AI-based forecasting of ethanol fermentation using yeast morphological data

Kaori Itto-Nakama 1,1 Shun Watanabe, 2 Naoko Kondo, 1 Shinsuke Ohnuki 1,1 Ryota Kikuchi, 2,3 Toru Nakamura, 4 Wataru Ogasawara, 5 Ken Kasahara, 2,∗ and Yoshikazu Ohya 1,6,∗

1Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba, Japan; 2Chitose Laboratory Corp., Biotechnology Research Center, Miyamae-ku, Kawasaki, Kanagawa, Japan; 3Circular Bioeconomy Development, Office of Society Academia Collaboration for Innovation, Kyoto University, Kitashirakawa Oiwake-cho, Sakyo-ku, Kyoto, Japan; 4NRI System Techno Ltd., Hodogaya-ku, Yokohama, Kanagawa, Japan; 5Department of Bioengineering, Nagaoka University of Technology, Nagaoka, Niigata, Japan; and 6Collaborative Research Institute for Innovative Microbiology, The University of Tokyo, Bunkyo-ku, Tokyo, Japan

∗Correspondence: Ken Kasahara, kasahara@chitose-bio.com; Yoshikazu Ohya, ohya@edu.k.u-tokyo.ac.jp

ABSTRACT

Several industries require getting information of products as soon as possible during fermentation. However, the trade-off between sensing speed and data quantity presents challenges for forecasting fermentation product yields. In this study, we tried to develop AI models to forecast ethanol yields in yeast fermentation cultures, using cell morphological data. Our platform involves the quick acquisition of yeast morphological images using a nonstaining protocol, extraction of high-dimensional morphological data using image processing software, and forecasting of ethanol yields via supervised machine learning. We found that the neural network algorithm produced the best performance, which had a coefficient of determination of >0.9 even at 30 and 60 min in the future. The model was validated using test data collected using the CalMorph-PC(10) system, which enables rapid image acquisition within 10 min. AI-based forecasting of product yields based on cell morphology will facilitate the management and stable production of desired biocommodities.

Graphical Abstract

Diagram of AI-based forecasting of ethanol fermentation using yeast morphological data.
Fermentation technology contributes greatly to modern society by producing a multitude of compounds that are essential to human life, including beverages, foods, nutrients, supplements, medicine, bioethanol, chemicals, and polymer materials (Paul et al. 2010; Terefe and Augustin 2019). For both newly developed and conventional fermentation products, efficient and stable product synthesis is desired to reduce production costs and environmental pollution (Akinsemolu 2018). This requires in-depth knowledge of fermentation processes and microbial systems. Effective management and monitoring of the fermentation process are essential elements (Schügerl 2001; Mandenius 2004; Mears et al. 2017).

The budding yeast Saccharomyces cerevisiae is an excellent model system to investigate the factors affecting the efficiency and performance of fermentation processes (Pretorius, Toit and Rensburg 2003; Kavšcek et al. 2015). Yeast ethanol fermentation is one of the oldest and most important biochemical processes involving various factors. Since the process control is generally considered as a prerequisite to determine the quality of the final product, there is a long history for improvement of the bioprocess and many in-line or off-line applications have been built to monitor the culture condition (Vojinović, Cabral and Fonseca 2006; Walker and Walker 2018). Initial glucose concentration, pH value, osmolarity and temperature affect the ethanol production capacity of S. cerevisiae; by optimizing these parameters, the ethanol fermentation activity of yeast can be enhanced (Lin et al. 2012). It is also necessary to more accurately detect the optimal moment when the alcoholic fermentation process reaches the critical performance threshold and needs to be interrupted. Therefore, rapid monitoring and modeling of biomass and ethanol production facilitates on-site checks of process and product quality. Ethanol production can be modeled using a modified version of the Gompertz model (Phuocoepthim et al. 2017). Although monitoring the extracellular environment is sufficient to build a simple fermentation model, this does not provide enough information to model the entire complex fermentation process and cellular conditions and to forecast the fermentation process. Omics studies such as transcriptomics, metabolomics, and phenomics can provide large quantities of data related to cellular metabolism, enabling detailed monitoring of the fermentation process (Zuzuarregui et al. 2006; Weckx, Kerrebroeck and Vyust 2019). Fermentation takes place under stress conditions, and many stress-related genes are induced during fermentation (Pérez-Torrado, Bruno-Bárcena and Matallana 2005). However, a genome-wide transcriptome analysis of yeast revealed that the genes induced by fermentation differ from those related to general stress responses (Marks et al. 2008). Transcriptomics data have been employed to predict ethanol yields in yeast using linear regression (Guo and Feng 2016). Furthermore, transcriptome engineering has been applied to ethanol-producing yeast to improve biofuel production (Michael et al. 2016). In a recent quantitative multilomics study, proteomics, transcriptomics, metabolomics, phenomics, and phospho-proteomics data were combined to yield useful insights into gene regulation metabolic status during yeast cell proliferation (Campbell et al. 2020). However, a fundamental limitation to combined omics approaches is that it takes an extensive period of time, often longer than a month, to acquire and analyze data. Therefore, the acquisition and analysis of large omics datasets for on-site fermentation management is not feasible.

