Odena A. Semi-supervised learning with generative adversarial networks. arXiv 2016;1606.01583.
[54] Zhai X, Oliver A, Kolesnikov A, Beyer L. S4L: self-supervised semi-supervised learning. Proc IEEE Int Conf Comput Vis 2019:1476–85.

[55] Noroozi M, Vinjimoor A, Favaro P, Pirsiavash H. Boosting self-supervised learning via knowledge transfer. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2018:9359–67.

[56] Laine S, Aila T. Temporal ensembling for semi-supervised learning. arXiv 2016;1610.02242.