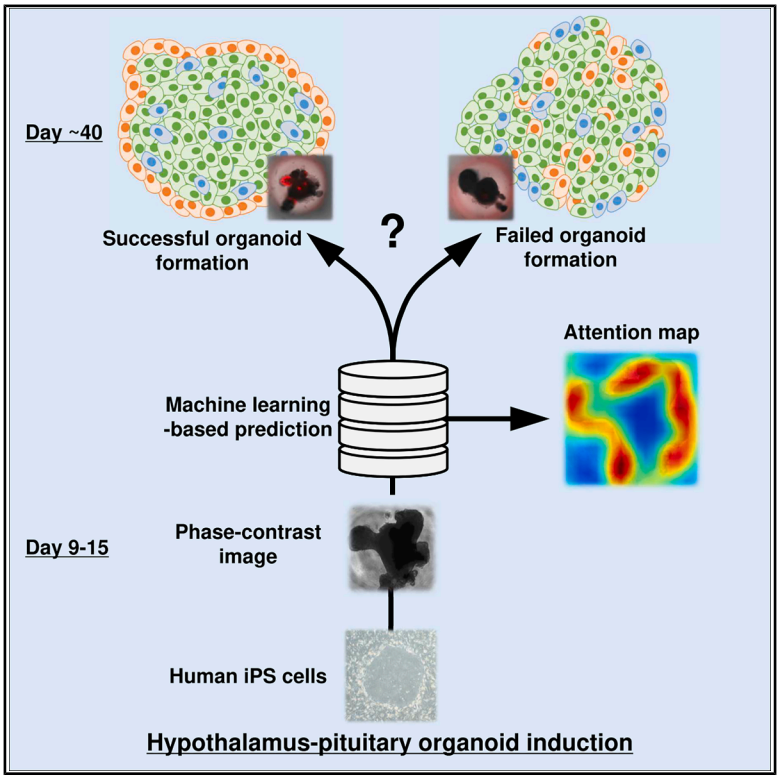
Cell Reports
Methods

Article

Prediction of the hypothalamus-pituitary organoid formation using machine learning

Graphical abstract



Authors

Ryusaku Matsumoto, Hidetaka Suga, Yutaka Takahashi, Takashi Aoi, Takuya Yamamoto

Correspondence

r.matsumoto@cira.kyoto-u.ac.jp (R.M.), takuya@cira.kyoto-u.ac.jp (T.Y.)

In brief

Matsumoto et al. establish a machine learning-based model capable of predicting the successful hypothalamus-pituitary organoid induction from human iPSCs based solely on phase-contrast images captured at early timing of differentiation.

Highlights

- Machine learning model can predict pituitary organoid formation from phase images
- The model outperforms human researchers and can be applied to other cell lines
- The model makes predictions based on morphological features of the organoid surface
- The morphological change mirrors the cellular localization pattern within the organoid

Matsumoto et al., 2025, Cell Reports Methods 5, 101119
August 18, 2025 © 2025 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.crmeth.2025.101119

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

Prediction of the hypothalamus-pituitary organoid formation using machine learning

Ryusaku Matsumoto,[1,2,3,*] Hidetaka Suga,[4] Yutaka Takahashi,[2,5] Takashi Aoi,[3] and Takuya Yamamoto[1,6,7,8,*]
[1]Center for iPS Cell Research and Application, Kyoto University, Kyoto 606-8507, Japan
[2]Division of Diabetes and Endocrinology, Department of Internal Medicine, Kobe University Graduate School of Medicine, Kobe, Hyogo 650-0017, Japan
[3]Division of Stem Cell Medicine, Kobe University Graduate School of Medicine, Kobe, Hyogo 650-0017, Japan
[4]Department of Endocrinology and Diabetes, Nagoya University Graduate School of Medicine, Nagoya, Aichi 466-8560, Japan
[5]Department of Diabetes and Endocrinology, Nara Medical University, Kashihara, Nara 634-8522, Japan
[6]Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto 606-8501, Japan
[7]Medical-Risk Avoidance Based on iPS Cells Team, RIKEN Center for Advanced Intelligence Project (AIP), Tokyo 103-0027, Japan
[8]Lead contact
*Correspondence: r.matsumoto@cira.kyoto-u.ac.jp (R.M.), takuya@cira.kyoto-u.ac.jp (T.Y.)
https://doi.org/10.1016/j.crmeth.2025.101119

MOTIVATION Organoid induction methods from pluripotent stem cells typically involve prolonged culture periods and often suffer from unstable induction efficiency, which hinders experimental productivity and widespread application. To address these problems, we developed a machine learning-based model capable of predicting the induction efficiency of the hypothalamus-pituitary organoids using only phase-contrast images of the early stages of their differentiation.

SUMMARY

Multi-cellular organoids are self-assembly aggregates that mimic biological functions and developmental processes of many tissue types *in vitro*. They are widely employed for disease modeling and functional studies. Hypothalamus-pituitary organoids can be generated through differentiation induction from pluripotent stem cells. However, their maturation is time consuming and labor intensive, and the quality of the resulting organoids can vary. Here, we developed a machine learning model capable of accurately predicting the successful generation of high-quality hypothalamus-pituitary organoids based solely on phase-contrast images captured during the early stage of differentiation. The model achieved an accuracy of 79% using images from organoids on day 9 to predict pituitary cell differentiation at day 40. Moreover, the computational approach identified the shape of the organoid surface as a critical determining factor that significantly affected the prediction. This model can help to enhance the efficiency of organoid induction experiments and illuminate the molecular mechanisms involved in hypothalamus-pituitary differentiation.

INTRODUCTION

The hypothalamus-pituitary axis is pivotal to multiple biological functions, including growth, reproduction, and the immune response. However, its regulatory mechanisms and the pathogenesis of the disorders affecting its functions remain largely unknown. Specifically, the lack of reliable human experimental models has impeded the analyses of this axis.

Recent advances in cell and tissue culture technologies have enabled the successful production and maintenance of *in vitro* self-organized organ-like structures known as organoids. Culture methods for various organoids, such as those representing the colon, pancreas, brain, and liver, have been described.[1–4] These technologies have begun to be utilized in disease models, drug testing, and regenerative medicine. Hypothalamus-pituitary organoids have been developed from mouse and human pluripotent stem cells (hPSCs),[5,6] which facilitate studies on the development and disorders of the human hypothalamus-pituitary axis. This technology also serves as a potential therapeutic tissue source for patients with hypopituitarism.[7,8] However, this differentiation method exhibits relatively high variability in induction efficiency across experiment batches and among wells, resulting in considerable heterogeneity in the quality of the induced hypothalamus-pituitary organoids.[9,10] In addition, because it typically takes over 2 months to induce pituitary hormone-producing cells,[6,9] the variability in organoid quality creates significant time and monetary costs for researchers. Consequently, along with developing a reproducible induction method,

This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

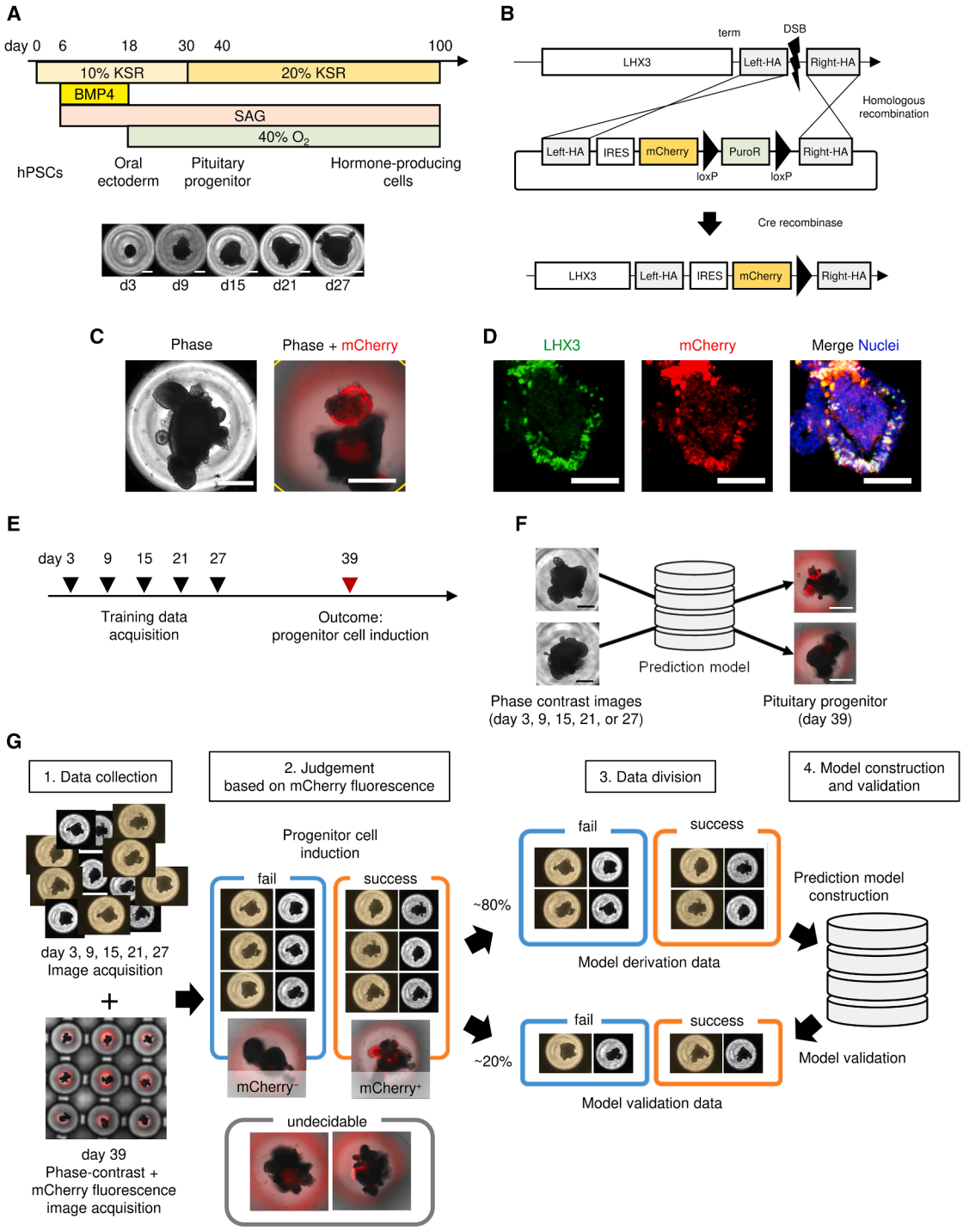


Figure 1. Establishment of a prediction model for the induction of hypothalamus-pituitary organoids from hiPSCs

(A) Overview of the methods used to induce hypothalamus-pituitary organoids from hPSCs. Scale bar represents 500 μm.

(B) Establishment of reporter iPSCs for LHX3^{+} pituitary progenitor cells using CRISPR-Cas9.

(C) Phase-contrast and mCherry fluorescence images of hypothalamus-pituitary organoid induced from established reporter iPSCs. Scale bar represents 500 μm.

(D) Immunostaining confirmed the co-localization of anti-LHX3 staining with mCherry fluorescence on day 39. Scale bar represents 100 μm.