There is a clear trade-off between sensing speed and the amount of data acquired when selecting a method to examine fermentation processes. Monitoring the extracellular environment of yeast fermentation cultures is rapid (seconds), but this does not provide the depth of metabolic insight provided by omics studies. Omics and Raman spectral data are well suited for training artificial intelligence (AI) programs using deep learning (jiang et al. 2020), but it is highly challenging to acquire omics data in real time. In this study, we instead focused on yeast cell morphology to overcome this problem. Yeast morphology can be described as phenome data composed of more than 500 features (Ohya et al. 2005). Indeed, yeast cell morphology is a sensitive readout that is rich in information and is related to gene expression; the individual deletion of >3500 nonessential genes results in distinct morphological phenotypes that are directly related to the specific gene functions (Suzuki et al. 2018). Furthermore, dynamic morphological changes occur in bottom-fermenting yeast throughout the fermentation process (Ohnuki et al. 2014). Metabolically active yeast have distinct morphological features, suggesting that morphological data can be used in management (Suzuki et al. 2018). Yeast morphological analyses can take several days due to staining and image processing, but nonstaining image analysis protocols can shorten the experimental and analytical period required to obtain the morphological data (Ohnuki, Nagami and Ohya 2009).

In this study, we used cell morphology as convolutional data to forecast fermentation products during fermentation. Using neural network, a supervised machine learning technique (Vlassides, Ferrier and Block 2001), we forecasted the ethanol yield at given and future time points. We used yeast morphology, which reflects yeast activity, to construct a management system from which our AI program can learn and then evaluate throughout the fermentation process.

Materials and methods
Culture condition
Saccharomyces cerevisiae industrial strain ATCC24702 stock cultures from −80 °C glycerol stocks were inoculated onto YPD plates (Sherman 1991) and grown at 25 °C for 3 days. Cells from the YPD plate were transferred to Erlenmeyer flasks containing 30 mL YPD and grown at 30 °C overnight on a shaker (Personal Lt-10F, TAITEC, Saitama, Japan) at 120 m−1. During precultivation, 25 mL of overnight liquid culture was used to inoculate Erlenmeyer flasks containing 125 mL YPD and grown to 1.2 × 10⁶ cells/mL under the same conditions. For the experimental culture, cells from the precultivation were transferred to 2-L Sakaguchi flasks containing 500 mL of YPD and grown at 30 °C on a rotary shaker (Bio Shaker BR-180FL, TAITEC, Saitama, Japan) with 100 m−1 agitation. The experimental culture was initiated at a cell concentration of 0.5-1.0 × 10⁶ cells/mL. Aliquots of 5 mL were collected every 30 min over a 9-h period, then once more after 24 h. Cell samples (1 mL) were used to monitor growth and to the measure glucose and ethanol levels in the culture media, while 4 mL of cells were fixed with 3.7% (v/v) formaldehyde and analyzed by microscope. Growth was monitored by
measuring the optical density (OD) at a wavelength of 600 nm (BioSpectrometer, Eppendorf, Hamburg, Germany) and by counting the cells (CellDrop BF, Wilmington, DE, USA). The AI model was created using the training data obtained with CalMorph-PC 5 times by performing independent experiments using the same culture equipment and same yeast colonies, on the same collection date, but culturing them in different Sakaguchi flasks and preparing the medium separately. A more robust AI model was created using the training data 18 times by performing independent experiments using the different yeast colonies, different glucose concentrations ranging from 1.8 to 2.4%, different temperatures from 28 to 34 °C, and on different collection dates. The AI model was tested using the 5 test data replicates obtained with CalMorph-PC(10) by performing out-of-time validation using the different yeast colonies and on a later collection date.

