INTRODUCTION

In microbial bioprocesses, yeast extract is commonly used as a source of nitrogen, vitamins, and minerals. Yeast extract is a complex raw material usually produced from baker’s or brewer’s yeast through autolysis or chemical digestion (Bekatorou et al., 2006; Ferreira et al., 2010). It is also used as supplemental material in serum-free media for mammalian cell culture and human immunoglobulin production (Hu et al., 2015; Mosser et al., 2013). The composition varies among lots and brands because of its complex substrates, uncontrolled fermentation conditions during yeast cultivation, and variation of downstream processes during manufacturing (Christ & Blank, 2019). This variation results in compositional differences and often causes inconsistent fermentation performances in microbial processes. If this occurs, laboratory testing or screening of many yeast extracts is performed to determine the most promising extract suitable for large-scale use.
Recombinant protein expression in *Escherichia coli* is an important technology used in heterologous protein production (Sørensen & Mortensen, 2005). When producing recombinant proteins with an *E. coli* protein expression system, yeast extract is often added to increase enzymatic activity and protein production (Li et al., 1990; Mohammadi et al., 2018; Nancib et al., 1991). In some cases, other raw materials, such as sugar canes molasses and corn steep liquor, have been used in addition to yeast extract to increase heterologous protein production (Ye et al., 2010). The experimental design of a protein expression experiment can be used to optimize the medium composition (Ye et al., 2010). However, the variation in raw material composition is often ignored when optimizing medium components in the laboratory. Povin et al. developed an automated turbidimetric system to screen yeast extract for the growth of *Lactobacillus plantarum* (Povin et al., 1997). Near-infrared (NIR) spectroscopy has been applied to investigate the effects of yeast extract composition on recombinant protein production (Kaspro et al., 1998). In the case of mammalian cell cultivation, a combination of spectroscopy and chemometrics has been used for the characterization of raw materials in media (Trunfo et al., 2017). NIR is useful for real-time monitoring and quality checking of microbial cultivation (Vann & Sheppard, 2017). However, this method does not provide feedback information for optimizing the media because the NIR spectra do not sufficiently capture the composition of the media. In previous studies, we successfully used metabolomics-based approaches with non-targeted analyses via gas chromatography-mass spectrometry (GC-MS) and machine learning to estimate the effect of yeast extract on microbial growth (Konishi, 2020; Tachibana et al., 2019; Watanabe et al., 2019). We demonstrated that 165 peaks were observed using GC-MS when *E. coli* was cultivated in 24 different medium compositions with 6 different yeast extracts. The data fit well to the partial least squares regression (PLS) model with reasonable accuracies. Because they are important medium components, the PLS model estimated several amino acids, and some of these amino acids were found to influence *E. coli* growth in validation experiments (Tachibana et al., 2019). This approach was also applied to bioethanol production. In the model fitting of PLS and DNN, (Konishi, 2020; Watanabe et al., 2019) the volatile components of hydrolysates derived from lignocellulosic biomass served as independent variables, and ethanol and cell yields served as dependent variables. However, this metabolomics approach has never been applied to heterologous protein production by *E. coli* mutants.

In general, PLS and its modified methods, such as orthogonal projections to latent structures and soft independent modeling of class analogy, are used in metabolomics studies (Bylesjö et al., 2007). DNN is a powerful tool for analyzing datasets derived from biological systems. However, it appears to be inapplicable to metabolomics studies because it is difficult to identify the contributing factors, and it is expensive and time-consuming to obtain data to construct a satisfactory machine learning model. Date and Kikuchi reported the use of DNN with a mean decreased accuracy based on a permutation algorithm that achieved higher classification accuracy than random forest regression (RF) and PLS and identified important variables (Date & Kikuchi, 2018).

In this study, we applied a DNN-mediated metabolomics approach to improve the estimation of the effects of raw materials during microbial cultivation on foreign protein production by *E. coli* using heterologous GFP expression in *E. coli* with different yeast extract compositions. The PLS, RF, neural networks (NN), and DNN models were compared based on the degree of model fitting, and significant variations were estimated by a mean increase errors (MIE) calculation based on a permutation algorithm.

