COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

PERISCOPE-Opt: Machine learning-based prediction of optimal fermentation conditions and yields of recombinant periplasmic protein expressed in *Escherichia coli*

Kulandai Arockia Rajesh Packiam [a,1], Chien Wei Ooi [a,b], Fuyi Li [c], Shutao Mei [d], Beng Ti Tey [a,b], Huey Fang Ong [e], Jiangning Song [d,f,*], Ramakrishnan Nagasundara Ramanan [a,*]

[a] *Chemical Engineering Discipline, School of Engineering, Monash University Malaysia, Jalan Lagoon Selatan, 47500 Bandar Sunway, Malaysia*
[b] *Advanced Engineering Platform, Monash University Malaysia, Jalan Lagoon Selatan, 47500 Bandar Sunway, Selangor, Malaysia*
[c] *Department of Microbiology and Immunology, The Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Victoria 3010, Australia*
[d] *Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Victoria 3800, Australia*
[e] *School of Information Technology, Monash University Malaysia, Jalan Lagoon Selatan, 47500 Bandar Sunway, Malaysia*
[f] *Monash Centre for Data Science, Faculty of Information Technoology, Monash University, Victoria 3800, Australia*

ARTICLE INFO

Article history:
Received 9 April 2022
Received in revised form 1 June 2022
Accepted 1 June 2022
Available online 3 June 2022

Keywords:
Optimization
Machine learning
Recombinant protein production
Periplasmic expression
Prediction model
Escherichia coli

ABSTRACT

Optimization of the fermentation process for recombinant protein production (RPP) is often resource-intensive. Machine learning (ML) approaches are helpful in minimizing the experimentations and find vast applications in RPP. However, these ML-based tools primarily focus on features with respect to amino-acid-sequence, ruling out the influence of fermentation process conditions. The present study combines the features derived from fermentation process conditions with that from amino acid-sequence to construct an ML-based model that predicts the maximal protein yields and the corresponding fermentation conditions for the expression of target recombinant protein in the *Escherichia coli* periplasm. Two sets of XGBoost classifiers were employed in the first stage to classify the expression levels of the target protein as high (>50 mg/L), medium (between 0.5 and 50 mg/L), or low (<0.5 mg/L). The second-stage framework consisted of three regression models involving support vector machines and random forest to predict the expression yields corresponding to each expression-level-class. Independent tests showed that the predictor achieved an overall average accuracy of 75% and a Pearson coefficient correlation of 0.91 for the correctly classified instances. Therefore, our model offers a reliable substitution of numerous trial-and-error experiments to identify the optimal fermentation conditions and yield for RPP. It is also implemented as an open-access webserver, PERISCOPE-Opt (http://periscope-opt.erc.monash.edu).

1. Introduction

Recombinant protein production (RPP) is a noteworthy biotechnological technique that finds rising applications in various sectors such as healthcare, detergents, food industry, and, most importantly, in research and development [1,2]. *Escherichia coli* (*E. coli*) is considered an ideal host for RPP because it offers numerous advantages, including simple nutritional requirements, faster cel-

Abbreviations: AUC, area under the curve; CfsSubsetEval, Correlation-based Forward Selection Subset Evaluator; ClassifierSubsetEval, Classifier Subset Evaluator; CV, cross-validation; *E. coli*, *Escherichia coli*; FC1, Feature Category 1; FC2, Feature Category 2; FC3, Feature Category 3; FC4, Feature Category 4; IPTG, isopropyl β-D-1-thiogalactopyranoside; LOOCV, Leave-one-out cross-validation; MAE, mean absolute error; MCC, Mathew correlation coefficient; ML, machine learning; MLR, machine learning in R; OD, optical density at 600 nm; PCC, Pearson correlation coefficient; pI, isoelectric point; RF, random forest; RFR, RF regression; RFR-High, RFR for high; RFR-Medium, RFR for medium; RMSE, root mean squared error; RPP, Recombinant protein production; RSM, response surface methodology; SMOTE, Synthetic Minority Oversampling Technique; SP, signal peptides; SVM, support vector machines; SVR, SVM regression; SVR-Low, SVR for class: "low"; XGB, XGBoost.

[*] Corresponding authors.

E-mail addresses: Jiangning.Song@monash.edu (J. Song), Ramanan@monash.edu, upmram@gmail.com (R.N. Ramanan).

[1] Present address: Computer Science & Engineering Discipline, Indian Institute of Technology Gandhinagar, Gandhinagar, Gujarat 382 055, India.

https://doi.org/10.1016/j.csbj.2022.06.006
2001-0370/

lular growth, and easiness in achieving high cell densities [3,4]. During RPP, the protein of interest can be directed towards the periplasmic space of *E. coli* by the use of a short amino acid sequence called signal peptides (SP). The periplasmic expression of recombinant proteins is preferred because the periplasmic space provides an oxidized environment that improves protein folding, especifically for proteins containing di-sulphide bonds. Moreover, the target proteins can be selectively recovered from the periplasmic space using the milder cell disruption steps that avoid the release of cytoplasmic content to the processing fluid [5–8]. The increasing demands for recombinant proteins have driven the necessity of optimizing various fermentation process parameters to achieve the maximal RPP. Despite tremendous works devoted to the aspects of RPP and technological advancement, achieving the high yields of recombinant proteins remains a challenge. Moreover, optimization of RPP involves tedious, costly, and time-consuming experiments to identify the optimal fermentation conditions which are specific to each type of protein [3].

Machine learning (ML) techniques have emerged as a game-changer in many areas of research, including the field of biotechnology. Biotechnological processes such as RPP can be deciphered using ML-based prediction tools, which are developed using the available data to provide a rough estimate of the unknown biological responses. For instance, there are several notable ML-based prediction tools for various RPP-based applications, including the prediction of protein solubility, protein folding rates, and protein expression yields. PROSO II [9], ccSOL Omics [10], Protein–Sol [11], DeepSol [12], PaRSnIP [13], SoDoPE [14] and SolTranNet [15] are among the ML-based tools that predict the protein solubility with high accuracies. Similarly, the well-known ML-based prediction tools for determining the protein folding rates are K-Fold [16], Pred-PFR [17], PRORATE [18], and SeqRate [19]. Additionally, ESPRESSO [20] and Periscope [21] are the advanced ML tools that can predict the protein expression yields in the cytoplasm and periplasm of *E. coli*, respectively; both tools incorporate the concepts of previously developed models used in the determination of protein solubility and folding rate. The above-mentioned ML-based tools were developed mainly by using the key features associated with the amino acid sequences. This is because the amino acid sequence primarily influences the protein solubility and folding rates, which in turn affects many facets of RPP [3,21]. A limited number of studies have also considered features from other aspects, such as gene sequences [22] and host strains and/or vectors [23]. Nevertheless, the ML models based on amino-acid sequence do not necessarily provide a clear notion of the optimized condition during RPP because the models exclude features related to the fermentation process conditions, which play vital roles in increasing the yields of RPP.

The present study aimed to derive a global ML-based model capable of predicting the optimal protein yield and fermentation process conditions for a target recombinant protein to be expressed in the periplasmic space of *E. coli*. Two sets of XGBoost (XGB) classifiers were employed in the first stage to classify the target protein into high (>50 mg/L), medium (between 0.5 and 50 mg/L) or low (<0.5 mg/L) expression levels. In accordance with the classified level of protein expression, the predictions of optimal fermentation conditions and yields were then attained using three sets of regression models based on support vector machine (SVM) or random forest (RF). This ensembled model was developed by integrating data from an existing bioinformatics tool, Periscope, and the data from literature and in-house experiments; in total, 84 protein-types and 103 SP-protein combinations were used in this study. 11,985 features were initially extracted by considering all important factors associated with the amino acid sequence and fermentation process condition. Then, the extraction of key feature-subsets using stepwise feature selection method was performed. The resultant robust model gives a good estimate of the maximal amount of recombinant protein and the fermentation conditions responsible for the optimal protein expression. Ultimately, our prediction tool could minimize the time spent on trial-and-error experiments to attain high yields of recombinant proteins.

2. Methods

2.1. 1. Generation of datasets

The data used in the present study include the amino acid sequence of the recombinant protein expressed in the periplasm of *E. coli*, the corresponding protein expression yield measurable in milligrams per litre, and the parameters of the fermentation process conditions. The data were collected from: i) an existing prediction model, namely Periscope [21]; ii) a literature search using popular search engines such as Scopus, Google Scholar, and PubMed; and (iii) our in-house experimental findings (**Tables S1 and S2**). These data were extracted from the research articles fulfilling the following criteria: i) *E. coli* strain as the host and *lac* promoters for expression; ii) heterologous protein expression in the periplasm; iii) SP at the N-terminus; iv) batch fermentation at shake flask scale, and finally, v) neither involving any genetic modification of the host strain nor including any co-expression vectors. On the whole, the present study comprised 461 datasets (collection of data) for 84 proteins and 103 SP-protein combinations. The sequence redundancy of the 103 SP-protein combinations was removed using the CD-HIT suite [24] at 90% of the sequence similarity threshold. Soluble protein expression level and protein expression yield were chosen as the response variables for the classification and regression tasks, respectively. The independent test datasets involved (i) data from ten proteins in which their amino acid sequences are completely unknown/unseen to the model, and (ii) data from eighteen proteins in which their amino acid sequences are known but the optimal fermentation conditions are unseen to the model. The latter data correspond to the optimal fermentation conditions reported in the studies dealing with the statistical optimization of recombinant proteins based on response surface methodology (RSM). The independent test datasets allow the verification of the global optimization scheme as well as the validation of the prediction performance of the model on the unknown amino acid sequences. The test datasets were further split manually to ensure the equal distributions of high, low, and medium instances as well as different sizes of proteins in the case of unseen amino acid sequences. All the data used in the current study are available as the supplementary information in this article.