Microscopy and image processing

The fixation times differed between the CalMorph-PC protocol (75 min) and the CalMorph-PC(10) protocol (1 min). In the CalMorph-PC protocol, yeast cells cultured in YPD were mixed with formaldehyde (final concentration of 3.7%) and potassium phosphate buffer (final concentration of 100 mM [pH 6.5]), and then incubated for 30 min at 30 °C. The cells were then collected by centrifugation (1940 × g, 5 min) and further incubated in potassium phosphate buffer containing 4% formaldehyde for 45 min. For the CalMorph-PC(10) protocol, yeast cells cultured in YPD were mixed with formaldehyde (final concentration, 3.7%) and incubated for 1 min. The cells were then collected by centrifugation (1940 × g, 5 min) and suspended in PBS buffer (prepared from PBS tablets [Takara Bio Inc., Shiga, Japan]). We needed to use fixed cells, because unfixed cells would continue the cell proliferation during microscopy operation. Specimens fixed using the CalMorph-PC and CalMorph-PC(10) protocols were observed under an Axio Imager phase-contrast microscope equipped with a Plan-Apochromat 100/1.4 oil lens (Carl Zeiss, Oberkochen, Germany), a CoolSNAP HQ cooled charge-coupled device camera (Roper Scientific Photometrics, Tucson, AZ, USA), and AxioVision software (Carl Zeiss). The phase-contrast images were processed by a Java-based program (23), and then analyzed using CalMorph (ver. 1.3).

Analytical methods

Glucose and ethanol in the culture medium were measured with the Glucose CII-test Wako (FUJIFILM Wako Pure Chemical Corp., Osaka, Japan) and F-kit (J. K. International, Tokyo, Japan), respectively. The mean and SD were calculated using data collected from 4 independent assays. Representative data for 5 biologically independent experiments are shown.

PCA of morphological data obtained by CalMorph-PC

We performed a PCA analysis using the CalMorph-PC data to examine yeast morphological changes over time. First, we normalized the CalMorph-PC data (5 experimental replicates) by determining the most appropriate probability distribution for each of the 31 parameters (Figure S1) and applying a one-way ANOVA using a generalized linear model. For the one-way ANOVA, we assumed that the morphology was evenly distributed among 20 different time points (n = 5). The mean and variance at time zero were calculated using maximum likelihood estimation, and then used as references to calculate Z values; the differences from the reference points were estimated by a one-way ANOVA and converted to Z values using a Wald test.

Next, we performed a PCA with the normalized data, using the morphological measurements as the observations and the time points as the variables. The cumulative contribution ratio of PC1-PC4 was >80% (Figure S2). The mean and variance (SD) of each time point was calculated, and the time change of each PC score was plotted. To identify the morphological features represented by PC1-PC4, the PC load was first determined using an uncorrelated test (P < .05, after Bonferroni correction). Parameters with absolute PC load values of >0.7 were considered to be significantly correlated with the corresponding PC (Figure S3). Parameters with maximum absolute load values for each PC were considered to be representative morphological features.