### 2 | MATERIALS AND METHODS

#### 2.1 | Microorganisms and chemicals

*Escherichia coli BL21(DE3)pLysS* (Invitrogen) was purchased from Thermo Fisher Scientific Japan. The pRSET-EmGFP bacterial expression vector was purchased from Thermo Fisher Scientific Japan. The vector pRSET-EmGFP was introduced to *E. coli BL21(DE3)pLysS* using the standard method in the users’ manual. The strain was cultivated in LB broth that included 10 g/L Bacto® Tryptone (Becton Dickinson and Co. [BD] Japan), 5 g/L Bacto® yeast extract (BD), and 10 g/L NaCl. The culture was incubated overnight at 37°C with shaking at 200 rpm. This was the inoculum used in the experiments. The culture broth was stored as frozen stocks with 30% glycerol in a deep freezer at −80°C. Experimental-grade yeast extracts were purchased from BD, Millipore Sigma Japan (Tokyo, Japan), Kyokuto Pharmaceutical Industrial Co. Ltd., and Nihon Pharmaceutical Co. Ltd. (Tokyo, Japan), and referred to as E1, E2, E3, and E4, respectively. Manufacturing-grade yeast extracts were provided by Oriental Yeast Co. Ltd., and Nihon Paper, and named M1, M2, M3, and M4.

#### 2.2 | GC-MS

To identify the hydrophilic components of the yeast extract, non-targeted GC-MS analyses were performed after trimethylsilylation according to a previous report (Tachibana et al., 2019). Each 5.0 g/L yeast extract sample (E1, E2, E3, E4, M1, M2, M3, and M4) and mixed samples (E1-E4, E2-E4, E3-E4, E4-M1, E4-M2, E4-M3, E1-M3, E2-M3, E3-M3, M3-M1, M3-M2, and M3-M4) were prepared and autoclaved at 121°C for 20 min. The sample (100 µl) was combined with 20 mg/ml of ribitol (60 µl). Then, 900 µl of water, methanol, and chloroform at a ratio of 1:2.5:1, respectively, was added. After extraction with thorough mixing, the tubes were centrifuged at 4°C for 5 min at 16,000×g. The top water phase (600 µl) was transferred into new tubes, dried partly by a centrifuge evaporator, and freeze-dried by a lyophilizer. Methoxyamine chloride (20 mg/ml in pyridine) was added to the lyophilized samples and incubated at 30°C for 90 min. After the incubation, N-methyl-N-(trimethylsilyl)trifluoroacetamide
was added and the mixture was incubated at 37°C for 30 min. The samples were then introduced into the GC-MS system.

The Agilent GC-MS system, 7980B and 5977A MSD was used with an HP-5 ms UI column (30 m × φ 0.25 mm ×fim thickness 0.25 μm). The instrument conditions were set as described previously (Tachibana et al., 2019). Peaks were obtained from total ion chromatograms using the deconvolution program in MassHunter software (Agilent Technology). The peak area was normalized by the internal standard (ribitol) peak. Peak annotation was performed with support from the NIST14 database.

### 2.3 Culture conditions

The frozen stocks (100 µl) were inoculated into 50 ml LB broth with 50 mg/L ampicillin and 35 mg/L chloramphenicol at 37°C for 9 h as a seed culture. To evaluate the effects of yeast extract supplementation on the yields of cells and GFP, 5.0 g/L yeast extract was added to M9 minimal salt medium composed of 12.0 g/L Na2HPO4, 6.0 g/L KH2PO4, 1.0 g/L NaCl, 2.0 g/L NH4Cl, 0.5 g/L MgSO4, 7H2O, 4.0 g/L glucose, 30 mg/L CaCl2·H2O, 20 mg/L thiamin hydrochloride, 50 mg/L ampicillin, and 35 mg/L chloramphenicol. The seed culture (1 ml, OD660 of approximately 5) was transferred into 50 ml M9 supplemented media in a 500-ml baffled Erlenmeyer flask and incubated at 37°C for 12 h at 200 rpm in an orbital shaker (G-BR-200, Taitec Co. Ltd).

Three hours after inoculation, 1 mM IPTG was added to induce GFP expression. Cell growth was monitored by measuring the turbidity at 660 nm using a spectrophotometer (V-630, JASCO Corporation). GFP expression levels were measured by a spectrofluorometer with a double monochromator and a micro drop sample holder (FP-8300, JASCO Corporation). For GFP quantification, the excitation and detection wavelengths were set at 487 and 509 nm, respectively. The fluorescence intensities at these wavelengths were used to quantify the GFP yields. Five microliters of diluted culture broth were measured using spectrofluoroscopy. The measurements were performed in at least triplicate after sampling at 0, 3, 6, 9, and 12 h.

### 2.4 Machine learning

The values of GFP intensities were decreased by five orders of magnitude before being evaluated by machine learning to adjust them to a suitable range for analysis using the DNN algorithm. In all machine learning algorithms except for principal component analysis (PCA), data from the E1 yeast extract were used for doubled validation calculations. The remaining data were separated into learning and test datasets with random cross-validation (85:15). PCA, PLS, and RF were performed on the Python 3.6 platform using the scikit-learn library (Buitinck et al., 2013). The number of components for the PLS models was set at 6. For RF, the parameters were set as the following: max_depth, 10; max_features, 6; max_leaf_nodes, none; n_estimators, 300; random_state, 2525; in case estimate cell yields and max_depth, 5; max_features, 169; n_estimators, 50; random_state, 2525; in case of GFP yield. The parameters were set after searching for the optimal parameters using the grid search function.

