OXFORD

Briefings in Bioinformatics, 2022, **23(2)**, 1–16
https://doi.org/10.1093/bib/bbac040
Problem Solving Protocol

POSREG: proteomic signature discovered by simultaneously optimizing its reproducibility and generalizability

Fengcheng Li†, Ying Zhou†, Ying Zhang, Jiayi Yin, Yunqing Qiu, Jianqing Gao and Feng Zhu

Corresponding author: Feng Zhu, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China. E-mail: zhufeng@zju.edu.cn; prof.zhufeng@gmail.com
†These authors contributed equally as co-first authors

Abstract

Mass spectrometry-based proteomic technique has become indispensable in current exploration of complex and dynamic biological processes. Instrument development has largely ensured the effective production of proteomic data, which necessitates commensurate advances in statistical framework to discover the optimal proteomic signature. Current framework mainly emphasizes the generalizability of the identified signature in predicting the independent data but neglects the reproducibility among signatures identified from independently repeated trials on different sub-dataset. These problems seriously restricted the wide application of the proteomic technique in molecular biology and other related directions. Thus, it is crucial to enable the generalizable and reproducible discovery of the proteomic signature with the subsequent indication of phenotype association. However, no such tool has been developed and available yet. Herein, an online tool, POSREG, was therefore constructed to identify the optimal signature for a set of proteomic data. It works by (i) identifying the proteomic signature of good reproducibility and aggregating them to ensemble feature ranking by ensemble learning, (ii) assessing the generalizability of ensemble feature ranking to acquire the optimal signature and (iii) indicating the phenotype association of discovered signature. POSREG is unique in its capacity of discovering the proteomic signature by simultaneously optimizing its reproducibility and generalizability. It is now accessible free of charge without any registration or login requirement at https://idrblab.org/posreg/

Keywords: feature selection, OMIC study, diagnostic accuracy, robustness, ensemble learning

Introduction

Proteomics based on mass spectrometry and other technologies is currently indispensable for researchers exploring complex dynamic biological processes [1–3]. The developments of relative instruments that underpin proteomics technology (such as data-independent acquisition) also go a long way to ensuring an effective production of proteomic data [4–7]. Therefore, commensurate advance in the statistical framework is necessitated for finding the sets of proteomic features that are truly significant in the biological process, which are so-called proteomic signatures [8, 9]. In such a context, feature selection (FS) emerged as a strategy for selecting key features and is playing an increasingly important role in the analysis of proteomic data [10]. A variety of FS methods have been developed and widely used in proteomics studies [11, 12] to train classifiers with better performance under given training sets, so that generalizability is widely regarded as a criterion to evaluate the performance of the selected signature [13, 14].

However, the current FS methods mainly emphasize the generalizability of the identified signature in predicting independent datasets [15] but neglect the reproducibility among signatures discovered from different sub-datasets [16]. Therefore, these current FS methods are usually sensitive to the perturbations in training datasets [17, 18], which leads to low overlap among signatures discovered from the different training sub-datasets generated from the same origin dataset and thus seriously restricted the extensive application of proteomics in molecular biology and other directions [19]. A practical

Ying Zhou is a PhD candidate of the Zhejiang University School of Medicine First Affiliated Hospital, China. She is interested in proteomics and bioinformatics.
Fengcheng Li, Ying Zhang and Jiayi Yin are PhD candidates of the College of Pharmaceutical Sciences in Zhejiang University, China. They are interested in bioinformatics.
Yunqing Qiu is a professor of First Affiliated Hospital in Zhejiang University, China. His research group is interested in precision medicine, diagnosis and treatment of liver disease and system biology.
Jianqing Gao is a professor of the College of Pharmaceutical Sciences in Zhejiang University, China. His research group has been working in drug delivery and molecular biology.
Feng Zhu is a professor at College of Pharmaceutical Sciences in Zhejiang University, China. His research laboratory (https://idrblab.org/) has been working in the fields of bioinformatics, OMIC-based drug discovery, system biology and medicinal chemistry.
Received: November 30, 2021. **Revised:** January 21, 2022. **Accepted:** January 27, 2022

proteomic signature should not only be generalizable but also reproducible [20–22], in other words, it should not only have good predictive performance in independent dataset but also should be stable regardless of the noise arising from measurement variability and biological differences [11]. To realize reproducible FS thus enhance the reliability and practicality of FS, reproducibility has thus been proposed as an equally important criterion as classification accuracy [23]. Moreover, the ensemble feature selection (Ensemble-FS) strategy has also been proven efficient in generating robust signature compared with typical FS methods [24–28]. This strategy is conducted by generating multiple signatures using different training sub-datasets (homogeneous) or FS methods (heterogeneous) and subsequently combing them into an ensemble signature [25, 29, 30]. Due to their capacities of enhancing FS reproducibility, the integration of Ensemble-FS and reproducibility evaluation is key for achieving better tradeoff between generalizability and reproducibility.

Currently, some powerful tools are available for biomarker analysis or FS (such as MetaboAnalyst [31] and MinE-RFE [32]), but the majority of them were developed only based on one single FS method and evaluated the FS solely on generalizability [31–35]. There is also one online tool called EFS that provides a heterogeneous ensemble of eight FS methods for binary classification studies and calculates the importance weight for each method in the ensemble [36]. However, the available tools do not provide any quantitative assessment for generalizability or reproducibility to demonstrate its superiority, nor does it provide any phenotype interpretation of the resulting signature. Therefore, it is essential to enable the generalizable and reproducible discovery of the proteomic signature with a subsequent indication of its phenotype association.

