


Evaluation of machine learning models on protein level inference from prioritized RNA features

Wenjian Xu [†], Haochen He[†], Zhengguang Guo and Wei Li

Corresponding authors: Zhengguang Guo, Core Facility of Instruments, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, School of Basic Medicine, Peking Union Medical College, 5 Dong Dan San Tiao, Beijing 100005, China. E-mail: guozhengguang@ibms.pumc.edu.cn; Wei Li, Beijing Key Laboratory for Genetics of Birth Defects, Beijing Pediatric Research Institute; MOE Key Laboratory of Major Diseases in Children; Rare Disease Center, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing 100045, China. E-mail: liwei@bch.com.cn

[†]These authors contributed equally to this work.

Abstract

The parallel measurement of transcriptome and proteome revealed unmatched profiles. Since proteomic analysis is more expensive and challenging than transcriptomic analysis, the question of how to use messenger RNA (mRNA) expression data to predict protein level is extremely important. Here, we comprehensively evaluated 13 machine learning models on inferring protein expression levels using RNA expression profile. A total of 20 proteogenomic datasets from three mainstream proteomic platforms with >2500 samples of 13 human tissues were collected for model evaluation. Our results highlighted that the appropriate feature selection methods combined with classical machine learning models could achieve excellent predictive performance. The voting ensemble model outperformed other candidate models across datasets. Adding the mRNA proxy model to the regression model further improved the prediction performance. The dataset and gene characteristics could affect the prediction performance. Finally, we applied the model to the brain transcriptome of cerebral cortex regions to infer the protein profile for better understanding the functional characteristics of the brain regions. This benchmarking work not only provides useful hints on the inherent correlation between transcriptome and proteome, but also has practical value of the transcriptome-based prediction of protein expression levels.

Keywords: proteogenomics, feature selection, prediction, transcription

Introduction

The correlation between proteins and their messenger RNA (mRNA) abundances has been a fundamental question [1, 2]. Although the mRNA sequence determines the protein sequence according to the central dogma, mRNA abundance and the corresponding protein abundance is not simply correlated, especially in different scenarios [3]. Protein levels are largely determined by coding mRNA under steady-state conditions. However, the protein level during dynamic transitions is influenced by many factors, including protein translation rate affected by synthesis capacity, protein's half-life affected by ubiquitin-proteasome pathway system, protein synthesis delay affected by ribosome saturation of high-abundance housekeeping proteins and the temporal and spatial distribution difference between mRNA and protein [1, 4]. Concurrent measurement of mRNA and protein genome-wide is also important for depicting the proteogenomic landscape of diseases, but protein quantification at high depth is more laborious, expensive and challenging than transcriptome quantification. Given the cost burden of proteomics, only about 20 proteogenomic datasets (matched proteome and transcriptome data from the same specimen) have been published

after years' extensive studies by National Cancer Institute (NCI) Clinical Proteomic Tumor Analysis Consortium (CPTAC), Chinese Human Proteome Project (CNHPP) and other groups [5–24] (Table 1). In these datasets, specimens at physiological conditions were collected at a single-time point and steady-state protein and RNA levels were quantified by proteomic and bulk transcriptome profiling. Therefore, protein levels in this work are the average level of many cells of tissue samples. These proteogenomic cohorts have repeatedly shown low mRNA–protein correlations (median value 0.2–0.4) and large intergene variability. For example, the metabolism-related gene showed high correlations, while ribosome genes showed low or negative correlations.

Low correlation between protein and mRNA levels indicates a complicated network of gene regulation driving all genes' expression. Since RNA quantification of a specimen is more routine and stable at higher throughput platforms, inferring protein level from RNAs computationally is intriguing and beneficial for the research of proteogenomics. This would also enable the mining of the mRNA expression data from large discovery cohorts without matched proteomic measurement. Therefore, CPTAC launched a community-based

Wenjian Xu is a research assistant at the Beijing Children's Hospital.

Haochen He is a research assistant at the Beijing Institute of Radiation Medicine.

Zhengguang Guo is an associate professor at the Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences.

Wei Li is a professor at the Beijing Children's Hospital, Capital Medical University.

Received: November 1, 2021. Revised: February 16, 2022. Accepted: February 23, 2022

© The Author(s) 2022. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Table 1. Overview of 20 proteogenomic datasets used in this work

Group	Dataset name	Source	Year	Project	Specimen source	Tissue	Transcriptome-proteome matched sample	Tumor sample	Normal sample
Label-free	CO_labelfree	[5]	2014	CPTAC	Colon and rectal cancer	Intestine	86	96	0
	BN_labelfree	[9]	2017	BrainSpan	Healthy brain	Brain	71	0	71
	PR_labelfree	[6]	2019	None	prostate cancer	Prostate	65	65	0
	LV_labelfree	[8]	2019	CNHPP	Hepatocellular Carcinoma	Liver	62	35	35
TMT	LU_labelfree	[7]	2020	CNHPP	Lung cancer	Lung	76	51	49
	LV_tmt	[10]	2019	None	Hepatocellular carcinoma	Liver	318	159	159
	CO_tmt	[14]	2019	CPTAC	Colon cancer	Intestine	95	110	0
	RC_tmt	[15]	2019	CPTAC	Renal cell carcinoma	Kidney	185	110	84
	PBN_tmt	[18]	2020	CPTAC	Pediatric brain cancer	Brain	188	218	0
	BR_tmt	[24]	2020	CPTAC	Breast cancer	Breast	122	134	0
	EC_tmt	[11]	2020	CPTAC	Endometrial carcinoma	Uterus	115	95	49
	LU_1_tmt	[12]	2020	CPTAC	Lung cancer	Lung	211	110	101
	LU_2_tmt	[13]	2020	None	Lung cancer	Lung	89	90	90
	LU_3_tmt	[22]	2021	CPTAC	Lung squamous cell carcinoma	Lung	202	108	99
	HN_tmt	[17]	2021	CPTAC	Head and neck squamous cell carcinoma	Mucosa	151	108	66
	PA_tmt	[23]	2021	CPTAC	Pancreatic cancer	Pancrea	140	140	67
	BN_tmt	[16]	2021	CPTAC	Glioblastoma	Brain	108	99	10
	iTRAQ	BR_itraq	[19]	2016	CPTAC	Breast cancer	Breast	77	77
OV_itraq		[20]	2016	CPTAC	Ovarian cancer	Ovary	119	169	0
GC_itraq		[21]	2019	CPTAC	Gastric cancer	Stomach	80	80	80

'challenge' in 2017, aiming to improve the predictability of protein levels from transcriptome and reported top-performing models after an unbiased assessment on two cancer datasets [19, 20, 25]. The challenge's top models work well for a subset of proteins, albeit still far from perfect-predicting protein level globally. The best performing method, described in Li's study [26], is a Random Forest Regression (RFR) model using all RNA features supplemented with mRNA level as a proxy. The authors found using all features was better than using selected features by protein abundances [26]. Li's method is abbreviated as teamHL&YG in this study. The challenge also reports four other models including the CPTAC baseline model based on Elastic Net (baselineEN) using all RNA features, an RFR model using prioritized RNA features with mRNA proxy (teamHYU), an ensemble model using gene copy number, prioritized RNA features and other gene metadata (teamDEARGENpg) and a least absolute shrinkage and selection operator (LASSO) model using gene copy number and prioritized RNA features (teamDMIS_PTG). Models derived from the challenge serve as a valuable resource and basis for further improvement.

To further improve performance on this task, several influencing factors should be considered. First, collecting more independent datasets would help optimize models. To the best of our knowledge, inferring protein levels from RNA expression profiles has not been benchmarked comprehensively in more tissues for all three mainstream proteomic platforms. Therefore, a large up-to-date collection of high-quality datasets is in great demand. Second, feature selection is very important

for machine learning with limited training samples. Selecting features by prior information, such as RNA abundances, gene network or pathway knowledge, was utilized by three of the top four models, although it seemed unnecessary for the RFR model [25, 26]. Specifically, the correlation-based feature selection method was used in one of the top four models, which prioritizes RNAs by Pearson's correlation with target protein groups (instead of a single protein) across all the data. A similar correlation-based feature selection was also used in our work on blood-based tissue gene expression prediction, where features with the highest absolute cosine similarity (ACS) were selected for each target [27]. Third, the performance of ensemble models is preferred but linear and nonlinear regression models were not significantly different [25]. Therefore, the lack of comprehensive benchmarking of protein level inference performance of different machine learning models, feature numbers and independent datasets from different tissues, leaves users without indications as to which prediction method could achieve optimal performance.

Here, we first curated a collection of publicly accessible datasets from three mainstream proteomic platforms. We improved the transcriptome-based protein level prediction performance model by introducing feature selection into classical regression models. To validate our method with independent datasets, we benchmarked the existing methods, our improved models, other ensemble models and Neural Network (NN) models. Moreover, we proposed the voting ensemble of six models which performed superior in most

benchmarking datasets. We then analyzed the influencing factors of protein's predictability. Finally, we applied the model to infer protein profiles of the cerebral cortex regions using the available brain transcriptome dataset. In summary, this study illustrated the practical value of the transcriptome-based protein level prediction on a wide range of tissues.

Materials and Methods

Dataset collection and preprocessing

The datasets were collected from proteogenomic cohorts (Table 1) and preprocessed using a unified pipeline, including gene ID mapping, NA value removal, normalization and standardization (Supplementary Methods). After preprocessing, we constructed 20 datasets consisting of 2560 tumor/control samples of 13 human tissues, such as colon and rectum, prostate, lung, liver and brain tissues.