NN and DNN were coded in Python 3.6 using TensorFlow 1.15.0 and the Keras library ([Abadi et al., 2015]) in all cases, the input shape was set for 205 parameters. To estimate the final yield, the output shape was set to 1 parameter, cell yield, or GFP. For time-course estimation, the output shape was set for 5 parameters corresponding to the sampling time for each cell growth and GFP sample. Conventional NN was composed of a single hidden layer with 100 units of hyperbolic tangent (tanh) activations. The network was constructed with fully connected networks. HeNormal class was used as a kernel weight initializer. Activations of output layers were set to linear. Adam algorithms were applied to the optimizer with the default setting of the Keras library. Learning was carried out to minimize the mean squared error (MSE) (eq 1). The time of training was set at 3,000. Model checkpoint functions were record weights of the model with minimal MSE:

\[
\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]

where n indicates the number of input variables, \(\hat{y}_i\) indicates the measurement values of dependent variables, and \(y_i\) indicates the estimated values of the dependent variables by the constructed model.

DNN were constructed with 4 hidden layers (200, 100, 50, and 20 units) and tanh activations. The evaluation layer was activated by ReLU formulation. The units were set as “single” for estimating the final yields of cell growth and GFP production and as “5 units” for estimating the time courses of cell growth or GFP production, which correspond to each sampling time. The number of training times was set at 10,000. The other DNN parameters corresponded to those of NN.

MIE calculations were performed with reference to the MDA calculation reported by Date and Kikuchi (Date & Kikuchi, 2018) For the MSE calculation, the values in a variable were randomly rearranged among the input data, which was called a permutation, and the rearranged data matrices were evaluated by the constructed DNN model. The model loss obtained by the permutations was compared with the model loss determined by the MSE calculation. In the calculation, a relatively small influence on MSE means that the constructed model was rarely influenced by the variable. However, a relatively large influence on MSE means that the constructed model was significantly affected by the variable. Based on this criterion, the MIE can evaluate the importance of the variables in the constructed DNN model. In this study, permutations were repeated 60 times for each variable, and the average MSE for each variable calculated from the rearranged matrices was used as representative importance.

To evaluate the effect of the top 20 important variables estimated by DNN-MIE, a sensitivity analysis was performed to estimate cell growth and GFP yield while varying only a single important variable in the E4 yeast composition. Here, the dependent variables were estimated by the DNN model for the time courses of cell growth (Figure 3j) and GFP production (Figure 3i), when the single
important variables were sequentially varied from a maximum to minimum value in the matrix of E4 compositions.

A personal computer (PC) equipped with a graphic processing unit was used for the calculations. The PC spec. OS: Ubuntu 16.04 LTS; CPU: Intel Core i7-8700 (3.2–4.6 GHz / 6 cores / 12 threads / 12MB cache); memory: DDR-2666 32 GB; GPU: NVIDIA GeForce GTX 1080Ti 11GB.

2.5 Validation by cultivation with adding important components

To validate the estimation by DNN, E. coli EmGFP were cultivated in the basal medium containing 0.05 g/L of an important component as estimated by DNN. The experiments were performed in triplicate. The yield of cells and GFP were evaluated after 9 h of cultivation, and the yield fold changes were calculated by normalizing these yields in reference to the control cultivation. The significance of these values was evaluated by F tests and t-tests (p < 0.05).

3 RESULTS

3.1 Composition of yeast extract

GC-MS detected 205 peaks from trimethylsilylated compounds associated with yeast extract. The compounds included 50 amino acids and their derivatives, 17 saccharides, 7 sugar alcohols, 20 organic acids, 6 nucleotides, 7 fatty acids, 66 miscellaneous compounds, and 32 unknown compounds, as annotated by the NIST14 database. Figure 1 indicates the score plots for the compositions of yeast extracts based on PCA. The contribution ratios of PC1 and PC2 were 29.7% and 14.0%, respectively. Extract samples E1, E2, E3, and E4 were plotted right-down on the score plot. M1 and M2 were plotted right-up, and M3 and M4 were left side. Therefore, the sole yeast extract samples were separated into at least three clusters. Each mixed yeast extract sample was plotted at intermediated places. The data were summarized in a data matrix (in supplemental data) that was used for the machine learning analyses.