In our research, an online tool, POSREG, was constructed to identify the signature from a given set of proteomic data using comprehensive assessment from both generalizable and reproducible perspectives. This tool works by (a) identifying various signatures of good reproducibility based on their relative-weighted consistency (CW_{rel}) and aggregating them into the ensemble feature rank using ensemble learning; (b) assessing the generalizability of ensemble feature rank to acquire optimal signature by area under the curve (AUC)-based golden section search and (c) assisting users to indicate the phenotype association of the acquired optimal signature by providing gene ontology (GO) enrichment. With the increasingly accumulated concern about reproducibility [37] and phenotype association [37], the POSREG is unique for its capacity in comprehensively identifying optimal signature from both generalizable and reproducible perspectives and thus expected to be popular in proteomics and precision medicine [38–43]. The POSREG is accessible without login requirement at https://idrblab.org/posreg/

Materials and methods

Benchmark datasets collected and analyzed in this study

To evaluate the performance of POSREG and prove the superiority of its underlying algorithm, the proteomics datasets available in the PRIDE database [44], ProteomeXchange [45] and iProx [46] were fully reviewed. Seven benchmark proteomics datasets with at least 20 samples from PRIDE [44] were finally collected and further analyzed as case studies in this work according to the following criteria: (1) datasets should be comparative proteomic studies; (2) datasets should encompass a broad biological research orientation; (3) the sample size for each group (control and case) in a study should be six at least and the sample size should be at least 20. These benchmarks were labeled as PXD000672, PXD002882, PXD003972, PXD004880, PXD005144, PXD006129 and PXD008840. The detailed descriptions of them are established in Table 1. To facilitate the directly using these benchmark datasets to conduct their analysis without pretreatment, pretreated benchmark datasets PXD005144 and PXD003972 are provided in Supplemental Information [47–49].

FS methods employed and analyzed in this study

FS methods are commonly categorized into filter, wrapper and embedded types [50]. The filter methods only pick up the intrinsic characteristics of the features, whereas the wrapper and embedded methods iteratively consider the classification performance of the features in specific models [51]. Although the wrapper and embedded methods are supposed to give better performance than the filter, the filtering methods are usually faster for calculation and the resulting signatures are also more universally applicable to different machine learning models [52, 53]. Moreover, a set of features with significance ranking (output of filter method) is more suitable for ensemble learning than a feature subset with no priority (output of the wrapper and embedded methods) [50]. In summary, the filter method is suitable for ensemble learning based on onerous repeating computation and is therefore adopted in POSREG.

To make POSREG applicable to most common situations, nine different filter FS methods based on varied feature searching and scoring theories were employed and analyzed in POSREG, which contained univariate filter methods (fold change analysis, Wilcoxon rank-sum test, etc.) and multivariate filter methods (correlation-based method, entropy-based filter, etc.). The categories and the brief introductions of these FS methods are demonstrated in Table 2. Furthermore, the detailed description of these nine FS methods which depicted their requirement of data distribution and structure are provided in Supplementary Method S1.

Table 1. Seven benchmark proteomics datasets were collected and analyzed in this study.

Dataset ID	References	Data acquisition	No. of features	Description of samples		
PXD000672	*Nat Med*. 21:407-13, 2015	DIA	3132	12 renal cell carcinoma samples from 6 patients	versus	12 healthy samples from 6 individuals
PXD002882	*Nat Commun*. 7:13419, 2016	DDA	4169	21 samples from Crohn's disease patients	versus	10 samples from healthy individuals
PXD003972	*Cell Rep*. 18:3219-3226, 2017	DIA	901	20 samples from 4 GRB2OST knock-in mice	versus	20 samples from 4 different GRB2WT mice
PXD004880	*Sci Rep*. 7:14818, 2017	DIA	5540	18 samples from Down syndrome patients	versus	18 samples from healthy individuals
PXD005144	*Cancer Med*. 6:1738-1751, 2017	DDA	653	66 tumor samples from 22 pancreatic cancer patients	versus	36 samples from 12 pancreatitis patients
PXD006129	*Cell Host Microbe*. 23:27-40, 2018	DDA	3243	15 samples from western-style diet-fed mice	versus	14 samples from chow diet-fed mice
PXD008840	*Nat Commun*. 9:1012, 2018	DDA	5439	84 tumor samples from gastric cancer patients	versus	84 normal tissues from the gastric cancer patients

FA: formaldehyde; $GRB2^{OST}$: GRB2 tagged with a One-STrEP-tag (OST); $GRB2^{WT}$: wild type GRB2. The dataset ID starting with PXD or IPX indicated that the corresponding dataset was collected from the Proteomics Identification Database (PRIDE) [44] or integrated Proteome resources (iProx) [46].

Table 2. Brief introduction of FS methods employed and analyzed in this study

FS method (Abbreviation)	Type	Brief introduction
CFS	Multivariate filter	Evaluate feature subset based on the prediction ability of each feature in it and the correlation between them [126].
Entropy-based filters (ENTROPY)	Multivariate filter	Select features based on the contribution of information related to class variables. Compensate for information gain bias [127].
FC	Univariate filter	Select features that have large differences between the control and case groups. Calculate FC by the ratio of mean intensities of proteins between the two groups [128].
LMEB	Univariate filter	Evaluate the differential abundance of features by drawing a volcano plot, which measures the differentially accumulated features based on fold changes and *t* statistics [129].
PLS-DA	Multivariate filter	Predict variables that maximize differences among predetermined samples. Infer classification of unclassified sample groups based on the calibration set with known class distribution [130].
ReliefF (REF)	Multivariate filter	Estimate attributes based on the degree of value differentiation between near instances [131].
Significance analysis of microarrays (SAM)	Univariate filter	Score each gene based on the change in gene expression relative to the standard deviation of repeated measurements [132].
Univariate *t*-test (*t*-test)	Univariate Filter	Rank features based on *P*-values. Features with a *P*-value <0.05 are considered to be significant [20].
Wilcoxon Rank-sum test (Wilcox)	Univariate Filter	Use magnitude-based ranks to establish the significant difference between the two groups. The significant difference shows when the ranks of the two groups are significantly separated [133].