(E) Phase-contrast images of organoids were obtained on days 3, 9, 15, 21, and 27 of induction. Then, pituitary progenitor cell induction was evaluated by detecting mCherry fluorescence on day 39.

(F) A schematic diagram of the prediction model for pituitary progenitor cells generated from only phase-contrast images of the organoids. Scale bar represents 500 μm.

(legend continued on next page)

it is beneficial to establish a means to predict organoid quality early in the induction process.

Machine learning (ML), a subfield of artificial intelligence, involves the development of algorithms that enable computers to perform specific tasks without requiring explicit programming. ML has been extensively refined and applied to numerous tasks across various fields, such as image or object identification, speech-to-text transcription, data analysis, and chatbot inquiry responses.[11] Supervised learning is a subtype of ML that utilizes a set of data with known classifications to train an algorithm to construct a model that fits the training data,[12] which, in turn, can predict the classification of unknown data. In cell biology, this approach has been used to develop classification models for various purposes, such as predicting cell types from only phase-contrast images,[13] distinguishing cells from the iPSCs of patients and healthy volunteers,[14] and analyzing cellular morphological changes during cellular senescence.[15] Several studies have already applied ML to organoid technologies to predict induction efficiencies from PSCs.[16,17] However, training and evaluation data were collected concurrently in those studies, thus limiting their predictive applicability. A separate study reported a model developed for predicting future retinal organoid induction efficiency from mouse embryonic stem cells (ESCs) based on early images.[18] Specifically, this study employed training data from day 5 to predict outcomes on day 9. While this represents progress, the relatively short interval between training and prediction limits its utility for applications such as hypothalamus-pituitary organoids, which require long-term culture periods for differentiation, as described above.

In this study, we constructed a prediction model of hypothalamus-pituitary organoid quality determined as the capacity to generate pituitary progenitor cells by training an algorithm with phase-contrast images of organoids during an early phase in the induction process. The model could predict organoid quality 1 month before their maturation with an accuracy of 79% and an area under the curve (AUC) value of 0.83 using only images from the early stages of organoid differentiation. Moreover, our detailed analysis has revealed that morphological changes of the organoid surface at the early phase of induction are highly associated with organoid quality after maturation, suggesting that a prediction model is also beneficial in understanding the molecular basis for ensuring the quality. This method promises to substantially reduce both material and labor costs in hypothalamus-pituitary organoid studies and will significantly help advance this research field. Furthermore, our approach described here is potentially applicable to other types of organoids, leading to a broad impact across related fields.

RESULTS

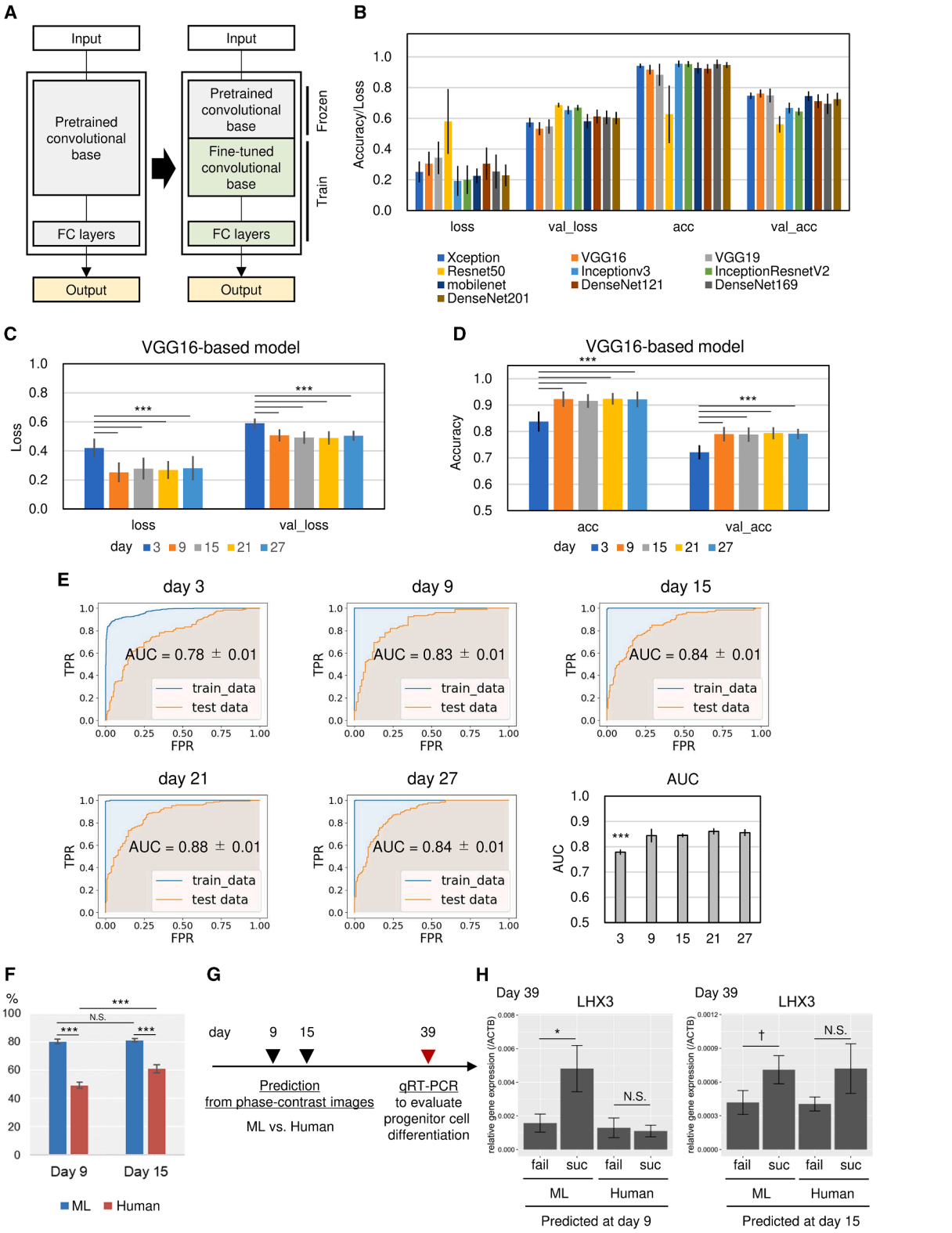
The hypothalamus-pituitary organoid induction method can recapitulate the interaction between the hypothalamus and oral ectoderm, which is essential for pituitary development in rodent models.[19] During the induction of human hypothalamus-pituitary organoids, it takes approximately 3 months for the organoids to acquire hormone-producing cells (Figure 1A). The most crucial event in the induction process is the differentiation of pituitary progenitor cells from the oral ectoderm, which requires interaction with the hypothalamus.[5,20] It has been shown that once sufficient differentiation into pituitary progenitor cells is successful, high-quality organoids capable of hormone secretion can be produced.[6] However, differentiation into pituitary progenitor cells still takes about 40 days. This study aimed to construct a model that can predict the successful development of high-quality organoids at a stage before the onset of marker expression.

Data collection

First, we established a reporter iPSC line for the pituitary progenitor cell marker *LHX3* to monitor pituitary progenitor cell differentiation. An IRES-mCherry sequence was inserted downstream of the *LHX3* locus in hiPSCs by CRISPR-Cas9 (Figure 1B) so that the fluorescence signal could be detected on the surface of the induced organoids without immunostaining (Figure 1C). Next, we confirmed that LHX3 and mCherry signals co-localized in organoids (Figure 1D). In addition, to assess whether the insertion of the reporter gene into iPSCs affects the morphology of the induced hypothalamus-pituitary organoids, we compared typical longitudinal images of organoids generated from LHX3-reporter iPSCs and parental iPSCs during differentiation (Figure S1). No apparent morphological differences were observed between the two groups, indicating that LHX3-mCherry iPSCs can monitor LHX3 expression without affecting organoid morphology.

To prepare training data, we performed 17 independent induction experiments using the established reporter iPSC line in a 96-well plate. Phase-contrast images of the induced organoids were obtained on days 3, 9, 15, 21, and 27 of induction, and mCherry fluorescence images were acquired on day 39 (Figures 1A and 1E). To build a prediction model for later LHX3 expression based on early phase-contrast images (Figure 1F), we manually classified each well as "fail," "success," or "undecidable" according to the absence, presence, or ambiguity of mCherry fluorescence in day 39 images (Figure 1G). "Undecidable" refers to cases where it was difficult to distinguish between reporter fluorescence and non-specific autofluorescence. We have excluded the "undecidable" images from further model constructions. Across the 17 induction experiments, an average of 53% ± 5% of the induced organoids was LHX3 positive ($LHX3^+$). Finally, the number of phase-contrast images used for model construction are as follows: day 3, 1,316 images; day 9, 935 images; day 15, 1,407 images; day 21, 1,407 images; and day 27, 1,393 images. Phase-contrast images were systematically preprocessed before model construction (Figure S2A).

(G) Overview of the development of the prediction model. (1) Phase-contrast images of organoids were obtained for prediction samples, and mCherry fluorescence images were obtained on day 39 to evaluate the pituitary progenitor cell induction. (2) The obtained phase-contrast images were labeled as "fail" or "success" based on the absence or presence of mCherry fluorescence, respectively. (3 and 4) The labeled data were divided into two groups: ~80% of the data were used for model derivation and ~20% were used for model validation.



(legend on next page)

CellPress
OPEN ACCESS

After preprocessing, 80% of the input images were used for model construction and 20% for model validation (Figure 1G).

Model construction

Using the obtained phase-contrast images, we established a convolutional neural network (CNN) model to attempt to predict LHX3$^+$ pituitary progenitor cell differentiation (Figure 1F). First, images collected on day 15 of the induction process were used for preliminary model constructions; however, when tested with the validation data, the accuracy was only approximately 0.6 (Figure S2B). This result may have been due to the paucity of our dataset and the similarity between "fail" and "success" data (Figure S3). We thus employed a model construction technique, fine-tuning, to improve the prediction accuracy (Figure 2A). Ten pre-trained models trained on ImageNet data (https://www.image-net.org/) (Xception, VGG16, VGG19, Resnet50, InceptionV3, InceptionResNetV2, Mobilenet, Densenet121, Densenet169, and Densenet201) were fine-tuned using our dataset. The prediction accuracy of the developed models was variable, and the VGG16-based model showed the best performance, followed by the VGG19-based model (Figures 2B and S2C), although the difference did not reach statistical significance. The validation data accuracy values of VGG16- and VGG19-based models were 0.76 ± 0.03 and 0.75 ± 0.05, respectively. Therefore, VGG16 was selected for further investigation, as the two models showed no statistical difference. The VGG16-based model was subsequently trained using phase-contrast images of organoids from datasets collected on days 3, 9, 15, 21, and 27 (Figures 2C and 2D). Data from days 9–27 showed comparable prediction accuracy, whereas data from day 3 showed lower accuracy (Figures 2C and 2D). Receiver operating characteristic (ROC) curve analyses of the VGG16 and VGG19 data also showed comparable AUC values for each dataset from days 9–27, but day 3 data yielded lower AUC values (Figures 2E and S2D). Because it is beneficial to predict the differentiation efficiency of organoids at an earlier stage, we focused on the day 9 and day 15 models (validation data accuracy of both day 9- and day 15-based models was 0.79 ± 0.03) (Figure 2D).

