Modeling the effects of media formulated with various yeast extracts on heterologous protein production by *Escherichia coli* using machine learning

Seiga Tachibana[1], Tai-Ying Chiou[1], and Masaaki Konishi[1]

[1]Kitami Institute of Technology

October 31, 2020

Abstract

In microbial manufacturing, yeast extract is an important component of growth media. The production of heterologous proteins is often varied because of yeast extract composition. To identify why this reduces protein production, the effects of yeast extract compositions on the growth and green fluorescent protein (GFP) production of engineered *Escherichia coli* were investigated using a deep neural network (DNN)-mediated metabolomics approach. We observed 205 peaks from various yeast extracts using gas chromatography-mass spectrometry. Principal component analyses of the peaks identified at least three different clusters. Using 20 different compositions of yeast extract in M9 media, the yields of cells and GFP in the yeast extract-containing media were higher than those in the control without yeast extract by approximately 3.0–5.0 fold and 1.5–2.0 fold, respectively. We compared machine learning models and found that DNN best fit the data. To estimate the importance of each variable, we performed DNN with a mean increase error calculation based on a permutation algorithm. This method identified the significant components of yeast extract. DNN learning with varying numbers of input variables provided numbers of the significant components. The influence of specific components on cell growth and GFP production was confirmed with a validation cultivation.

1 Introduction

In microbial bioprocesses, yeast extract is commonly used as 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.[1,2] It is also used as a supplemental material in serum-free media for mammalian cell culture and human immunoglobulin production.[3,4]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.[5] 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.[6] When producing recombinant proteins with the *E. coli* protein expression system, yeast extract is often added to increase enzymatic activity and protein production.[7,8,9] In some cases, other raw materials, such as sugar cane molasses and corn steep liquor, have been used in addition to yeast extract to increase heterologous protein production.[10] The experimental design of a protein expression experiment can optimize the medium composition.[10] However, the variation in raw material composition is often ignored when optimizing medium components in the laboratory. Porvin et al. developed an automated tubidimetric system to screen yeast extract for growth of *Lactobacillus plantarum* .[11] Near-infrared (NIR) spectroscopy has been applied to investigate the effects of yeast extract composition on recombinant protein production.[12]

In the case of mammalian cell cultivation, a combination of spectroscopy and chemometrics has been used for the characterization of raw materials in media.[13] NIR is useful for real-time monitoring and quality checking of microbial cultivation.[14]However, this method no longer provides feedback information for optimizing 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.[15,16,17] 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.[15] This approach was also applied to bioethanol production. In the model fitting of PLS and DNN,[16,17] 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 modelling of class analogy, are used in metabolomics studies.[18] 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. 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.[19]

In this study, we applied a DNN-mediated metabolomics approach to improve 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 (Tokyo, 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 Dikinson and Co. (BD) Japan, Tokyo, 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 -80degC. Experimental-grade yeast extracts were purchased from BD, Millipore Sigma Japan (Tokyo, Japan), Kyokuto Pharmaceutical Industrial Co. Ltd. (Tokyo, Japan), and Nihon Pharmaceutical Co. Ltd. (Tokyo, Japan), and referred to as E1, E2, E3, and E4, respectively. Manufacturing-grade yeast extracts were provided by manufacturers including Oriental Yeast Co. Ltd. (Tokyo, Japan), Nihon Paper (Tokyo, Japan), and named as M1, M2, M3, and M4.

2.2 GC-MS

To identify the hydrophilic components of yeast extract, non-targeted GC-MS analyses were performed after trimethylsilylation according to a previous report.[15] 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 121degC for 20 min. The sample (100 μL) was combined with 20 mg/mL ribitol (60 μL). Then, 900 μL of water, methanol, and chloroform at a ratio of 1:2.5:1, respectively, were added. After extraction with thorough mixing, the tubes were centrifuged at 4°C for 5 min with 16,000×g . The top water phases (600 μL) were 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 a HP-5 ms UI column (30 m × 0.25 mm × firm thickness 0.25 μm). The instrument conditions were set as described previously.[15]Peaks were obtained from total ion chromatograms using the decombolution program in MassHunter software (Agilent Technology, CA, USA). 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 Na_2HPO_6, 6.0 g/L KH_2PO_4, 1.0 g/L NaCl, 2.0 g/L NH_4Cl, 0.5 g/L $MgSO_4$[?]$7H_2O$, 4.0 g/L glucose, 30 mg/L $CaCl_2$[?]H_2O, 20 mg/L thiamin hydrochloride, 50 mg/L ampicillin, and 35 mg/L chloramphenicol. The seed culture (1 mL, OD_{660} of approximately 5) was transferred into 50 mL of media in a 500 mL baffled Erlenmeyer flask and incubated at 37degC for 12 h at 200 rpm in an orbital shaker (G[?]BR-200, Taitec Co. Ltd., Tokyo, Japan). 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, Tokyo, Japan). GFP expression levels were measured by a spectrofluorometer with a doubled monochrometer and a micro drop sample holder (FP-8300, JASCO Corporation, Tokyo, Japan). 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 represent 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. In all machine learning algorithms except for principal component analysis (PCA), data from the E1 yeast extract was 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.[20] 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 1.5 and the Keras library (*https://keras.io/*).[21]In all cases, the input shape was set for 205 parameters. To estimate the final yield, the output shape was a single 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 times of training was set at 3,000. Model check point functions were record weights of the model with minimal MSE.