Metrics used to grope and evaluate the optimal signature

POSREG comprehensively used two types of well-established metrics in the process of identifying proteomic signatures optimal in both terms of reproducibility and generalizability.

Metrics Type I. Reproducibility of Multiple Signatures Identified from Different Data Subsets.

Experts working on the discovery of predictive proteomic biomarkers have always been plagued by the difficulty of reproducing their research results, even with the same input dataset and FS method, which directly constrained the practicability and reliability of their identified biomarkers [54–57]. To increase the confidence of domain experts in their research findings and identified biomarkers, reproducibility has thus become an equally important criterion as diagnostic accuracy [23, 58, 59]. A series of metrics based on the distinct underlying theory, including Jaccard's index [60], Percentage of overlapping Gene [61], Pearson's correlation coefficient [62], Weighted Consistency [63] and so on, have thus been proposed for reproducibility evaluation. Nevertheless, most of them are susceptible to the size of feature subsets so that they are unsuitable for the reproducibility evaluation and comparison of feature subsets with different sizes.

The CW_{rel} [63] was proposed based on weighted consistency, it is calculated based on multiple signatures, it counts the occurrence times of each feature in every single set of signatures and the total occurrence times of all features in all signatures, then uses the specific ratio of these two to represent the overall robustness [28, 55, 64–66]. The detailed description of its statistical calculation is further demonstrated in Supplementary Method S2. CW_{rel} satisfied the property of randomness correction and is thus empowered to avoid the 'subset-size-bias problem' [63]. Therefore, POSREG introduced the CW_{rel} to compare the reproducibility among signatures with different sizes of feature subsets in the real-time process of optimal signature discovery.

Metrics Type II. Diagnostic Accuracy of Classification Model Built on Identified Signatures.

The prime goal of FS is to identify a series of truly significant markers, which could be employed to describe the biological differences [67]. This prime goal demands the identified markers to be generally applicable to the data not involved in FS, which is generally called the generalizability of FS [68]. And the major way of validating the generalizability of identified markers is to evaluate the diagnostic accuracy of classification models built on these markers in an independent dataset [69–71]. Therefore, the receiver operating characteristic (ROC) analysis and the AUC metrics were introduced in POSREG to assess the diagnostic accuracy of the classifier constructed based on the identified signature.

The generalizability of FS method was assessed by a 5-fold nested cross-validation (CV) using the following steps. First, the original data were split into 5-fold, each fold was iteratively selected as a test set. Second, for each outer iteration, the remaining data were further split into 4-fold, each fold was iteratively selected as a validation set and the left folds were training set. Third, for each inner iteration, the training set was adopted for FS and model training with different parameters, and the validation set was used to assess the quality of this model. Fourth, the best model of each inner loop was selected and was evaluated on the test set of each outer loop by the AUC value calculated using the ROC and AUC function in the *R* packages pROC [72]. Finally, the generalizability is calculated by averaging the AUC values of all 5-fold of the outer loop.

Nested CV ensuring the unbiased assessment of generalizability

CV is a well-established technique for assessing the generalizability of FS [73]. This technique divided the dataset into n parts, picked out one part to assess the generalizability, and the left n-1 parts were used to perform FS and build a classifier [73]. This process would be repeated n times until every part had ever been used for assessing generalizability [73]. The final generalizability is the average of all folds [73]. However, due to the extensive experimental costs and serious technique limitations [74], the 'small sample size' problem was reported to be one of the bottlenecks in current proteomic studies, which were typically <100 [75]. The performance estimations of the 'small sample size' study by n-fold CV were reported to be overoptimistic due to the excessive variances and biased results [76, 77]. So far, several strategies are developed to address the small sample size problem [76, 77]. Among them, a strategy named 'nested CV' was proposed as an effective way of giving unbiased performance estimations for the dataset of not only large but also relatively small sample size, which was thus adopted in this study to assess the generalizability [78, 79]. According to the original publications of 'nested CV' [78, 79], the studies of ≥20 samples can achieve unbiased performance estimation. Thus, it is recommended to analyze the dataset of sufficient samples (≥20) by POSREG, and the analytical results for the dataset of fewer than 20 samples should be considered with caution.

The nested CV also split the datasets into n folds and one portion of data was iteratively picked out as a test set for generalizability assessment, which formed the outer loop. The difference with typical n-fold CV is that the remaining n-1 parts are further iteratively split into training set and validation set for FS and parameter tuning, which formed the inner loop. For iterations of the inner loops, new models developed on different feature sets and parameters were validated by validation sets, and the best performed model will be selected to be evaluated on the test set of each outer loop. The final generalizability is the average of all n estimates in outer loops [78, 79]. As shown in Figure 1, this strategy first set apart a test set (*Test 1*) before training the models, and left this test set not being used in modeling and FS. Second, the remaining data (*Remaining 1*) were iteratively split into Train and Validation sets for training and validating the model. Third, the performances of model construction were assessed based on the Test 1 data that were set apart at the beginning. Finally, the above processes were repeated by another four times via setting apart four additional test sets (*Test n*, n = 2,3,4,5), and those five test sets were independent among each other (without overlap among any test sets). All in all, as shown in Figure 1, both modeling and FS were integrated into the CV process, and the testing sets were not used during this process.