2.2. Feature extraction

A total of 11,985 initial features were extracted and further classified into four major categories of feature. Since RPP is primarily regulated by the amino acid sequence of a protein, we focused on the features that can be extracted directly from the amino acid sequences, as well as features like physico-chemical and structural properties derived indirectly using these amino acid sequences. Amongst these features, Feature Category 1 (FC1) constitutes general features such as the length of the protein, occurrences of each type of the 20 amino acid residues, and the maximum number of consecutive identical amino acid residues. FC1 also included other features such as the occurrences and the maximum number of consecutive amino acid residues with similar physico-chemical properties. Lastly, the structural features like molecular weight, isoelectric point (pI), net charge, solubility, protein folding rate, and helix/sheet propensity were also added to FC1 (refer to Tables S3-1 to S3-3 for details). Because of the high dimensionality, the occurrences of each of the dipeptides were separately grouped into Feature Category 2 (FC2), as shown in Table S3-4. There is a possibility that the influence of each feature calculated from the amino acid sequences not only arises from the occurrences of respective residues but also due to the occurrence frequencies for a given protein length (i.e., the numbers of amino acid residues). To examine this possibility, we considered both occurrences as well as occurrence frequencies (standardized by the length of the protein) as two separate features. With these additional features, FC1 and FC2 consist of 149 and 800 features, respectively. All the interactions between each of the two features from FC1 resulted in the derivation of 11,026 interactive features classified under Feature Category 3 (FC3). Finally, Feature Category 4 (FC4) encompasses 10 features extracted from the fermentation process conditions. The extensive features in FC4 include cell density (measured as the optical density at 600 nm), inducer concentration upon induction of protein expression, post-induction temperature and time, and all the six interactions between these features (Table S3-5). In order to avoid any potential biases resulted from different levels of the dataset, normalization of the data was performed using Equation (1). All features except those in FC3 and interactive features from FC4 were normalized since these features were calculated using the normalized FC1 and FC4 datasets, respectively.

$$x_{N,i,n} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{1}$$

where, $x_{N,i,n}$ and x_i are the normalized and actual values, respectively, of the feature x for the i-th protein, while x_{min} and x_{max} are the minimum and maximum actual values of the feature x amongst all the n proteins.

Cell density differs with respect to the type of fermentation media used, and accordingly, the cell densities resulting from different types of fermentation media were normalized separately. Similarly, the *lac* promoters could be induced by either isopropyl β-D-1-thiogalactopyranoside (IPTG) or lactose, and therefore, each type of inducer was normalized individually. As there is no direct relation reported between media or inducers, normalizations were performed using Equation (1) for all the data corresponding to each of the media and inducer type, and the normalized values for all the data were used directly for further study. The results of prediction of the optimized fermentation conditions assume utilizations of the most commonly used fermentation medium (Luria Bertani broth) and inducer (IPTG) as the default parameters.

2.3. Software packages and algorithms

The open-source software packages, Weka 3.8 [25] and R [26], were used in this study because they offer a wide range of algorithms for data-preprocessing, feature selection, and ML tasks. The ML tasks such as feature selection and benchmarking experiments were performed using Weka. Furthermore, the models were trained and evaluated using R with the machine learning in R (MLR) package [27]. A filter-based method, Correlation-based Forward Selection Subset Evaluator (CfsSubsetEval), was used in the selection of initial features. CfsSubsetEval selects the feature subsets on the basis of their high correlations to the response variable against their weak inter-correlation. Additionally, a wrapper-based method known as Classifier Subset Evaluator (ClassifierSubsetEval) was used in the final step of feature selection where the merits of the features was obtained by evaluating each feature subset in conjunction with the classification or regression algorithm; albeit being a time-consuming approach, ClassifierSubsetEval function could generate a reliable selection outcome. Three different algorithms, namely SVM, XGB, and RF, were used in both classification and regression tasks because they have been widely employed in bioinformatics for protein-based prediction tasks. These three algorithms were extensively tested and evaluated in the stages of feature selection, model training and independent test. Eventually, the most appropriate algorithm was selected for the specific tasks of classification and regression.

2.4. Feature selection

In this study, the important features were selected for both classification and the regression tasks based on a stepwise feature selection strategy, which includes: (i) Features were selected from both FC2 and FC3 using the CfsSubsetEval method along with the search method, Best First. (ii) The numbers of the selected FC3 features were further reduced using CfsSubsetEval and Greedy Stepwise methods. The 'Generate ranking' option was set to true, and the number of features to be retained was fixed as ten. (iii) Finally, the key optimal features were selected from all the features from FC1 and FC4 along with previously selected features from FC2 (Step 1) and FC3 (Step 2) using ClassifierSubsetEval. All the three classification and regression algorithms used in conjunction with ClassifierSubsetEval resulted in the selection of different key features and of varying numbers. The final selections of features as well as the training algorithm were conducted in consideration of the performance of the trained model together with the numbers and nature of the selected features. Furthermore, the relative importance of the selected features was evaluated according to the previously reported strategy [21], which is briefly described here: Each of the selected features was removed one at a time until all the selected features were removed completely; subsequently, the models were trained using the best-performing algorithms, and the changes in performance measures were computed and compared.

2.5. Training and evaluation of model

The three ML algorithms, SVM, XGB, and RF, were employed in both classification and regression tasks. Fine-tuning of the hyperparameters was not performed because the default values of these parameters in the MLR package gave a better result. Based on the performances of the given algorithms on the training datasets, the best-performing algorithm was chosen for further development of the model. Several widely-used performance metrics, including accuracy, error rate, precision, recall, F-measure, Mathew correlation coefficient (MCC), and area under the curve (AUC), were used in the performance evaluation of the classification models [see Equations (2 – 7)]. Similarly, Pearson correlation coefficient (PCC), mean absolute error (MAE), and root mean squared error (RMSE) were also calculated and used for the assessment of regression models [Equations (8 – 10)] [21]:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{2}$$

$$Error\ rate = \frac{FP + FN}{TP + FN + FP + TN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

$$PCC = \frac{n\sum(y_{Pred} \times y_{Act}) - \sum y_{Act}\sum y_{Pred}}{\sqrt{\left[n\sum {y_{Act}}^2 - \left(\sum y_{Act}\right)^2\right]\left[n\sum {y_{Pred}}^2 - \left(\sum y_{Pred}\right)^2\right]}} \tag{8}$$

$$MAE = \frac{1}{n}\sum |y_{Act} - y_{Pred}| \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum(y_{Act} - y_{Pred})^2} \tag{10}$$

where *TP*, *TN*, *FP* and *FN* represent the numbers of true positives, true negatives, false positives, and false negatives, respectively. y_{Act} is the actual value of the protein expression yield; y_{Pred} is the predicted protein expression yield; *n* is the number of instances used in the prediction.

The predictive performance of the models was assessed using internal cross-validation (CV) test where the whole dataset was split into either training dataset or internal testing dataset, based on the method utilized for CV. Leave-one-out cross-validation (LOOCV) test was chosen as a CV method for the model assessment owing to the lower number of datasets used in both classification and regression tasks. In LOOCV, the model was trained with all but leaving out one dataset for internal testing, and the whole training process was repeated until all the datasets had been internally tested. The performance metrics were averaged for all the cases during LOOCV. Subsequently, the model was validated using a set of unseen instances, i.e., independent test dataset, and the corresponding performance of the model was correlated to the prediction ability of the developed model.

2.6. Webserver implementation

The prediction model has been implemented as an online webserver, PERISCOPE-Opt (https://periscope-opt.erc.monash.edu/), to provide users with easy access to the model and its predictions. Based on the amino acid sequences of SP and protein provided as inputs, the proposed model predicts the optimal fermentation conditions corresponding to the maximal yield of recombinant periplasmic protein. Reactjs framework was used in the web implementation, processing the user-defined input data and then returning the outcomes of the model predictions. To eliminate the potential impact of service disruption from other external web tools linked to our model, PERISCOPE-Opt requires the users to manually key in the data generated from the relevant external web tools. A few web tools retrieving the necessary protein information were suggested and can be assessed via the web interface.

3. Results

3.1. Computational framework of the proposed optimization model

The proposed prediction model (Fig. 1) was constructed as a two-stage architecture with the following components: i) two sets of XGB classifiers in the first stage to classify the expression level of the given protein as low, medium or high class, and ii) three sets of regression models trained using the algorithms (SVM and RF) to quantify the protein yield with respect to each class in the second stage. In the training of classification model, we adopted two strategies to address the issue of data imbalance arising from the unequal distribution of the three classes. Firstly, two binary classifiers were utilized in a way that the first classifier categorized the given data into the majority class (medium) and both the minority classes (high and low) together, followed by the classification of the minority classes using the second classifier. Secondly, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to generate additional dummy data in the minority classes during both classification tasks (**Table S4**). Therefore, in the first stage, XGB–Classifier 1 categorized the input of amino acid sequence into either the class of medium-level expression or the class of non-medium-level expression. If the input falls into the class of non-medium-level expression, then XGB–Classifier 2 would further classify the input as the class of low-level expression or the class of high-level expression. Based on the predicted class generated in the first stage, one of the three regression models, i.e., SVM regression (SVR) for low expression data (“SVR-Low”), RF regression (RFR) for medium (“RFR-Medium”) or high (“RFR-High”) were employed in the second stage to predict the expression yield. The proposed model was further employed in the computation of expression levels and yields for 180 combinations consisting of various levels of process features [namely, optical density at 600 nm (OD) (0.4, 0.7, 1.0), IPTG (0.1, 0.5, 1.0 mM) as inducer, temperature (20, 25, 30, 37 °C) and time (4, 8, 12, 16, 24 h)] for the given input of amino acid sequence using R programming. Then, it classified each of the combinations into the respective classes and quantified the corresponding expression yields. Based on the resulting expression yields, we are able to arrive at the top ten values for the optimal periplasmic expression yields of recombinant proteins and the respective fermentation conditions.