3.2 Cultivation

The cultivation results are summarized in Figures A1 and A2. Figure A1 indicates the time courses of cell growth as OD660. Figure A2 demonstrates the time courses of GFP fluorescence intensities. In the control experiment using M9 minimal medium (Figure A1U and Figure A2U), the cell growth was weak and the final yield was 1.11 ± 0.25. GFP production was also weakly induced, and the final GFP yield was 1.83 × 10^4 ± 1.52 × 10^2 after 9 h. All the yeast extracts stimulated cell growth and GFP expression. Cell growth and GFP drastically increased between 2 h and 4 h after inoculation, after which the curves plateaued or decreased slightly. The fold changes in growth after adding yeast extracts were between 2.72 and 4.50, and E3 was the best enhancer. The fold changes in GFP were between 1.62 and 2.84, and the best enhancer was E4. Experimental-grade yeast extracts tended to stimulate more cell growth and GFP production than manufacturing-grade yeast extracts. Interestingly, mixing both experimental-grade and manufacturing-grade yeast extracts increased cell growth and GFP production.

3.3 Comparing machine learning algorithms

Figure 2 shows boxplots of MSEs for the training data (MSE_{train}), crossed validation (MSE_{val}), and doubled validation (MSE_{test}) between different machine learning architectures. Learning calculations were carried out ten times for each machine learning. To summarize the results of the model fitting, MSE_{train}, MSE_{test} and MSE_{val} were observed as the smallest values in DNN in the calculated machine learning architectures and were observed as the higher values in the others.

Figure 3 shows plots of the measured and predicted values of the best model for each machine learning analysis. For the PLS model, the coefficients of determination for the training data (R^2_{train}) were 0.961 and 0.958 for cell growth and GFP yields, respectively. The coefficients of determination for the test data (R^2_{test}), which can be also defined as Q^2 in a metabolomics analysis, were 0.815 and 0.852, respectively (Figure 3a and e). The coefficients of determination in the cross-validation seemed to be sufficient in general metabolome analyses (Watanabe et al., 2019). However, the predicted values were severely varied in the test data and the validation data. RF showed similar R^2_{train} values to PLS, and higher R^2_{test} values than PLS, with lowered MSE_{train} and MSE_{test} values but large MSE_{val} values (Figure 3b and f). This indicates that RF overfit the training data and test data, which suggests a poor forecasting ability.
for the extrapolation data. The NN fit the training data and test data similar to RF, and the MSE values were one order of magnitude smaller than those of RF (Figure 3c and g). This means that the NN model can forecast extrapolation data. DNN demonstrated very high coefficients of determination and low MSE values using all data (Figure 3d and h). In the case of multivalent outputs using time-course data, the data were excellently fitted to DNN (Figure 3j and l), which were preferred to those of RF (Figure 3i and k).

### 3.4 Important variables

To identify the important variables, MIE were applied to the DNN models using multivalent output models (Figure 3j and l). Figure 4 indicates the top 20 most important variables based on the MIE calculations. Glycerol, phosphate, glutamic acid (Glu), and trehalose or maltose indicated high average MSE in the case of both cell growth and GFP production. Several of the amino acids observed were representatives of important components. To examine the numerical values of the important variables, we recalculated the relationships between the number of input variables, ordered by significance, and the MSE for each learning was indicated (Figure A3). The MSE decreased along with an increase in the number of variables and converged minimal values. The results meant that the thresholds of the importance values for cell growth and GFP were 18 and 15, respectively. According to the sensitivity analysis of the important variables, the M4 yeast extract used for a basal medium containing Glu, trehalose or maltose, isoleucine (Ile), lysine (Lys), phosphate, glycine (Gly), and aspartic acid (Asp) were predicted to increase cell yields by less than 30% (Figure A4).
Simultaneously, Glu, glycerol, phosphate, and Lys were estimated to increase the yield of GFP (Figure A5). Interestingly, almost all the important variables over the thresholds were estimated to exert a slight effect on the cell growth and GFP yields.

### 3.5 Validation by supplemental cultivation

We performed validation cultivations to confirm the estimation of important components by DNN-MIE. Figure 5 demonstrates the results of these validation experiments. Glu, maltose, Alanine (Ala), phenylalanine (Phe), Ile, trehalose, Lys, Asp, phosphate, Gly, and sorbitol significantly increased GFP yields, and leucine (Leu), Serine (Ser), threonine (Thr), asparagine (Asn), valine (Val), glycerol, and tyrosine (Tyr) significantly decreased GFP yields. In particular, because Glu increased the GFP yields to 112.9% ± 4.0% and the cell yields to 104.8% ± 1.5%, this amino acid was predicted to be the most important variable. Maltose stimulated the GFP yields by 106.5% ± 1.2% and the cell yields by 104.9% ± 1.1%. Asp induced less than 2% of the cell and GFP yields. Sorbitol showed no influence on the cell and GFP yields. The final pH was between 6.44 and 6.54 in all cases.