Ensemble-FS for aggregating multiple proteomic signatures

Ensemble learning was proposed based on the proverb 'two heads are better than one', which combined multiple models to obtain better performance than a single one [80]. Ensemble learning was initially popular only among classifications and has gradually been found to be also efficient for improving other machine learning disciplines, such as FS [81]. Ensemble-FS combined the output of several feature selectors (generated from different methods or training datasets) to form an ensemble feature ranking. The place of each feature in ensemble ranking is jointly determined by its previous rankings

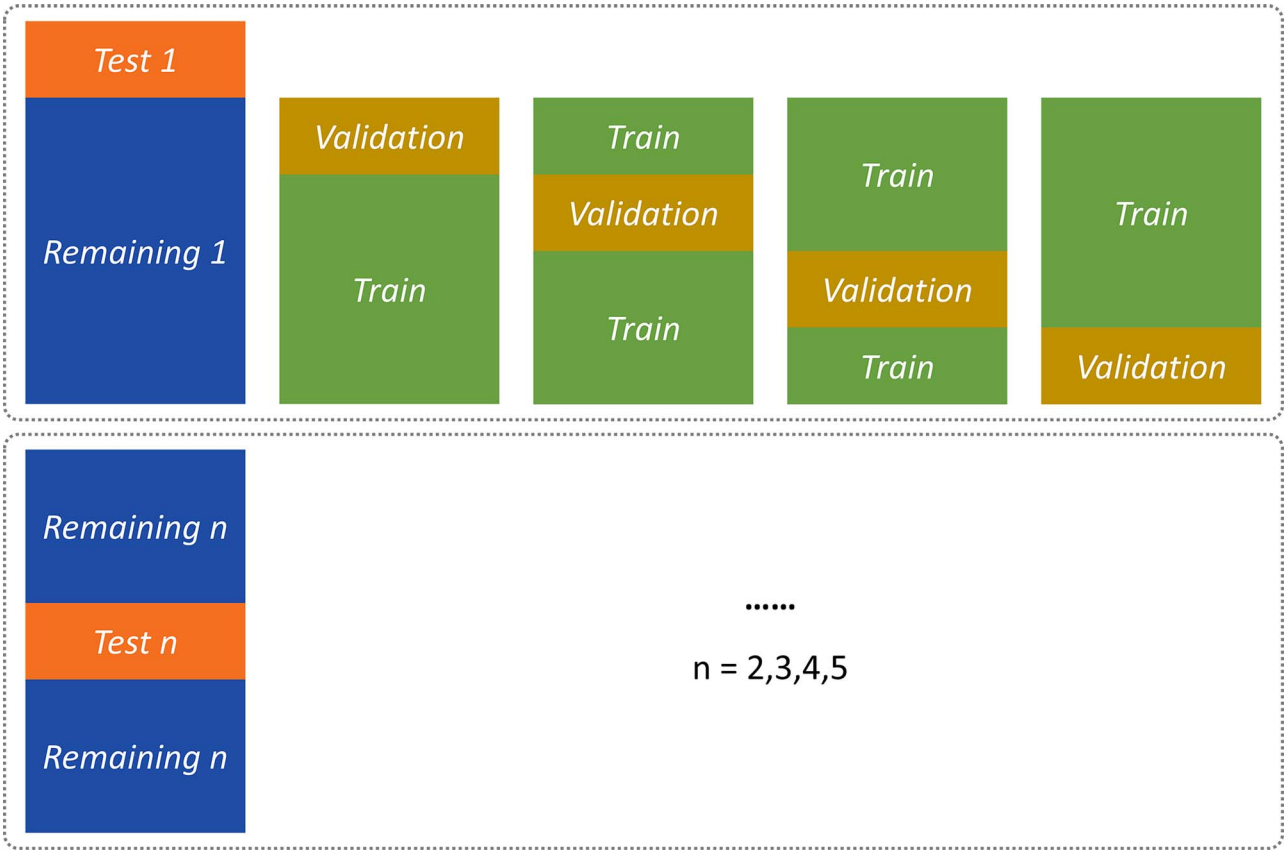

Figure 1. Schematic diagram of the nested CV.

[81]. Ensemble learning was introduced to aggregate multiple proteomic signatures generated in the process of reproducibility evaluation into ensemble feature ranking. The homogeneously distributed ensemble, which integrated multiple feature list generated using the same FS method and different training datasets, is provided in POSREG. Six ensemble methods including arithmetic mean, geometric mean, median, min, robust rank aggregation (RRA) and Stuart were provided using aggregateRanks function in *R* packages RobustRankAggreg [82].

AUC-based search for acquiring optimal signature of high accuracy

Given that the performance of the classifier is strongly influenced by the number of features [83], how many features should be added to the training set to achieve the most accurate classifier is a frequently encountered issue [84]. Due to the constrained computation resources, it is impractical to assess the accuracy of every possible combination of proteomic features using the exhaustive method [84]. Thus, Liu's group proposed an iterative golden-section search method based on 5-fold AUC to approximate the optimal size of features of high accuracy [85]. The golden-section search algorithm is a classic algorithm for finding the extreme of a single-variable function, the rationale behind this approach is to successively narrow down the range of search intervals inside which the extremum is believed to exist [86]. Supposed that AUC is the function of feature size, then this function could only be a unimodal function (which has a single optimum in the domain of definition) or a monotonically increasing function (in which the dependent variable increases with the independent variable in the domain of definition) if the features were added into signature in order of feature significance ranking [85]. Under such supposition, the golden-section search algorithm could be adopted to find the optimal signature with the highest AUC [85].