3.2. Features selected for model construction

Our preliminary analysis indicated that the conventional methods of feature selection employing various WEKA-based algorithms resulted in the extraction of significant features mostly from FC2 and FC3 but with a very few or no features derived from FC1 and FC4 (results not shown). This outcome could be mainly due to the higher number of features in FC2 (800 numbers) and FC3 (11,026 numbers) than in FC1 (149 numbers) and FC4 (10 numbers), leading to the higher occurrence of their selection. Therefore, to reduce the bias caused by the high dimensionality of features, we applied a stepwise feature selection strategy to both the classification and the regression tasks. Our initial steps of feature selection adopted a filter-based algorithm, CfsSubsetEval, for the identification of main features from FC2 and FC3. The important FC2 and FC3 features were then combined with all the features from FC1 and FC4; a wrapper-based method, ClassifierSubsetEval, was used to further select key features. The selected features for both classification and regression tasks consists of features mostly from FC1 and FC4, a few within FC3 and none from FC2 (**Tables 1 and S5**). The features selected based on the amino acid sequences are occurrences and occurrence frequencies (i.e., occurrences per unit length of protein) of amino acids such as glutamic acid (Occ_E, OF_E), valine (Occ_V), sulfur (OF_S), phenylalanine (OF_F) and methionine (OF_M), and that of the maximum consecutive alanine (MNC_A, OF_MNC_A) and cysteine (OF_MNC_C). Apart from that, the key features based on the physico-chemical properties are occurrence frequency of aromatic (OF_Aromatic), aliphatic (OF_Aliphatic) and hydrophilic residues (OF_Hphil_ESG and OF_Hphil_KD calculated using ESG and KD methods). Other key features with respect to the structural properties that were identified to be vital for either classification or regression tasks are expected number of amino acids in transmembrane helices (Expno_AA_TM), the ratio of the helix to sheet propensity (Helix_to_Sheet_PHD), coil propensity (Coil_PHD) and solubility score (Pred_Sol). For the features related to fermentation process condition, temperature and OD×Time are the two most significant features to be considered in the classification task, while almost all the other process-condition features were significant in the regression task.

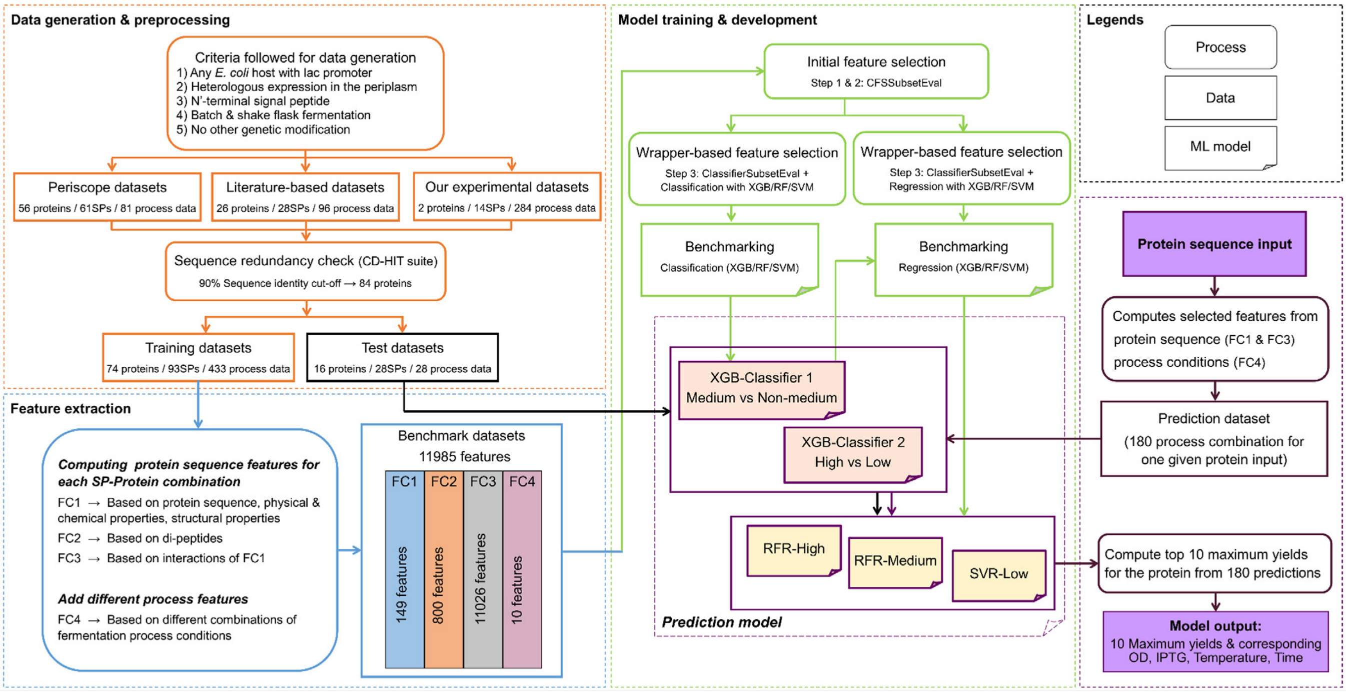

Data generation & preprocessing
Criteria followed for data generation
1) Any E. coli host with lac promoter
2) Heterologous expression in the periplasm
3) N'-terminal signal peptide
4) Batch & shake flask fermentation
5) No other genetic modification
Periscope datasets
56 proteins / 61SPs / 81 process data
Literature-based datasets
26 proteins / 28SPs / 96 process data
Our experimental datasets
2 proteins / 14SPs / 284 process data
Sequence redundancy check (CD-HIT suite)
90% Sequence identity cut-off → 84 proteins
Training datasets
74 proteins / 93SPs / 433 process data
Test datasets
16 proteins / 28SPs / 28 process data
Feature extraction
Computing protein sequence features for each SP-Protein combination
FC1 → Based on protein sequence, physical & chemical properties, structural properties
FC2 → Based on di-peptides
FC3 → Based on interactions of FC1
Add different process features
FC4 → Based on different combinations of fermentation process conditions
Benchmark datasets
11985 features
FC1
FC2
FC3
FC4
149 features
800 features
11026 features
10 features
Model training & development
Initial feature selection
Step 1 & 2: CFSSubsetEval
Wrapper-based feature selection
Step 3: ClassifierSubsetEval + Classification with XGB/RF/SVM
Wrapper-based feature selection
Step 3: ClassifierSubsetEval + Regression with XGB/RF/SVM
Benchmarking
Classification (XGB/RF/SVM)
Benchmarking
Regression (XGB/RF/SVM)
XGB-Classifier 1
Medium vs Non-medium
XGB-Classifier 2
High vs Low
RFR-High
RFR-Medium
SVR-Low
Prediction model
Legends
Process
Data
ML model
Protein sequence input
Computes selected features from protein sequence (FC1 & FC3) process conditions (FC4)
Prediction dataset
(180 process combination for one given protein input)
Compute top 10 maximum yields for the protein from 180 predictions
Model output:
10 Maximum yields & corresponding OD, IPTG, Temperature, Time

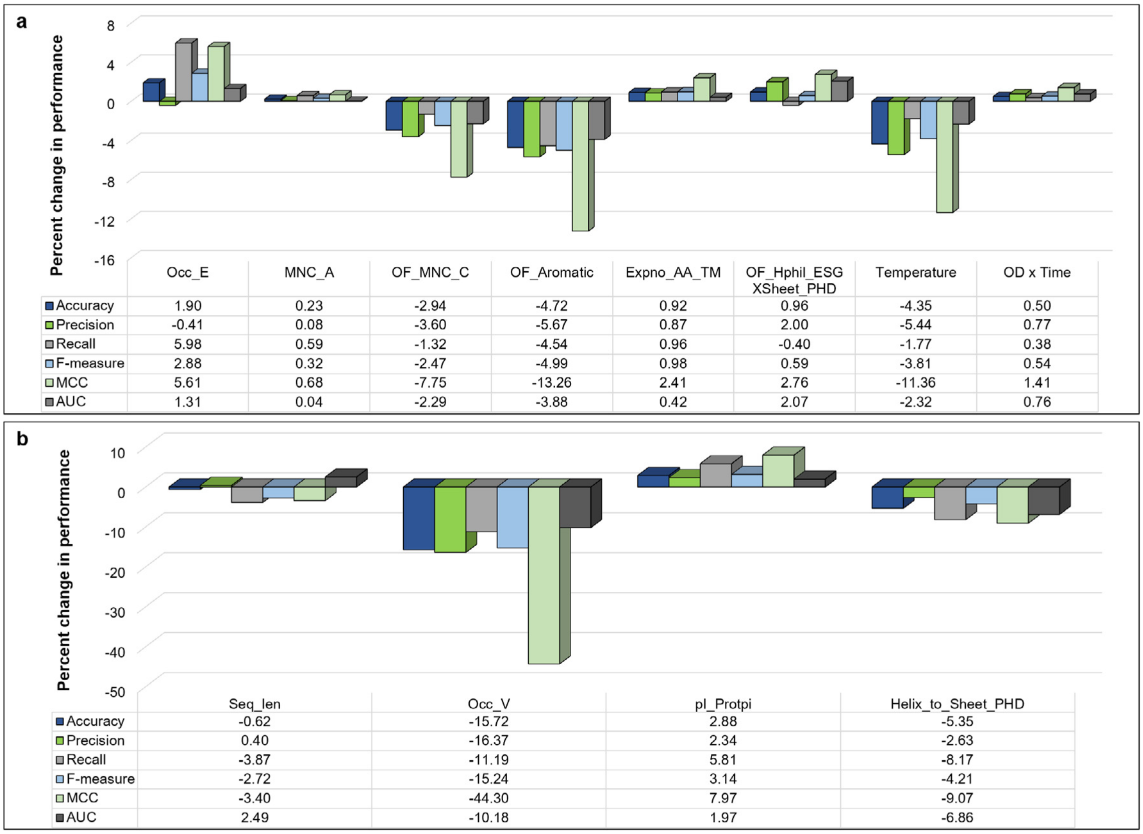
Fig. 1. Framework of the proposed prediction model. Low: yield is<0.5 mg/L, Medium: yield is between 0.5 and 50 mg/L, High: yield is higher than 50 mg/L. Non-medium refers to both High and Low together.