Comparison with human predictions and utility on other iPSC lines

To investigate whether the constructed model is superior to human prediction of successful organoid differentiation, we compared the predictive accuracy for data not employed in model constructions between the ML model and 4 human researchers with extensive experience (>100 times) of the induction method. As a result, the ML model outperformed human researchers for both day 9 and day 15 datasets (Figure 2F). Whereas human researchers showed worse prediction accuracy for images from day 9 than day 15, the ML model showed comparable prediction accuracy. These results showed that the ML model was able to predict differentiation efficiency earlier during the induction process than experienced human researchers, demonstrating the superiority of ML over human assessments.

We also examined whether the insertion of the *LHX3* reporter influences the prediction results. Organoids were induced from the parental non-reporter iPSC line and evaluated on days 9 and 15 both by ML and a human researcher. On day 39, quantitative reverse-transcription PCR (RT-qPCR) was used to examine the expression level of *LHX3* in the organoids as a marker of pituitary progenitor cell differentiation (Figure 2G). Organoids categorized as "success" by the ML model on days 9 and 15 had higher expression levels of *LHX3* than those classified as "fail." By contrast, a trained researcher could predict successful organoid induction only on day 15 (Figure 2H), demonstrating that the constructed model works not only with a non-reporter iPSC line but also outperforms human prediction. Moreover, we evaluated the utility of the developed model on additional iPSC lines that were different from the iPSC line used for its construction (201B7). Organoids were induced from two additional iPSC lines (HC06 and 409B2) and classified

Figure 2. Development of a hypothalamus-pituitary organoid prediction model

(A) Schematic diagram of the fine-tuning of pre-trained models. The first one-third of the convolutional layers were frozen, and the remaining two-thirds and fully connected (FC) layers were retrained using our dataset.

(B) Accuracy and loss of pre-trained models fine-tuned with day 15 organoid images (n = 10). ANOVA followed by Tukey's post hoc test showed the Resnet50-based model with lower accuracy and higher loss than other models. The VGG16-based model showed the best prediction accuracy and the lowest loss for test data, followed by the VGG19-based model. Val_loss, loss for validation dataset images; val_acc, accuracy for validation dataset images. Data are expressed as mean ± SD.

(C) Loss of the VGG16 model fine-tuned with images from days 3, 9, 15, 21, or 27 of the induction process. Day 9, 15, 21, and 27 models showed comparable loss, whereas the day 3 model was higher (n = 20–50). Data are expressed as mean ± SD.

(D) Accuracy of the VGG16 model fine-tuned using organoid images from days 3, 9, 15, 21, or 27 of the induction process. Day 9, 15, 21, and 27 models showed comparable prediction accuracy, whereas the day 3 model was lower (n = 20–50). Data are expressed as mean ± SD.

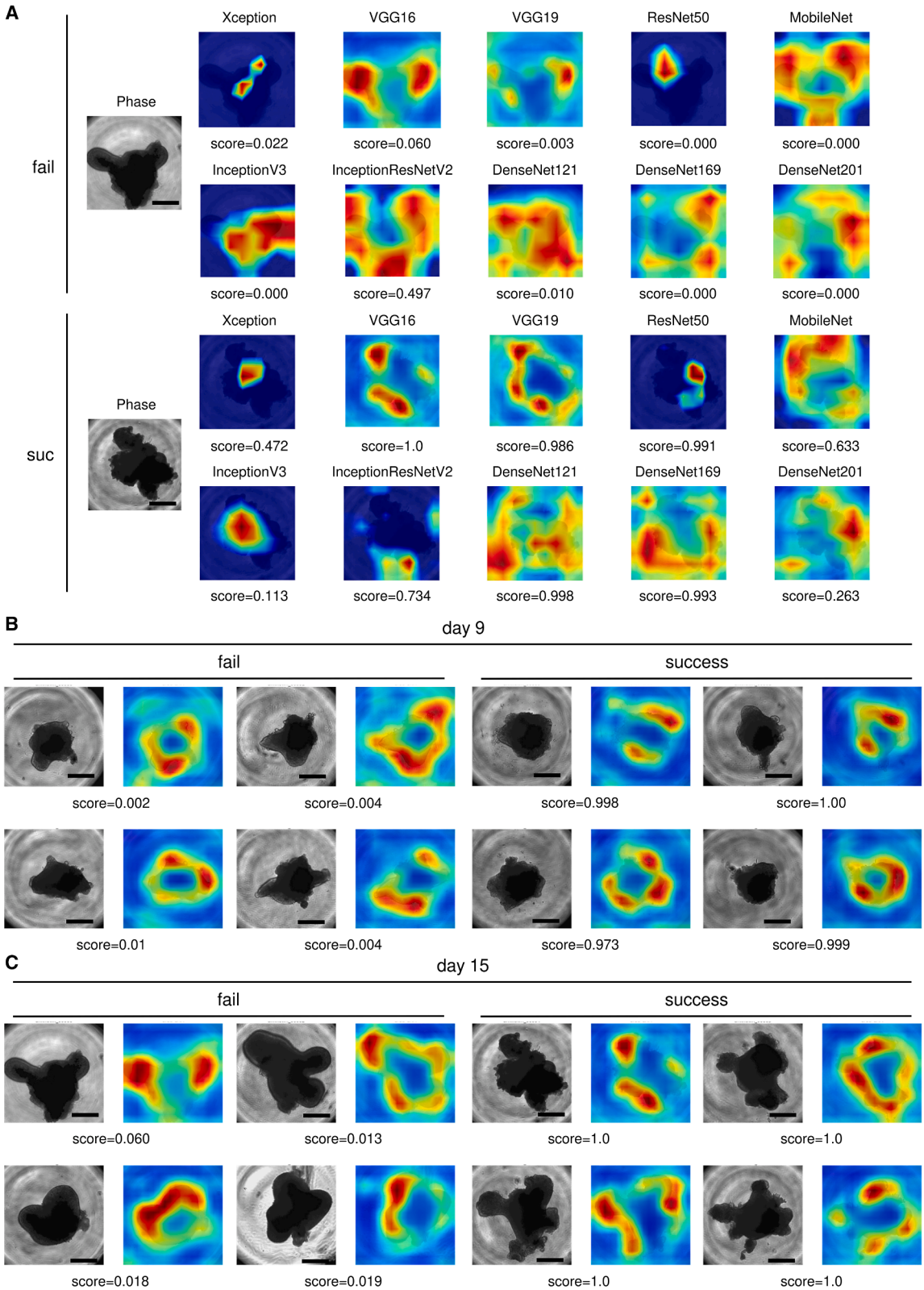
(E) ROC curves and AUC values of the VGG16 model fine-tuned with organoid images from days 3, 9, 15, 21, or 27. Day 9, 15, 21, and 27 models had comparable AUC values, whereas the day 3 model was low (n = 15). TPR, true positive rate; FPR, false positive rate. Data are expressed as means ± SD.

(F) Comparison of the prediction accuracy of ML and 4 human researchers (n = 3). Three hundred organoid images from days 9 and 15 were evaluated. ML outperformed human researchers for data from both days 9 and 15. Although human researchers could not predict the induction efficiency for data from day 9, the ML model showed comparable prediction accuracy for data from both days 9 and 15. Data are expressed as mean ± SEM.

(G) Overview of ML and human predictions of organoid induction outcomes from non-reporter iPSCs. Both ML model and human experimenter predicted "fail" or "success" using organoid images from days 9 and 15, with samples collected on day 39 for RT-qPCR analysis of the actual differentiation efficiency into pituitary progenitor cells.

(H) Comparison of the accuracy of ML and human predictions, as assessed by evaluating the mRNA expression level of the pituitary progenitor marker *LHX3* (day 9, n = 6; day 15, n = 5–6). On day 9, ML showed better prediction accuracy than the human experimenter, whereas, on day 15, ML and the human showed almost comparable prediction accuracy. Data are expressed as mean ± SEM.

†$p < 0.1$, *$p < 0.5$, and ***$p < 0.001$.



(legend on next page)

as "fail" or "success" on days 9 and 15 using the model (Figure S4A). No apparent morphological differences were observed in the induced organoids (Figure S4B). Subsequently, *LHX3* expression was assessed by RT-qPCR on day 39. In HC06, RT-qPCR showed that *LHX3* expression was higher in organoids classified as "success" than those classified as "fail" on both days 9 and 15 (Figure S4C). Similarly, in 409B2, there was a trend toward increased *LHX3* expression in "success" samples compared to "fail" samples, although the difference did not reach statistical significance. These results suggest that, despite some variation among cell lines, the prediction model developed in this study can be applied to various iPSC lines.

Identification of the regions that contributed to the model's prediction

Despite the beneficial performance of our CNN-based prediction model, it is unclear how the model classifies the images into predefined categories. Understanding whether the constructed model utilizes the organoid state as a predictive factor is crucial for assessing the biological validity of the model. Several approaches have been proposed to investigate how CNNs perform classifications.[21,22] We utilized the gradient-weighted class activation mapping Grad-CAM method to generate attention maps highlighting the parts of input images important to the classification.[23] First, we generated attention maps of 10 developed fine-tuned models for typical "fail" and "success" images from day 15 (Figure 3A). These heatmaps showed that the organoid surfaces were critical for the classification in the models with better prediction accuracy, such as VGG16 and VGG19 (Figures 3A and 3B). By contrast, models with poor prediction accuracy, such as ResNet50, InceptionV3, and InceptionResNetV2, failed to recognize organoid morphologies (Figure 3A). Representative Grad-CAM heatmaps of the VGG16-based model merged with input images from days 9 and 15 are shown in Figures 3B and 3C. These heatmaps confirmed that the surfaces of induced organoids played a critical role in distinguishing between the "fail" and "success" categories on both days 9 and 15 (Figures 3B and 3C). Next, we investigated the morphological characteristics of regions highlighted by Grad-CAM in "fail" and "success" organoids. In day 9 "fail" images, Grad-CAM attention maps indicated that large budding areas with smooth or extensively rough surfaces contributed to "fail" predictions (Figure 4A). Immunostaining of cell-type-specific markers suggested that large budding areas with smooth surfaces were associated with differentiation into neural tissue lacking an oral ectoderm layer, while extensively rough surfaces composed of non-specific cells that were negative for representative marker genes of cell types were typically found in the organoid and exhibited non-specific autofluorescence (Figure 4A). Conversely, typical features of day 9 "success" images included small budding areas and slightly rough surfaces (Figure 4B). Immunostaining revealed that these surface features of "success" organoids were associated with PITX1$^+$ oral ectoderm layers (Figure 4B). In day 15 "fail" images, Grad-CAM highlighted extensively dark budding areas with smooth surfaces (Figure 4C). These areas were inferred to represent hypothalamus or neural tissues lacking oral ectoderm layers (Figure 4C), similar to the observations on day 9. In day 15 "success" images, slightly rough surfaces and duplex structures in the budding areas were critical to "success" predictions (Figure 4D). These duplex structures were associated with the hypothalamus and oral ectoderm layers (Figure 4D). These observations suggested that the developed ML model identified cell types within organoids by recognizing morphological changes in phase-contrast images. Consistent with previous reports, these findings also indicate that oral ectoderm induction and its interaction with hypothalamic tissues are vital to the differentiation of LHX3$^+$ pituitary progenitor cells.