POSREG used the basic idea of golden-section search, optimized algorithm and implemented it in *R* language. The AUC-based golden-section search was conducted after the generation of ensemble feature ranking, it iteratively selected feature subsets with different sizes according to the golden ratio to build classifiers and evaluate AUC separately, then continuously narrowed the range of possible feature size based on AUC value and finally finds the optimal signature with the highest AUC. The detailed description of the AUC-based golden-section search for acquiring optimal signature with the highest accuracy is further demonstrated in Supplementary Method S3.

Phenotype association indication based on signature enrichment analysis

Proteomic signature determined in a proteomic study should be directly related to the phenotype (preferably

as upstream as possible) and plays a real role in the phenotype as opposed to merely being correlated [69]. GO resource provides computable knowledge about the function of gene and gene products and is extensively adopted for the analysis of omics-related data [87–91]. To help users intuitively understand the phenotype association of acquired proteomic signatures, enrichment analysis of selected proteomic signatures can be performed in POSREG using the enrichGO function of the *R* package clusterProfiler [92].

To measure the level of phenotype association, all features in the identified proteomic signature are first enriched based on their involved biological process, cellular component, molecular function or all terms. Then, a bubble chart displaying top30 GO terms with the least *P*-values was plotted by *R* package ggplot2 [93] to better visualization of the enrichment result [94–97]. Finally, the users can relate these enriched terms to their studied phenotype, and therefore comprehend the relevance of the identified features to their studied phenotype. On the one hand, signature enrichment analysis in POSREG could be instructive for studies where phenotypic relationships are still unclear. On the other hand, this additional function could provide some bidirectional validation for researches with established phenotype association.

Webserver implementation and requirements for input file format

POSREG is developed based on the operating system of CentOS Linux v7.4.1708, which is configured with Apache HTTP web server v2.4.6 and Apache Tomcat servlet container. The main web interface is constructed using the *R* package shiny v0.13.1 and the webserver was deployed on the shiny-server v1.4.1.759 (*R* v3.4.1). Apart from Shiny package, POSREG also employed various other *R* packages in the background processing, including affy [98], clusterProfiler [92], coin [99], DOSE [100], ggplot2 [93], mixOmics [101], pathview [102], pcaMethods [103], pROC [72], RobustRankAggreg [82] and so on. Both the official and mirror sites of POSREG are accessible to all users without any login requirement by most commonly used web browsers including Google Chrome, Mozilla Firefox, Safari, Microsoft Edge and Internet Explorer (10 or later). And the source code of POSREG enabling the assessment on a local computer is also provided, users only need to configure the *R* environment, RStudio software and install the corresponding *R* packages using the packages we provide.

POSREG is capable to handle datasets in commonly used formats including txt, xlsx, tab-delimited and csv. The row of the input file should be samples and the column of the input file should be features, respectively. In particular, the first row of the input file should be the feature name, the first column must be the sample name and the second column indicates the class label (case or control) of each sample, it is important to note that the name of the second column must be exactly 'Class'. In addition, if the user needs to perform pathway enrichment, the feature name of the input file must be annotated to UniProt ID or ENTREZID.

Results and discussion

Validating the feasibility of using CWrel for FS reproducibility evaluation

Researchers dedicated to biomarkers discovery have always been focused on discovering efficient signatures that can precisely reflect the biological difference [52] but ignored the reproducibility of their proposed signatures [104]. This leads to the problem of low reproducibility of signatures proposed by different research groups for the same research issue, even though they all achieved good prediction performance [55]. To ensure the stability of identified features and ultimately enhance their practicality, the metrics CW_{rel} was applied to assess the reproducibility in the pipeline of POSREG.

As demonstrated in Materials and Methods and Supplementary Methods S1, CW_{rel}'s unique trait of avoiding the 'subset-size-bias problem' gives it the ability to compare the robustness between proteomic signatures of different feature sizes [63]. Therefore, it is feasible to use CW_{rel} as a reproducibility assessment metric to find the most robust feature size. To comprehensively validate the feasibility of using CW_{rel} for reproducibility assessment, the benchmark dataset PXD000672 [105] was analyzed as an example. For each of the nine FS methods of POSREG, (i) firstly 50 sub-datasets were randomly selected from the benchmark dataset using stratified sampling with put-back, the sampling process is sample-wise and half of the samples from the control group and case group was randomly selected each time; (ii) and then these sub-datasets were analyzed using this particular FS method to generate 50 feature rankings; (iii) after that numerous of feature subsets with different feature sizes from top 1% to top 50% of total feature amount were divided from these feature rankings; (iv) lastly, the feature subsets with same feature size were collected to calculate the CW_{rel} value under particular feature size. These four preceding steps make up one single independent replicated trial which can illustrate the trend of CW_{rel} with the proportion of selected features. It is worth mentioning that although the random sampling method adopted here was stratified sampling with put-back, a conditional statement was set to ensure the difference among different sub-datasets (with at least one distinct sample in both control and case groups). In other words, based on the random stratified sampling and conditional statement, it was guaranteed that those sampled sub-datasets were different from each other, which could thus be adopted to assess the CW_{rel}.