Comparison of morphological data obtained by CalMorph-PC(10) and CalMorph-PC

CalMorph-PC(10) morphological data (3 replicates) were normalized using the same method as for CalMorph-PC. The mean and variance at time zero of CalMorph-PC (5 replicates) were used as references to calculate Z values; the differences from the reference points were estimated by a one-way ANOVA and converted to Z values by a Wald test. The Z values for the CalMorph-PC and CalMorph-PC(10) morphological data were combined and subjected to PCA. The cumulative contribution ratio of PC1-PC3 was almost 80% (Figure S4). To evaluate the similarity of the morphological data obtained from CalMorph-PC and CalMorph-PC(10), each PC load was normalized by t-value, and Pearson product moment correlation coefficients were calculated.

We also analyzed the morphological parameters assigned to the same PC pattern. Parameters with similar maximum PC loads were considered to have the same PC pattern (Figure S5). Direct comparisons of the time-dependent changes for each parameter are shown in Figure S6.

Creation of the training dataset

The morphological data acquired by CalMorph-PC and the corresponding alcohol concentrations were combined for each time series of the culture set to form a morphology–ethanol dataset. For fair learning, training and verification data are usually assigned by random sampling, but if random sampling were performed for each data point in this analysis, there is a risk that the past will be predicted using future information. Therefore, the dataset was divided for each culture unit and leave-one-out cross-validation was performed.

Model training

An AI model and a robust AI model were created using morphology–ethanol datasets from 5 and 18 independent experimental replicates, respectively. The algorithm used a Gradient-Boosting Decision Tree (GBDT) and a neural network. The xgblinear method (Chen and Guestrin 2016) of the xgboost package in R was used for GBDT and the netnet method (Ripley and Venables 2016) of the nnet package in R was used for the neural network. The caret package (Kuhn 2008) in R was used to manage machine learning. During training, the data were first quantile normalized, the standardized data were trained, and the training was completed after confirming that the loss function took the minimum value. As the loss function, the sum of the fitting criterion and weight decay, and the root mean square error (RMSE) was
used for the neural network and GBDT, respectively. The training and validation data were allocated in a 4:1 ratio, and 5-fold cross-validation was performed to generate predictions. The coefficient of determination, RMSE, and MAE (Mean Absolute Error) were then calculated by comparison with the actual amount of ethanol, and the average value of each was used as the evaluation index of the model.

Hyperparameter tuning

Hyperparameter tuning of the neural network and GBDT was performed manually using the functions in the caret package. For the neural network and GBDT, “tuneGrid = expand.grid (size = c (1:10), decay = seq (0.1, 1, 0.1),” and “tuneLength = 4” were used. To avoid overfitting, decay (weight decay) was used.

Model prediction of new test data

Test data obtained with CalMorph-PC(10) were input to the robust AI model to forecast the amount of ethanol. The coefficient of determination, RMSE, and MAE were calculated by comparing these results with the actual amount of ethanol, and the average value of each was taken to evaluate the model for the new test data.

Results

Platform for AI-based forecasting with yeast morphological data

The platform developed in this study acquires morphological images of the budding yeast S. cerevisiae using a nonstaining protocol, extracts high-dimensional morphological data using image processing software specialized for budding yeast, and then forecasts ethanol yield using supervised machine learning (Figure 1).

We used CalMorph-PC, an established nonstaining image acquisition protocol (Ohnuki, Nogami and Ohya 2009), to obtain the yeast morphological data (Figure 1a). In previous studies, we acquired yeast morphological data using a triple-staining protocol and the CalMorph program (Ohya et al. 2015; Okada and Ohya 2016). The triple-staining CalMorph protocol requires at least 3 days for analysis, whereas unstained cells can be directly analyzed using phase-contrast microscopy within 1 day using CalMorph-PC. The triple-stained images provided information regarding the cell wall, actin and nuclear DNA, and 501 morphological parameters could be extracted from these data. However, only 31 morphological parameters could be extracted from the unstained phase-contrast images. The CalMorph program was applied and adapted to create the CalMorph-PC program, in which the phase-contrast images were preprocessed using a Java-based program (Ohnuki et al. 2014) before digital image acquisition with CalMorph (Figure 1a).