Each dataset was split for modeling and evaluation using 5-fold cross-validation (CV). In each of the five iterations, the dataset (X, Y) was randomly partitioned into 80% for training (denoted as $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$, where i is sample id) and 20% for testing. More details were described in Supplementary Methods.

The prediction model

We illustrated the modified predictive model in a flow chart, in which per-protein feature selection steps were shown (Figure 1A). We formulated a multitask regression model for inferring protein level using RNA profile of the same sample. Mean absolute error (MAE), root mean squared error (RMSE) and Pearson's correlation coefficient (PCC, r) were used to evaluate the predictive performance of a target protein. Candidate machine learning models were linear regression (LR), LASSO, ridge regression (Ridge), Huber regression (HR), support vector regression (SVR), RFR, three NN models with 2–3 hidden layers (NN1, NN2 and NN3) and four ensemble methods (Bagging, Boosting, Stacking and Voting).

Three existing methods baselineEN, teamHYU and teamHL&YG were implemented and compared to our models in this study. BaselineEN is an elastic net model, whose regularization hyperparameter is optimized with 5-fold CV. TeamHYU used both RFR using prioritized RNA features and mRNA proxy model which directly used RNA level as predicted protein level. If gene predictive performance of RFR is better in 5-fold CV within the training set, teamHYU outputs RFR predict value. Otherwise, teamHYU outputs mRNA proxy value. TeamHL&YG has three components: mRNA proxy model, RFR model trained on one dataset and pan-cancer RFR model trained on combined dataset. No optimization was needed for the mRNA proxy model. The pan-cancer model was trained on combined isobaric tags for relative and absolute quantification (iTRAQ) datasets of two tissues because of the limited sample size of the

CPTAC challenge. We decided not to include the pan-cancer model for two reasons: first, the combination of tandem mass tag (TMT) datasets is not feasible because each dataset has its own inner control; second, recent cohorts are using more samples so that low sample size is not a problem. Therefore, improving performance requires efforts in building a regression model and incorporation of proxy model into regression. Briefly, we benchmarked four feature selection methods followed by 13 regression models against baselineEN and proposed the best candidate model; next, the mixing ratio of proxy and regression model was optimized within the range of (1:0, 9:1, 7:1, ..., 1:7, 1:9, 0:1); the final 'regression+proxy' model was compared to baselineEN, teamHYU and teamHL&YG.

The other two existing methods of teamDEARGENpg and teamDMIS_PTG were not comparable to ours because they also use DNA copy number variation as part of input features. Implementation details of the models were described in Supplementary Methods.

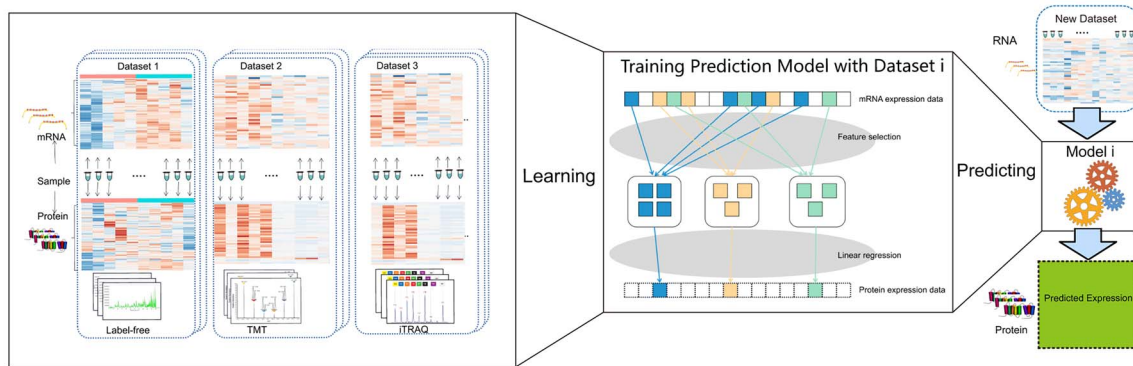
Feature selection

The feature length of most linearly correlated RNAs is a crucial hyperparameter. Assume vector $\mathbf{y}_{(p)}$ represents the expression value of target protein p in all N training samples and vector $\mathbf{x}_{(R)}$ represents the expression value feature RNAR in all N training samples. We quantify the ACS between feature vector $\mathbf{x}_{(R)}$ and target vector $\mathbf{y}_{(p)}$. R is considered highly correlated with p when their ACS value is close to 1. We can prioritize arbitrary S features for protein p by ACS approach and, therefore, reduce feature dimension as desired. To determine optimal feature length, we evaluated the models' protein level prediction performance by CV on the training set with a series of feature lengths $S \in [10, 20, 50, \dots, 5000]$. For all target proteins, when the model trained with reduced feature sets have the smallest average MAE, smallest average RMSE and the largest average r in the serial experiments, the value S is the optimal feature length S^o for this specific dataset. Select top S^o features to construct the final dataset $\{\mathbf{x}'_{i(p)}, \mathbf{y}_{i(p)}\}_{i=1}^N$, where $\mathbf{x}'_i \in \mathcal{R}^{S^o}$, $S^o \ll N$. Three other feature selection methods were compared to ACS, which are raw cosine similarity, Spearman's correlation and random selection. The teamHYU method can be viewed as another control method for ACS feature selection because it includes feature selection by neighbor relations on the PPI network and Pearson's correlation value.

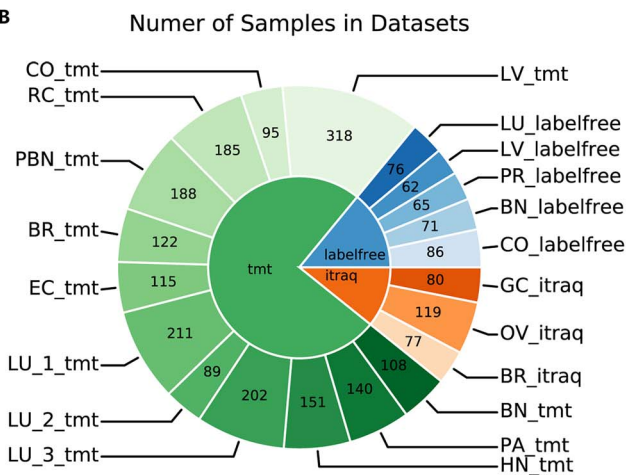
Weighted voting ensemble model

To test whether an ensemble improves the prediction performance, a voting ensemble with customized weights (Voting-wt) was defined as weighted mean of six classical models. These voting ensembles had their own optimal feature length on each dataset. The average metric correlation (r) of all proteins in each optimal model was calculated using 5-fold CV on the training set. The ascending

A



B



C

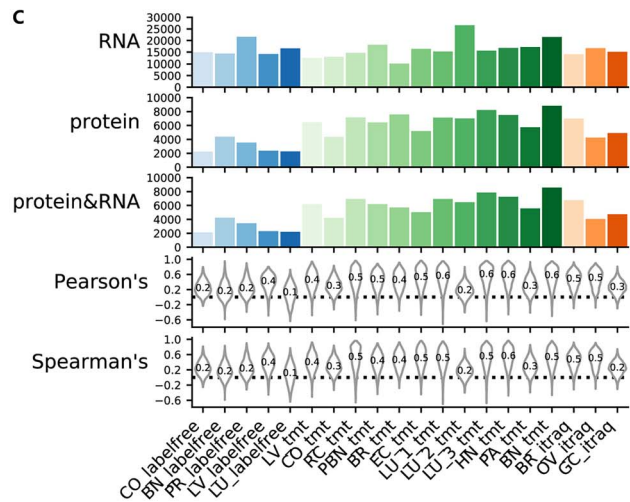


Figure 1. Overview of flow chart and benchmarking datasets. **(A)** Flow chart of the enhanced predictive model with feature selection module. The datasets of transcriptome-proteome matched samples from Labelfree, TMT and iTRAQ platform are used to train dataset-specific models. The trained model can be used for protein level inference. The summary of proteogenomic cohort datasets used in this work is shown in **(B)** and **(C)**. **(B)** Pie plot: the area of section is proportional to sample size; the color indicates dataset and its platform. **(C)** Bar plots show the number of RNA, protein and overlapping genes. Violin plots show RNA-protein correlations by Pearson's and Spearman's correlation coefficient.

rank of r was used to derive the models' weight. Specifically, the sum of model weights was 1, so the weight from the best to the worst model was $\frac{6}{21}$, $\frac{5}{21}$, $\frac{4}{21}$, $\frac{3}{21}$, $\frac{2}{21}$ and $\frac{1}{21}$, respectively. Then, retrain models with optimal feature length and new weights on the whole training set and evaluate on the hold out test set. The predicted values from the six models were combined by corresponding weight as the predicted output of Voting-wt. The analysis was exclusively performed on the large datasets which have >150 samples.

Dataset complexity

We adapted the concept of data complexity from single-cell RNAseq and implemented the computation of data complexity locally (Supplementary Methods). The average expression of every gene for each cell population in a dataset was calculated to represent the prototype of the cell population in the full gene space before describing the complexity of a dataset in ref. [28]. In this study, samples were directly treated as prototypes of the data matrix. Full features were used to measure the complexity of RNA/protein matrix of each dataset. Briefly,

the gene expression profile of samples was standardized by gene. Next, pairwise Pearson's correlation between samples was calculated. For each sample, the maximum correlation to another sample was recorded. Finally, the mean value of these per sample maximum correlations was taken to describe the complexity of matrix. The dataset complexity value was collected in both RNA and protein levels.