As illustrated in Figure 2, the aforementioned independent replicated trial was repeated 50 times under each of all nine FS methods in POSREG. On the one hand, these repeats turned out to have broadly consistent trends in

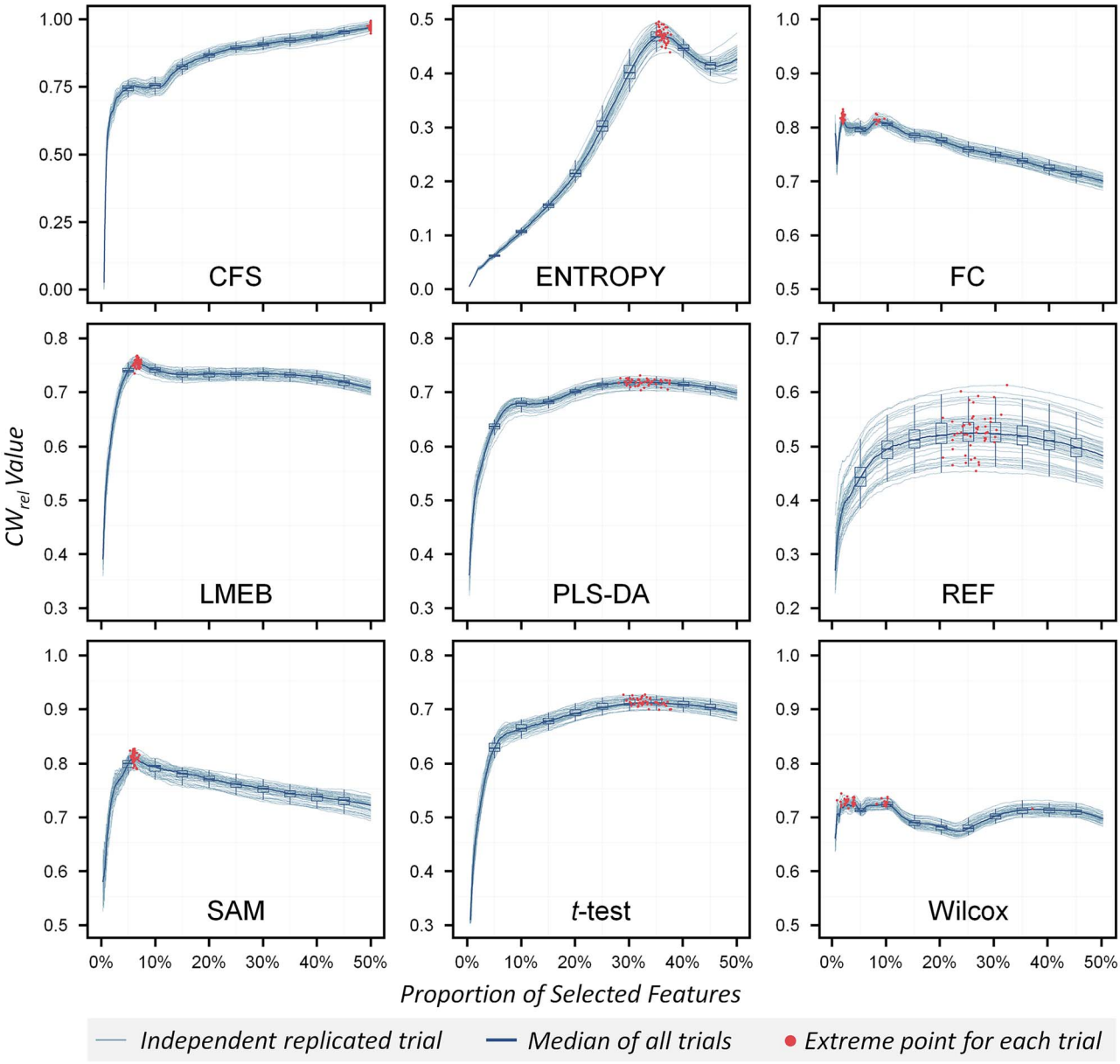

Figure 2. The level of stability of CW_{rel}-based FS reproducibility evaluation was assessed by comparing the trends in CW_{rel} between different independent replicated trials in benchmark dataset PXD000672 [105]. The x-axis in each sub-figure denoted the proportion of features selected into CW_{rel} calculation, and the y-axis denoted the value of CW_{rel}. The trend in CW_{rel} of each independent replicated trial was represented by a thin light-blue line and the median value of all replicated trials was connected and drawn as a thick dark-blue line. The maximum points of all 50 repeats were marked red dots.

CWrel (denoted by thin light-blue solid lines) and the maximum points of these trials (denoted by red solid dots) always occurred at the close coordinates for one particular FS method. On the other hand, the curves of CW_{rel} for different methods tend to trend differently and their maximum points occurred in different coordinates. These results indicate that CW_{rel} is a stable and robust metric for FS reproducibility evaluation and the CW_{rel}-based FS reproducibility evaluation is required for choosing appropriate methods and feature subset sizes because the reproducibility varies between FS methods and the proportion of selected features.