CalMorph automatically extracted high-dimensional morphological parameters from the digital yeast cell images within minutes (Figure 1b). Initially, we examined the 33 CalMorph parameters that were related to cell shape, such as mother cell size, bud cell size, ratio of budded cells, and bud neck size (as summarized in Figure S1). However, two parameters related to luminescence were omitted from the analysis, because the phase-contrast images provided no morphological information regarding luminescence. Therefore, our program performed 31-dimensional morphological analyses.

At the final stage of our platform, time-dependent ethanol production was forecasted using yeast morphological data obtained with CalMorph-PC (Figure 1c). An AI-prediction model was created using the 31-dimensional morphological data as the explanatory variable and alcohol production as the objective variable. To forecast ethanol production in our model system, yeast industrial strain ATCC24702 was grown and cells were collected every 30 min for a total of 9 h and once more after 24 h to obtain morphological information and to measure the amount of ethanol produced at each time point. The AI model was created using these reference data 5 times by performing independent experiments, on the same collection date, but culturing in different flasks and preparing medium separately. This brought the total number of data points of the morphology–ethanol dataset to 3200 ((31 + 1) × 20 × 5). We then created and
Dynamic changes in yeast morphology during fermentation

In our previous study, we found that the cell wall, actin, and nucleus morphology changed dynamically during cell proliferation in bottom-fermenting brewer's yeast (Ohnuki et al. 2014). In this study, we first investigated how the cell shape changed during the fermentation process using the nonstaining image acquisition protocol. To determine the morphological patterns, we performed a principal component analysis (PCA) with the morphological data, in which the morphological measurements were classed as observations and the different time points were classed as variables. The morphological parameters that correlated with the same principal components (PCs) changed in a similar pattern over time (Raychaudhuri, Stuart and Altman 2000; Ohnuki et al. 2014). Using 5 replicates of 31-dimensional morphological data, 4 PCs were detected when the cumulative contribution ratio reached 80% (Figure S2).

We found that the morphological traits related to each PC displayed similar changes over time (Figure 2). Therefore, we focused on the morphological parameters that correlated with each PC (Figure S3). The value of PC1, represented by C11-2 (bud cell size), fluctuated after the start of cultivation, and then stabilized (Figure 3a, b). The change in PC2, represented by C12-1 (mother cell outline length), was transient with the start of the logarithmic growth phase (Figure 3a and c). The values in PC3, represented by C120 (small bud ratio), decreased during the logarithmic growth phase (Figure 3a and d). The change in PC4, represented by C121 (medium bud ratio), occurred later in the growth period compared to the change in PC3 (Figure 3a and e). These findings indicate that the yeast cell shape changed dynamically and in accordance with growth and ethanol production phase.

Accuracy of forecasting by reference data

An AI-prediction model was created to forecast alcohol production, using 31-dimensional morphological data as the explanatory variable. Fermentation is a highly complex process in which metabolites are produced and nutrients are consumed simultaneously. Various experimental factors can influence the accurate acquisition of morphological data during the yeast fermentation process, which can render traditional linear models of fermentation imprecise and unreliable. Therefore, we employed 2 nonlinear algorithms, a neural network and GBDT, to build and test models of the yeast fermentation process.

We created the models using reference data including the morphological data acquired by CalMorph-PC and corresponding alcohol concentrations from 5 independent experimental replicates. The coefficient of determination was then calculated using the leave-one-out method. The neural network algorithm produced an $R^2$ value of 0.92 (RMSE = 0.12), while the GBDT algorithm generated an $R^2$ value of 0.70 (RMSE = 0.20) (Table 1). We then applied the models to forecast the amount of alcohol that would be produced 30 and 60 min after morphological data acquisition. The neural network model forecasting values were strongly correlated with the reference values, with $R^2 = 0.91$ and $R^2 = 0.92$ for 30 and 60 min after image acquisition, respectively (Table 1). We also examined forecasting performance during the fermentation process. The results showed that the forecasting performance of the neural network model was remarkably high for the middle stage of the yeast fermentation culture (Figure S7). These results suggest that our AI model can be used to forecast the alcohol yields.