Gene characteristics

Ten characteristics of genes were included as potential effect factors on protein's predictability, which are posttranslational modification (PTM), gene relation to human disease, gene expression tissue specificity, protein complex membership, protein subcellular localization, protein functional class, protein half-life, gene essentiality, gene length and protein relative abundance compared to other genes. Gene and protein characteristics were retrieved from public data source, including OMIM (version 2020-07-30, <https://www.omim.org>) [29], Human Protein Atlas (version 2021-02-24, <https://www.proteinatlas.org>) [30], iPTMnet (version 6.0,

<https://research.bioinformatics.udel.edu/iptmnet/>) [31], peptide turnover dataset [32], CORUM (version 3.0 <http://mips.helmholtz-muenchen.de/corum/>) [33] and Bartha's review study [34]. We preprocessed the categorical variables and assigned a unique category for each gene at every characteristic level (Supplementary Methods). For the other numerical factors, we sort the genes by the value of the factor and split them into two groups, labeled as long/short or high/low group, respectively. Each gene was assigned to one major group in every characteristic level for convenience. Any genes without any annotation in a characteristic level were omitted from the specific analysis.

Implementation and visualization

Pearson's and Spearman's correlation coefficients were implemented by `scipy.stats` from `scipy` v1.3 [35] in Python v3.7. Prediction models were implemented with `scikit-learn` v0.21 [36] and NN library `scorch` (<https://skorch.readthedocs.io/en/stable/>) using default parameters unless otherwise specified in Supplementary Methods. After optimal feature length was determined, the runtime of feature selection, model training and model test were benchmarked on Linux servers (AMD EPYC 7742 64-Core CPU 2.25GHz, 1 Tb RAM) (Supplementary Table S1). The plots were visualized using `Matplotlib` v3.1 [37] in Python, `ggplot2` v3.3.3 [38] and `ggpubr` v0.4.0 in R v4.0.3. In figures with significance marks, statistical significance was indicated by the following convention: * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$, **** $P \leq 0.0001$. Principal component analysis (PCA) was performed in SIMCA v15.02 (Umetrics, Sweden) statistical software.

Results

Overview of curated benchmarking datasets

To analyze protein prediction comprehensively, we curated a total of 20 datasets consisting of 2560 tumor/control samples of 13 human tissues from cohort publications (Figure 1B). Twelve datasets used TMT-labeled quantitative proteomics, three used iTRAQ quantification, while the other five used label-free quantification. Therefore, the datasets are representative of protein expression prediction challenges. The TMT datasets consist of more samples than the other two groups. 7000–13 000 proteins were identified in each dataset before preprocessing and 2200–8800 high-quality proteins were retained after preprocessing (Figure 1C). More proteins were detected in TMT and iTRAQ profiling datasets than in label-free ones. The detected RNAs were > 10 000 in every dataset. On average, the datasets had 5339 proteins with mRNA levels available (Figure 1C). The median Pearson's/Spearman's correlation coefficients were 0.12–0.60 and 0.10–0.55, respectively. The protein–RNA correlation coefficients of the dataset did not strongly correlate with sample size, protein number and RNA number.

Benchmark of regression models and feature selection methods on curated datasets

In our previous work on gene expression inference tasks, we observed that a small set of selected blood RNA features linearly correlated with a target gene expressed in tissues could result in good predictive performances [27]. The ACS was used to prioritize features for blood-based tissue gene expression prediction models. Since protein prediction tasks are similar to those gene expression inference settings, it is worth testing whether ACS could improve protein prediction performance. As a preexperiment, we adopted ACS to quantitate how well one RNA is linearly correlated with the protein on the same data sets [25, 26], namely 'OV_itraq_itraq' and 'BR_itraq' in benchmarking datasets. We calculated pairwise ACS between all possible RNA–protein pairs using the training set split. For each protein, we took 10–5000 prioritized RNA features by ACS to construct a series of low-dimensional RNA feature sets. Then, we fit the RFR model (used by winner team HL&YG) and compared its predictive performance across different low-dimensional feature sets and the full feature set using a 5-fold CV scheme. In metric r and at least one of the errors, RFR with ACS selected features perform better than that with all features at some 'sweet point' feature length. No improvements were observed with randomly selected features (Supplementary Figure S1). RFR using ACS prioritized features is a better choice than no feature selection in these two iTRAQ datasets.

We next investigated the effect of feature selection on the performance of nine classical regression models, which are LR, LASSO, Ridge, HR, SVR, RFR, NN1, NN2 and NN3. On each of the 20 datasets, model fitting and evaluation were conducted after RNA feature prioritization by ACS and three other feature selection methods, respectively. As shown in column 6 in Supplementary Figures S2–S4 and Supplementary Table S2, the RFR model performed the best at about 200 optimal features on 18 datasets out of 20. The results of the three performance metrics showed that the overall performances of the RFR model did not change dramatically with feature lengths near optimal points. When feature length decreased from the maximum all to the minimum 10, the performances increased first and decreased later. Comparing four feature selection methods, ACS improved RFR performance more than the other three methods in more than half datasets (14/20, 12/20, 12/20 in terms of r , RMSE, MAE, respectively) (column 6 in Supplementary Figures S2–S5). When random features were used as parallel controls, RFR performances descended all the way when feature length decreased.

We extended the analysis to nine regression models and consistently observed the comparative advantage of ACS (columns 1~9 in Supplementary Figure S5). Random features were worse than prioritized features at almost any feature length. We next analyzed nine classical regression models with ACS prioritized features

at different feature length (Supplementary Figures S2–S4). The linear models (LR, Ridge and HR) descended to the lowest performance around 100 prioritized features and the optimal performances were achieved at either >5000 or <20 optimal features. The performance of Lasso was insensitive to feature number compared to that of LR, HR and Ridge models. SVR with a nonlinear kernel had a similar performance trend to RFR. NN1~NN3 models achieved optimal performance at 1000–5000 features. Interestingly, the optimal performance of nonlinear models was quite comparable to the linear models, suggesting that linear and nonlinear models have their respective advantages. Moreover, performance with <20 features was comparable to that of the optimal feature set. These results on benchmarking datasets confirmed the necessity of feature selection for protein level inference models.

We will analyze how much were the models improved using prioritized features by ACS and optimal feature length (Figure 2A). All nine models outperformed baselineEN in more datasets with ACS prioritized features than themselves with all features (Figure 2B). However, the beneficial effect of ACS seemed to be dataset dependent. The median performance of optimal RFR was the best among the nine models and still not better than baselineEN (Figure 2C). The relatively good performance of RFR in our analysis explained its popular usage in existing methods of teamHYU and teamHL&YG.

Performance of ensemble model

The weighted mean ensemble is a type of voting ensemble. They combine the predictions from multiple models proportionally to each model's capability. Voting ensemble of multiple models (teamHL&YG, teamHYU, teamDEARGENpg, teamDMIS_PTG) where the average metric (r) of all proteins in each model as model's voting weight were superior over individual ones in two tissues [25]. Next, we investigated Voting/Stacking ensemble models of six classical models (LR, LASSO, Ridge, HR, SVR, RFR), Boosting ensemble of 50 (default) decision tree regressors (DTRs), Bagging ensemble of 10 (default) DTRs with curated benchmarking datasets. The ensemble models were built with their own optimal ACS feature sets. Then, we compared the performances of a total of 13 models: m1–m6: six classical models, m7–m9: three NNs, m10: Boosting, m11: Bagging, m12: Stacking ensemble of m1–m6, and m13: Voting ensemble of m1–m6.

As shown in Figure 2 and Supplementary Figures S2–S4, the upper and lower limits of four ensemble model performances depend strongly on the dataset in use. All models generally performed better on TMT datasets than on iTRAQ and label-free datasets (Figure 2A). BaselineEN performs the best in six datasets (Figure 2B). The Voting/Stacking ensemble achieved better overall r than baselineEN in 12/20 and 9/20 datasets, respectively. RFR with optimal feature sets was the third-best model which performs better than baselineEN in 9/20 datasets (Figure 2B). The median performance of Voting ensemble was the best among 13 models in terms of r , RMSE and

MAE (Figure 2C). Compared to baselineEN, Voting ensemble was better in terms of r , but not in terms of RMSE and MAE. We next investigated voting ensemble with customized weights (Voting-wt) of six classical models on five largest datasets. Voting-wt was only improved marginally in 3/5 datasets (Supplementary Figure S6).

Performance improvement by adding proxy model

The above benchmarking analysis of prediction models was focused on the regression model (Figure 2 and Supplementary Figures S1–S6). We next evaluated how a proxy model, a simple but essential component of the original teamHYU and teamHL&YG method, affects the performance of 13 models. Only those common genes with both protein and RNA measurements were included in this proxy model analysis (the number of common genes shown in Figure 1C). As shown in Figure 3 and Supplementary Table S3, the optimal regression model performed better in 15 datasets and the proxy model performed better in 5 datasets. When the proxy model was mixed with the regression model at a range of ratios, every mixture model gained performance improvement (Figure 3A, B and Supplementary Figures S7 and S8). The top-performing mixture ratio of proxy to regression was 1:3 for most models on average (Figure 3C). Since regression model accounts for a larger proportion, we named the final mixture as regress + proxy. The regress+proxy models were 15% better than regression themselves in terms of r on average (Figure 3D). These results together with the previous literature showed that the mixture of regression model and proxy model can reach a better performance than either model.