The same analyses were carried on the other five benchmark datasets in Figure 3. Particularly, these benchmarks were also analyzed by all nine FS methods in POSREG, and the resulting maximum points of their 50 times independent replicated trials are denoted by small hollow dots in different colors. On the one hand, as shown in Figure 3, under the same FS method, the proportion of selected features where CW_{rel} reaches its maximum varies widely across datasets. Take the Wilcoxon Rank-sum Test (Wilcox) method as an example (whose maximum points are denoted by pink hollow dots). It was found in the results of some benchmarks (PXD000672, PXD003972, PXD005144) that CW_{rel} reached the maximum value with a very small proportion of selected features, whereas the other benchmarks (PXD002882, PXD006129, PXD008840) reached the maximum CW_{rel} with a medium proportion of feature. On the other hand, the reproducibility of the

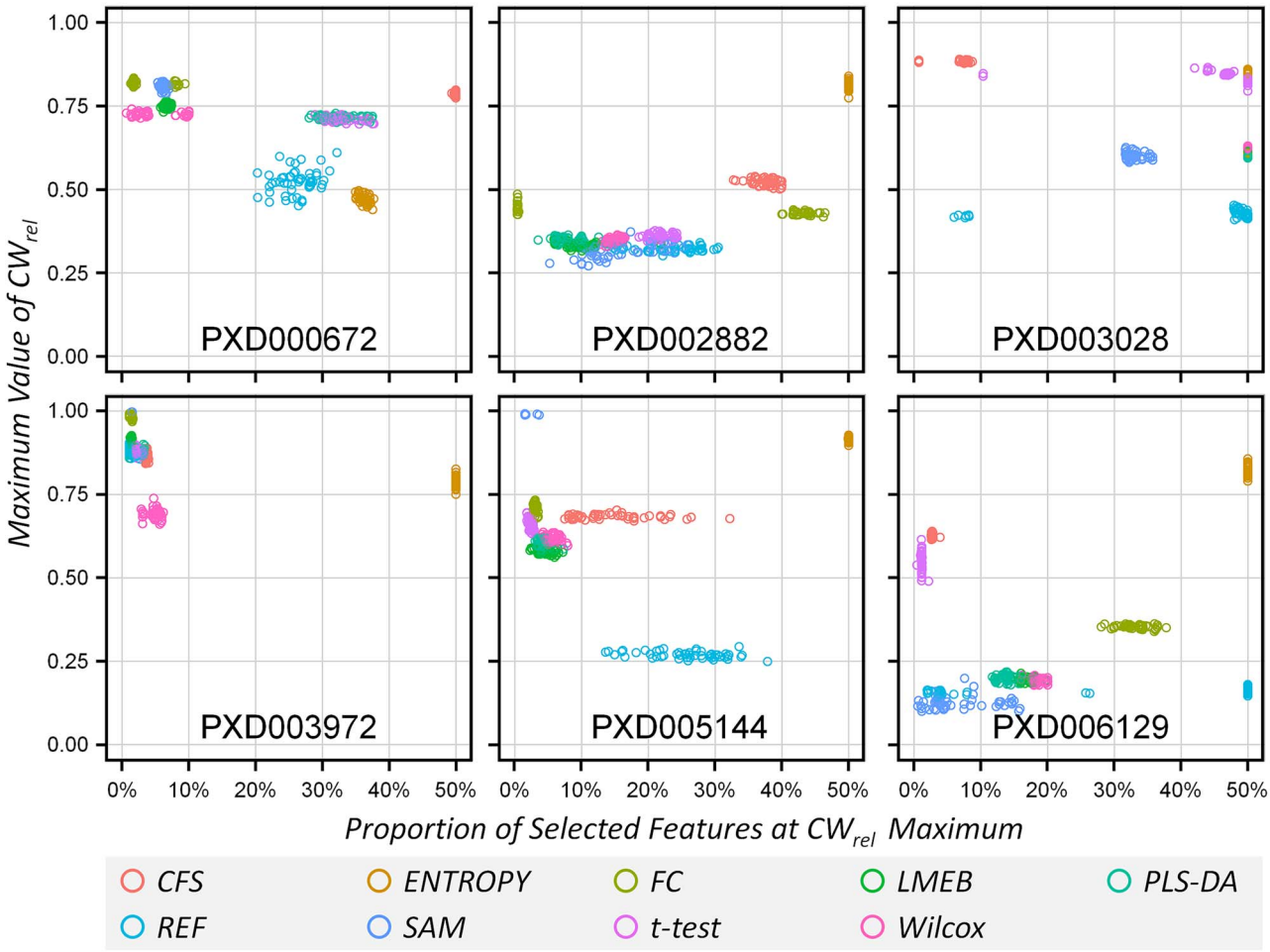

Figure 3. The scatter plot for maximum points of the CW_{rel}-proportion of selected feature curve generated from CW_{rel}-based FS reproducibility evaluation for six benchmark datasets under nine FS methods. The x-axis in each sub-figure denoted the proportion of features selected where CW_{rel} reaches its maximum value, and the y-axis denoted the maximum value of CW_{rel}. All the maximum points were denoted by hollow dots, whereas different colors showed different methods.

same FS method also varied considerably across different datasets. Again taking the Wilcox method as an example, if we use 0.5 as a cutoff, the maximum value of CW_{rel} is regarded as high in the results of some benchmarks (PXD000672, PXD003972, PXD005144, PXD008840) and low in the results of other benchmarks (PXD002882, PXD006129). Therefore, the optimal FS method in terms of reproducibility varies from data to data, and it is essential to perform a CWrel-based FS reproducibility evaluation for choosing an appropriate FS method.

Enhancing the reproducibility by maximizing CW_{rel} and ensemble learning

Denote the proportion of selected features when CW_{rel} reaches its maximum value with $PSF_{\max}(CW_{rel})$, it is not hard to discover from Figures 2 and 3 that $PSF_{\max}(CW_{rel})$ is an intermediate size in most cases. That is to say, for the stability of FS, the number of features is not the less the better, nor the more the better [106]. This indicated that a specific FS method has a clear limit of power of recognizing features [107], which means it considers all of the top $PSF_{\max}(CW_{rel})$ features to be entirely significant for classification when dealing with a specific set of data [108]. If the FS method was adopted to select a fewer proportion of features than$PSF_{\max}(CW_{rel})$, the stability of FS method would decrease because of the difficulty in choosing between the top $PSF_{\max}(CW_{rel})$ significant features. And vice versa, if the proportion of features need to be selected is more than$PSF_{\max}(CW_{rel})$, the stability of FS will also be reduced because the excess part will be randomly selected among these redundant features that are considered unimportant. Therefore, FS reproducibility could be enhanced if top $PSF_{\max}(CW_{rel})$ features are adopted for downstream analysis to maximize CW_{rel}.