Quality and quantity of training data

When constructing a similar prediction model, controlling the quality and quantity of training data during data collection is required for building a good prediction model. Therefore, we investigated the necessary image size and the number of training samples required for model construction. To study the effects of image size, we resized original images from day 15 into 50 × 50, 100 × 100, 200 × 200, 300 × 300, 400 × 400, and 600 × 600 pixels and fine-tuned VGG16 with these training images (Figure S5A). As expected, the larger the training images, the higher the prediction accuracy of developed prediction models (Figures S5B and S5C). However, as expected, larger images take longer for model construction (Figure S5D). Given that accuracy reached a plateau at around 200 × 200 or 300 × 300 pixels (Figures S5B and S5C), it suggested that training images with such resolutions are required for model construction to evaluate such organoid models.

Next, to investigate the effects of training dataset size on prediction performance and time required for model construction, VGG16 was fine-tuned with 100, 200, 400, 600, 800, and 1,000 samples of our day 15 training data. As expected, increasing the amount of training data resulted in lower loss and higher prediction accuracy (Figures S5E and S5F). Similar to image size, more training samples required more time for model construction (Figure S5G), with prediction accuracy plateauing at 600–800 training samples (Figure S5F).

Lastly, we investigated the effects of focal position within organoids during imaging on prediction outcomes. In this study, the focus was manually adjusted on the periphery of organoids. To determine the impact of focal position on prediction performance, we captured organoid images after shifting the objective lens between +1,500 and −1,500 μm from the original position at 500 μm intervals. We calculated prediction scores and generated attention maps from these images (Figures S6A and S6B). When the focal position was shifted, Grad-CAM failed to recognize the surface of organoids (Figures S6A and S6B), and both

Figure 3. Grad-CAM-mediated visualization of the input image regions that were important for the decision-making of the developed prediction model

(A) Phase-contrast images and Grad-CAM heatmaps generated by applying fine-tuned models to "fail" and "success" data from day 15.

(B and C) Phase-contrast images and Grad-CAM heatmaps generated by applying VGG16-based model to "fail" and "success" data from days (B) and 15 (C). The organoid surface was identified as important regions contributing to the prediction under all conditions. Scale bar represents 500 μm.

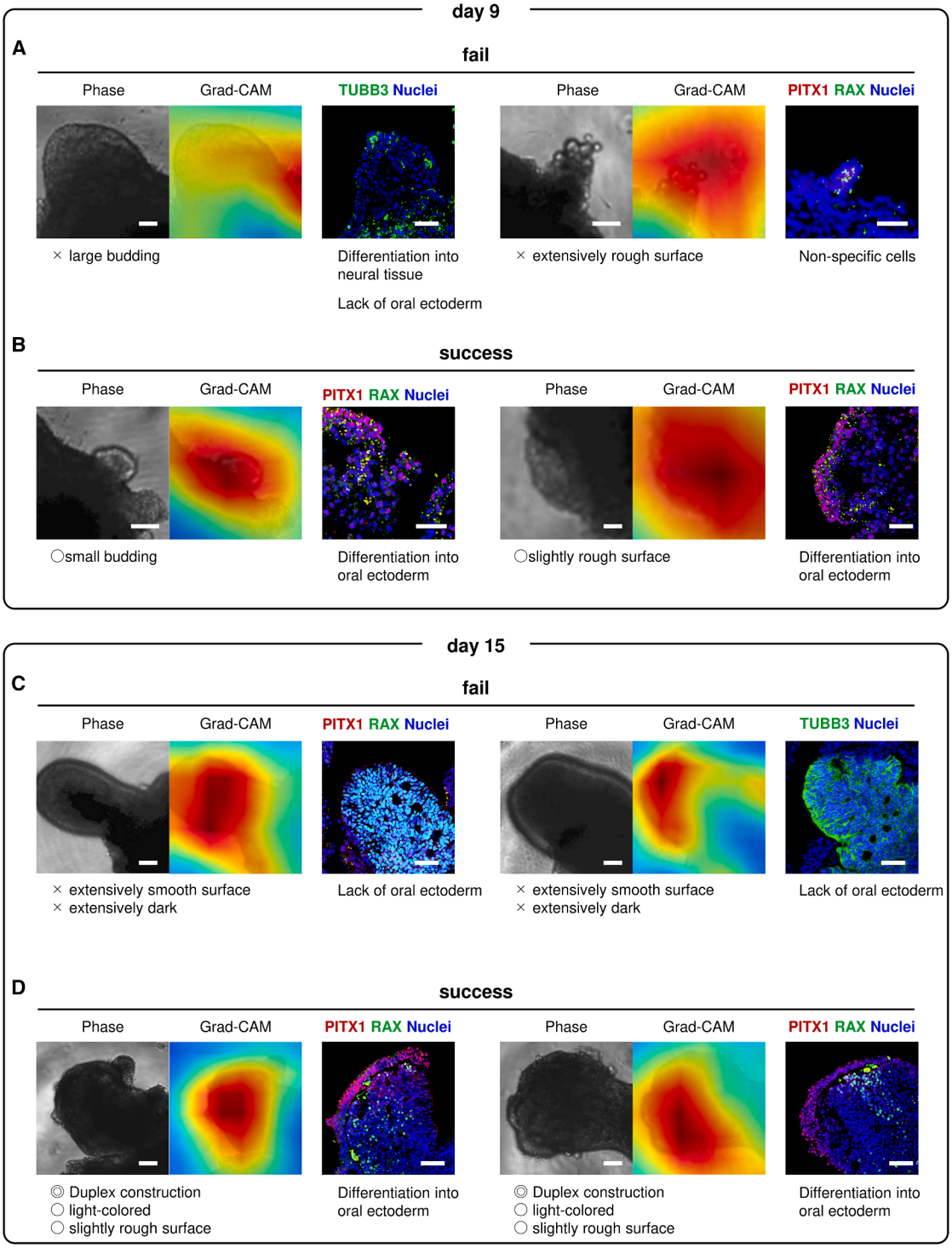


Figure 4. Morphological features contributing to predictions, as indicated by Grad-CAM attention maps

Left two images show morphological characteristics of highlighted regions by Grad-CAM in "fail" and "success" images from days 9 (A) and (B) and 15 (C) and (D). On day 9, "large budding" and "too rough surface" of organoids were associated with "fail" inductions. By contrast, "small budding" and "slightly rough surface" were associated with "success" inductions. On day 15, "extensively smooth surface" and "extensively dark" features were associated with "fail" inductions. By contrast, "duplex construction," "light-colored," and "slightly rough surface" of organoids were associated with "success" inductions. Right image shows the immunostaining of cell-type-specific markers in regions with similar morphological characteristics. These immunostained samples are not identical to phase-contrast image samples. Scale bar represents 100 μm.

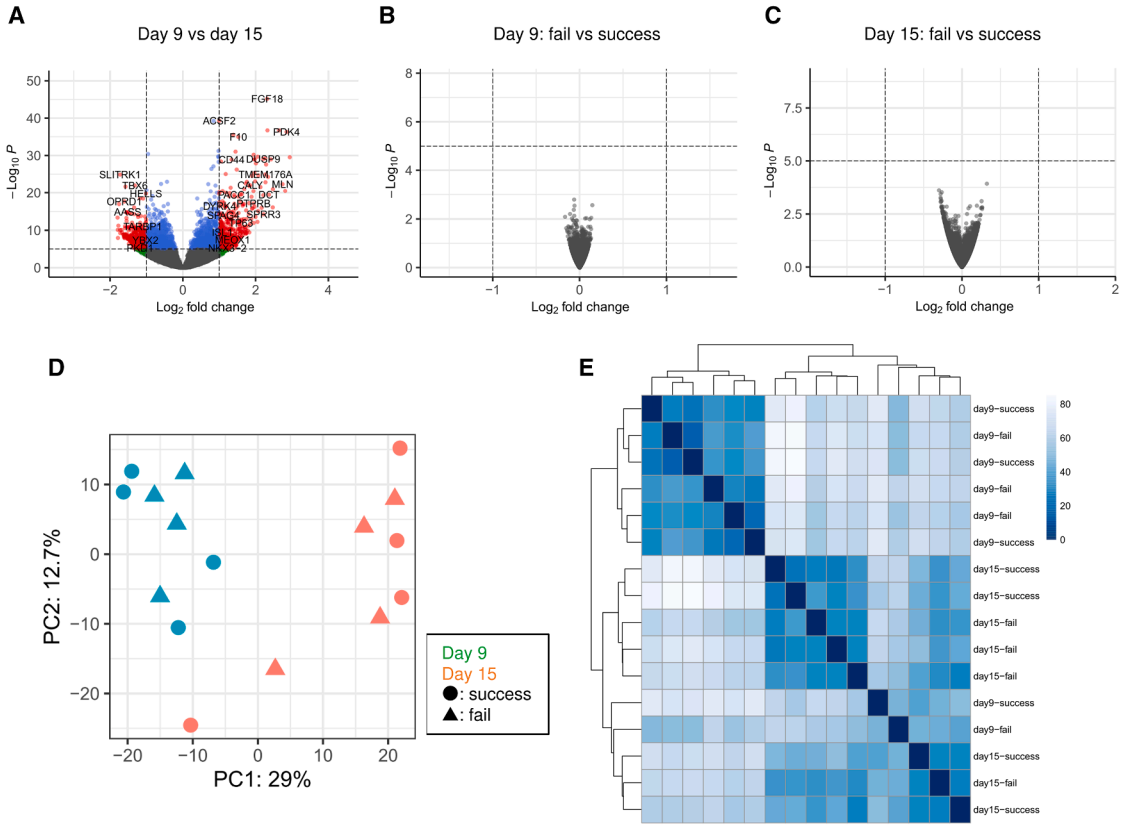


Figure 5. Comparisons of "fail" and "success" hypothalamus-pituitary organoids

(A) Volcano plot comparing hypothalamus-pituitary organoids from days 9 and 15.

(B and C) Volcano plots comparing "fail" and "success" organoids from days 9 (B) and 15 (C).

(D) Principal-component analysis of "fail" and "success" organoids from days 9 and 15.

(E) Hierarchical clustering analysis of "fail" and "success" organoids from days 9 and 15. Sample-to-sample distances calculated by the *dist* function in the R package pheatmap are shown as a heatmap.