Evaluation of final models and existing models

We next evaluated whether final regress+proxy models perform better than existing models. Voting+proxy at ratio 1:3 ranked first overall and maintains the best performance in 10/20, 5/20, 5/20 datasets in terms of r , MAE and RMSE, respectively (Figure 4A and Supplementary Table S4). We here noticed that rankings of models by three performance metrics were sometimes different though generally consistent. For example, Stacking+proxy ranked first in 4/20, 4/20, 4/20 datasets; teamHL&YG ranked first in 1/20, 1/20, 1/20 datasets; baselineEN ranked first in 2/20, 3/20, 3/20 datasets (Figure 4A and Supplementary Figure S9). BaselineEN (without proxy) seemed to be a simple and competitive method. Figure 4C showed that the median performance of Voting+proxy across 20 datasets was the best among all models in three metrics.

Moreover, gene-wise comparison of model performances confirmed the superiority of Voting+proxy over existing models in 13/20 datasets (Table 2). We then compared the performance of all models on training splits and the test splits (Supplementary Table S5). The range of r was 0.7–0.9 in training fit and 0.2–0.7 in test fit. The results suggested that all models, including

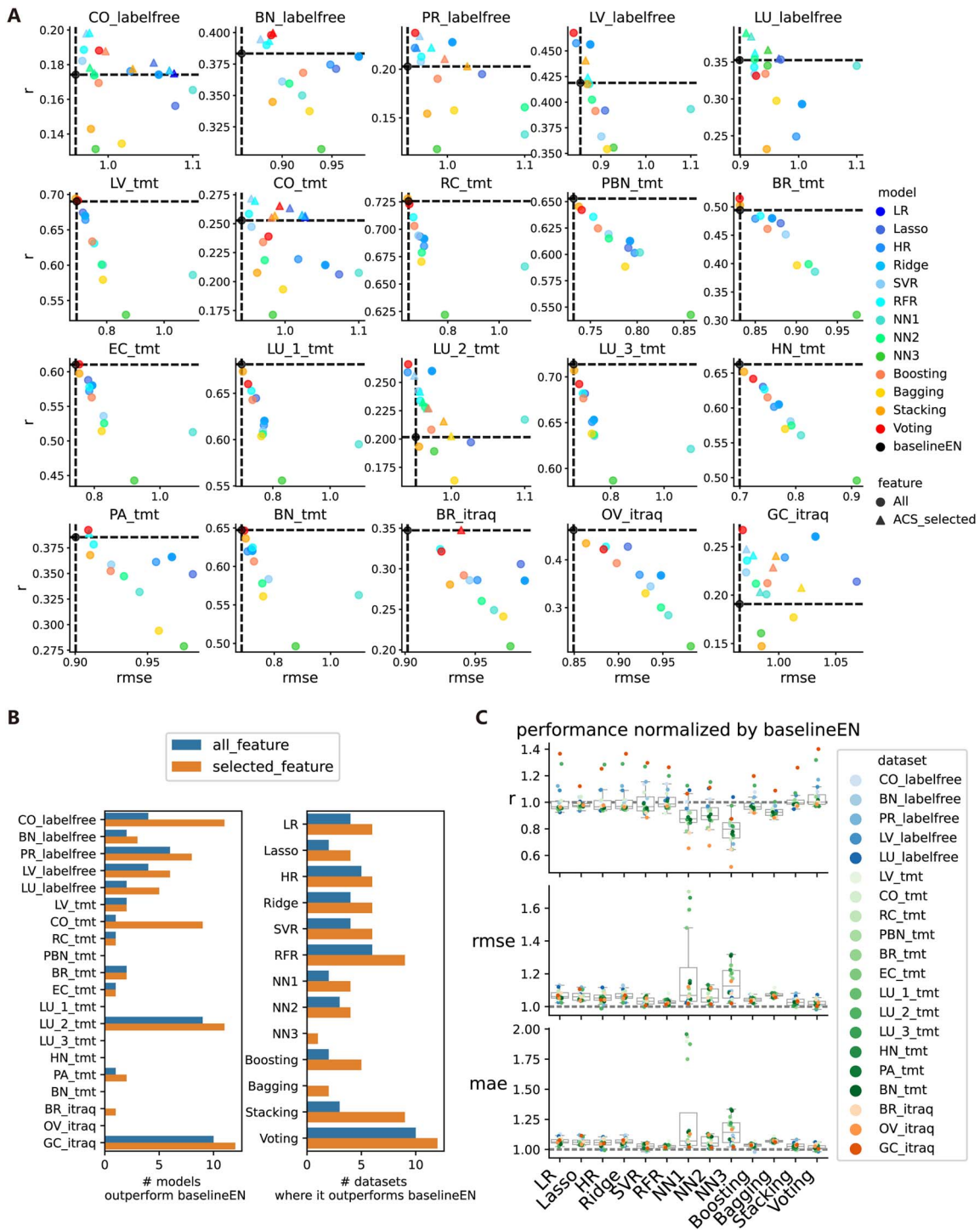


Figure 2. Evaluation of the effect of ACS feature selection on 20 datasets using baselineEN and 13 regression models. (A) Scatter plots show the performance metric r and RMSE. Marker shape indicates with (triangles) or without (circles) feature selection. All triangles indicate the models outperform baselineEN with ACS prioritized features. (B) Left panel, the number of models outperform baselineEN in each dataset; right panel, the number of datasets that each model outperforms baselineEN. (C) Relative performance comparison. Model performance metrics are normalized by that of baselineEN in the same dataset. Good models are supposed to have $r > 1$ and RMSE/MAE < 1 .

both ours and existing ones, showed a certain degree of overfitting in benchmarking datasets.

The relationship between overall predictability and data sets characteristics

Since the remarkable variation of protein inference performance across benchmarking datasets, we investigated

what characteristics of datasets affect the overall performances with the weighted mean model. Multiple characteristics of datasets were collected, including sample size, RNA number, protein number, number of tissue-specific genes, gene-level mRNA–protein correlations, the complexity of RNA and protein matrices and proteomic platform (Supplementary Table S6).

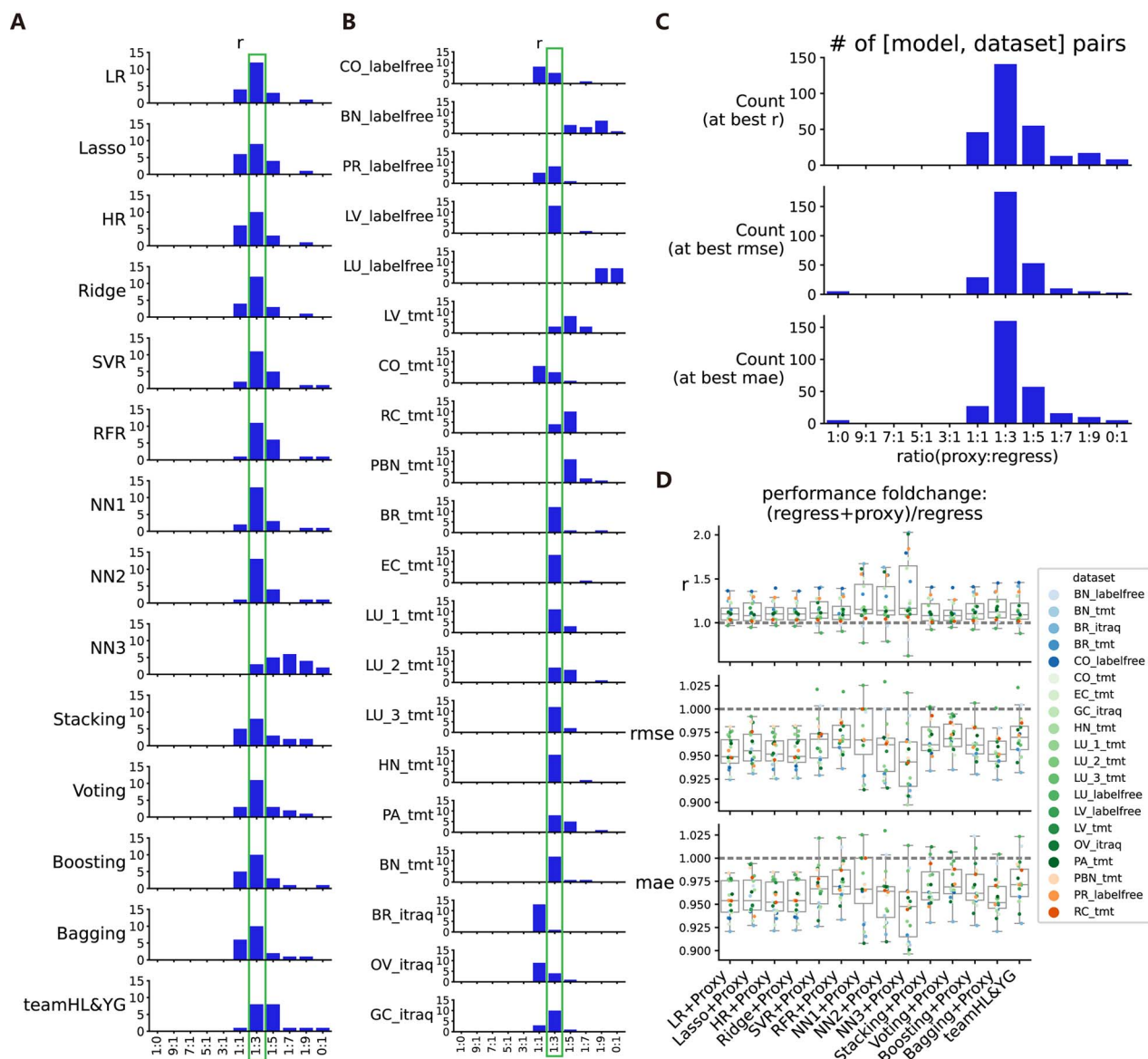


Figure 3. Performance improvement by adding proxy to 13 regression model and teamHL&YG. Bar plots show the distribution of top-performing mixing ratio (proxy:regression) for all [model, dataset] pairs in terms of r : (A) each model on all 20 datasets and (B) each dataset with all 14 models. (C) The best ratio distribution of any models and datasets in terms of three metrics. (D) Relative performance comparison before and after adding proxy. Model performance fold change are calculated for models in the same dataset. Good models are supposed to have $r > 1$ and RMSE/MAE < 1 .