$$MSE \;=\; \frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i\right)^2$$

(eq. 1), where n indicates the number of input variables,y_i indicates the measurement values of dependent

variables, and y_i indicates the estimate 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 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[19] For the MSE calculation, the values in a variable were randomly rearranged among the input data, which was called 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 variables. Based on the 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 a representative importance.

To evaluate the effect of the important variables, a sensitivity analysis was performed to estimate the cell growths and GFP yields while varying only a single important variable in the yeast composition.

A personal computer (PC) equipped with a graphic processing unit were used for the calculations. PC Spec. OS: Ubuntu 16.04LTS, CPU: Intel Core i7-8700 (3.2-4.6 GHz / 6 cores / 12 threads / 12MB cash), 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 14.0% and 9.0%, 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 that was used for the machine learning analyses.

3.2 Cultivation

The cultivation results are summarized in **Figures S1and S2** . **Figure S1** indicates the time courses of cell growth as OD_{660}. **Figure S2** demonstrates the time courses of GFP fluorescence intensities. In the control experiment using M9 minimal medium (Figure S1U and Figure S2U), 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 $1.83\times10^4\pm1.52\times10^2$ after 9 h. All of the yeast extracts stimulated cell growth and GFP expression. Cell growth and GFP drastically increased between 2 h and 4 h after inoculation, and then 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 the 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_{test}), and doubled validation (MSE_{val}) between different machine learning architectures. Leaning calculations were carried out ten times in each machine learning. For estimating cell yields, the MSE_{train} of PLS, RF, NN, and DNN recorded 1.10×10^{-2} +- 9.9×10^{-4}, 1.62×10^{-3} +- 1.12×10^{-3}, 2.05×10^{-3} +- 1.40×10^{-3}, and 7.30×10^{-4} +- 9.00×10^{-4} as the means +- standard deviations, respectively. The MSE_{test} of PLS, RF, NN, and DNN recorded 7.58×10^{-2} +- 2.09×10^{-2}, 9.78×10^{-2} +- 7.10×10^{-2}, 3.67×10^{-2} +- 2.33×10^{-2}, and 4.90×10^{-3} +- 5.30×10^{-3}, respectively. The MSE_{val} of PLS, RF, NN, and DNN recorded 9.84×10^{-1} +- 5.57×10^{-1}, 5.70×10^{-1} +- 8.87×10^{-2}, 1.16×10^{-2} +- 1.29×10^{-2}, and 3.43×10^{-3} +- 3.37×10^{-3}, respectively. For estimating the GFP yields, the MSE_{train} of PLS, RF, NN, and DNN were 1.66×10^{-2} +- 2.85×10^{-3}, 1.08×10^{-2} +- 6.70×10^{-4}, 2.84×10^{-2} +- 3.43×10^{-2}, and 6.26×10^{-3} +- 1.31×10^{-2}, respectively. The MSE_{test} of PLS, RF, NN, and DNN were 7.04×10^{-1} +- 1.23×10^{-1}, 6.96×10^{-2} +- 4.48×10^{-2}, 4.12×10^{-2} +- 2.30×10^{-2}, and 8.70×10^{-3} +- 7.70×10^{-3}, respectively. The MSE_{val} of PLS, RF, NN, and DNN were 8.28×10^{-1} +- 5.25×10^{-1}, 3.55×10^{-1} +- 4.36×10^{-2}, 6.10×10^{-2} +- 2.67×10^{-2}, and 9.69×10^{-3} +- 1.31×10^{-2}, respectively. 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 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 (**Figures 3A and 3E**). The coefficients of determination in the cross-validation seemed to be sufficient in general metabolome analyses.[16] 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 **(Figures 3B and 3F**). This indicates that RF overfit the train data and test data, which suggests a poor forecasting ability for the extrapolation data. The NN fit the train data and test data similar to RF, and the MSE_{val} values were one order of magnitude smaller than those of RF (Figures 3C and 3G). This means that the NN model can forecast extrapolation data. DNN demonstrated very high coefficients of determination and low MSE values using all data (**Figures 3D and 3H**). In the case of multivalent outputs using time course data, the data were excellently fitted to DNN (**Figures 3J and 3L**), which were preferred to those of RF (**Figures 3I and 3K**).