3.3. Assessment of feature importance

To further assess the relative importance of the selected features for the two classification and three regression tasks, we trained the models by removing one feature at a time until all were considered. The respective changes in the predictive performance were measured by benchmarking the above models against the model with all the selected key features for each of the tasks (Figs. 2 and 3). OF_Aromatic, OF_MNC_C, and temperature were found to be the most important features for XGB–Classifier 1, as the removal of these features led to the drastic decreases in MCC and accuracy by 8–13% and 3–5%, respectively. Similarly, the removal of Occ_V or Helix_to_Sheet_PHD affected the performance of XGB–Classifier 2 drastically, as seen from the steep decreases in accuracies by 5–16% and in MCC by 4–15%, respectively. Seq_len can also be considered an important feature because its elimination impacted XGB–Classifier 2 negatively, resulting in the decreases of the recall, F-measure and MCC by 3–4%. For the regression task, the feature "temperature" was found to be crucial because its removal resulted in the increases in MAEs and RMSEs by 5–10% and 6–9%, respectively. Additionally, in the cases of SVR-Low and RFR-High, a slight decrease (0.5–1.5%) in PCC was observed after the elimination of features "OD" and IPTG. Strikingly, OD×Temp, OD×Time, IPTG×Temp and Temp×Time are the process-interaction features deemed to be significant for the regression models "RFR-Medium" and "RFR-High". The impact of these process-interaction features on the performance of regression tasks is also substantiated by removing each of the above-mentioned process-interaction features during the model training, which has resulted in 0.1–4% decrease in PCC and 0.1–9% increament in MAEs or RMSEs. Similarly, the absence of features such as OF_Hphil_ESG and OF_DmE resulted in the decreases in PCC, and elevated MAEs and RMSEs. Contrastingly, the interactions between process features played a minor role for SVR-Low, suggesting that these process interactions, when controlled by fine-tuning process parameters, can yield a higher level of recombinant protein expression. Other than these pertinent features, FC3 interactive features "Occ_N×MNC_Y" and "Occ_Y×OF_DmE" showed a major decline (17%) in PCC when they were removed from the training datasets. In addition, other features such as Occ_E, OF_Hphil_ESG×Sheet_PHD, Coil_PHD, OF_Aliphatic, OD, OF_Aromatic, Pred_Sol, MNC_A, Expno_AA_TM, and OD×Time had minor effects on the predictive performance after their removal during model training; nonetheless, we retained these features because they improved the performance.

3.4. Performance of classification and regression tasks

Both classification tasks, namely i) classification of medium and non-medium classes, and ii) subsequent classification of high and low classes, were benchmarked with the three well-known and widely used algorithms: XGB, RF, and SVM. XGB was found to outperform both SVM and RF in both classification tasks (Fig. 4**A**). In the internal validation tests, the average accuracies of XGB–Classifiers 1 and 2 are 76.45% and 77.27%, respectively, while their respective average accuracies in the independent tests are 82.14% and 85.71%. In both cases of classification tasks, the performance measures such as precision, recall, F-measure and AUC were also found to be above 0.75 while the MCC was around 0.5 (Table 2). Similarly, regression tasks were benchmarked using the similar algorithms (i.e., XGB, RF, and SVM) coupled with LOOCV for the three classes - high, low and medium (Fig. 4**B**). In the internal cross validation, the PCCs of three regression models were the highest, i.e., SVR-Low (0.83), RFR-Medium (0.90) and RFR-High (0.87). The PCCs of these regression models in the independent testing were also higher than their respective counterparts. As shown in Table 3, the lower values of MAEs and RMSEs of the three chosen regression models are in good agreement with the range of expression yields of each class. The values of PCC, MAE, and RMSE of SVR-Low, RFR-Medium and RFR-High suggest that the developed regression model can predict the protein expression yields with greater accuracy and reliability. The MAE and RMSE values for SVR-Low



	Occ_E	MNC_A	OF_MNC_C	OF_Aromatic	Expno_AA_TM	OF_Hphil_ESG XSheet_PHD	Temperature	OD x Time
Accuracy	1.90	0.23	-2.94	-4.72	0.92	0.96	-4.35	0.50
Precision	-0.41	0.08	-3.60	-5.67	0.87	2.00	-5.44	0.77
Recall	5.98	0.59	-1.32	-4.54	0.96	-0.40	-1.77	0.38
F-measure	2.88	0.32	-2.47	-4.99	0.98	0.59	-3.81	0.54
MCC	5.61	0.68	-7.75	-13.26	2.41	2.76	-11.36	1.41
AUC	1.31	0.04	-2.29	-3.88	0.42	2.07	-2.32	0.76

	Seq_len	Occ_V	pI_Protpi	Helix_to_Sheet_PHD
Accuracy	-0.62	-15.72	2.88	-5.35
Precision	0.40	-16.37	2.34	-2.63
Recall	-3.87	-11.19	5.81	-8.17
F-measure	-2.72	-15.24	3.14	-4.21
MCC	-3.40	-44.30	7.97	-9.07
AUC	2.49	-10.18	1.97	-6.86

Fig. 2. Feature importance for a) XGB Classifier 1 b) XGB Classifier 2. Performance of the model has been evaluated using ten times 10-fold cross validation (100 experiments).

remained too low (0.06 and 0.09, respectively), while the values of these measures increased drastically for RFR-Medium (59–65 times that of SVR-Low) and RFR-High (12–13 times compared to RFR-Medium) (Table 3). Such an increase in MAE and RMSE values is common due to the high orders of the ranges within the expression-yield levels in each of these classes.

3.5. Predictive performance of the developed model

Independent test validation gave an overall classification accuracy of 75% for the 28 unseen instances; the prediction results showed that 21 instances, including the six unseen proteins, were correctly classified and are close to the real experimental values (Table 4). Similarly, the model predicted the expression yield of the correctly classified instances with a high PCC of 0.91. Regardless of the misclassification, the PCC for the prediction of all the 28 instances remained substantially high (0.80). The MAE and RMSE values for the correctly classified instances were found to be remarkably low, i.e., 22.38 and 62.07, respectively.

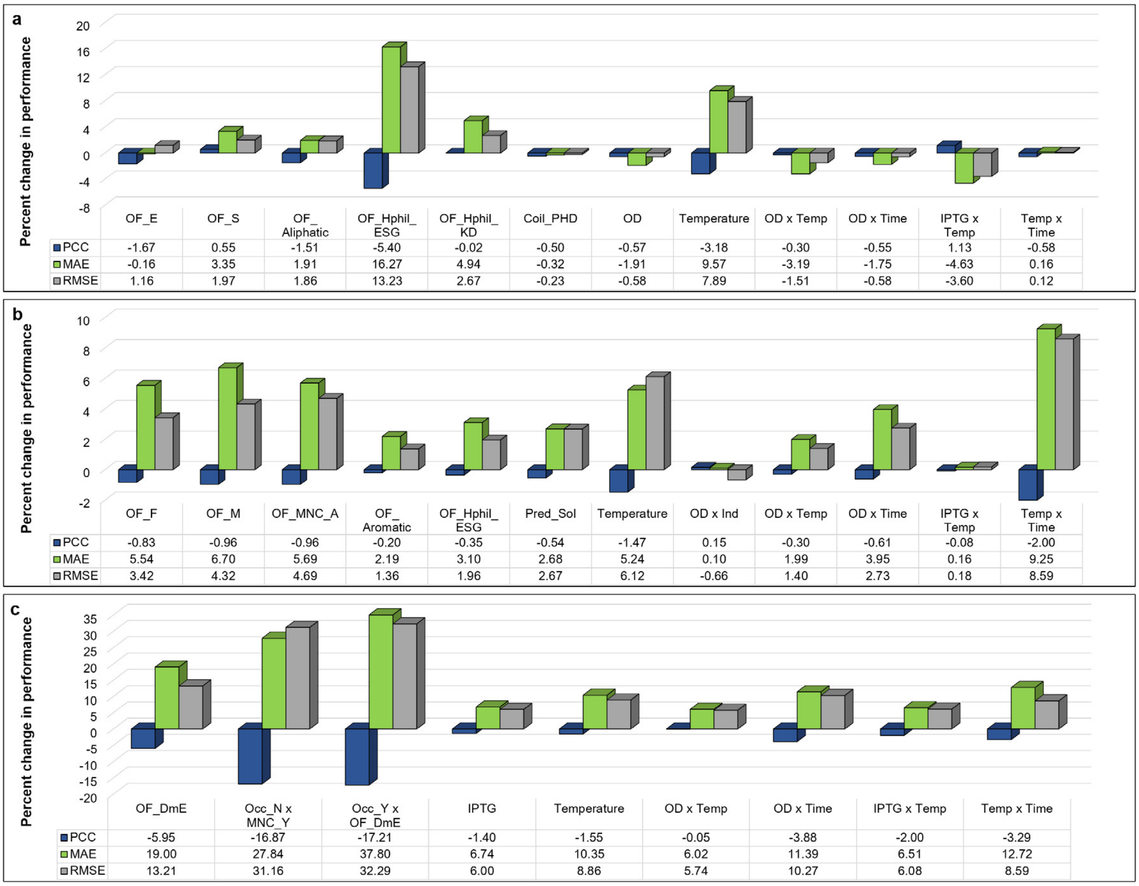
3.6. Prediction of the maximal RPP using the proposed model

For a target protein, the proposed model predicted the top ten protein yields and the corresponding fermentation process conditions. We evaluated all the 28 independent test datasets and presented the most optimal yields and fermentation conditions (Table 5). To further evaluate the prediction performance of the proposed model, we compared the predictions of the top ten optimal yields and fermentation conditions for the proteins used in our experiments with the predicted yields at the given conditions using the respective statistical regression equations (Tables S6-1 to S6-14). The predicted results were in close agreement with each other, highlighting that our model mimics the individual RSM-based regression models and has an additional advantage of extending the predictions to any recombinant protein.