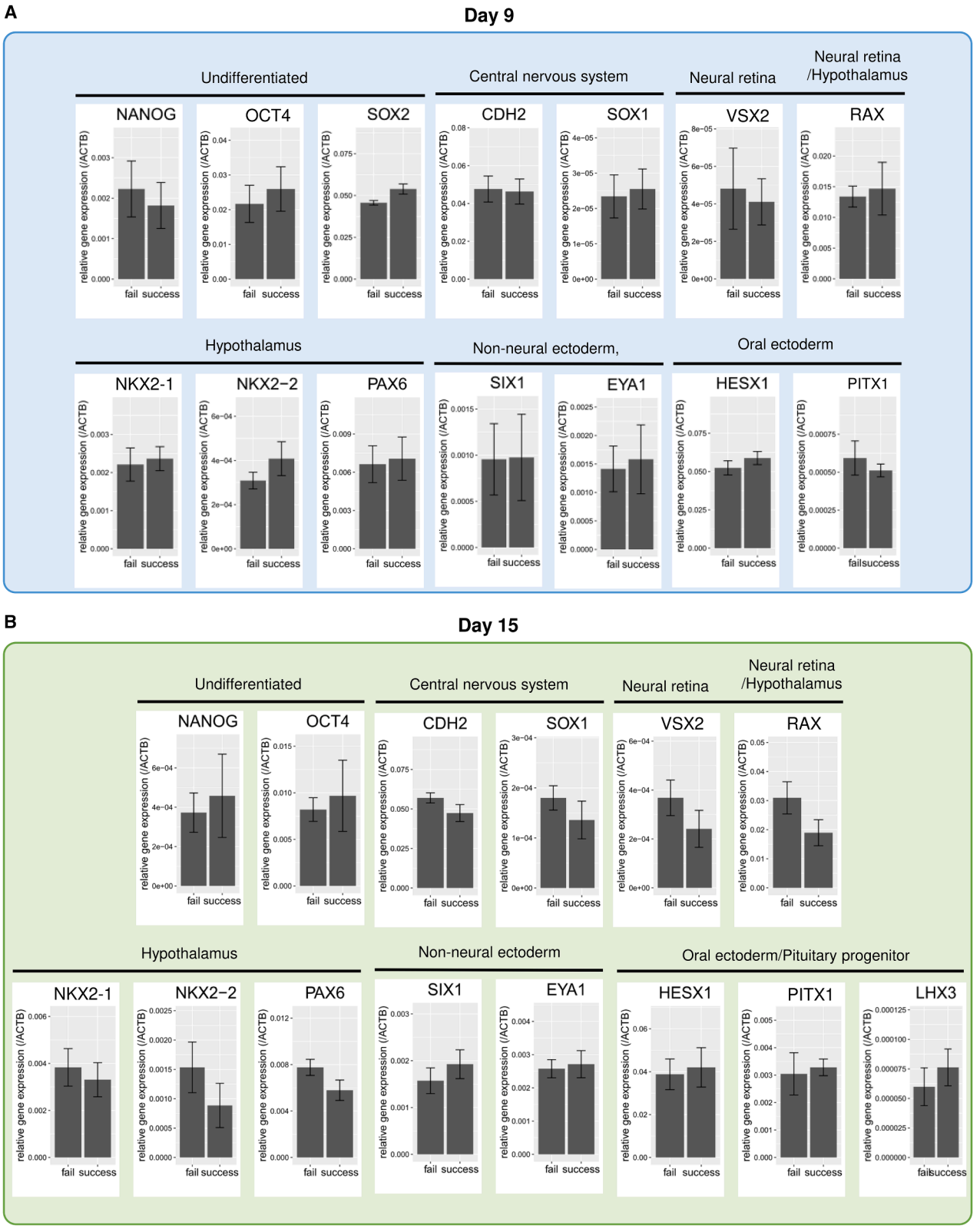
organoids predicted as "success" and "fail" when focused on the center were consistently predicted as "fail" under shifted focal positions (Figures S6C and S6D). This trend became more pronounced with increasing deviation of the focal distance from the center (Figure S6C).

These results will be helpful when constructing a prediction model for other types of organoids. In particular, the focal position may impact prediction outcomes significantly, thus necessitating careful data collection.

Analysis of "fail" and "success" organoids

Although Grad-CAM visualization demonstrated organoid topography as crucial for progenitor cell prediction, the molecular basis underlying these morphological differences was unclear. To address this, we performed RNA sequencing (RNA-seq) on "fail" and "success" organoids to identify differences in their transcriptional profiles. A comparison of day 9 and day 15 organoids revealed significant transcriptional differences (Figure 5A; Table S1). Gene Ontology (GO) analysis of the top 200 upregulated genes during induction revealed enrichment of gene sets related to embryonic development, including those associated with camera-type eye, forebrain, and digestive system development (Figure S7A; Table S2). Notably, genes with the GO term "gland development" were also enriched (Figure S7A), encompassing genes known to play critical roles in pituitary gland development, such as *PITX1*, *SHH*, *WNT1*, and *WNT10A*. These results suggest that our pituitary organoid induction process recapitulates key aspects of the pituitary developmental process.

A comparison between "fail" and "success" organoids from days 9 and 15 indicated similar gene expression profiles, as shown in volcano plots (Figures 5B and 5C; Tables S3 and S4). In the principal-component analysis (PCA) and hierarchical clustering analysis, while samples from days 9 and 15 formed distinct clusters, "fail" and "success" organoids did not separate within each time point (Figures 5D and 5E). We also evaluated the normalized counts of representative marker genes for different types of cells, including undifferentiated cells, non-neural ectoderm, oral ectoderm, hypothalamus, neural, and neural retina cells, which are known to be present in the organoids.[6] Marker genes for undifferentiated (*NANOG*, *POU5F1*, and *SOX2*) and early pituitary (*HESX1*) markers were downregulated during the induction (Figure S11B). Central nervous system (*CDH2* and *SOX1*), cranial placode (*SIX1* and *EYA1*), neural retina (*RAX*,



(legend on next page)

OTX2, and *VSX2*), ventral hypothalamus (*NKX2-1* and *NKX2-2*), and oral ectoderm (*PITX1* and *PITX2*) cells were upregulated (Figure S7B). However, a dorsal hypothalamus marker (*PAX6*) showed no change (Figure S7B). Importantly, no significant differences were observed among these marker genes between "fail" and "success" organoids on either day 9 or 15. This lack of significant differences in gene expression was further confirmed by RT-qPCR (Figures 6A and 6B). Additionally, our immunostaining analysis showed that PITX1$^+$ oral ectoderm cells were induced on the organoid surface on days 9 and 15 (Figures 7A and 7B) and RAX$^+$/NKX2-1$^+$ ventral hypothalamus cells were induced inside organoids (Figure 7B). This structure was similar in "fail" and "success" organoids. Notably, however, we observed that cell populations not directly related to pituitary development, such as TUBB3$^+$ neural cells and VSX2$^+$ neural retina cells, were more frequently localized to the surface of "fail" organoids compared to "success" organoids on day 15, although this change was not apparent on day 9 (Figures 7C and 7D). Hot areas identified in attention maps might correspond to these regions (Figure 4C). These findings imply that the spatial organization of specific cell types within organoids may play a critical role in successful induction into pituitary cells. Our RT-qPCR analysis also showed that central nervous system (*CDH2* and *SOX1*) and neural retina (*RAX* and *VSX2*) markers tended to be higher in "fail" organoids than "success" organoids on day 15 (Figure 6B), but these differences did not reach statistical significance. In summary, subtle differences, including the composition and spatial organization of cell populations on the surface of early organoids, as identified by the attention map, may contribute to the generation of high-quality organoids. These implications highlight the potential for the attention map to recognize subtle features that influence future organoid quality.

DISCUSSION

In this study, the ML-based model for morphological imaging was capable of predicting the efficiency of hypothalamus-pituitary organoid differentiation during the early stages of induction. Generally, fewer than 100 organoid images are collected in a typical organoid induction experiment using 96-well plates, thus making it difficult to prepare tens of thousands of datasets for model training. Here, even with limited training samples, we found that fine-tuning a pre-trained image recognition model could improve prediction accuracy. Among the 10 pre-trained models tested, some models, such as VGG16, VGG19, and mobilenet, showed better prediction accuracy for our dataset, whereas other models, such as Resnet50, InceptionV3, and InceptionResnetV2, showed poor performance (Figure 2B). However, the optimal model for fine-tuning or transfer learning is supposed to vary depending on several factors, including the training dataset, sample size, CNN model structure, and training settings. The established model could predict the differentiation efficiency into pituitary progenitor cells 1 month in advance in a non-invasive manner using only phase-contrast images at early culture stages. The prediction models in this study were designed to output binary classifications ("fail" and "success"). However, depending on experimental needs, the cutoff threshold can be adjusted to modify sensitivity and specificity or to include an "undecidable" category for a ternary output (e.g., scores 0–0.33 = fail, 0.33–0.66 = undecidable, and 0.66–1.0 = success). This model is expected to not only reduce labor and experimental costs of organoid induction by improving the efficiency of studies but could also be applied to many other applications, such as the optimization of culture conditions, the selection of cell lines with higher differentiation efficiency, and the identification of factors involved in hypothalamus-pituitary differentiation. In addition, we developed prediction models using the learning images obtained by different microscopes and exposure settings after the normalization described above. If we used learning data captured in a single condition, prediction efficiency would be improved. However, we intentionally utilized standardized images from diverse settings to ensure that our models could be generalized to function effectively across various conditions, including those beyond our own experimental setup. Although studies have previously utilized ML to predict cell types and differentiation efficiency, these predictions were typically made using data collected at almost the same time point as the outcomes being predicted. In contrast, this study demonstrated the potential of an ML-based approach to predict future differentiation efficiency, thus addressing the challenge of long-term forecasting.

CNN models are potent tools for classification, such as image recognition, and are now widely utilized in cell biology. However, understanding how CNN models classify subjects or datasets remains challenging. Recently, explainable artificial intelligence (XAI) approaches, including Grad-CAM, have become standard methods to analyze the basis for CNN-based classification. Here, we found that specific features of the organoid surface were important contributors to the prediction of a developing organoid as "fail" or "success," suggesting that the formation of the oral ectoderm outer layer is critical to the successful induction of hypothalamus-pituitary organoids. Indeed, immunostaining revealed that some "fail" organoids lacked oral ectoderm on the organoid surface (Figure 7D), consistent with our observations that neural lineage marker genes (*CDH2*, *SOX1*, *VSX2*, and *RAX*) were downregulated, as oral ectoderm lineage marker genes (*SIX1*, *EYA1*, *HESX1*, *PITX1*, and *LHX3*) were upregulated in "success" organoids compared to "fail" organoids, despite not reaching statistical significance (Figure 6B).

Our ML-based model could predict the induction efficiency with sufficient accuracy from day 9 of organoid development.

Figure 6. Cellular compositions of "fail" and "success" hypothalamus-pituitary organoids

The expression levels of undifferentiated cell (*NANOG*, *OCT4*, and *SOX2*), central nervous system (*CDH2* and *SOX1*), neural retina (*VSX2* and *RAX*), hypothalamus (*RAX*, *NKX2-1*, *NKX2-2*, and *PAX6*), non-neural ectoderm (*SIX1* and *EYA1*), oral ectoderm (*HESX1* and *PITX1*), and pituitary progenitor (*LHX3*) markers on days 9 (A) and 15 (B) (n = 4). On day 9, marker gene expression was comparable between "fail" and "success" organoids. By contrast, on day 15, central nervous system, neural retina, and hypothalamus markers were upregulated in "fail" organoids compared to "success" organoids, although the levels of all other markers were comparable. Data are expressed as means ± SEM.