### 4 DISCUSSION

In this study, we evaluated the application of machine learning algorithms as a method to examine the composition profiles of various yeast extracts and their effect on GFP heterogeneous protein production by *E. coli*. According to the GC-MS profiling of yeast extracts, a variety of compositions were observed (Figure 1). Using 20 different compositions of yeast extracts, the yields of cell growth and GFP production in *E. coli* varied between 3.05 ± 0.04 and 5.00 ± 0.23 and between $2.55 \times 10^5 \pm 4.13 \times 10^3$ and $4.86 \times 10^5 \pm 4.17 \times 10^3$. 

**Figure 3** Measured and predicted values by each machine learning algorithm. (a), PLS model for cell yields; (b) RF model for cell yields; (c) NN model for cell yields; (d), DNN model for cell yields; (e) PLS model for GFP yields; (f) RF model for GFP yields; (g) NN model for GFP yields; (h) DNN model for GFP yields; (i), RF model for time courses of cell growth; (j) DNN model for time courses of cell growth; (k) RF model for time courses of GFP; (l) DNN model for time courses of GFP. Symbols: yellow circles, training data; red circles, test data; blue circles, validation data.
studies (Date & Kikuchi, 2018; Konishi, 2020). In particular, MSE than the losses of PLS. This trend has been observed in previous data tended to fit the algorithms with smaller estimating losses based on the comparison of algorithms (Figure 2 and Figure 3).

medium components, RF, NN, and DNN were applied to the present (Kimura et al., 2018; Tachibana et al., 2019; Watanabe et al., 2019). extract compositions via a metabolomics approach. PCA and PLS have been frequently applied to metabolomics approaches (Bylesjö et al., 2007; Date & Kikuchi, 2018; Kimura et al., 2018; Tian et al., 2018). However, the PLS algorithm did not fit the experimental data as well as the other algorithms, although the coefficients of determination ($R^2_{\text{learn}}$ and $R^2_{\text{test}}$; synonym $Q^2$) were generally sufficient (Kimura et al., 2018; Tachibana et al., 2019; Watanabe et al., 2019). To improve the estimation of the cultivation results from the medium components, RF, NN, and DNN were applied to the present data based on the comparison of algorithms (Figure 2 and Figure 3). The data tended to fit the algorithms with smaller estimating losses than the losses of PLS. This trend has been observed in previous studies (Date & Kikuchi, 2018; Konishi, 2020). In particular, $\text{MSE}_{\text{NN}}$ decreased in the case of NN. This means that NN can avoid overfitting the training data. DNN showed smaller losses than NN, and it was the best model for estimating cultivation results. The described DNN structure may not be the best model for the present data because the DNN structures can be further arranged. In addition, there is a limited amount of experimental data, and this limited dataset may affect the DNN model. However, the strategies using DNN algorithms improve the model accuracies in comparison with PLS. In general, it is difficult to calculate the important variables via DNN algorithms. In this study, the important variables can be estimated by DNN-MIL using permutation algorithms. Glu, Asp, trehalose or maltose, glycerol, and phosphate were estimated to be the important components for GFP production (Figure 4). Furthermore, analyses of the relationships between the input variables gave the top 18 and 15 most important variables that dominated the estimating accuracies, for cell and GFP yields, respectively. Indeed, adding additional Glu at 0.05 g/L increased the GFP yield by 12.9% when M4 yeast extract was used as a component of the production medium (Figure 5). These results demonstrate that DNN-MIL can calculate the features of yeast extract compositions for GFP production. However, the sensitivity analyses (Figures A4 and A5) estimated that the important variables were found by DNN-MIL and that the analyses determined less of an influence on the cell and GFP yields. We believe the differences were caused by the difference in input data. This was because the important variables were calculated using a global dataset of all yeast extracts used by DNN-MIL, while the sensitivity analyses were performed for individual specific yeast extracts (M4). These results show that individual important variables may weakly influence cellular activities such as growth and expressing foreign proteins in basal yeast extracts. These effects may vary among different brands and lots of yeast extract. Although glycerol was estimated to increase cell and GFP yields in the case of M4 yeast extract, the yields of cells and GFP were significantly decreased in the experimental validation (Figure 5). This difference in the results between the sensitivity analysis and the experimental validation was
observed. Thus, the risk of false positives or negatives using estimations made by machine learning is still a concern.