4. Discussion

In the present work, we have developed a robust ML-based tool that is capable of predicting the top-ten maximal protein expression yield and their fermentation process conditions for the expression of recombinant proteins in the periplasm of *E. coli*. Importantly, we have combined the key features from both amino acid sequence and fermentation process to gain a better understanding of the important determinants of the recombinant protein expression and to construct a precise model that yields good predictions. Our results demonstrate that the developed prediction model offers greater predictability and reliability as to the experimental findings. The primary reason behind the successful prediction by the optimization model is the strength and diversity of the datasets used in the development of the model. Another notable factor is the appropriate selection of the feature-subsets that represent each of the models precisely.

The screening of important and meaningful features was made possible by the vast number of features extracted from literature and the use of the stepwise feature selection strategy. The diverse sets of the selected features (e.g., amino acid composition, physicochemical and structural properties), together with the features related to the fermentation process, were processed by the stepwise feature selection strategy to yield a more meaningful prediction result. Our feature selection strategy revealed that the features based on amino acid sequence seem to play a vital role in the clas-



	OF_E	OF_S	OF_ Aliphatic	OF_Hphil_ ESG	OF_Hphil_ KD	Coil_PHD	OD	Temperature	OD x Temp	OD x Time	IPTG x Temp	Temp x Time
■ PCC	-1.67	0.55	-1.51	-5.40	-0.02	-0.50	-0.57	-3.18	-0.30	-0.55	1.13	-0.58
■ MAE	-0.16	3.35	1.91	16.27	4.94	-0.32	-1.91	9.57	-3.19	-1.75	-4.63	0.16
■ RMSE	1.16	1.97	1.86	13.23	2.67	-0.23	-0.58	7.89	-1.51	-0.58	-3.60	0.12

	OF_F	OF_M	OF_MNC_A	OF_ Aromatic	OF_Hphil_ ESG	Pred_Sol	Temperature	OD x Ind	OD x Temp	OD x Time	IPTG x Temp	Temp x Time
■ PCC	-0.83	-0.96	-0.96	-0.20	-0.35	-0.54	-1.47	0.15	-0.30	-0.61	-0.08	-2.00
■ MAE	5.54	6.70	5.69	2.19	3.10	2.68	5.24	0.10	1.99	3.95	0.16	9.25
■ RMSE	3.42	4.32	4.69	1.36	1.96	2.67	6.12	-0.66	1.40	2.73	0.18	8.59

	OF_DmE	Occ_N x MNC_Y	Occ_Y x OF_DmE	IPTG	Temperature	OD x Temp	OD x Time	IPTG x Temp	Temp x Time
■ PCC	-5.95	-16.87	-17.21	-1.40	-1.55	-0.05	-3.88	-2.00	-3.29
■ MAE	19.00	27.84	37.80	6.74	10.35	6.02	11.39	6.51	12.72
■ RMSE	13.21	31.16	32.29	6.00	8.86	5.74	10.27	6.08	8.59

Fig. 3. Feature importance for a) SVR-Low b) RFR-Medium c) RFR-High. Performance of the model has been evaluated using ten times 10-fold cross validation (100 experiments).

sification of "low" and "high" classes, hinting that the expression yield of a protein is completely dependent on its amino acid sequence. Therefore, based on the amino acid sequence, the expression of a recombinant protein is defined as "low", "medium" or "high", and a further fine-tuning of the process parameters can result in the substantial amounts of protein expression yields within that particular class. Apart from that, out of the selected features, the occurrence of valine (Occ_V), occurrence frequencies of aromatic residues (OF_Aromatic) and hydrophilic residues (OF_Hphil_ESG), difference in the occurrence frequency of aspartic acid minus glutamic acid residues (OF_DmE), the ratio of the helix to sheet propensity (Helix_to_Sheet_PHD), temperature (Temp) and the interaction between temperature and time (Temp×Time) seemed to be highly relevant and significant. Past studies corroborated the importance of the selected features as well. For instance, the probability of expressing the protein in soluble form was inversely correlated to the size of protein [28], thereby hinting that Seq_len is an essential feature. Besides, the composition of amino acid was found to be a critical factor inducing the metabolic stress during RPP in *E. coli* [29]; hence, the expression of recombinant protein can be improved by adjusting the amino acids composition [30]. The present study revealed specifically that the occurrences and occurrence frequencies of amino acids such as Occ_E, Occ_V, OF_E, OF_S, OF_F, OF_M, MNC_A, OF_ MNC_A, and OF_MNC_C are the significant factors for the soluble protein expression in the periplasm of *E. coli*. Similarly, the occurrences of the hydrophilic residues, i.e., proline (P), tyrosine (Y), histidine (H), glutamine (Q) and asparagine (N) seemed to be key determinants in SVR-Low and RFR-Medium in this study. The protein solubility was proven to be affected by the presence of hydrophilic amino acids in the protein, which may in turn influence the expression levels of recombinant protein in *E. coli* [30,31]. Trevino et al. (2007) showed that the solubility of ribonuclease from *Streptomyces aureofaciens* (RNase Sa) was enhanced by the presence of amino acid residues such as aspartic acid (D), glutamic acid (E) and serine (S) in the protein sequence as compared to the other hydrophilic residues such as asparagine (N), glutamine (Q), and threonine (T) [32]. Therefore, the occurrence frequency of aspartic acid residues minus glutamic acid residues (OF_DmE) being a significant feature in RFR-High was validated. Although pI has been identified as a key feature in the development of XGB-Classifier 1, it is shown that pI does not affect the protein solubility or expression of mammalian proteins in *E. coli* [33].

Combining fermentation-process-based features with amino-acid-sequence-based features is a highly beneficial initiative. For instance, a previous study revealed that the recombinant expression of insoluble proteins is highly correlated with temperature and fermentation time [34]. Our results showed that Temp×Time interaction feature is a significant process feature. Also, it is well known that a lower cell density could result in a lower expression yield [35], while different concentrations of inducer have varying levels of influence on the yields of recombinant protein [36]. Our findings supported these facts and demonstrated the importance of OD and IPTG features in the development of regression models (SVR-Low and RFR-High). Furthermore, existing literature on the statistical optimization of RPP indicated that the interactions between the process features are significant in soluble protein expression [7,37–42]. Our models (XGB Classifier 1, SVR-Low,