As shown in the left column of Figure 5 and Supplementary Figure S10, total protein and tissue-specific gene number of datasets affected model performances but RNA number did not, which indicates that the depth of protein profiling experiment is an important influencing factor. Noticeable performance difference between proteomic platforms was observed, which agrees with previous results (Figures 4 and 5). The complexity of RNA and protein matrices were both positively correlated with model performances (middle column of Figure 5 and Supplementary Figure S10). The performance of datasets was improved when the sample size increased and when the mRNA-protein correlation coefficient increased (right column of Figure 5 and Supplementary Figure S10). These results indicated that the optimal performances vary across datasets might

be explained by the differences of proteomic datasets characteristics.

The relationship between protein predictability and gene characteristics

As shown in Table 2, there was considerable variation in protein prediction performance at gene level. We want to investigate whether the prediction performance is significantly different in the various subset of protein, so we use gene characteristics to group and compare the proteins. Multiple factors may influence the prediction performance of genes, such as protein complex membership and protein half-life [26]. To the best of our knowledge, gene characteristics and protein predictability have not been investigated side-by-side across multi-tissue and multi-platform datasets. Using Voting ensemble mixed

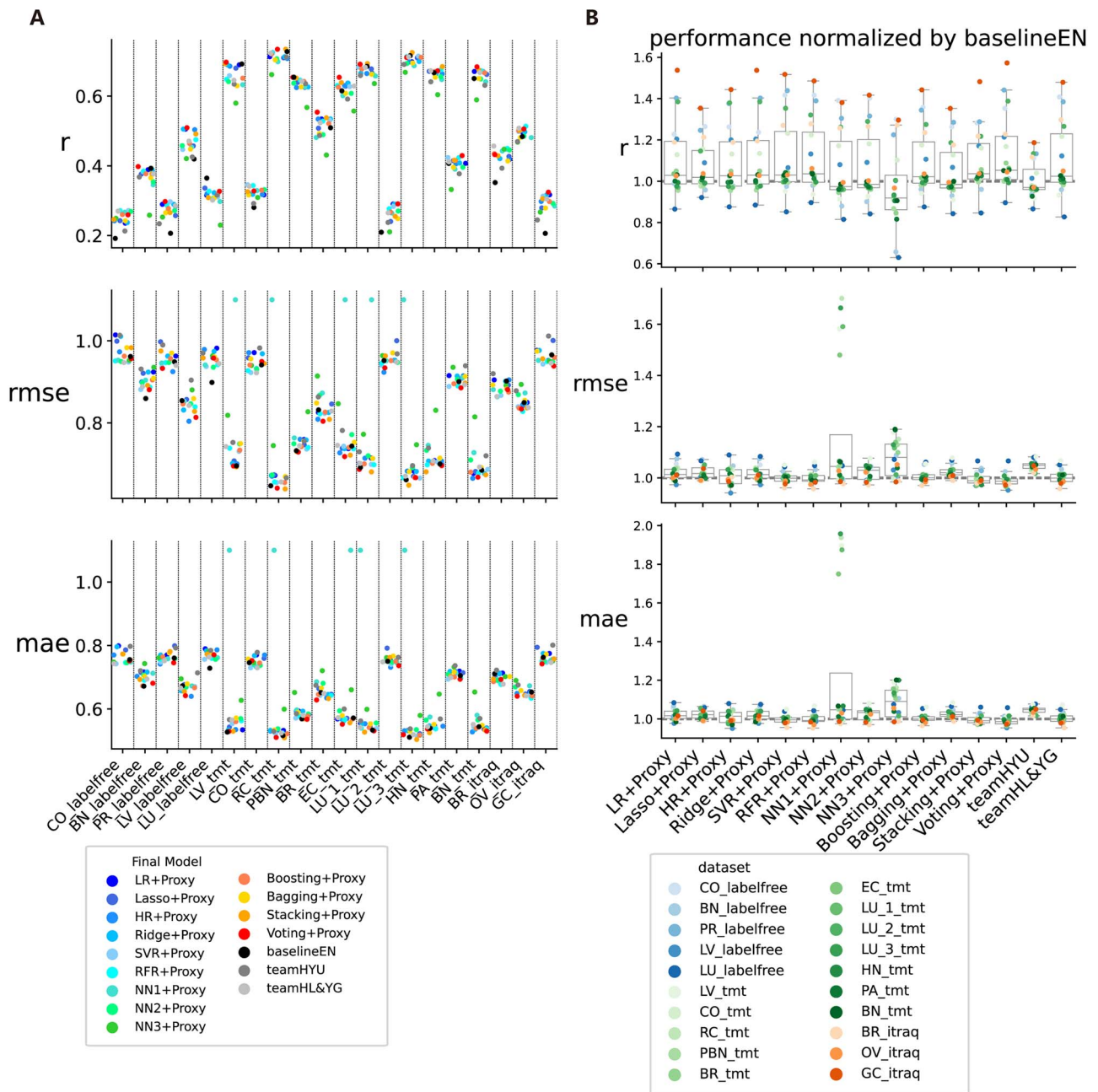


Figure 4. Performance of three existing models and 13 optimized regression+proxy models on 20 datasets. (A) Model performance in three metrics: (top panel) r , (middle panel) RMSE and (bottom panel) MAE. Dot colors indicate models; horizontal axis represents datasets. Good models are supposed to be higher in top panel and lower in middle/bottom panels. (B) Relative performance comparison. Model performance metrics are normalized by that of baselineEN in the same dataset. Good models are supposed to have $r > 1$ and RMSE/MAE < 1 .

with proxy as a representative model, we investigated the potential effects of gene characteristics on protein's predictability.

Briefly, genes outside protein complexes were better predicted than genes belonging to protein complexes (Figure 6A and Supplementary Figure S11A). Genes with high tissue-specific expression were better predicted than non-tissue-specific genes (Figure 6B and Supplementary Figure S11B). Long half-life proteins were better predicted than short half-life proteins (Figure 6C and Supplementary Figure S11C). Genes with long peptide length were better predicted than genes

with short peptide length (Figure 6D and Supplementary Figure S12A). The two factors, gene relation to human disease and protein subcellular localization, did not affect protein predictability in most datasets (Figure 6E-F and Supplementary Figure S12B, C). Protein abundance was positively correlated with protein predictability in label-free datasets (Figure 6G and Supplementary Figure S13A). Consistent with Yang's study [25], ribosomal proteins were worse predicted than other protein functional classes (Figure 6H and Supplementary Figure S13B). Proteins modifiable by sumoylation, phosphorylation or glycosylation were

Table 2. Model performance comparison per protein

	Model performance in metric <i>r</i>				P-value of Wilcoxon signed-rank test		
	BaselineEN	TeamHYU	TeamHL&YG	Voting+Proxy	H1: Voting+Proxy >baselineEN	H1: Voting+Proxy >teamHYU	H1: Voting+Proxy >teamHL&YG
CO_labelfree	0.169 ± 0.217	0.213 ± 0.195	0.270 ± 0.180	0.259 ± 0.187	3.1e-168	2.1e-68	1
BN_labelfree	0.383 ± 0.268	0.368 ± 0.254	0.377 ± 0.243	0.398 ± 0.242	1.1e-12	3.7e-66	2.5e-84
PR_labelfree	0.201 ± 0.238	0.234 ± 0.229	0.286 ± 0.211	0.298 ± 0.213	4e-270	2.4e-167	3.6e-23
LV_labelfree	0.419 ± 0.213	0.425 ± 0.202	0.471 ± 0.180	0.509 ± 0.164	1.2e-218	1.5e-209	1.1e-162
LU_labelfree	0.351 ± 0.237	0.316 ± 0.218	0.302 ± 0.231	0.327 ± 0.214	1	1.4e-08	4.7e-39
LV_tmt	0.691 ± 0.150	0.634 ± 0.168	0.646 ± 0.162	0.697 ± 0.137	3.5e-45	0	0
CO_tmt	0.251 ± 0.226	0.292 ± 0.199	0.342 ± 0.180	0.328 ± 0.192	0	6.7e-116	1
RC_tmt	0.726 ± 0.171	0.712 ± 0.178	0.723 ± 0.173	0.734 ± 0.162	2.5e-91	2.7e-273	6.7e-115
PBN_tmt	0.653 ± 0.153	0.636 ± 0.152	0.642 ± 0.150	0.654 ± 0.147	5.2e-05	1.2e-185	4.4e-180
BR_tmt	0.497 ± 0.213	0.487 ± 0.196	0.524 ± 0.188	0.554 ± 0.175	0	0	0
EC_tmt	0.611 ± 0.206	0.591 ± 0.199	0.629 ± 0.188	0.654 ± 0.175	0	0	1.7e-253
LU_1_tmt	0.682 ± 0.187	0.659 ± 0.186	0.682 ± 0.174	0.691 ± 0.172	1.6e-83	0	3.6e-132
LU_2_tmt	0.201 ± 0.196	0.238 ± 0.187	0.265 ± 0.168	0.291 ± 0.167	0	7e-269	7.6e-259
LU_3_tmt	0.714 ± 0.176	0.691 ± 0.179	0.713 ± 0.169	0.711 ± 0.169	1	0	1
HN_tmt	0.663 ± 0.203	0.641 ± 0.194	0.673 ± 0.180	0.672 ± 0.185	4e-55	0	0.039
PA_tmt	0.389 ± 0.191	0.377 ± 0.171	0.412 ± 0.164	0.429 ± 0.161	2.4e-225	0	4.3e-140
BN_tmt	0.647 ± 0.189	0.631 ± 0.180	0.668 ± 0.165	0.684 ± 0.160	0	0	3.5e-306
BR_itraq	0.348 ± 0.228	0.393 ± 0.223	0.457 ± 0.191	0.432 ± 0.206	0	2.7e-159	1
OV_itraq	0.464 ± 0.231	0.469 ± 0.205	0.508 ± 0.189	0.506 ± 0.197	6.2e-228	1.1e-174	1
GC_itraq	0.190 ± 0.195	0.245 ± 0.165	0.305 ± 0.150	0.324 ± 0.148	0	0	2.1e-65

Best performing model for each dataset is indicated in bold.