Shortening of fixation time for rapid monitoring

Yeast morphological data can be acquired on the same day using CalMorph-PC. However, to complete realistic monitoring until the next measurement time, the cell images must be obtained within minutes, not hours. Therefore, we developed the CalMorph-PC(10) protocol, which can rapidly acquire phase-contrast yeast image data. For the CalMorph-PC(10) protocol, we shortened the fixation time from 75 min to 1 min (Figure 4).

This change enabled the rapid acquisition of morphological data for AI forecasting. This is not a real-time monitoring, but it does provide sufficient morphological data before the next measurement.

We investigated the time-dependent morphological changes in the morphological dataset obtained with the CalMorph-PC(10) protocol by comparing them to CalMorph-PC datasets. Morphological data obtained using the 2 protocols were combined and subjected to PCA (Figure S4). We compared the patterns of morphological change over time between the 2 methods. To achieve this, we first investigated the pattern similarity by considering each PC as an index. We found significant correlation between the 31 morphological outputs of CalMorph-PC(10) and CalMorph-PC ($R = 0.929; P < .01$, noncorrelation test; Figure 5), which indicated that the patterns were similar.

Next, we determined which morphological parameters were closely related to each PC, and compared them between CalMorph-PC(10) and CalMorph-PC. We found that 17 of the 20 parameters examined correlated with the same PC and exhibited similar time courses in both CalMorph-PC(10) and CalMorph-PC (Figure S5; Figure S6, red squares). These findings indicate that CalMorph-PC and CalMorph-PC(10) can be used to monitor morphological pattern changes over time mostly in a similar way.

CalMorph-PC and CalMorph-PC(10) emphasize accurate morphological change and measurement speed, respectively. To determine how accurately CalMorph-PC extracts the morphological information compared with CalMorph-PC(10), we examined the variance between experiments. We compared 5 replicated datasets obtained with the 2 protocols. We found that the dispersion between the experiments was significantly larger with CalMorph-PC(10) than with CalMorph-PC ($P < .01$, Mann–Whitney U-test, Figure S8). Based on these results, we concluded that CalMorph-PC is more suitable for the accurate extraction of morphological changes than CalMorph-PC(10).

Accuracy of forecasting by test data

To test the accuracy of the AI model, we used the morphological data rapidly acquired by CalMorph-PC(10) as test data. For this purpose, we reconstructed a robust model using 18 experimental replicates, including the 5 replicates already used as reference data. Then, the AI model was tested using the 5 replicates of test data obtained with CalMorph-PC(10) by planning an out-of-time validation using different yeast colonies and performing the analysis on a later collection date. The neural network model produced an $R^2$ value of 0.83 (RMSE = 0.22). The model was used to forecast the alcohol yields at 30 and 60 min after image data acquisition; the neural network model forecasting was reliable, with $R^2 = 0.90$ and $R^2 = 0.88$, respectively. Taken together, these
findings indicate that our predictive AI model is applicable to forecast the alcohol yields in the near future using morphological test data acquired during ethanol fermentation.

**Discussion**

In this study, we developed an AI model that could successfully forecast the amount of ethanol produced during yeast fermentation using omics data, i.e., 31-dimensional yeast morphological data obtained with CalMorph-PC(10). We validated the AI model using morphological data acquired in parallel with fermentation. The amounts of ethanol produced at the time of image acquisition, as well as at 30 and 60 min in the near future, were forecasted with high accuracy using our pretrained neural network model. Thus, CalMorph-PC(10) is suited for forecasting the fermented product in the near future. In addition, high-dimensional yeast morphology data can be used to monitor the vitality of yeast cells, allowing us to obtain important information on a yeast culture during fermentation, such as estimation of the reproduction rate of yeast populations, assessment of the
AI-based forecasting using yeast morphological data

physiological state of yeast cells, and evaluation of the fermentation conditions (Ohnuki et al. 2014). Thus, this study provides additional evidence showing the importance of morphology data during fermentation. AI-based forecasting based on cell morphology will enable reliable management and promote stable production, potentially rendering this program highly valuable to the fermentation industry (Kondakci and Zhou 2017).