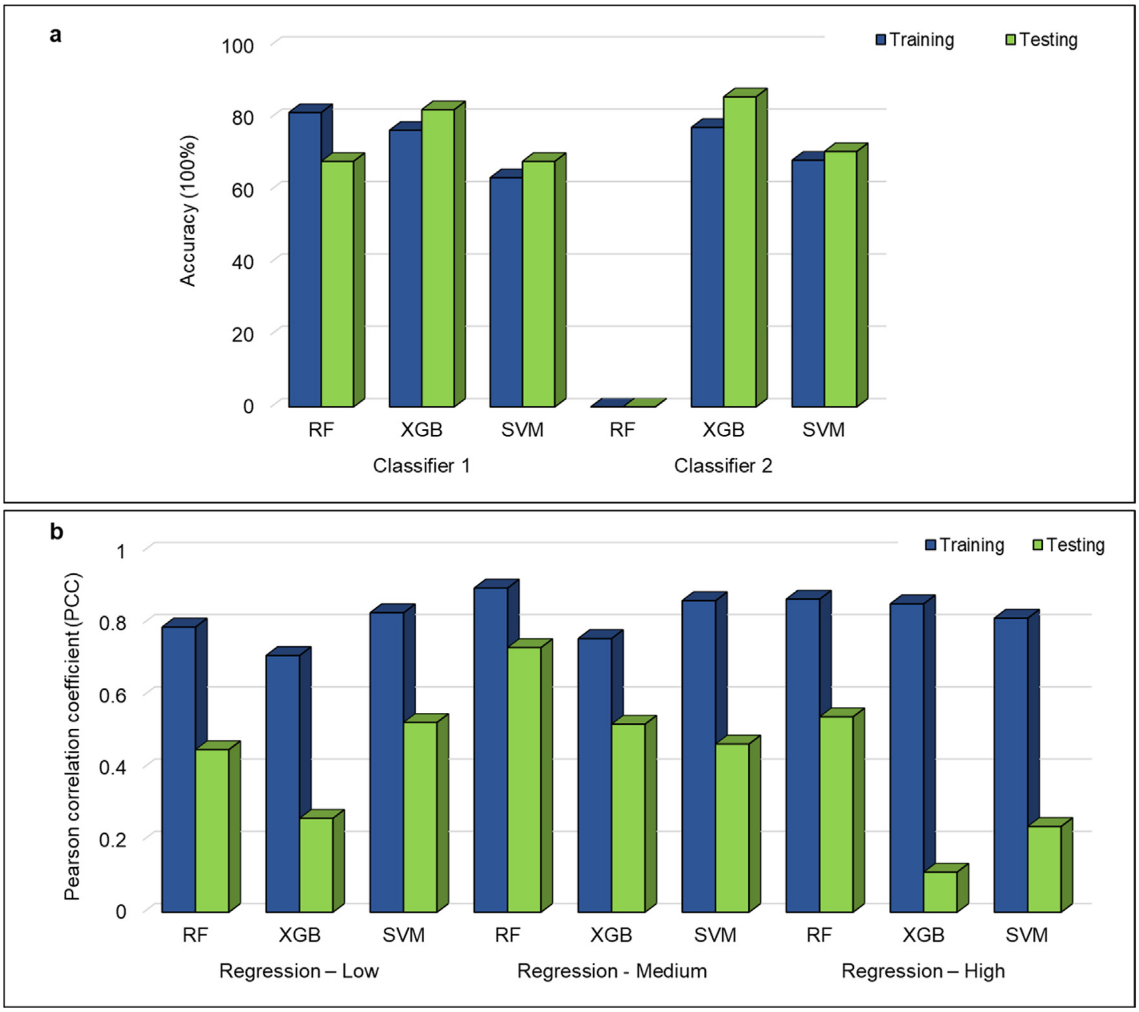


Fig. 4. Benchmarking of the performance of different algorithms. a) Classification tasks for both training and testing datasets b) Regression tasks for both training and testing datasets.

Table 1
Selected features for prediction model.

Feature category	Classification models		Regression models		
	XGB–Classifier 1	**XGB–Classifier 2**	**SVR–Low**	**RFR–Medium**	**RFR–High**
FC1	Occ_E MNC_A OF_MNC_C OF_Aromatic Expno_AA_TM	Seq_len Occ_V pI_Protpi Helix_to_Sheet_PHD	OF_E OF_S OF_Aliphatic OF_Hphil_ESG OF_Hphil_KD Coil_PHD	OF_F OF_M OF_MNC_A OF_Aromatic OF_Hphil_ESG Pred_Sol	OF_DmE
FC2	–	–	–	–	–
FC3	OF_Hphil_ESG×Sheet_PHD	–	–	–	Occ_N×MNC_Y Occ_Y×OF_DmE
FC4	Temperature ODxTime	–	OD Temperature OD×Temp OD×Time IPTG×Temp Temp×Time	Temperature OD×IPTG OD×Temp OD×Time IPTG×Temp Temp×Time	IPTG Temperature OD×Temp OD×Time IPTG×Temp Temp×Time
Selected features	8	4	12	12	9

RFR-Medium, and RFR-High) confirmed that the interactions among the process features (cell density, inducer concentration, post-induction time, and temperature) contributed significantly to the prediction of the expression yields. Further investigation of the interactions between these process features in governing the expression of the recombinant protein may be fruitful in gaining a better understanding of the keys facets to achieve high yields of recombinant protein.

At the given fermentation conditions, the predicted expression yields of proteins for the correctly classified instances were found to be closely matching to the actual values given in the independent test datasets used (Table 4). For example, the predicted

Table 2
Classification Task – Benchmarking with three algorithms.

Algorithm	Classifier 1			Classifier 2		
	RF	**XGB**	**SVM**	**RF***	**XGB**	**SVM**
Selected number of features	4	8	16	1	4	6
Accuracy (%)	81.36	76.45	63.35	–	77.27	68.18
Error rate (%)	18.63	23.55	36.65	–	22.73	31.82
Precision	0.814	0.764	0.636	–	0.776	0.682
Recall	0.814	0.764	0.634	–	0.773	0.682
F-measure	0.813	0.764	0.629	–	0.773	0.682
MCC	0.626	0.527	0.267	–	0.549	0.362
AUC	0.913	0.788	0.643	–	0.791	0.747

Performance of the model has been evaluated using leave-one-out cross validation (LOOCV).

* Since there is only one key feature selected, further model training is neither essential nor meaningful in this case.

Table 3
Regression task – Benchmarking with three algorithms.

Algorithm	Regression – Low			Regression – Medium	Regression – High				
	RF	**XGB**	**SVM**	**RF**	**XGB**	**SVM**	**RF**	**XGB**	**SVM**
Selected number of features	5	5	12	12	12	14	9	6	10
Pearson Correlation Coefficient (PCC)	0.7891	0.7103	0.8288	0.8971	0.7574	0.8623	0.8664	0.8534	0.8137
Mean Absolute Error (MAE)	0.0738	0.2759	0.0623	3.6673	8.8066	4.6152	47.4097	114.973	47.0493
Root Mean Squared Error (RMSE)	0.0944	0.2993	0.0887	5.7796	13.5347	6.6317	76.694	163.13	90.4669

Performance of the model has been evaluated using leave-one-out cross validation (LOOCV).

Table 4
Predicted yields at the given experimental conditions.

No	SP-protein combination	Process conditions				Actual expression		Predicted expression	
		OD (au)	**IPTG (mM)**	**Temperature (°C)**	**Time (h)**	**Yield (mg/L)**	**Level**	**Yield (mg/L)**	**Level**
1	pel-B-eGFP	0.5	0.1	18	4	4.7	M	3.0	M
2	Cex-eGFP	0.7	0.5	27	4	4.8	M	0.3	*L
3	ompA-eGFP	1	0.5	18	4	2.8	M	2.0	M
4	ompC-eGFP	0.7	0.5	18	4	5.7	M	2.6	M
5	Lpp-eGFP	0.4	0.1	18	4	0.6	M	1.6	M
6	DmsA-eGFP	1	1	18	4	80.1	H	20.7	M
7	MdoD-eGFP	0.4	0.5	27	4	14.2	M	6.2	M
8	pel-B-TMT	0.4	0.5	28	4	53.2	H	27.1	*M
9	Cex-TMT	0.4	1	38	4	137.5	H	120.8	H
10	ompA-TMT	–	–	–	4	0.0	L	0.0	L
11	ompC-TMT	–	–	–	4	0.0	L	0.0	L
12	Lpp-TMT	0.7	1	18	4	95.5	H	116.5	H
13	DmsA-TMT	0.4	0.7	28	4	482.2	H	175.3	H
14	MdoD-TMT	0.4	1	38	4	24.5	M	9.0	M
15	pelB-IFN	4 (TB)	0.05	25	14	0.4	L	0.1	L
16	pelB-VEGFR2-D3	1	1	37	20	2.0	M	4.1	M
17	pho-rhES	0.6	0.3	25	13.57	2.2	M	1.8	M
18	modspA-CALB	1	12.5%(L)	24	15 h	234.0	H	126.5	H
19	MBP-6 × His-U24	0.5–1.0	0.3	18	16	2.8	M	3.0	M
20	Pel-B-SynVNAR-A6	0.5	0.1	18	21	27.0	M	7.5	M
21	modBlaasp-hAct A	0.6	1	37	8	150.0	H	0.0	*L
22	CusF-GFP	0.5	0.1	12	25	8.0	M	5.2	M
23	ecotin-HArbd	0.6–0.8	0.4–1	30	‘8–10	10.0	M	13.3	M
24	mBiP-scFv	0.5	0.2	30	5	115.0	H	7.3	*M
25	pelB-scFv-dmOKT3	0.8	0.1	22–24	18–20	0.2	L	13.3	*M
26	stII-vtPA	0.5	1	30	6	0.2	L	0.1	L
27	LTIIb-B-CT-B	0.3	0.02	37	6	190.0	H	8.3	*M
28	pelB-rPA	0.7	1	24	21	0.0	L	0.2	L

The misclassified instances are represented by an asterisk (*). TB – Terrific broth (medium); L – Lactose (inducer).

expression yields of the recombinant proteins pel-B-eGFP, ompA-eGFP, MBP-6 × His-U24, and ecotin-HArbd (classified as medium expression) as well as ompA-TMT, ompC-TMT and stII-vtPA (classified as low expression) resembled the actual expression levels of periplasmic proteins in *E. coli* closely. Only a few instances, such as the predicted expression yields of cex-TMT and Lpp-TMT (classified as high expression), showed slight variation from the actual values of expression yields reported in the literature, while the predicted expression yields of DmsA-TMT and modspA-CALB showed moderate deviations from their actual expression yields. However, the expression yields as predicted from the misclassified instances varied tremendously, particularly for those instances of High-class of protein expression being misclassified into either low or medium class. For example, a high deviation in the predicted expression yields was noted in the cases of modBlaasp-hAct A, mBiP-scFv, and LTIIb-B-CT-B, which were misclassified as

Table 5
Predicted maximal predicted yields and the corresponding fermentation conditions.