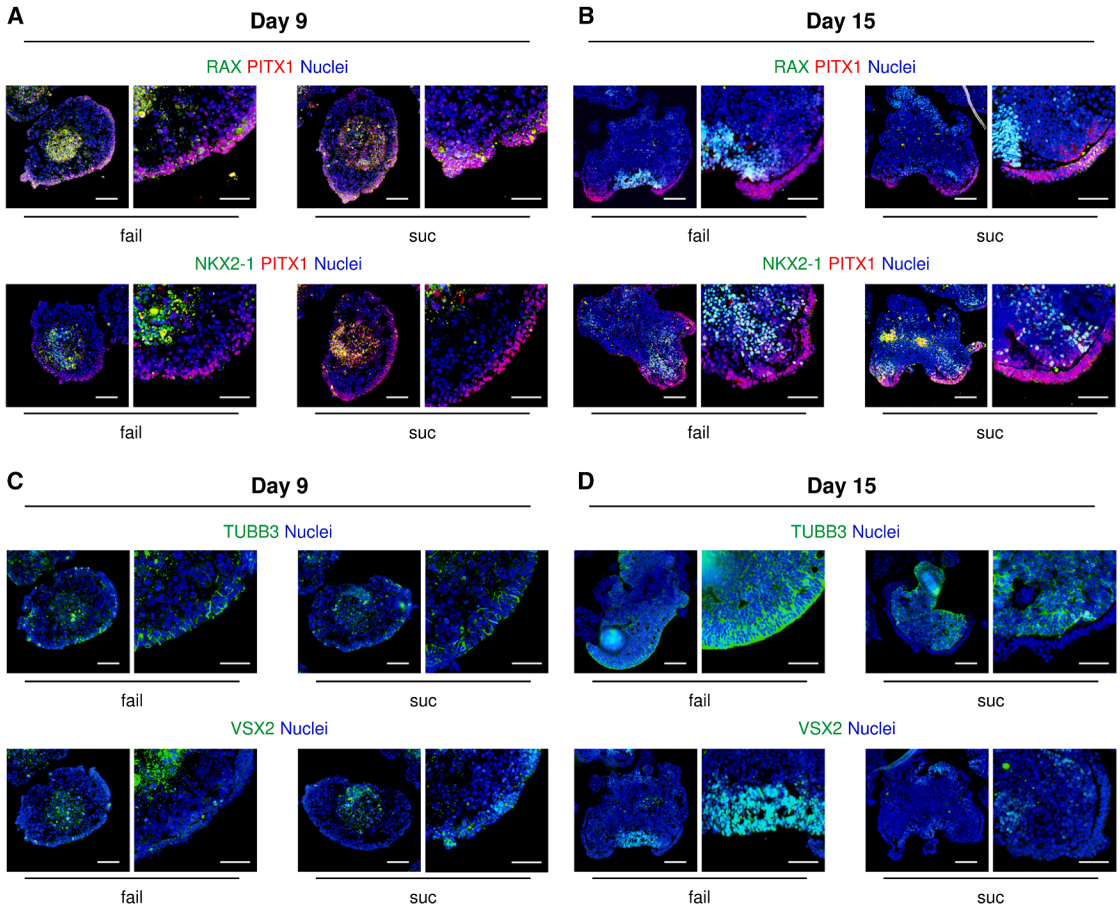


Figure 7. Spatial localization of cell types within hypothalamus-pituitary organoids

(A and B) Immunostaining of hypothalamus (RAX and NKX2-1) and oral ectoderm markers (PITX1) on day 9 (A) and 15 (B) organoids. On day 9, the outer layer of the organoid differentiated into oral ectoderm cells, whereas on day 15, the inside of the organoid differentiated into hypothalamus cells. There were no significant differences between "fail" and "success" organoids. Scale bar represents 100 μm.

(C and D) Immunostaining of a central nervous system (TUBB3) and neural retina markers (VSX2) on days 9 (C) and 15 (D) in labeled organoids. At day 15, the expression levels were increased on the surface of organoids predicted to "fail," while there were no significant differences between their levels in "success" and "fail" on day 9. Scale bar represents 100 μm.

However, model accuracy was insufficient on day 3, at the embryoid body formation stage (Figure S1). This finding is likely related to the fact that BMP4 and SAG were added from day 6 onwards to drive the development of hypothalamic and pituitary cells during the organoid induction process (Figure 1A). Prediction on day 3 was difficult because organoids were still at the stage of embryoid body formation. Although there was no statistically significant difference, on day 15, "fail" organoids showed higher expression of neural genes, and "success" organoids tended to have increased markers of the oral ectoderm lineage (Figure 6B). However, differences in gene expression were unclear on day 9 (Figure 6A). Nonetheless, our prediction model demonstrated comparable accuracy in forecasting pituitary differentiation on days 9 and 15 based on observable morphological differences, indicating that morphological changes precede lineage specifications in this organoid system. During embryonic pituitary development, interactions between the hypothalamus and oral ectoderm are required for pituitary cell specification.[19] This induction method for hypothalamus-pituitary organoids recapitulates these tissue interactions, highlighting the importance of spatial relationships between cell types for successful induction. Hence, a balance of induced hypothalamus and oral ectoderm cells and their interactions at the tissue level might contribute to the induction process. Our immunostaining experiments for central nervous system and neural retina markers (Figures 7C and 7D) imply that the mis-localization of these cell populations, which are not directly involved in hypothalamus-pituitary development, limits pituitary induction. Our developed models might predict these cell-type-specific localizations from morphological changes, especially on the organoid surface.

In conclusion, the model described here could predict the differentiation efficiency of hypothalamus-pituitary organoids at an early stage of induction and will improve the efficiency and quality control of organoid induction experiments. In the future, a more accurate model will be generated by accumulating data

from multiple cell lines generated at various institutions. In addition, we believe this approach is widely applicable to other organoid cultures and has the potential to contribute to improved and automated culture systems.

Limitations of the study

We note several limitations in this study. First, the developed model was derived from a single reporter cell line. However, this model can be used for non-reporter parental and other cell lines (Figures 2H and S4), highlighting the broad utility of our developed model. Second, although organoids are three-dimensional structures, our prediction models were constructed based on two-dimensional images. This approach was chosen to utilize only two-dimensional phase-contrast images to ensure that our model could be easily applied to standard laboratory environments. However, prediction models based on three-dimensional images should provide better prediction accuracy, although additional data collection and model construction are required. These training data were captured with manual focusing; however, since the focal position within organoids affected prediction outcomes (Figure S6), more objective and reproducible imaging methods would be preferable. Third, although Grad-CAM analyses showed that morphological differences on the organoid surfaces were important features for prediction, it remains unclear what these morphological changes truly represent and how they contribute to hypothalamus-pituitary development. Immunostaining experiments and corresponding attention map images imply that cellular composition and spatial organization within the organoids are critical for proper differentiation. However, organoids are highly heterogeneous, containing various cell types beyond the oral ectoderm, making it difficult to identify conclusively the cell components essential for the induction process. Lastly, we performed transcript-level quantification (RNA-seq and RT-qPCR) to assess differentiation efficiency. However, quantitative assessment at the protein level remains insufficient. Hypothalamus-pituitary organoids are heterogeneous cell aggregates composed of multiple cell types, and their cellular composition, size, and shape exhibit substantial variability across experimental batches and between wells. Furthermore, the inner organoid regions contain a large number of dead cells due to hypoxia, resulting in considerable non-specific autofluorescence. These features make quantitative evaluation methods, such as cell counting, technically challenging.

RESOURCE AVAILABILITY

Lead contact

Requests for information and resources should be directed to the lead contact, Takuya Yamamoto (takuya@cira.kyoto-u.ac.jp).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All RNA-seq data of hypothalamus-pituitary organoids predicted to fail or succeed on day 9 or day 15 are available from Gene Expression Omnibus (NCBI) with the accession number GEO: GSE301692.
- Scripts used in this study and the h5 file of developed model weights were deposited to Zenodo: https://zenodo.org/records/15453478.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank the members of the Yamamoto laboratory for valuable discussions and K. Hui for editing the manuscript. We also thank T. Asano, M. Soen, M. Sakakibara, and K. Kanie for evaluating the induction efficiencies of pituitary organoids. This work was supported by a Grant-in-Aid for JSPS Fellows (21J01689 to R.M.), a grant for Promising Investigator from the Japan Endocrine Society (to R.M.), a grant from iPS Academia Japan (to R.M.), the Fellowship Program of Development of Young Researchers from the Center for iPS Cell Research and Application (to T.Y.), the JST FOREST Program (grant number JPMJFR206C, to T.Y.), JST CREST (grant number JPMJCR2023, to T.Y.), AMED (grant numbers JP19gm1310002, JP21bm0104001, JP21gm1310011, JP22bm1223002, JP23bm1323001, and JP24bm1123053, to T.Y.), and JSPS A3 Foresight Program (number JPJSA3F20230001, to T.Y.).

AUTHOR CONTRIBUTIONS

T.Y. supervised the entire work; H.S., Y.T., T.A., and T.Y. supervised the biological experiments; R.M. performed the biological experiments; R.M. constructed the prediction models; R.M. and T.Y. analyzed the data; and R.M. and T.Y. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Cell culture
- METHOD DETAILS
  - Generation of LHX3 reporter iPSCs
  - Hypothalamus-pituitary organoid induction
  - Imaging of induced organoids
  - Preprocessing of input images
  - Construction of a convolutional neural network (CNN)
  - Fine-tuning
  - Gradient-weighted class activation mapping (Grad-CAM)
  - Comparison of the predictive accuracy between machine learning model and human researchers
  - Immunostaining
  - Quantitative RT-PCR (qRT-PCR)
  - RNA sequencing analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2025.101119.