To achieve successful management of fermentation, large quantities of morphological data must be acquired and analyzed within a short time period. Using the CalMorph protocol, 501-dimensional morphological data can be obtained rapidly from triple-stained images (Ohya et al. 2005). However, it takes several days to fix, stain, and observe the cells in the CalMorph protocol, rendering this method unsuitable for rapid AI-based

Figure 3. Time courses of morphological features that correlated with each principal component. (a) Line graphs showing growth (OD$_{600}$; black line), glucose concentration (%; red line), and ethanol concentration (%; blue line). Measurements were taken every 30 min for 9 h. Data points represent the mean value of 5 replicates. Error bars represent SD. Representative phase-contrast images of yeast at 90-min stages are shown (top). The time courses of morphological features that correlated with each principal component are shown as bars (PC1, red; PC2, green; PC3, blue; PC4, gray). (b-e) Time-course changes of the typical morphological features correlated with PC1 (b; bud cell size), PC2 (c; mother cell outline length), PC3 (d; small bud ratio), and PC4 (e; medium bud ratio). Shaded areas represent 1 SD from the mean value.
In this study, we found that ethanol production can be accurately forecasted using our AI model. The model forecasted ethanol production at a given time and at 60 min with considerable accuracy, with a coefficient of determination \( R^2 > 0.9 \). Typically, AI prediction is deemed accurate when the coefficient of determination is \( R^2 > 0.7 \). (Dessai et al. 2008; Sewsynker, Kana and Lateef 2015; Sewsynker-Sukai, Faloye and Kana 2017), so our model was not overfitted with respect to the training data, because the test data had a coefficient of determination of 0.83-0.90. Note that our test data were acquired after the training data according to the out-of-time validation method. Thus, these results clearly demonstrate the utility of our AI model.

Of the 3 machine learning algorithms tested, the neural network model was the highest performing, followed by the multiple regression analysis, and then the GBDT algorithm (data not shown). Neural networks are well suited for forecasting and identifying patterns within a limited information space (Pasini 2015; Shaikhina and Khovanova 2017). However, one limitation of neural networks is that they are unable to forecast information that is not similar to the learned information. Specifically, we do not yet know if our model would be applicable under conditions that differ to the ones tested in this study. For example, it may be necessary to build a new model for fermentation processes using 20% glucose as the substrate (Gomar-Alba, Morcillo-Parrar and Del Olmo 2015; Scherer 2016). Research is currently underway in our laboratory on how the performance of the neural network -based forecasting depends on glucose concentration up to 20%.

**Acknowledgments**

We thank members of the Laboratory of Signal Transduction for helpful discussions. This work was supported by NEDO (W.O., K.K., and Y.O.).

**Supplementary material**

Supplementary material is available at *Bioscience, Biotechnology, and Biochemistry* online.
Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

Author contribution

Y.O. and K.K. conceived the study and designed experiments; K.I.-N. and N.K. performed experiments; S.W. build and test AI models; S.O. analyzed data and prepared figures; Y.O. wrote the draft; K.I.-N., N.K., S.W., and S.O. edited the manuscript; R.K., T.N., and W.O. interpreted results and edited the final version and manuscript; all authors reviewed the manuscript.