No	SP-protein combination	Experimental				Predicted							
		Process conditions				**Optimal expression**		**Process conditions**				**Optimal expression**	
		OD (au)	**IPTG (mM)**	**Temperature (°C)**	**Time (h)**	**Yield (mg/L)**	**Level**	**OD (au)**	**IPTG (mM)**	**Temperature (°C)**	**Time (h)**	**Yield (mg/L)**	**Level**
1	pel-B-eGFP	0.5	0.1	18	4	4.7	M	1	0.1	20	4	3.8	M
2	Cex-eGFP	0.7	0.5	27	4	4.8	M	0.7	0.5	25	16	4.6	M
3	ompA-eGFP	1	0.5	18	4	2.8	M	0.7	1	25	24	1.9	M
4	ompC-eGFP	0.7	0.5	18	4	5.7	M	0.7	0.5	25	4	4.3	M
5	Lpp-eGFP	0.4	0.1	18	4	0.6	M	0.7	0.5	30	24	1.9	M
6	DmsA-eGFP	1	1	18	4	80.1	H	1	0.5	30	4	38.4	*M
7	MdoD-eGFP	0.4	0.5	27	4	14.2	M	0.4	0.5	30	8	8.1	M
8	pel-B-TMT	0.4	0.5	28	4	53.2	H	1	0.5	30	4	31.9	*M
9	Cex-TMT	0.4	1	38	4	137.5	H	0.4	1	37	24	141.5	H
10	ompA-TMT	–	–	–	4	0.0	L	1	1	30	24	0.1	L
11	ompC-TMT	–	–	–	4	0.0	L	0.4	1	20	4	26.0	*M
12	Lpp-TMT	0.7	1	18	4	95.5	H	0.4	1	37	24	151.6	H
13	DmsA-TMT	0.4	0.7	28	4	482.2	H	0.4	0.5	30	8	268.3	H
14	MdoD-TMT	0.4	1	38	4	24.5	M	0.4	1	37	8	16.3	M

The misclassified instances are represented by an asterisk symbol (*).

low or medium classes of protein expression. This undesirable prediction outcome was caused by the very high orders of difference in the ranges of these three classes; hence, a poor classification accuracy of both XGB-Classifiers 1 and 2 eventually affects the performance of the overall prediction model. This issue was addressed by adopting three different regression models to cover a wide range of protein expression levels as categorized by the classification models. If the expression level of a target protein is classified correctly in the first stage of prediction, the expression yield of the target protein will be highly likely to be accurately predicted by the respective regression model.

The maximal protein expression yield was predicted by computing the protein expression yields under various combinations of fermentation process conditions and by selecting the top-ten maximal yields. A significant improvement in the prediction performance was noted when different combinations of fermentation process condition were included in the testing sets (Table 5). One of the previously misclassified instances, namely cex-eGFP, was correctly classified when different fermentation process combinations were considered during the testing; accordingly, the predicted expression yield (4.6 mg/L) was close to the actual levels (4.8 mg/L). Similarly, the predicted yield of cex-TMT expression (141.5 mg/L) approximately matched the actual protein expression levels (137.5 mg/L). Most of the predicted levels of protein expression were close to their actual levels, except for a few instances (e.g., DmsA-eGFP, ompC-TMT, DmsA-TMT and Lpp-TMT) showing slight deviation in the predicted protein expression levels (Table 5). Therefore, the predicted top-ten expression yields and the corresponding fermentation conditions suggested that the model predictions were similar to those achieved during experiments (Table 5 and Tables S6-1 to S6-6). For instance, the fermentation conditions corresponding to the maximal predicted yields of pel-B-eGFP, ompC-eGFP, and DmsA-TMT exactly match the experimental conditions that lead to the optimal yields of protein, while for the other instances, these fermentation conditions are almost similar to the optimal conditions as predicted by our model. Based on the supplementary tables (Tables S6-1 to S6-6), it is evident that process-level features play an important role in the expression of "high" class, while interactions within process-level features are significant in the "medium" expression class, which is also substantiated by the feature selection strategy (Table 1). Therefore, our prediction tool enables an easy optimization of RPP by suggesting (i) whether a particular target protein will be able to express in significant amounts and (ii) the ranges of fermentation parameters based on the predicted top-ten expression levels of target protein. These predictions could provide a good basis towards experiment design, by choosing an appropriate (i) target protein, (ii) selection of signal peptide and (iii) a set of fermentation conditions to start with, in an attempt to achieve the desired yields of recombinant proteins. Such insights will also be valuable in the subsequent optimization studies conducted to improve the yield and design of the industrial-scale RPP in *E. coli*.

The major limitation in model development for predicting protein expression yields is the scarcity of the availability of experimental results from the relevant studies. Although there are many reported studies about the production of periplasmic recombinant proteins by *E. coli*, all these data could not be considered in model development because of: i) the missing information related to the fermentation process conditions; ii) the lack of or irretrievable amino acid sequence; iii) the irrelevant scale of the fermentation process (micro-level or bioreactor level); iv) the non-quantifiable protein concentration (in mg/L); v) the vectors using promoters other than *lac* promoters, and hosts or vectors being modified as different genetic variants. Secondly, the data collected are prone to some degrees of variability contributed from the different protocols of fermentation and protein quantification used by researchers; for example, the scale of shake flask fermentation may add up to these variations. Finally, the variability due to specific host strains of *E. coli* and the corresponding vectors can vastly impact the recombinant protein expression yields. However, the majority of these general limitations have been addressed in the best possible ways during the development of PERISCOPE-Opt. For example, in spite of the available data being scarce, which becomes a trade-off for the prediction accuracy of the optimization model, the dataset generated for the development of the proposed model is robust because it consisted of a wide range of proteins (84 different types) and SP-protein combinations (103 different types). Further, the data incorporated proteins of all sizes (i.e., the smallest protein contained 80 amino acids while the biggest protein was of 668 amino acids long in size) along with a good number of instances corresponding to each class: high, low or medium. Next, the variability due to the data collected from various sources was kept to minimal by considering data generated using the shake flask fermentation so that the process conditions, including agitation and mixing, will be quite similar. Micro-scale and bioreactor-based fermentations were ruled out of present study as these methods may offer additional variabilities in process conditions compared to shake flask fermentations. Different schemes of protein downstream-processing, purification or quantifications can affect the expression yields of final protein but these recombinant protein yield data collected at the particular fermentation conditions are still comparable and relevant to our model development. Lastly, *lac*-operator based vectors were considered in the data collection to avoid biases in protein expression due to the uses of other types of vector. Genetically improved *E. coli* host strains such as those with additional molecular chaperones were avoided as these strains may exhibit additional variabilities, while the conventionally used *E. coli* host strains were assumed to express periplasmic protein similarly. The variability caused by the different conventional host strains being used for fermentation were not resolved as it is out of scope of the present study. Nevertheless, considering *E. coli* host strains and other vectors as a profound variable will be a good option as the addition of features with respect to the properties of *E. coli* host and vectors will be highly beneficial to the improvement of the prediction accuracy of the proposed model. Similarly, other aspects that can potentially help to improve the prediction accuracy of the developed model is the incorporation of other relevant features based on gene factors (codon bias and mRNA secondary structures). Apart from that, the incorporation of features corresponding to the experimentally-derived structural properties of proteins instead of using the sequence-derived and predicted structural properties will tend to improve the performance of the prediction model. The inclusion of these novel features in future works will serve as avenues to fine-tune the developed model for improved prediction.

5. Conclusions

The ML tools available for the protein-based applications generally consider features with respect to the amino acid sequence and are incapable of predicting the optimal conditions of RPP. Therefore, an ML-based model has been developed by combining the features from amino acid sequence and the fermentation process to predict the optimal yield as well as the corresponding fermentation conditions for the expression of a given recombinant protein in the periplasm of *E. coli*. Our proposed two-stage framework, PERISCOPE-Opt, successfully suggested the optimal recombinant protein yields matching closely with the reported experimental results. The recommended optimal yields and the corresponding fermentation conditions give an overall idea of the fermentation process for the expression of a target protein. PERISCOPE-Opt could serve as a powerful and reliable web tool that identifies the optimal fermentation conditions and RPP yield without reliance on the excessive rounds of trial-and-error experiments.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge Monash University Malaysia, Fundamental Research Grant Scheme (FRGS), Malaysia and Major Inter-Disciplinary Research (IDR) project, Monash University for providing the research support needed for this work. Kulandai Arockia Rajesh Packiam also, acknowledges the Vice-Chancellor's International Inter-campus PhD Travel Grant (2019) and the Department of Chemical Engineering International Research Mobility Grants (2019) awarded by Monash University to support the mobility during the current research and conference presentation of a part of the work.

Funding information

This study was funded by Fundamental Research Grant Scheme (FRGS), Malaysia [FRGS/1/2016/TK02/MUSM/02/3] and Major Inter-Disciplinary Research (IDR) project awarded by Monash University. Kulandai Arockia Rajesh Packiam was supported by Graduate Research Scholarships, Monash University Malaysia.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.06.006.

References

[1] Ahmadi MK, Pfeifer BA. Recent progress in therapeutic natural product biosynthesis using *Escherichia coli*. Curr Opin Biotechnol 2016;42:7–12. https://doi.org/10.1016/j.copbio.2016.02.010.

[2] Liu M, Feng X, Ding Y, Zhao G, Liu H, Xian M. Metabolic engineering of *Escherichia coli* to improve recombinant protein production. Appl Microbiol Biotechnol 2015;99:10367–77. https://doi.org/10.1007/s00253-015-6955-9.

[3] Packiam KAR, Ramanan RN, Ooi CW, Krishnaswamy L, Tey BT. Stepwise optimization of recombinant protein production in *Escherichia coli* utilizing computational and experimental approaches. Appl Microbiol Biotechnol 2020;104:3253–66. https://doi.org/10.1007/s00253-020-10454-w.

[4] Sandomenico A, Sivaccumar JP, Ruvo M. Evolution of *Escherichia coli* Expression System in Producing Antibody Recombinant Fragments. Int J Mol Sci 2020, Vol 21, Page 6324 2020;21:6324. https://doi.org/10.3390/IJMS21176324.