3.4 Important variables

To identify the important variables, MIE were applied to the DNN models using multivalent output models (**Figures 3J and 3L**).**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. In order 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_{test} each learning were indicated (**Figure S3**). The MSE_{test} 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 S4**). Simultaneously, Glu, glycerol, phosphate, and Lys were estimated to increase the yield of GFP (**Figure S5**). Interestingly, almost all of 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 2.55×10^4 +- $4.13 \times 10^3$and 4.86×10^4 +- 4.17×10^3, respectively (**Figures S1 and S2**). The differences in GFP and cell yields were associated with the composition profiles of the yeast extracts. Then, we applied machine learning algorithms to determine the relationship between the cultivation results and the yeast extract compositions via a metabolomics approach. PCA and PLS have been frequently applied to metabolomics approaches.[18,19,22,23] However, the PLS algorithm did not fit the experimental data as well as the other algorithms, although the coefficients of determination (R^2_{learn} and R^2_{test}, synonym Q^2) were sufficient in general.[15,16,23] 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.[17,19] In particular, MSE_{val} decreased in the case of NN. This means that NN can avoid overfitting to 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 to 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, the relationships between the number of input variables give top 18 and 15 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 S4 and S5**) 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 each individual important variable 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 were 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), Asn, Asp, Gln, 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.[24] In this study, Glu and Asp, but not Asn, increased the expression of GFP. These results may indicate that*E. coli* behaviors in rich medium were varied compared with its activity in the basal media and standard culture conditions. Kurmar et al. also reported that 20 mixed amino acids with chemically defined media increased recombinant peptide production by 40% in *E. coli*BL21 (DE3).[25] Generally, the addition of amino acids to growth medium can influence *E. coli* protein expression. In rich medium, *E. coli* cells grow faster, and 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.[26,27] However, the effects of individual amino acids in 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 S5), but these components reduced the yields in the actual validating cultivation (Figure 5). The other components which 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 instrumentational 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.[19]

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 applicable to general metabolomics.

Acknowledgments

This research was partly supported by NEDO project (P20011) of METI, Japan.

Conflict of interest

The authors declare no financial of commercial conflict of interest. We thank Korin Albert, PhD, from Edanz Group (https://en-author-services.edanzgroup.com/ac) for editing a draft of this manuscript.

Data Availability Statement

The data that support the findings of this study are mainly available in the supplementary materials of this article. Additional data are available upon request.

Author Contributions

Seiga Tachibana : conceptualization, methodology, and investigation; **Chiou Tai-Ying** : writing, re-

viewing, and editing the manuscript; **Masaaki Konishi** : writing the original draft of the manuscript, visualization, supervision, and project administration.

5 References

[1] A. Bekatorou, C. Psarianos, A. A. Koutinas, *Food Technol. Biotechnol.* **2006** , *44* , 407.

[2] I. Ferreira, O. Pinho, E. Vieira, J. Tavarela, *Trend Food Sci.* **2010** , *21* , 77.

[3] D. Hu, Y. Sun, X. Liu, J. Liu, X. Zhang, L. Zhao, H. Wang, W. S. Tan, L. Fan, *Appl. Microbiol. Biotechnol.* **2015** ,*99* , 8429.

[4] M. Mosser, I. Chevalot, E. Olmos, F. Blanchard, R. Kapel, E. Oriol, I. Marc, A. Marc, *Cytotechnology* **2013** , *65* , 629.

[5] J. J. Christ, L. M. Blank, L.M., *FEMS Yeast Res.***2019** , *19* : foz011. DOI: 10.1093/femsyr/foz011.

[6] H. P. Sorensen, K. K. Mortensen, *J. Biotechnol.***2005** , *115* , 113.

[7] N. Nancib, C. Branlant, J. Boudrant, *J. Ind. Microbiol.***1991** , *8* , 165.

[8] X. L. Li, J. W. Robbin, K. B. Taylor, *J. Ind. Microbiol.***1990** , *5* , 165.

[9] F. Mohammadi, N. Nezafat, A. Berenijian, M. Negahdaripour, M. Zamani, M. B. Ghoshoon, M. H. Horowvat, S. Hemmati, Y. Ghasemi,*Curr. Pham. Biotechnol.* **2018** , *19* , 856.

[10] Q. Ye, X. Li, M. Yan, H. Cao, L. Xu, Y. Xhang, Y. Chen, J. Xiong, P. Ouyang, H. Ying, *Appl. Microbiol. Biotechnol.***2010** , *87* , 517.

[11] J. Povin, E. Fonchy, J. Conway, C. P. Champagne, *J. Microbiol. Methods* **1997** , *29* , 153.

[12] P. R. Kasprow, A. J. Lange, D. J. Kirwan, *Biotechnol. Prog.* **1998** , *14* , 318.