Received: May 13, 2024
Revised: February 20, 2025
Accepted: July 8, 2025
Published: August 4, 2025

REFERENCES

1. Sato, T., Stange, D.E., Ferrante, M., Vries, R.G.J., Van Es, J.H., Van den Brink, S., Van Houdt, W.J., Pronk, A., Van Gorp, J., Siersema, P.D., and Clevers, H. (2011). Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. Gastroenterology *141*, 1762–1772. https://doi.org/10.1053/j.gastro.2011.07.050.
2. Boj, S.F., Hwang, C.I., Baker, L.A., Chio, I.I.C., Engle, D.D., Corbo, V., Jager, M., Ponz-Sarvise, M., Tiriac, H., Spector, M.S., et al. (2015). Organoid models of human and mouse ductal pancreatic cancer. Cell *160*, 324–338. https://doi.org/10.1016/j.cell.2014.12.021.
3. Huch, M., Gehart, H., van Boxtel, R., Hamer, K., Blokzijl, F., Verstegen, M.M.A., Ellis, E., van Wenum, M., Fuchs, S.A., de Ligt, J., et al. (2015). Long-term culture of genome-stable bipotent stem cells from adult human liver. Cell *160*, 299–312. https://doi.org/10.1016/j.cell.2014.11.050.
4. Lancaster, M.A., Renner, M., Martin, C.-A., Wenzel, D., Bicknell, L.S., Hurles, M.E., Homfray, T., Penninger, J.M., Jackson, A.P., and Knoblich, J.A. (2013). Cerebral organoids model human brain development and microcephaly. Nature *501*, 373–379. https://doi.org/10.1038/nature12517.
5. Suga, H., Kadoshima, T., Minaguchi, M., Ohgushi, M., Soen, M., Nakano, T., Takata, N., Wataya, T., Muguruma, K., Miyoshi, H., et al. (2011). Self-formation of functional adenohypophysis in three-dimensional culture. Nature *480*, 57–62. https://doi.org/10.1038/nature10637.
6. Ozone, C., Suga, H., Eiraku, M., Kadoshima, T., Yonemura, S., Takata, N., Oiso, Y., Tsuji, T., and Sasai, Y. (2016). Functional anterior pituitary generated in self-organizing culture of human embryonic stem cells. Nat. Commun. *7*, 10351. https://doi.org/10.1038/ncomms10351.
7. Suga, H. (2015). Differentiation of Pluripotent Stem Cells into Hypothalamic and Pituitary Cells. Neuroendocrinology *101*, 18–24. https://doi.org/10.1159/000369821.
8. Matsumoto, R., and Takahashi, Y. (2021). Human pituitary development and application of iPSCs for pituitary disease. Cell. Mol. Life Sci. *78*, 2069–2079. https://doi.org/10.1007/s00018-020-03692-8.
9. Kasai, T., Suga, H., Sakakibara, M., Ozone, C., Matsumoto, R., Kano, M., Mitsumoto, K., Ogawa, K., Kodani, Y., Nagasaki, H., et al. (2020). Hypothalamic Contribution to Pituitary Functions Is Recapitulated In Vitro Using 3D-Cultured Human iPS Cells. Cell Rep. *30*, 18–24.e5. https://doi.org/10.1016/j.celrep.2019.12.009.
10. Ochiai, H., Suga, H., Yamada, T., Sakakibara, M., Kasai, T., Ozone, C., Ogawa, K., Goto, M., Banno, R., Tsunekawa, S., et al. (2015). BMP4 and FGF strongly induce differentiation of mouse ES cells into oral ectoderm. Stem Cell Res. *15*, 290–298. https://doi.org/10.1016/j.scr.2015.06.011.
11. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444. https://doi.org/10.1038/nature14539.
12. Norvig, P., and Russell, S. (2020). Artificial Intelligence: A Modern Approach (Pearson Education).
13. Christiansen, E.M., Yang, S.J., Ando, D.M., Javaherian, A., Skibinski, G., Lipnick, S., Mount, E., O'Neil, A., Shah, K., Lee, A.K., et al. (2018). In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images. Cell *173*, 792–803. https://doi.org/10.1016/j.cell.2018.03.040.
14. Imamura, K., Yada, Y., Izumi, Y., Morita, M., Kawata, A., Arisato, T., Nagahashi, A., Enami, T., Tsukita, K., Kawakami, H., et al. (2021). Prediction Model of Amyotrophic Lateral Sclerosis by Deep Learning with Patient Induced Pluripotent Stem Cells. Ann. Neurol. *89*, 1226–1233. https://doi.org/10.1002/ana.26047.
15. Kusumoto, D., Seki, T., Sawada, H., Kunitomi, A., Katsuki, T., Kimura, M., Ito, S., Komuro, J., Hashimoto, H., Fukuda, K., and Yuasa, S. (2021). Anti-senescent drug screening by deep learning-based morphology senescence scoring. Nat. Commun. *12*, 257. https://doi.org/10.1038/s41467-020-20213-0.
16. Park, K., Lee, J.Y., Lee, S.Y., Jeong, I., Park, S.Y., Kim, J.W., Nam, S.A., Kim, H.W., Kim, Y.K., and Lee, S. (2023). Deep learning predicts the differentiation of kidney organoids derived from human induced pluripotent stem cells. Kidney Res. Clin. Pract. *42*, 75–85. https://doi.org/10.23876/j.krcp.22.017.
17. Lim, M.H., Shin, S., Park, K., Park, J., Kim, S.W., Basurrah, M.A., Lee, S., and Kim, D.H. (2023). Deep Learning Model for Predicting Airway Organoid Differentiation. Tissue Eng. Regen. Med. *20*, 1109–1117. https://doi.org/10.1007/s13770-023-00563-8.
18. Kegeles, E., Naumov, A., Karpulevich, E.A., Volchkov, P., and Baranov, P. (2020). Convolutional Neural Networks Can Predict Retinal Differentiation in Retinal Organoids. Front. Cell. Neurosci. *14*, 171. https://doi.org/10.3389/fncel.2020.00171.
19. Zhu, X., Gleiberman, A.S., and Rosenfeld, M.G. (2007). Molecular Physiology of Pituitary Development: Signaling and Transcriptional Networks. Physiol. Rev. *87*, 933–963. https://doi.org/10.1152/physrev.00006.2006.
20. Matsumoto, R., Suga, H., Aoi, T., Bando, H., Fukuoka, H., Iguchi, G., Narumi, S., Hasegawa, T., Muguruma, K., Ogawa, W., and Takahashi, Y. (2020). Congenital pituitary hypoplasia model demonstrates hypothalamic OTX2 regulation of pituitary progenitor cells. J. Clin. Investig. *130*, 641–654. https://doi.org/10.1172/JCI127378.
21. Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. Preprint at arXiv. https://doi.org/10.48550/arXiv.1312.6034.
22. Zeiler, M.D., and Fergus, R. (2014). In Visualizing and Understanding Convolutional Networks. held in Cham, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds. (Springer International Publishing), pp. 818–833.
23. Ramprasaath, R., Selvaraju, M.C., Das, A., Vedantam, R., Devi, P., and Batra, D. (2017). Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV) (IEEE), pp. 618–626.
24. Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861–872. https://doi.org/10.1016/j.cell.2007.11.019.
25. Ishida, Y., Kawakami, H., Kitajima, H., Nishiyama, A., Sasai, Y., Inoue, H., and Muguruma, K. (2016). Vulnerability of Purkinje Cells Generated from Spinocerebellar Ataxia Type 6 Patient-Derived iPSCs. Cell Rep. *17*, 1482–1490. https://doi.org/10.1016/j.celrep.2016.10.026.
26. Okita, K., Matsumura, Y., Sato, Y., Okada, A., Morizane, A., Okamoto, S., Hong, H., Nakagawa, M., Tanabe, K., Tezuka, K.-i., et al. (2011). A more efficient method to generate integration-free human iPS cells. Nat. Methods *8*, 409–412. http://www.nature.com/nmeth/journal/v8/n5/abs/nmeth.1591.html.
27. Haeussler, M., Schönig, K., Eckert, H., Eschstruth, A., Mianné, J., Renaud, J.-B., Schneider-Maunoury, S., Shkumatava, A., Teboul, L., Kent, J., et al. (2016). Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. Genome Biol. *17*, 148. https://doi.org/10.1186/s13059-016-1012-2.
28. Pelletier, J., and Sonenberg, N. (1988). Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. Nature *334*, 320–325. https://doi.org/10.1038/334320a0.
29. Gullt, A., and Pal, S. (2017). Deep Learning with Keras (Packt Publishing Ltd)).
30. Mishra, B., Singh, N., and Singh, R. (2014). Master-slave Group Based Model for Co-ordinator Selection, an Improvement of Bully Algorithm (IEEE), pp. 457–460.
31. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization (IEEE), pp. 618–626.
32. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast

universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

33. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. j. *17*, 10. https://doi.org/10.14806/ej.17.1.200.
34. Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*, D766–D773. https://doi.org/10.1093/nar/gky955.
35. Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169. https://doi.org/10.1093/bioinformatics/btu638.
36. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.
37. Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit polyclonal anti-LHX3	Millipore	Cat# AB3202; RRID: AB_2281220
Mouse monoclonal anti-TUBB3	Biolegend	Cat# MMS-435P; RRID: AB_2313773
Mouse monoclonal anti-VSX2	Santa Cruz	Cat# sc-365519; RRID: AB_10842442
Mouse monoclonal anti-RAX	Santa Cruz	Cat# sc-271889; RRID: AB_10708730
Rabbit polyclonal anti-PITX1	SIGMA-ALDRICH	Cat# HPA008743; RRID: AB_1855413
Mouse monoclonal anti-NKX2.1	Millipore	Cat# MAB5460; RRID: AB_571072
Donkey anti-Mouse IgG Alexa Fluor 488	Thermo Fisher Scientific	Cat# A-21202; RRID: AB_141607
Donkey anti-Rabbit IgG Alexa Fluor 546	Thermo Fisher Scientific	Cat# A-10040; RRID: AB_2534016
Goat anti-Rat IgG Alexa Fluor 647	Thermo Fisher Scientific	Cat# A-21247; RRID: AB_141778
Chemicals, peptides, and recombinant proteins		
DMEM/F12	Sigma-Aldrich	D6421
KnockOut Serum Replacement	Thermo Fisher Scientific	10828028
GlutaMAX	Thermo Fisher Scientific	35050061
non-essential amino acids	Sigma-Aldrich	M7145
2-mercaptoethanol	Sigma-Aldrich	M3148
basic fibroblast growth factor	FUJIFILM	060–04543
Accutase	Nacalai Tesque	12679–54
Lipofectamine Stem reagent	Thermo Fisher Scientific	STEM00001
puromycin	InvivoGen	ant-pr-1
0.25% Trypsin	Thermo Fisher Scientific	15090046
collagenase type IV	Thermo Fisher Scientific	17104019
CaCl2	Nacalai Tesque	16973–64
Y-27632	FUJIFILM	034–24024
Recombinant human BMP4	R&D systems	314-BP
SAG	Cayman Chemical	11914
4% paraformaldehyde	Nacalai Tesque	09154–85
O.C.T. compound	Sakura Finetek	09154–85
RNeasy Mini kit	Qiagen	74104
TB Green Premix mix Ex Taq II	TaKaRa	RR820S
DMEM/F12	Sigma-Aldrich	D6421
Critical commercial assays		
ReverTra Ace qPCR RT kit with gDNA Remover	TOYOBO	FSQ-301
TruSeq Standard mRNA library prep kit	Illumina	20040532
Deposited data		
RNA-seq data of hypothalamus-pituitary organoids predicted to fail or succeed at day 9 or 15	This paper	GSE301692
Experimental models: Cell lines		
201B7	RIKEN BRC	HPS0063
409B2	RIKEN BRC	HPS0076
HC06	In house iPSC from Dr. Keiko Muguruma	https://doi.org/10.1016/j.celrep.2016.10.026.