[5] Kaur JJ, Kumar A, Kaur JJ. Strategies for optimization of heterologous protein expression in *E. coli*: Roadblocks and reinforcements. Int J Biol Macromol 2018;106:803–22. https://doi.org/10.1016/J.IJBIOMAC.2017.08.080.

[6] Huleani S, Roberts MR, Beales L, Papaioannou EH. *Escherichia coli* as an antibody expression host for the production of diagnostic proteins: significance and expression. Https://DoiOrg/101080/0738855120211967871 2021. https://doi.org/10.1080/07388551.2021.1967871.

[7] Rostami N, Goharrizi LY. Cloning, Expression, and Purification of the Human Synthetic Survivin Protein in *Escherichia coli* Using Response Surface Methodology (RSM). Mol Biotechnol 2021:1–11. https://doi.org/10.1007/S12033-021-00399-4/FIGURES/6.

[8] Rigi G, Rostami A, Ghomi H, Ahmadian G, Mirbagheri VS, Jeiranikhameneh M, et al. Optimization of expression, purification and secretion of functional recombinant human growth hormone in *Escherichia coli* using modified staphylococcal protein a signal peptide. BMC Biotechnol 2021;21:1–18. https://doi.org/10.1186/S12896-021-00701-X/TABLES/5.

[9] Smialowski P, Doose G, Torkler P, Kaufmann S, Frishman D. PROSO II - A new method for protein solubility prediction. FEBS J 2012;279:2192–200. https://doi.org/10.1111/j.1742-4658.2012.08603.x.

[10] Agostini F, Cirillo D, Livi CM, Delli Ponti R, Tartaglia GG. ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in *Escherichia coli*. Bioinformatics 2014;30:2975–7. https://doi.org/10.1093/bioinformatics/btu420.

[11] Hebditch M, Carballo-Amador MA, Charonis S, Curtis R, Warwicker J. Protein-Sol: a web tool for predicting protein solubility from sequence. Bioinformatics 2017;33:3098–100. https://doi.org/10.1093/bioinformatics/btx345.

[12] Khurana S, Rawi R, Kunji K, Chuang G-Y, Bensmail H, Mall R. DeepSol: a deep learning framework for sequence-based protein solubility prediction. Bioinformatics 2018;34:2605–13. https://doi.org/10.1093/bioinformatics/bty166.

[13] Rawi R, Mall R, Kunji K, Shen C-H-H, Kwong PD, Chuang G-Y-Y. PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine. Bioinformatics 2018;34:1092–8.

[14] Bhandari BK, Gardner PP, Lim CS. Solubility-Weighted Index: fast and accurate prediction of protein solubility. Bioinformatics 2020;36:4691–8. https://doi.org/10.1093/BIOINFORMATICS/BTAA578.

[15] Francoeur PG, Koes DR. SolTranNet-A Machine Learning Tool for Fast Aqueous Solubility Prediction. J Chem Inf Model 2021;61:2530–6. https://doi.org/10.1021/ACS.JCIM.1C00331/SUPPL_FILE/CI1C00331_SI_001.PDF.

[16] Capriotti E, Casadio R. K-Fold: a tool for the prediction of the protein folding kinetic order and rate. Bioinformatics 2007;23:385–6. https://doi.org/10.1093/bioinformatics/btl610.

[17] Shen H-B, Song J-N, Chou K-C. Prediction of protein folding rates from primary sequence by fusing multiple sequential features. J Biomed Sci Eng 2009;2:136–43. https://doi.org/10.4236/jbise.2009.23024.

[18] Song J, Takemoto K, Shen H, Tan H, Gromiha MM, Akutsu T. Prediction of protein folding rates from structural topology and complex network properties. IPSJ Trans Bioinforma 2010;3:40–53. https://doi.org/10.2197/ipsjtbio.3.40.

[19] Lin GN, Wang Z, Xu D, Cheng J. SeqRate: sequence-based protein folding type classification and rates prediction. BMC Bioinf 2010;11:S1. https://doi.org/10.1186/1471-2105-11-S3-S1.

[20] Hirose S, Noguchi T. ESPRESSO: A system for estimating protein expression and solubility in protein expression systems. Proteomics 2013;13:1444–56. https://doi.org/10.1002/pmic.201200175.

[21] Chang CCH, Li C, Webb GI, Tey B, Song J, Ramanan RN. Periscope: quantitative prediction of soluble protein expression in the periplasm of *Escherichia coli*. Sci Rep 2016;6:21844. https://doi.org/10.1038/srep21844.

[22] Habibi N, Mohd Hashim SZ, Norouzi A, Shamsir MS, Samian R. Prediction of recombinant protein overexpression in *Escherichia coli* using a machine learning based model (RPOLP). Comput Biol Med 2015;66:330–6. https://doi.org/10.1016/j.compbiomed.2015.09.015.

[23] Chan W-C, Liang P-H, Shih Y-P, Yang U-C, Lin W, Hsu C-N. Learning to predict expression efficacy of vectors in recombinant protein production. BMC Bioinf 2010;11(Suppl 1):S21. https://doi.org/10.1186/1471-2105-11-S1-S21.

[24] Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 2010;26:680–2. https://doi.org/10.1093/bioinformatics/btq003.

[25] Frank E, Hall MA, Witten IH, Kaufmann M. WEKA Workbench Online Appendix for ''Data Mining: Practical Machine Learning Tools and Techniques.'' Fourth Edition, 2016. Morgan Kaufmann; 2016.

[26] R Core Team. R: A Language and Environment for Statistical Computing 2017.

[27] Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, et al. {mlr}: Machine Learning in R. J Mach Learn Res 2016;17:1–5.

[28] Francis DM, Page R. Strategies to optimize protein expression in *E. coli*. Curr Protoc Protein Sci 2010:1–29. https://doi.org/10.1002/0471140864.ps0524s61.

[29] Bonomo J, Gill RT. Amino acid content of recombinant proteins influences the metabolic burden response. Biotechnol Bioeng 2005;90:116–26. https://doi.org/10.1002/bit.20436.

[30] Wei H, Fang M, Wan M, Wang H, Zhang P, Hu X, et al. Influence of hydrophilic amino acids and GC-content on expression of recombinant proteins used in vaccines against foot-and-mouth disease virus in *Escherichia coli*. Biotechnol Lett 2014;36:723–9. https://doi.org/10.1007/s10529-013-1421-1.

[31] Matsui D, Nakano S, Dadashipour M, Asano Y. Rational identification of aggregation hotspots based on secondary structure and amino acid hydrophobicity. Sci Rep 2017;7:1–12. https://doi.org/10.1038/s41598-017-09749-2.

[32] Trevino SR, Scholtz JM, Pace CN. Amino Acid Contribution to Protein Solubility: Asp, Glu, and Ser Contribute more Favorably than the other Hydrophilic Amino Acids in RNase Sa. J Mol Biol 2007;366:449–60. https://doi.org/10.1016/j.jmb.2006.10.026.

[33] Dyson MR, Shadbolt SP, Vincent KJ, Perera RL, McCafferty J. Production of soluble mammalian proteins in *Escherichia coli*: Identification of protein features that correlate with successful expression. BMC Biotechnol 2004;4:32. https://doi.org/10.1186/1472-6750-4-32.

[34] Gutiérrez-González M, Farías C, Tello S, Pérez-Etcheverry D, Romero A, Zúñiga R, et al. Optimization of culture conditions for the expression of three different insoluble proteins in *Escherichia coli*. Sci Rep 2019;9:1–11. https://doi.org/10.1038/s41598-019-53200-7.

[35] Rosano GL, Ceccarelli EA. Recombinant protein expression in *Escherichia coli*: Advances and challenges. Front Microbiol 2014;5:1–17. https://doi.org/10.3389/fmicb.2014.00172.

[36] Marschall L, Sagmeister P, Herwig C. Tunable recombinant protein expression in *E. coli*: enabler for continuous processing? Appl Microbiol Biotechnol 2016;100:5719–28. https://doi.org/10.1007/s00253-016-7550-4.

[37] Pan H, Xie Z, Bao W, Zhang J. Optimization of culture conditions to enhance cis-epoxysuccinate hydrolase production in *Escherichia coli* by response surface methodology. Biochem Eng J 2008;42:133–8. https://doi.org/10.1016/j.bej.2008.06.007.

[38] Azaman SNA, Ramakrishnan NR, Tan JS, Rahim RA, Abdullah MP, Ariff AB. Optimization of an induction strategy for improving interferon-α2b production in the periplasm of *Escherichia coli* using response surface methodology. Biotechnol Appl Biochem 2010;56:141–50. https://doi.org/10.1042/BA20100104.

[39] Papaneophytou CP, Kontopidis GA. Optimization of TNF-α overexpression in *Escherichia coli* using response surface methodology: Purification of the protein and oligomerization studies. Protein Expr Purif 2012;86:35–44. https://doi.org/10.1016/j.pep.2012.09.002.

[40] Papaneophytou CP, Rinotas V, Douni E, Kontopidis G. A statistical approach for optimization of RANKL overexpression in *Escherichia coli*: Purification and characterization of the protein. Protein Expr Purif 2013;90:9–19. https://doi.org/10.1016/j.pep.2013.04.005.

[41] Papaneophytou C, Kontopidis G. A comparison of statistical approaches used for the optimization of soluble protein expression in *Escherichia coli*. Protein Expr Purif 2016;120:126–37. https://doi.org/10.1016/j.pep.2015.12.014.

[42] Rigi G, Mohammadi SG, Arjomand MR, Ahmadian G, Noghabi KA. Optimization of extracellular truncated staphylococcal protein A expression in *Escherichia coli* BL21 (DE3). Biotechnol Appl Biochem 2014;61:217–25. https://doi.org/10.1002/bab.1157.