[13] N. Trunfio, H. Lee, J. Starkey, C. Agarabi, J. Liu, S. Yoon,*Biotechnol. Prog.* **2017** , *33* , 1127.

[14] L. Vann, J. Sheppard, *J. Ind. Microbiol. Biotechnol.***2017** , *44* , 1589.

[15] S. Tachibana, K. Watanabe, M. Konishi, *J. Biosci. Bioeng.* **2019** , *128* , 468.

[16] K. Watanabe, S. Tachibana, M. Konishi, *Bioresour. Technol.* **2019** , *281* , 260.

[17] M. Konishi, *J. Biosci. Bioeng.* **2020** ,*129* , 723 729.

[18] M. Bylesjo, M. Rantalainen, O. Cloarec, J. K. Nicholson, E. Holmes, J. Trygg, *J. Chemometrics* **2007** , *20* , 341.

[19] Y. Date, J. Kikuchi, *J. Anal. Chem.* **2018** ,*90* , 1805.

[20] L. Buitinck, G. Louppe, M. Bolondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. Vanderplas, A. Joly, B. Holt, G. Varoquaux, *arXiv***2013** , 1309.0238v1 https://arxiv.org/abs/1309.0238.

[21] M. Abadi, A. Agarwal. P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, R. Jozefowicz, Y. Jia, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, M. Schuster, R. Monga, S. Moore, M. Murray, C. Olah, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, *arXiv* **2015** , 1603.04467*https://arxiv.org/abs/1603.04467*

[22] X. Tian, Q. Yu, D. Yao, L. Shao, Z. Liang, F. Jia, X. Ji., T. Hui, R. Dai, *Front Microbiol.* **2018** , 9. 2936.*https://www.frontiersin.org/articles/10.3389/fmicb.2018.02936/full*

[23] K. Kimura, T. Inaoka, K. Yamamoto, *J. Biosci. Bioeng.***2018** , *126* , 611.

[24] D. C. Chow, M. R. Deher, K. Trabbic-Carlson, A. Chikoti, A.,*Biotechnol. Prog.* **2006** , *22* , 638 646.

[25] J. Kumar, A. S. Cauhan, R. L. Shah, J. A. Gupta, A. S. Rathore,*Biotechnol. Bioeng.* **2020** , *117* , 2420.

[26] H. Tao, C. Bausch, C. Richmond, F. R. Blattner, T. Conway, T.,*J. Bacteriol.* **1999** , *181* , 6425 6440.

[27] A. Baez, A. Kumar, A. K. Sharma, E. D. Anderson, J. Shiloach,*N. Biotechnol.* **2019** , *25* , 120 128.

Figure legends

Figure 1. PCA plot of yeast extract compositions. The percentages in brackets indicate the contributions of each component. Symbols: blue circles, E1; green circles, E2; red circles, E3; yellow circles, E4; blue triangles, M1; green triangles, M2; red triangles, M3; yellow triangles, M4; blue squares, mixture of E1 and E4; green squares, mixture of E2 and E4; red squares, mixture of E3 and E4; yellow squares, mixture of E4 and M1; blue diamonds, mixture of E4 and M2; green diamonds, mixture of E1 and M3; red diamonds, mixture of E2 and E3; yellow diamonds, mixture of E4 and M3; blue reverse triangles, mixture of E3 and M3; green reverse triangles, mixture of M3 and M1; red reverse triangles, mixture of M3 and M2; yellow reverse triangles, mixture of M3 and M4.

Figure 2. Comparison of $\mathrm{MSE}_{\mathrm{train}}$, $\mathrm{MSE}_{\mathrm{test}}$, and $\mathrm{MSE}_{\mathrm{val}}$ for each machine learning algorithm. A, $\mathrm{MSE}_{\mathrm{train}}$ for estimating cell yields; B, $\mathrm{MSE}_{\mathrm{train}}$ for estimating GFP yields; C, $\mathrm{MSE}_{\mathrm{test}}$ for estimating cell yields; D, $\mathrm{MSE}_{\mathrm{test}}$ for estimating GFP yields; E, $\mathrm{MSE}_{\mathrm{val}}$ for estimating cell yields; F, $\mathrm{MSE}_{\mathrm{val}}$ for estimating GFP yields. The triangles indicate means, the dashed lines indicate medians, and the boxes indicate quantiles. Circles indicate outliers. Error bars indicate 1.5-fold standard deviations. The number of replication: $n = 10$.

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.

Figure 4. Top 20 most important variables calculated by DNN-MIL. A, cell growth; B, GFP expression. Red dashed lines indicate minimal values of the averaged MSE in all variables.

Figure 5. The results of the validating cultivation. Each component was added at 0.05 g/L in basal medium with M4 yeast extract. Significance: *, $0.01 < p$ [?] 0.05; **, p [?] 0.01. Error bars indicate standard deviations.