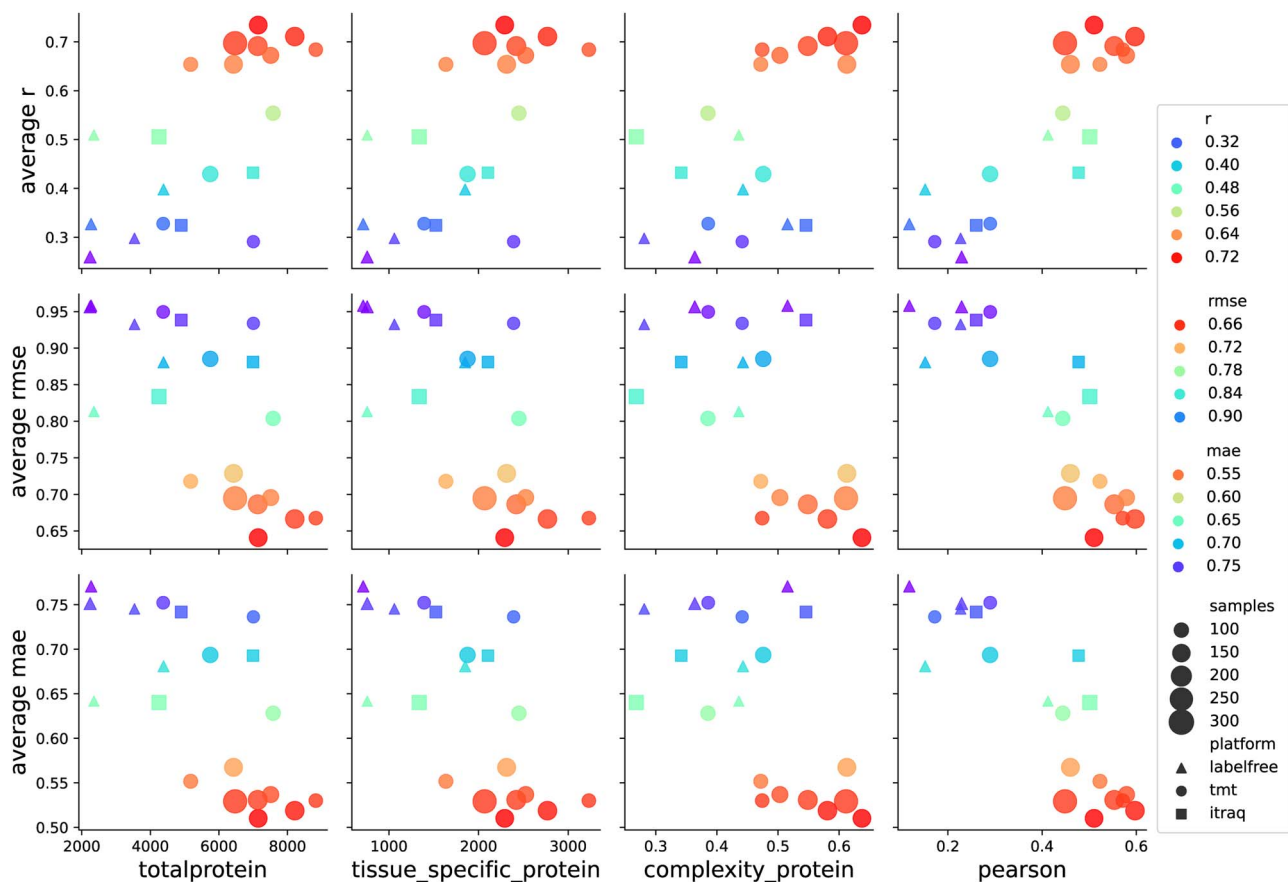


Figure 5. The relationship between predictability and characteristics of data sets. Dots represent datasets. Dot size represents sample size and dot shape represents platform. Scatter plots: vertical axis represents performance *r* (top row), RMSE (middle row) and MAE (bottom row); horizontal axis from left to right represents the number of proteins, number of tissue specific proteins, complexity of protein matrices and PCC between mRNA and protein.

better predicted than other PTM types (Figure 6I and Supplementary Figure S13C). Proteins of essential genes were worse predicted than those of nonessential genes (Figure 6J and Supplementary Figure S13D).

Validation of model performances on human brain atlas dataset

The Allen Institute for Brain Science created a comprehensive atlas of the human brain by transcriptome profiling hundreds of anatomically precise subdivisions, such as cerebral cortex regions parietal lobe (PL), frontal lobe (FL), occipital lobe (OL), temporal lobe (TL), cingulate gyrus (CgG), cerebellar cortex (CbCx) and other neurological nuclei [39, 40]. However, no similar proteome atlas is available. To predict brain region proteome expression, a prediction model was pretrained on the BN_labelfree dataset and take the Allen transcriptome dataset as input data (Supplementary Table S7). The predicted proteome was visualized in PCA plots (Figure 7, Supplementary Table S8). The agreement between samples' annotation and proteome/transcriptome profiles was also evaluated with three clustering performance metrics (Silhouette Coefficient, Davies-Bouldin score and Calinski & Harabasz score). At the measured transcriptome level, CbCx was far separated from the other six regions; Striatum (Str) and globus pallidus (GP) were also clearly grouped, but other regions were not be separated from each other. At the predicted proteome level, the obvious separation of CbCx, Str and GP was still preserved. Moreover, the proteome-based grouping of OL, PL, FL, TL and CgG substructures within the cerebral cortex was clearer than the transcriptome. OL, classic visual processing cortex, formed a compact cluster separated from other cerebral cortex regions (Figure 7C, D). The CgG region, involved in emotion processing and behavior regulation, is known to exhibit distinctive cytoarchitectural structure from other neocortex. Indeed, the CgG was more tightly grouped at the predicted proteome level. Similar phenomena were observed in samples from the second donor of Allen's dataset. Taken together, predicted proteome better reflected the cytoarchitectural and functional characteristics of brain regions than input transcriptome.

User guide on predicting protein profiles

Users are recommended to download our Github repository which contains a tiny training set (100 genes×200 samples) and script files. All 16 predictive models of this study were wrapped into demo.py. Users should run this demo as a first try to understand the format of the transcriptome input and predicted proteome output. Then, users can move to work on either the 20 benchmarking datasets (see Data availability) or in-house data. The prepacked models are highly adjustable and new models can be incorporated into the pipeline.

Conclusion and Discussion

In this study, we curated a collection of datasets with matched protein-RNA profiles, benchmarked machine learning models on inferring protein expression levels using RNA expression profile. Then, we proposed the weighted mean of six classical models as a new ensemble model for RNA-profile-based protein expression inference. We demonstrated that the Voting ensemble model outperformed other candidate models across most benchmarking datasets. Adding the proxy model to the new model would further improve the prediction performance. Our work would enable in-depth reanalysis of important biological samples with only transcriptome measurements available. Therefore, we applied the pretrained prediction model to the brain mRNA profile of cerebral cortex regions and showed the inferred protein profile better reflected the functional characteristics of brain regions than the RNA expression profile. This case study on a real-world dataset highlighted the potential benefit of computationally predicting protein expression and suggested that prediction models could complement transcriptome data.

Dataset characteristics would affect prediction performance. Large dataset size and complexity make the model capture the underlying pattern in more detail. Moreover, the positive correlation between protein number and dataset performance indicates that the proteomic data depth and quality are crucial and should be improved in future. The difference in these characteristics might explain the inferior performance (1) on label-free datasets compared to TMT datasets, and (2) in some tissues than others (Figure 6). Independent validation on more replicate datasets is necessary for mucosa (head-neck), ovarian and stomach tissues which have single dataset available until now.