(*Continued on next page*)

CellPress
OPEN ACCESS

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Oligonucleotides		
Oligonucleotides used in this work	This paper	see Table S5
Recombinant DNA		
pX459 v2.0	Addgene plasmid	#62988
Software and algorithms		
Scripts used in this study	This paper	https://zenodo.org/records/15453478
CRISPOR web tool	UCSC	https://crispor.gi.ucsc.edu/
Python 3.9.7	Python Software Foundation	https://www.python.org/
CUDA 11.0	NVIDIA Corporation	https://developer.nvidia.com/cuda-11.0-download-archive
cuDNN 8.0	NVIDIA Corporation	https://developer.nvidia.com/rdp/cudnn-archive
TensorFlow 2.4.1	Google	https://www.tensorflow.org/
Keras	Google	https://keras.io/
R (4.1.2 or 3.6.2)	The R Foundation	https://www.r-project.org/
cutadapt 4.6	Marcel Martin	https://cutadapt.readthedocs.io/en/stable/#
STAR 2.7.11a	Alex Dobin	https://github.com/alexdobin/STAR
htseq-count 2.0.5	the HTSeq team	https://htseq.readthedocs.io/en/master/index.html
DESeq2 1.42.0	Bioconductor	https://bioconductor.org/packages/
EnhancedVolcano	Kevin Blighe	https://github.com/kevinblighe/EnhancedVolcano
clusterProfiler	Bioconductor	https://bioconductor.org/packages/
Other		
Inverted system microscope	Olympus	IX71
Evident Olympus Objective	Olympus	UPLFLN4XPH
ECLIPSE Ti2 microscope	Nikon	Ti2
QuantStudio3	Thermo Fisher Scientific	QuantStudio3
TapeStation	Agilent Technologies	G2991BA 4200
NextSeq 2000	Illumina	NextSeq 2000

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Cell culture

Human induced pluripotent stem cells (hiPSCs) were cultured as previously described with some modifications.[24] In brief, hiPSCs were maintained on mitomycin C-treated feeder cells (SNL or mouse embryonic fibroblasts) in DMEM/F12 (Sigma-Aldrich) supplemented with 20% (v/v) KnockOut Serum Replacement (KSR, Thermo Fisher Scientific), 1× GlutaMAX (Thermo Fisher Scientific), 1× non-essential amino acids (Sigma-Aldrich), 80 μM 2-mercaptoethanol (Sigma-Aldrich), and 7.5 ng/mL basic fibroblast growth factor (FGF) (FUJIFILM). This study used the hiPSC line 201B7[24] for model derivation and HC06[25] and 409B2[26] for model validation (Figure S4).

METHOD DETAILS

Generation of LHX3 reporter iPSCs

Single guide RNA (sgRNA) sequences targeting the region downstream of *LHX3*, a pituitary progenitor marker, were designed using the CRISPOR web tool (http://crispor.tefor.net/).[27] The donor plasmid for homologous recombination was designed to insert the IRES-mCherry sequence[28] and the puromycin-resistance gene (PuroR) (Figure 1B). Briefly, hiPSCs were dissociated with Accutase (Nacalai Tesque) into single cells, and 3.75×10^5 cells were seeded into 6-well culture plates. Subsequently, cells were co-transfected with 1 μg of a Cas9-and sgRNA-expressing plasmid (pX459; Addgene, #62988) and 1 μg of donor plasmid including the LHX3-IRES-mCherry using Lipofectamine Stem reagent (Thermo Fisher Scientific). After selecting successful recombinants using puromycin

(InvivoGen), cells underwent limited dilution. Homologous recombination was confirmed by PCR and validated via PCR-directed sequencing. Subsequently, Cre recombinase was expressed transiently to remove the PuroR gene cassette, and Cre-loxP recombination was confirmed via PCR and validated by PCR-directed sequencing.

Hypothalamus-pituitary organoid induction

Hypothalamus-pituitary organoid induction was performed as previously described with some modifications[6,20] (Figure 1A). In brief, after removal of feeder cells with CTK solution (0.25% Trypsin (Invitrogen), 0.1 mg/mL collagenase type IV (Invitrogen), 10 mM $CaCl_2$ (Nacalai Tesque), and 20% KSR in PBS), hiPSCs were dissociated into single cells using Accutase (Nacalai Tesque). Subsequently, cells were plated into V-bottom low-attachment 96-well plates (Sumitomo Bakelite) at a density of 1.5-2.0$\times 10^4$ cells/well and were cultured in growth factor-free chemically defined medium (gfCDM) supplemented with 10% (v/v) KSR. A selective inhibitor of Rho-associated kinase, Y-27632 (20 μM; FUJIFILM), was added on days 0–3. From day 6 onwards, 5 nM BMP4 (R&D systems) and 2 μM SAG (Cayman Chemical) were included in the culture medium. From day 18 onwards, BMP4 was withdrawn, and organoids were cultured under 40% O_2 conditions. From day 30 onwards, organoids were maintained in gfCDM supplemented with 20% (v/v) KSR.

Imaging of induced organoids

Phase-contrast images of organoids were obtained with an inverted system microscope, IX71 (Olympus). All training dataset images were captured using a 4× microscope objective (UPLFLN4XPH, Olympus). Original image sizes were 1,600 × 1,200 or 2,048 × 2,048 pixels. Fluorescent images of immunostained samples were obtained using an ECLIPSE Ti2 microscope (Nikon).

Preprocessing of input images

Input phase-contrast images varied considerably in size, color, intensity, and contrast (Figure S2A). First, color images were converted to gray-scale images with gamma control using the OpenCV-Python package. Second, images were resized and trimmed to 300 × 300 pixels. Finally, the image intensity was averaged (average = 128, SD = 32) (Figure S2A).

Construction of a convolutional neural network (CNN)

The training was performed using multiple iterations, each with a batch size of 256 images, with a starting learning rate of 1×10^{-5} with Root Mean Squared Propagation optimization. During each iteration, the accuracy and loss were recorded until up to 400 iterations. Model weights were updated continuously according to the loss gradient. The aim was to minimize the loss function (categorical cross-entropy). The training was stopped when model loss ceased to decrease for 20 iterations.

The computing environment was as follows: CPU, Intel Xeon E5-2660 v4 2.00 GHz, E5-2697 v2 2.70 GHz, or E7-4870 2.40 GHz and GPGPU, NVIDIA Tesla K20m 5.0 GB or CPU, Intel Core i9-13900 and GPGPU, NVIDIA RTX A4500, with CUDA version 11.0 and cuDNN version 8.0. Calculations were performed using Python version 3.9.7 (http://www.python.org). The training was performed with TensorFlow version 2.4.1 and Keras's deep learning library.[29]

Fine-tuning

To improve the CNN classification accuracy, we employed a fine-tuning method.[30] We tried ten pre-trained models available in Keras (Xception, VGG16, VGG19, Resnet50, InceptionV3, InceptionResnetV2, Mobilenet, Densenet121, Densenet169, and Densenet201) and fine-tuned the pre-trained CNNs based on the dataset of training phase-contrast images. First, we constructed preliminary models to determine the extent of retraining required for the CNNs by varying the frozen ratio using our dataset. We found that the optimal configuration was achieved when approximately one-third of the layers were fixed. Then, the final set of fully linked layers was replaced with a new set of fully linked layers corresponding to two classes ("fail" and "success"). The first one-third of convolutional layers were frozen and their weights could not be changed. Subsequently, the remaining two-thirds of convolutional layers were trained with our dataset.

Gradient-weighted class activation mapping (Grad-CAM)

Grad-CAM[31] was performed to visualize the regions of input phase-contrast images contributing to the model's prediction output. Grad-CAM is a localization technique for CNN-based networks, which computes the gradients flowing into the final convolutional layer from a certain target into output heatmaps that highlight specific areas of interest. These areas could be interpreted as critical regions in an image that the network focused on to make its prediction.

To obtain a classification localization map L_{ij}^c (i and j are width and height, respectively), the gradient of the score y was computed based on the intensity A_{ij}^k of the convolutional layer, $\frac{\partial y}{\partial A_{ij}^k}$. By averaging the values over all pixels (global average-pooling), we calculated the weight factor w_k for the k feature map as follows:

$$w_k = \frac{1}{z} \sum_{ij} \frac{\partial y}{\partial A_{ij}^k}$$

CellPress
OPEN ACCESS

Consequently, the localization map L_{ij}^{c} was calculated as follows:

$$L_{ij}^{c} = \sum_{k} w_k \cdot A_{ij}^{k}$$

The obtained localization map is shown as a heatmap, in which red indicates a large contribution to the output, thus forming the main basis for judgment, while blue indicates a small contribution.

Comparison of the predictive accuracy between machine learning model and human researchers

To compare the predictive accuracy between the machine learning model and human researchers, we evaluated prediction accuracy for phase-contrast images from days 9 and 15, which were not used for prediction model construction. The machine learning models and 5 human researchers independently predicted LHX3^{+} pituitary progenitor cell differentiation at day 39 using 100 phase-contrast images not employed in the model construction.

Immunostaining

Organoids were fixed with 4% paraformaldehyde (Nacalai Tesque) for 30–60 min at room temperature. After overnight incubation in 20% sucrose, samples were embedded in O.C.T. compound (Sakura Finetek) and frozen at −80°C. Subsequently, frozen blocks were cryosectioned at 10 μm thickness. Samples were then permeabilized with 0.3% Triton X-100, blocked with 2% skimmed milk, and incubated with primary antibodies overnight at 4°C. Primary antibodies used in the study are listed in Table S5. Next, samples were washed with 0.05% Tween 20 and incubated with species-specific Alexa Fluor-conjugated secondary antibodies (Thermo Fisher Scientific) for 2 h at room temperature.

Quantitative RT-PCR (qRT-PCR)

Total RNA was extracted from 4 to 8 organoids using the RNeasy Mini kit (Qiagen). Subsequently, up to 500 ng of total RNA was reverse transcribed using ReverTra Ace qPCR RT kit (TOYOBO) and qRT-PCR was performed with QuantStudio3 (Thermo Fisher Scientific) using TB Green Premix mix Ex Taq (TaKaRa). Thermal cycling profiles were as follows: initial denaturation at 95°C for 2 min, followed by 40 cycles of denaturation at 95°C for 10 s, annealing at 60°C for 30 s, and extension at 72°C for 30 s. All samples were assayed in duplicate. *ACTB* was used as an internal control. Primer sequences are listed in Table S5.

RNA sequencing analysis

The quality of extracted RNA was checked using TapeStation (Agilent Technologies), and RNA samples with RIN value >7.0 were used for RNA-seq analysis. RNA-seq library construction was performed using an Illumina Stranded mRNA Prep Kit (Cat# 20040532, Illumina) according to the manufacturer's instructions. Obtained libraries were sequenced with NextSeq 2000 with paired-end mode (Illumina). Sequenced reads were mapped to the GRCh38 plus the reporter sequences using STAR v2.7.11a[32] after trimming the adaptor sequences with cutadapt v4.6.[33] GENCODE reference annotation (v44/Ensembl 110) was used for gene annotation.[34] The uniquely and properly mapped reads were used for further analysis. The raw counts were calculated using htseq-count ver. 2.0.5[35] with the GENCODE gtf file. The normalization of read counts was performed using DESeq2 v1.42.0 based on a negative binomial distribution.[36] A volcano plot was generated using the R package EnhancedVolcano (https://github.com/kevinblighe/EnhancedVolcano). PCA was performed using the R function prcomp. Heatmaps were generated using the R package pheatmap (https://github.com/raivokolde/pheatmap). Sample-to-sample distances were calculated using the *dist* function in the package. GO analysis was performed using the R package clusterProfiler.[37]

QUANTIFICATION AND STATISTICAL ANALYSIS

Data are expressed as means ± SD or SEM. Where appropriate, continuous data were compared with a Student's t test or Mann-Whitney U test. Multiple groups were compared by a one-way ANOVA with Tukey's post hoc multiple comparison. Statistical analysis was performed using R (version 3.6.2).