Glu, Ala, Phe, Ile, Lys, and Asp increased the cell and GFP yields, and Leu, Ser, Thr, Asn, Val, and Tyr decreased the cell and GFP yields (Figure 5). Chow et al. also reported that in recombinant *E. coli* BLR(DE3), Asp, Asn, Glu, and Glu increased the production of elastin-like polypeptides, which are recombinant peptide-based biopolymers that contain repetitive sequences enriched in Gly, Val, Pro, and Ala (Chow et al., 2006). In this study, Glu and Asp, but not Asn, increased the expression of GFP. These results may indicate that *E. coli* behaviors in the rich medium were varied compared with its activity in the basal media under standard culture conditions. Kumar et al. also reported that 20 mixed amino acids with chemically defined media increased recombinant peptide production by 40% in *E. coli* BL21 (DE3) (Kumar et al., 2020). Generally, the addition of amino acids to the growth medium can influence *E. coli* protein expression. In a rich medium, *E. coli* cells grow faster, and the expression of the majority of the translation apparatus genes is significantly elevated. This is consistent with known patterns of growth rate-dependent regulation and an increased rate of protein synthesis in rapidly growing cells. The behavior in minimal cells would be controlled by the biosynthesis of building blocks, such as de novo biosynthesis of amino acids and nucleotides (Baez et al., 2019; Tao et al., 1999). However, the effects of individual amino acids in a rich medium have not been sufficiently studied, and surprisingly, there is no common consensus today. Therefore, many engineers associated with industrial production are forced to screen for the best raw materials, such as different brands and lots of yeast extracts, because they have no information on the significant components in the raw materials. In this study, we demonstrated that the DNN-MIL algorithm can be applied to estimate the cell growth and GFP yield by a recombinant strain of *E. coli*, and it can predict the components that are most important for cell growth and GFP production. A part of this estimation was matched to the results of the validating cultivations with the additional components. In particular, Glu was estimated to be the most important variable in the DNN-MIL simulation. The GFP yield increased by 12.9% in the validating cultivation. These results imply that the DNN-MIL between compositions of raw materials, yields of cells, and heterologous protein production can provide promising information for the optimization of medium components and quality control. However, the DNN model may lead to fallacies because of the deviation of the learning dataset. Based on the sensitivity analysis, phosphate and glycerol were estimated to increase cell and GFP yields (Figure A5), but these components reduced the yields in the actual validating cultivation (Figure 5). The other components that could not be detected by GC-MS were ignored in the present study. These other components may affect the behaviors estimated by DNN-MIL. This weakness of the current strategy will be improved by enriching the datasets via increasing the numbers of raw materials and using additional instrumental analyses. To our knowledge, this is the first study to use a DNN-mediated approach for a regression model, although Date and Kikuchi have already demonstrated DNN-mediated metabolomics for a classification model (Date & Kikuchi, 2018).

In this study, only 20 experimental cases were used for calculating the machine learning models, and the amount of data may be insufficient for the DNN algorithm. We believe the model will be improved by increasing the variety of cultivation conditions and the number of medium compositions. However, it is advantageous to use the present machine learning method with a small dataset to decrease time-consuming and expensive data sampling although comprehensive trial-and-error cultivation may more easily lead to the optimal culture condition with or without machine learning. Therefore, to use machine learning mediated optimization in microbial cultivation, it will be generally important that a convenient DNN model is constructed using a small and reasonable dataset.

In conclusion, the GC-MS profiles of yeast extracts and cultivation yields of a heterologous protein fit best to the DNN algorithm. The MIL calculation based on a permutation algorithm identified the important variables that have the potential to enhance or reduce protein production and cell growth. The DNN-mediated omics-like analysis between media and cultivation can be applied to new strategies for optimizing medium compositions and for quality control of media components. In addition, DNN-mediated metabolomics approaches are also applicable to general metabolomics.

ACKNOWLEDGMENTS

This research was partly supported by the New Energy and Industrial Technology Development Organization (NEDO) project (Project code: P20011) of the Ministry of Economy, Trade and Industry (METI), Japan. We thank Korin Albert for editing a draft of this manuscript.

CONFLICT OF INTEREST

None declared.

AUTHOR CONTRIBUTIONS

Seiga Tachibana: Conceptualization (lead); Investigation (lead); Methodology (lead). Tai-Ying Chiou: Writing-original draft (equal); Writing-review & editing (equal). Masaaki Konishi: Project administration (lead); Supervision (lead); Visualization (lead); Writing-original draft (equal); Writing-review & editing (equal).

ETHICS STATEMENT

None required.