Gene characteristics would also affect prediction performance. First, ubiquitination-modified proteins showed lower prediction performance, while SUMO-modified proteins showed higher prediction performance. The ubiquitin-proteasome pathway system would lead to protein degradation, and modulate the protein half-life. SUMOylation is a partner of ubiquitination in protein stability regulation [41]. Second, genes outside protein complexes were better predicted than genes belonging to protein complexes. Protein forming complexes were more likely coregulated post-transcriptionally [42]. Ribosomal genes, also complex proteins, showed worse prediction than other genes, probably because ribosome genes showed low or negative correlations between mRNA and protein level [5, 20]. Third, the prediction models were more effective in predicting tissue-specific genes, providing the possibility to do protein biomarker analysis in disease-related tissues. Last, essential genes showed lower prediction performance. Gene essentiality could offer insights into biology, clinical genetics and drug development [34, 43]. Our results suggested experimental measurements

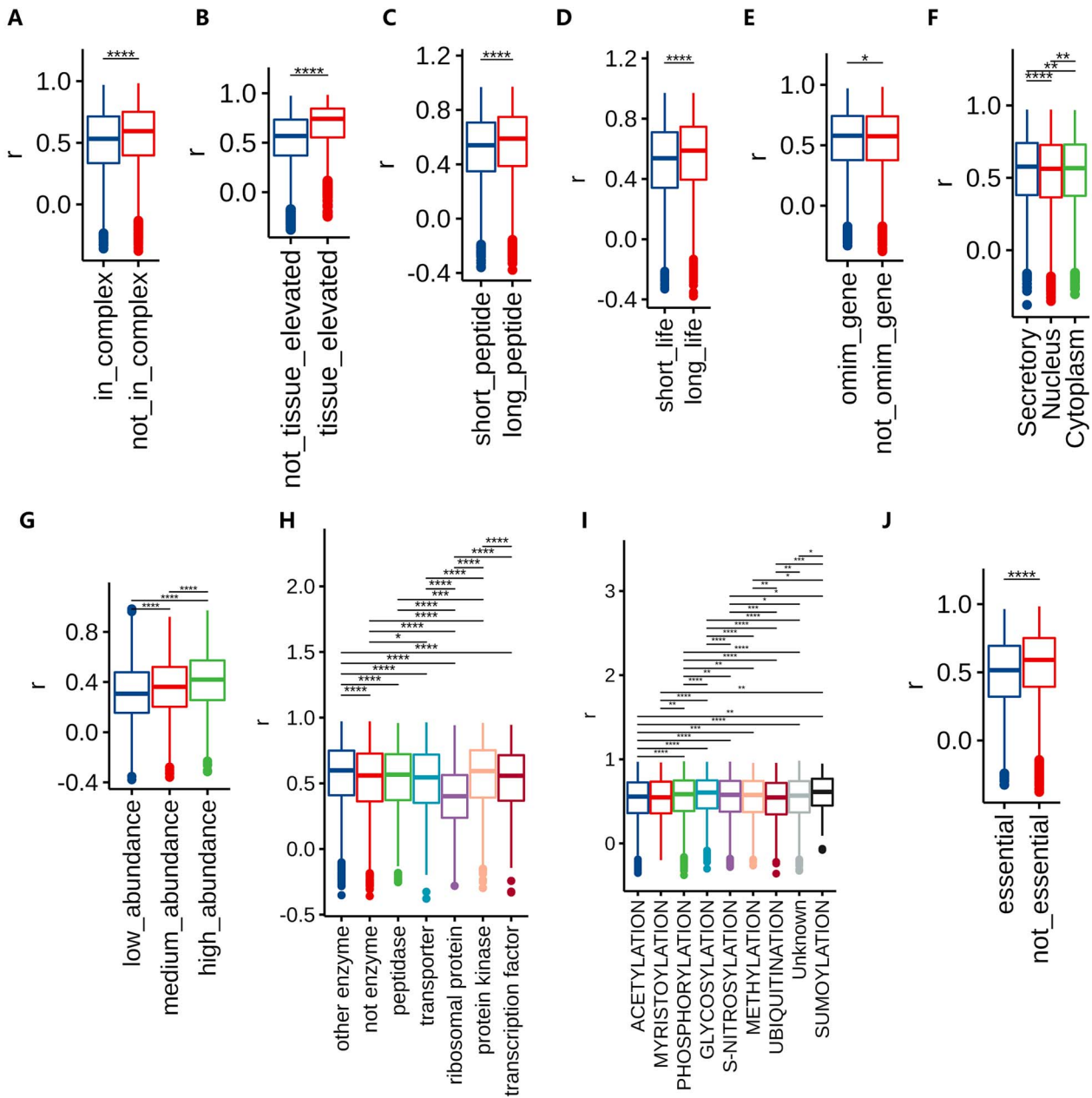


Figure 6. Gene characteristics affecting protein's predictability. Box plots show the distribution of performance metric r of gene groups side by side. Genes are grouped by (A) protein complex membership, (B) gene expression tissue specificity, (C) gene length, (D) protein half-life, (E) gene relation to human disease, (F) protein subcellular localization, (G) protein relative abundance compared to other genes, (H) protein functional class, (I) posttranslational modification and (J) gene essentiality.

were still necessary if essential genes are the key concerns.

Previous brain proteogenomic study of Carlyle et al. found that the protein data could amplify the cytoarchitectural and functional variation between brain regions than mRNA data [9]. Our previous proteogenomic study of the 29 Brodmann area in the human brain cerebral cortex also found that proteomic data could better reflect the functional parcellation of the brain cerebral cortex, such as Cg and OL [44]. In this study, we used the Allen transcriptome dataset of human brain to predict brain region proteome expression using machine learning models. The predicted proteome better reflected the

cytoarchitectural and functional characteristics of brain regions than input transcriptome, which was consistent with the findings of Carlyle and our previous studies. The prediction model showed good performances on the human brain transcriptomic dataset and might provide new insight on the molecular basis of brain functions.

Our work has several limitations. First, we handled missing values by removing all genes that contain NA values in datasets instead of using the same data imputation of Li's study [26]. We used minimal complete datasets only to reduce imputation-induced bias. Second, we didn't train a so-called pan-cancer model or trans-tissue model because the common genes between

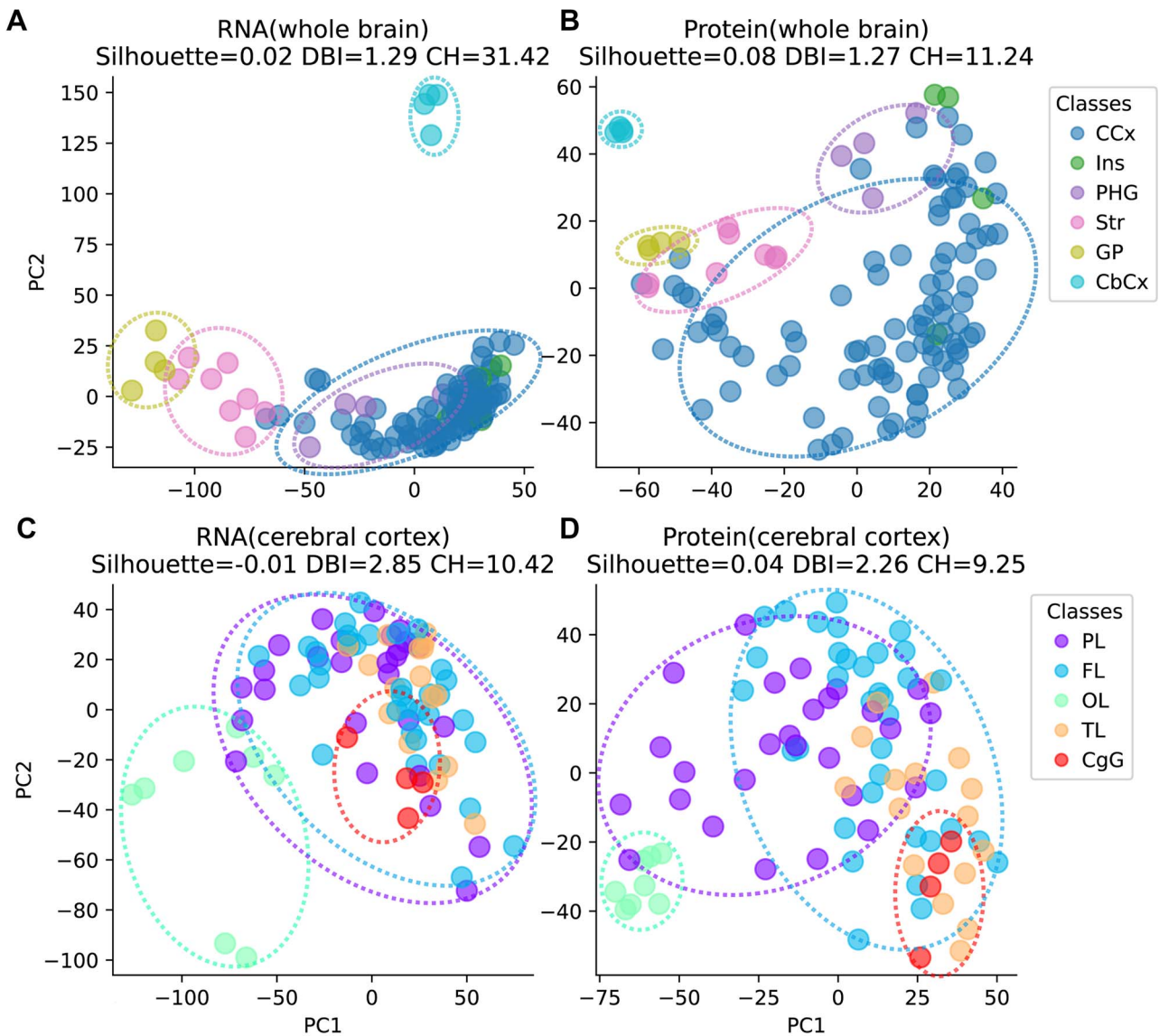


Figure 7. Predicted protein profiles recapitulate the brain structure. Row 1: PCA plots show the distance of samples computed from (A) RNA profiles (input, experimental data) and (B) protein profiles (output, computational inferred data) of whole brain. Row 2: PCA plots of (C) RNA profiles and (D) protein profiles of the cerebral cortex. Dots represent brain samples, and are colored by their brain region. Three clustering performance metrics are calculated (Supplementary Methods): Silhouette: Silhouette Coefficient (the best value is 1 and the worst value is -1), DBI, Davies-Bouldin score (lower values indicating better clustering) and CH, Calinski & Harabasz score (lower values indicating better clustering). CCx, Cerebral cortex; Ins, insula; PHG, parahippocampal gyrus; Str, striatum; GP, globus pallidus; CbCx, cerebellar cortex; PL, parietal lobe; FL, frontal lobe; OL, occipital lobe; TL, temporal lobe; CgG, cingulate gyrus.

datasets are very few. Third, protein–mRNA correlations in the original cohort publication of the datasets were not directly comparable to the values in our work because the preprocessing procedure was different. Fourth, it is worth noticing that RNA features from ACS-based feature selection were predictive variables of their target protein, but it does not necessarily indicate regulatory relationship between RNAs and target protein. Fifth, deep learning model and RFR have been used to predict 24 cell-surface protein abundances from transcriptome using single-cell multimodal omics data [45–47]. The sample size of benchmarking datasets is three orders of magnitude lower than single-cell datasets. It is probably why the deep learning models did not perform

well in this work. Finally, for ease of computation, we reimplemented ‘teamHL&YG’ model rather than using the original authors’ codes directly. The output of our reimplementations and original implementation were almost identical (Supplementary Table S9).