Funding

This study was supported by the New Energy and Industrial Technology Development Organization (NEDO), Japan, under Grant number JPNP19001.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Akinsemolu AA. The role of microorganisms in achieving the sustainable development goals. J Cleaner Prod 2018;182:139-55.
Belini VL, Suhr H, Wiedemann P. Online monitoring of the morphology of an industrial sugarcane biofuel yeast strain via in situ microscopy. J Microbiol Methods 2020;175:105973.
Campbell K, Westholm J, Kasvandik S et al. Building blocks are synthesized on demand during the yeast cell cycle. Proc Natl Acad Sci U S A 2020;117:7575-83.
Chen T, Guestarín C. Xgboost: a scalable tree boosting system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 13-17 August 2016:785-94, 2016. arXiv:1603.02754 [cs.LG].
Conacher CG, Naïdoo-Blassopoles RK, Rossouw D et al. Real-time monitoring of population dynamics and physical interactions in a synthetic yeast ecosystem by use of multicolour flow cytometry. Appl Microbiol Biotechnol 2020;104:5547-62.
Desai KM, Survase SA, Saudagar PS et al. Comparison of artificial neural network (ANN) and response surface methodology (RSM) in fermentation media optimization: case study of fermentative production of scleroglucan. Biochem Eng J 2008;41:266-73.
Gomar-Alba M, Morcillo-Parra MA, Del Olmo M. Response of yeast cells to high glucose involves molecular and physiological differences when compared to other osmostress conditions. FEMS Yeast Res 2015;15:1-14.
Guo W, Feng X. OM-FBA: integrate transcriptomics data with flux balance analysis to decipher the cell metabolism. PLoS One 2016;11:1-20.
Haandbæk N, Bürgel SC, Heer F et al. Characterization of subcellular morphology of single yeast cells using high frequency microfluidic impedance cytometer. Lab Chip 2014;14:369-77.
Jiang H, Xu W, Ding Y et al. Quantitative analysis of yeast fermentation process using Raman spectroscopy: comparison of CARS and VCFA for variable selection. Spectrochim Acta A Mol Biomol Spectrosc 2020;228:17781.
Kavšček M, Stražar M, Ćurk T et al. Yeast as a cell factory: current state and perspectives. Microb Cell Fact 2015;14:1-10.
Sewsynker Y, Kana EBG, Lateef A. Modelling of biohydrogen generation in microbial electrolysis cells (MECs) using a committee of artificial neural networks (ANNs). Biotechnol Biotechnol Equip 2015;29:1208-15.

Sewsynker-Sukai Y, Faloye F, Kana EBG. Artificial neural networks: an efficient tool for modelling and optimization of biofuel production (a mini review). Biotechnol Biotechnol Equip 2017;31:221-35.

Shaikhina T, Khovanova NA. Handling limited datasets with neural networks in medical applications: a small-data approach. Artif Intell Med 2017;75, 51-63.

Sherman F. Getting started with yeast. In: Guthrie C, Fink G. R. (eds), Methods Enzymol., Vol. 194. San Diego: Academic Press, 1991, 3-21.

Suzuki G, Wang Y, Kubo K et al. Global study of holistic morphological effectors in the budding yeast Saccharomyces cerevisiae. BMC Genomics 2018;19:1-14.

Vees CA, Veiter L, Sax F et al. A robust flow cytometry-based biomass monitoring tool enables rapid at-line characterization of S. cerevisiae physiology during continuous bioprocessing of spent sulfite liquor. Anal Bioanal Chem 2020;412:2137-49.

Vlassides S, Ferrier JG, Block DE. Using historical data for bioprocess optimization: modeling wine characteristics using artificial neural networks and archived process information. Biotechnol Bioeng 2001;73:55-68.

Vojinović V, Cabral JMS, Fonseca LP. Real-time bioprocess monitoring: Part I: In situ sensors. Sens Actuators, B 2006;114:1083-91.

Walker GM, Walker RSK. Chapter three—enhancing yeast alcoholic fermentations. Adv Appl Microbiol 2018;105:87-129.

Yu BY, Elbuenen C, Shen C et al. An integrated microfluidic device for the sorting of yeast cells using image processing. Sci Rep 2018;8:1-12.

Zakhartsev M, Reuss M. Cell size and morphological properties of yeast Saccharomyces cerevisiae in relation to growth temperature. FEMS Yeast Res 2018;18:1-16.

Zhuang S, Renault N, Archer I. A brief review on recent development of multidisciplinary engineering in fermentation of Saccharomyces cerevisiae. J Biotechnol 2021;339:32-41.

Zuzuarregui A, Monteoliva L, Gil C et al. Transcriptomic and proteomic approach for understanding the molecular basis of adaptation of Saccharomyces cerevisiae to wine fermentation. Appl Environ Microbiol 2006;72:836-47.