DATA AVAILABILITY STATEMENT

The data are provided in full in this paper except for the supplemental data (the DataMatrix for machine learning) which are available in the figshare repository at https://doi.org/10.6084/m9.figshare.14716263

ORCID

Masaaki Konishi https://orcid.org/0000-0002-0616-073X
REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jozefowicz, R., Jia, Y., Kaiser, L., Kudlur, M., … Zheng, X. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv, 1603.04467. https://arxiv.org/abs/1603.04467

Baez, A., Kumar, A., Sharma, A. K., Anderson, E. D., & Shiloach, J. (2019). Effect of amino acids on transcription and translation of key genes in E. coli K and B grown at a steady state in minimal medium. New Biotechnology, 49, 120–128. https://doi.org/10.1016/j.nbt.2018.10.004

Bekatorou, C., Psarianos, A., & Koutinas, A. (2006). Production of food grade yeasts. Food Technology and Biotechnology, 44, 407.

Buitinck, L., Louppe, G., Bolondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. arXiv, 1309.0238v1. https://arxiv.org/abs/1309.0238

Bylesjö, M., Rantalainen, M., Cloarec, O., Nicholson, J. K., Holmes, E., & Trygg, J. (2007). OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. Journal of Chemometrics, 20(8–10), 341–351.

Chow, D. C., Dreher, M. R., Trabbi-Carlson, K., & Chilkoti, A. (2006). Ultra-high expression of a thermally responsive recombinant fusion protein in E. coli. Biotechnology Progress, 22(3), 638–646. https://doi.org/10.1021/bp0503742

Christ, J. J., & Blank, L. M. (2019). Saccharomyces cerevisiae containing 28% polyphosphate and production of a polyphosphate-rich yeast extract thereof. FEMS Yeast Research, 19(3), foz011. https://doi.org/10.1093/femsyr/foz011

Date, Y., & Kikuchi, J. (2018). Application of a deep neural network to metabolomics studies and its performance in determining important variables. Analytical Chemistry, 90(3), 1805–1810. https://doi.org/10.1021/acs.analchem.7b03795

Ferreira, O., Pinho, E., & Vieira, J. T. (2010). Brewer’s Saccharomyces yeast biomass: characteristics and potential applications. Trends in Food Science and Technology, 21, 77.

Hu, D., Sun, Y., Liu, X., Liu, J., Zhang, X., Zhao, L., Wang, H., Tan, W. S., & Fan, L. (2015). Understanding the intracellular effects of yeast extract on the enhancement of Fc-fusion protein production in Chinese hamster ovary cell culture. Applied Microbiology and Biotechnology, 99(20), 8429–8440. https://doi.org/10.1007/s00218-015-6789-5

Kaspraw, P. R., Lange, A. J., & Kirwan, D. J. (1998). Correlation of fermentation yield with yeast extract composition as characterized by near-infrared spectroscopy, Biotechnology Progress, 14(2), 318–325. https://doi.org/10.1021/bp980001j

Kimura, K., Inaoka, T., & Yamamoto, K. (2018). Metabolome analysis of Escherichia coli ATCC25922 cells treated with high hydrostatic pressure at 400 and 600 MPa. Journal of Bioscience and Bioengineering, 126(5), 611–616. https://doi.org/10.1016/j.jbiosc.2018.05.007

Konishi, M. (2020). Bioethanol production estimated from volatile compositions in hydrolysates of lignocellulosic biomass by deep learning. Journal of Bioscience and Bioengineering, 129(6), 723–729. https://doi.org/10.1016/j.jbiosc.2020.01.006

Kumar, J., Chauhan, A. S., Shah, R. L., Gupta, J. A., & Rathore, A. S. (2020). Amino acid supplementation for enhancing recombinant protein production in E. coli. Biotechnology and Bioengineering, 117(8), 2420–2433. https://doi.org/10.1002/bit.27371

Li, X. L., Robbin, J. W., & Taylor, K. B. (1990). Pharmaceuticals from cultured algae. Journal of Industrial Microbiology, 5, 165.

Mohammadi, F., Nezafat, N., Berenjian, A., Naghadaripour, M., Zamani, M., Ghoshoon, M. B., Morowvat, M. H., Hemmati, S., & Ghasemi, Y. (2018). Extracellular production of a potent and chemically resistant nattokinase in immobilized Escherichia coli using response surface methodology. Current Pharmaceutical Biotechnology, 19(11), 856–868. https://doi.org/10.2174/13892010196618102115405

Mossor, M., Chevalot, I., Olfos, E., Blanchard, F., Kapel, R., Oriol, E., Marc, I., & Marc, A. (2013). Combination of yeast hydrolysates to improve CHO cell growth and IgG production. Cytotechnology, 63(4), 629–641. https://doi.org/10.1007/s10616-012-9519-1

Nancib, N., Brantlan, C., & Boudrant, J. (1991). Metabolic roles of peptone and yeast extract for the culture of a recombinant strain of Escherichia coli. Journal of Industrial Microbiology & Biotechnology, 8, 165.