In summary, this benchmarking work on transcriptome-based proteome prediction provides useful hints on how to optimize and use the models. This work will also help researchers understand the inherent correlation between transcriptome and proteome. With computationally estimated but large-scale proteogenomic characterization of tumor samples, insights on patient stratification and prognosis can be gained by analyzing cancer biology at both proteome and transcriptome levels.

Data Availability

All raw data were retrieved from openly accessible source. Preprocessed data sets generated in this study are available as Supplementary Tables and EBI BioStudies [48] under accession S-BSST733. The source code is shared at <https://github.com/xuwenjian85/GeneExpressPredProtein>. Further materials are available upon reasonable request to the corresponding authors.

Key Points

- We improved previous transcriptome-based protein level prediction model by introducing feature selection.
- We curated the largest collection of benchmarking datasets from three mainstream proteomic platforms, benchmarked our improved model and proposed a new ensemble model with superior performance.
- We analyzed the influencing factors of predictability at gene level and dataset level.
- We applied the model to brain transcriptome of cerebral cortex regions and showed the inferred protein profiles better reflect the functional characteristics of brain regions than RNA expression profile.

Supplementary Data

Supplementary data are available online at <https://academic.oup.com/bib>.

Acknowledgements

We would like to thank P.W. and N.T. for providing data preprocessing details. Most datasets used in this work were generated by the CPTAC, TCGA, BrainSpan and CNHPP. We would like to thank the projects for making data freely available to the public.

Funding

This work was supported by grants from the Beijing Natural Science Foundation (5214023); National Natural Science Foundation of China (31400669; 31830054), the CAMS special basic research fund for central public research institutes (2017PT310004).

References

1. Liu Y, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. *Cell* 2016;**165**:535–50.
2. Lahtvee PJ, Sanchez BJ, Smialowska A, et al. Absolute quantification of protein and mRNA abundances demonstrate variability in gene-specific translation efficiency in yeast. *Cell Syst* 2017;**4**:495–504 e5.
3. Fortelny N, Overall CM, Pavlidis P, et al. Can we predict protein from mRNA levels? *Nature* 2017;**547**:E19–20.
4. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* 2012;**13**:227–32.
5. Zhang B, Wang J, Wang X, et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* 2014;**513**:382–7.
6. Sinha A, Huang V, Livingstone J, et al. The Proteogenomic landscape of curable prostate cancer. *Cancer Cell* 2019;**35**:414–427 e6.
7. Xu JY, Zhang C, Wang X, et al. Integrative proteomic characterization of human lung adenocarcinoma. *Cell* 2020;**182**:245–261 e17.
8. Jiang Y, Sun A, Zhao Y, et al. Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. *Nature* 2019;**567**:257–61.
9. Carlyle BC, Kitchen RR, Kanyo JE, et al. A multiregional proteomic survey of the postnatal human brain. *Nat Neurosci* 2017;**20**:1787–95.
10. Gao Q, Zhu H, Dong L, et al. Integrated proteogenomic characterization of HBV-related hepatocellular carcinoma. *Cell* 2019;**179**:1240.
11. Dou Y, Kawaler EA, Cui Zhou D, et al. Proteogenomic characterization of endometrial carcinoma. *Cell* 2020;**180**:729–748 e26.
12. Gillette MA, Satpathy S, Cao S, et al. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell* 2020;**182**:200–225 e35.
13. Chen YJ, Roumeliotis TI, Chang YH, et al. Proteogenomics of non-smoking lung cancer in East Asia delineates molecular signatures of pathogenesis and progression. *Cell* 2020;**182**:226–244 e17.
14. Vasaiakar S, Huang C, Wang X, et al. Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. *Cell* 2019;**177**:1035–1049 e19.
15. Clark DJ, Dhanasekaran SM, Petralia F, et al. Integrated proteogenomic characterization of clear cell renal cell carcinoma. *Cell* 2019;**179**:964–983 e31.
16. Wang L-B, Karpova A, Gritsenko MA, et al. Proteogenomic and metabolomic characterization of human glioblastoma. *Cancer Cell* 2021;**39**:509–528.e20.
17. Huang C, Chen L, Savage SR, et al. Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. *Cancer Cell* 2021;**39**:361–379.e16.
18. Petralia F, Tignor N, Reva B, et al. Integrated proteogenomic characterization across major histological types of Pediatric brain cancer. *Cell* 2020;**183**:1962–1985.e31.
19. Mertins P, Mani DR, Ruggles KV, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 2016;**534**:55–62.
20. Zhang H, Liu T, Zhang Z, et al. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell* 2016;**166**:755–65.
21. Mun DG, Bhin J, Kim S, et al. Proteogenomic characterization of human early-onset gastric cancer. *Cancer Cell* 2019;**35**:111–124 e10.
22. Satpathy S, Krug K, Jean Beltran PM, et al. A proteogenomic portrait of lung squamous cell carcinoma. *Cell* 2021;**184**:4348–4371.e40.
23. Cao L, Huang C, Cui Zhou D, et al. Proteogenomic characterization of pancreatic ductal adenocarcinoma. *Cell* 2021;**184**:5031–5052.e26.
24. Krug K, Jaehnig EJ, Satpathy S, et al. Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy. *Cell* 2020;**183**:1436–1456.e31.
25. Yang M, Petralia F, Li Z, et al. Community assessment of the predictability of cancer protein and phosphoprotein levels from genomics and transcriptomics. *Cell Syst* 2020;**11**:186–195.e9.
26. Li H, Siddiqui O, Zhang H, et al. Joint learning improves protein abundance prediction in cancers. *BMC Biol* 2019;**17**:107.

27. Xu W, Liu X, Leng F, et al. Blood-based multi-tissue gene expression inference with Bayesian ridge regression. *Bioinformatics* 2020;**36**:3788–94.
28. Abdelaal T, Michielsen L, Cats D, et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol* 2019;**20**:194.
29. Amberger JS, Bocchini CA, Scott AF, et al. OMIM.Org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res* 2019;**47**:D1038–43.
30. Uhlén M, Fagerberg L, Hallström BM, et al. Proteomics. Tissue-based map of the human proteome. *Science* 2015;**347**:1260419.
31. Huang H, Arighi CN, Ross KE, et al. iPTMnet: an integrated resource for protein post-translational modification network discovery. *Nucleic Acid Res* 2018;**46**:D542–50.
32. Zecha J, Meng C, Zolg DP, et al. Peptide level turnover measurements enable the study of Proteoform dynamics. *Mol Cell Proteomics* 2018;**17**:974–92.
33. Giurgiu M, Reinhard J, Brauner B, et al. CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acid Res* 2019;**47**:D559–63.
34. Bartha I, di Iulio J, Venter JC, et al. Human gene essentiality. *Nat Rev Genet* 2018;**19**:51–62.
35. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat Method* 2020;**17**:261–72.
36. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;**12**:2825–30.
37. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng* 2007;**9**:90–5.
38. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Houston, Texas, USA: Springer, 2016.
39. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* 2012;**489**:391–9.
40. Shen EH, Overly CC, Jones AR. The Allen human brain atlas: comprehensive gene expression mapping of the human brain. *Trends Neurosci* 2012;**35**:711–4.
41. Chang H-M, Yeh ETH. SUMO: from bench to bedside. *Physiol Rev* 2020;**100**:1599–619.
42. Gonçalves E, Fragoulis A, Garcia-Alonso L, et al. Widespread post-transcriptional attenuation of genomic copy-number variation in cancer. *Cell Syst* 2017;**5**:386–398.e4.
43. Chen H, Zhang Z, Jiang S, et al. New insights on human essential genes based on integrated analysis and the construction of the HEGIAP web-based platform. *Brief Bioinform* 2020;**21**:1397–410.
44. Guo Z, Shao C, Zhang Y, et al. A global multiregional proteomic map of the human cerebral cortex. *Genom Proteom Bioinform* 2021;**S1672–0229**(21):00225–4.
45. Zhou Z, Ye C, Wang J, et al. Surface protein imputation from single cell transcriptomes by deep neural networks. *Nat Commun* 2020;**11**:651.
46. Xu F, Wang S, Dai X, et al. Ensemble learning models that predict surface protein abundance from single-cell multimodal omics data. *Methods* 2021;**189**:65–73.
47. Dai X, Xu F, Wang S, et al. PIKE-R2P: protein-protein interaction network-based knowledge embedding with graph neural network for single-cell RNA to protein prediction. *BMC Bioinform* 2021;**22**:139.
48. Sarkans U, Gostev M, Athar A, et al. The BioStudies database-one stop shop for all data supporting a life sciences study. *Nucleic Acid Res* 2018;**46**:D1266–70.