Povin, J., Fonchy, E., Conway, J., & Champagne, C. P. (1997). An automatic turbidimetric method to screen yeast extracts as fermentation nutrient ingredients. Journal of Microbiological Methods, 29(3), 153–160. https://doi.org/10.1016/S0167-7012(97)00032-8

Sørensen, H. P., & Mortensen, K. K. (2005). Advanced genetic strategies for recombinant protein expression in Escherichia coli. Journal of Biotechnology, 115(2), 113–128. https://doi.org/10.1016/j.jbiotec.2004.08.004

Tachibana, S., Watanabe, K., & Konishi, M. (2019). Estimating effects of yeast extract compositions on Escherichia coli growth by a metabolomics approach. Journal of Bioscience and Bioengineering, 128(4), 468–474. https://doi.org/10.1016/j.jbiosci.2019.03.012

Tao, H., Bausch, C., Richardson, C., Blattner, F. R., & Conway, T. (1999). Functional genomics: expression analysis of Escherichia coli growing on minimal and rich media. Journal of Bacteriology, 181(20), 6425–6440. https://doi.org/10.1128/JB.181.20.6425-6440.1999

Tian, X., Yu, Q., Yao, D., Shao, L., Liang, Z., Jia, F., Li, X., Hui, T., & Dai, R. (2018). New insights into the response of metabolome of Escherichia coli O157:H7 to Ohmic heating. Frontiers in Microbiology, 9, 2936. https://www.frontiersin.org/article/10.3389/fmicb.2018.02936/full

Trufnio, N., Lee, H., Starkey, J., Agarabi, C., Liu, J., & Yoon, S. (2017). Characterization of mammalian cell culture raw materials by combining spectroscopy and chemometrics. Biotechnology Progress, 33(4), 1127–1138. https://doi.org/10.1002/btpr.2480

Vann, L., & Sheppard, J. (2017). Use of near-infrared spectroscopy (NIRS) in the biopharmaceutical industry for real-time determination of critical process parameters and integration of advanced feedback control strategies using MIDUS control. Journal of Industrial Microbiology and Biotechnology, 44(12), 1589–1603. https://doi.org/10.1007/s10295-017-1984-2

Watanabe, K., Tachibana, S., & Konishi, M. (2019). Modeling growth and fermentation inhibition during bioethanol production using component profiles obtained by performing comprehensive targeted and non-targeted analyses. Bioresource Technology, 281, 260–268. https://doi.org/10.1016/j.biortech.2019.02.081

Ye, Q., Li, X., Yan, M., Cao, H., Xu, L., Xiang, Y., Chen, Y., Xiong, J., Ouyang, P., & Ying, H. (2010). High-level production of heterologous proteins using untreated cane molasses and corn steep liquor in Escherichia coli medium. Applied Microbiology and Biotechnology, 87, 517.

How to cite this article: Tachibana, S., Chiou, T.-Y., & Konishi, M. (2021). Machine learning modeling of the effects of media formulated with various yeast extracts on heterologous protein production by Escherichia coli. MicrobiologyOpen, 10, e1214. https://doi.org/10.1002/mbo3.1214
Figure A1  Time courses of cell growth in media using various yeast extract. A, E1; B, E2; C, E3; D, E4; E, M1; F, M2; G, M3; H, M4; I, E1-E4; J, E2-E4; K, E3-E4; L, E4-M1; M, E4-M2; N, E1-M3; O, E2-M3; P, E4-M3, Q, E3-M3; R, E3-M1; S, M2-M3; T, M3-M4; U, control without yeast extract. Error bars indicate standard deviation.
Figure A2. Time courses of fluorescence of GFP in media using various yeast extract. A, E1; B, E2; C, E3; D, E4; E, M1; F, M2; G, M3; H, M4; I, E1-E4; J, E2-E4; K, E3-E4; L, E4-M1; M, E4-M2; N, E1-M3; O, E2-M3; P, E4-M3; Q, E3-M3; R, E3-M1; S, M2-M3; T, M3-M4; U, control without yeast extract. Error bars indicate standard deviation.
Figure A3  The number of important variables vs MSE. Error bars indicate standard deviation.
Figure A4  Sensitivity analysis of cell yields in a medium using M4 yeast extract. Blue lines indicate growths (OD$_{600}$) estimated from the individual yeast extract compositions by a DNN model. The red dashed line indicates the concentration of the component in the E4 yeast extract.
Figure A5  Sensitivity analysis of GFP yields in a medium using M4 yeast extract. Blue lines indicate GFPs production estimated from the individual yeast extract compositions by a DNN model. The red dashed line indicates the concentration of the component in E4 yeast extract.