Maximizing the CW_{rel} value could only be used to determine the most stable feature subset size under a specific FS method [63]. Nevertheless, even at the most stable feature subset size, there will still be some differences in the multiple feature subsets picked out using the same FS method multiple times [85]. Therefore, how to comprehensively consider the different rankings of each feature in multiple feature subsets and derive a conclusive feature ranking from them was a problem [17]. Ensemble-FS has been proposed as a solution for

the aforementioned problem because of its ability to combine the output of multiple feature selectors into a more stable and efficient ensemble feature ranking [29]. Therefore, ensemble learning is introduced in POSREG to aggregate the top $PSF_{\max}(CW_{rel})$ features of multiple feature rankings generated during CW_{rel} calculation to form an ensemble feature ranking, which further enhanced the FS reproducibility on the basis of maximizing the CW_{rel}.

Determining the optimal signature by AUC-based golden section search

The ensemble feature ranking is generated with enhanced reproducibility by maximizing CW_{rel} and ensemble learning. However, the generalizability of the resulting ensemble feature ranking has not been assessed yet. Moreover, under some circumstances, the overall CW_{rel} is significantly low [109], so that even if the top $PSF_{\max}(CW_{rel})$ features are selected for ensemble, the resulting ensemble feature rank will potentially contain too many features [110]. As the performance of the classifier is strongly influenced by the number of features [83], the classifier built on too many features is insufficient and impractical [111]. Therefore, a rapid and sufficient generalizability assessment to determine the optimal proteomic signature based on the ensemble feature ranking was embedded into the POSREG workflow, which is the AUC-based golden section search. As discussed in Materials and Methods and Supplementary Method S4, the AUC-based golden section search can find the maximum of the single variable function (accuracy against feature size) with iteratively narrowing searching range [85], so that POSREG can not only assess the generalizability of selected features but also control the number of features to some extent through the procedure of AUC-based golden section search.

The capacity of POSREG in improving reproducibility and generalizability

To verify the superiority of POSREG in both reproducibility and generalizability perspectives, two benchmarks PXD005144 [112] and PXD008840 [113] were collected and assessed by both POSREG workflow and traditional FS workflows for comparison. As for the traditional FS, the most common way is to directly choose the top 50 or top 100 features to form the final signature, whereas some researchers choose to use the top 5% or top 10% of the total feature [114]. Thus, the POSREG workflow is compared with traditional FS workflows top 50, top 100, top 5% and top 10% simultaneously under two types of FS methods: univariate filter methods [represented by fold change (FC) and linear models and empirical Bayes (LMEB)] and multivariate filter methods [represented by correlation-based feature selection (CFS) and partial least squares discriminant analysis (PLS-DA)]. Each workflow was repeated 50 times with different samples produced by the bootstrap sampling and the comparison was based on the mean value of 50 repetitions to avoid the serendipity.

As shown in Figure 4**a**, the reproducibility was assessed by the mean value of CW_{rel} and drawn with an orange bar at the upper of mirrored bar plot, whereas the generalizability was assessed by the mean value of AUC and drawn with a blue bar at the lower of mirrored bar plot. POSREG workflow achieved higher performance in reproducibility and generalizability in most cases. This result is further corroborated by Figure 4**b** as the violin plot of POSREG is more concentrated than the other four groups and its median value is also in a higher position. To conclude, POSREG performed better in both CW_{rel} and AUC than traditional methods, which verified its capacity in improving both reproducibility and generalizability of filter FS methods.

Comparing POSREG with established wrapper and embedded FS techniques

POSREG workflow has demonstrated better reproducibility and generalizability than traditional filter methods in Figure 4, but its superiority or inferiority to the wrapper and embedded method is not yet known, and it is thus necessary to compare POSREG with established wrapper and embedded FS techniques. Random forest-recursive feature elimination (RF-RFE) [115] and least absolute shrinkage and selection operator (LASSO) [116] were chosen as representatives for the comparison with POSREG as well-established and common used embedded and wrapper FS techniques, respectively. Four benchmark datasets PXD003972 [117], PXD004880 [118], PXD005144 [112] and PXD008840 [113] were collected and assessed using RF-RFE, LASSO and POSREG workflow 50 times to generate 50 different signatures. Then the CWrel and AUC metrics of these signatures were calculated to represent the reproducibility and generalizability of the corresponding FS method. The assessing result is illustrated in Figure 5. As shown in Figure 5, the AUC values for all three methods are fairly good, but POSREG methods achieve higher CW_{rel} while getting comparable AUC values as RF-RFE and LASSO. From this, it can be seen that POSREG workflow has equivalent generalizability and superior reproducibility comparison to the established wrapper and embedded FS techniques.

Demonstrating the superiority of POSREG with the case study on PXD005144

To better demonstrate the superiority of POSREG, we compared the obtained biological finding for PXD005144 using POSREG and that from the corresponding published paper [112]. In the original publication of PXD005144, significantly different proteins between two groups were identified using three FS methods, and all these proteins were further compared with each other to get a list of 20 proteins that are common to all three methods [112].

To give a comparison between our results and the results provided in the original publication of PXD005144, the following steps were conducted: (a) FS

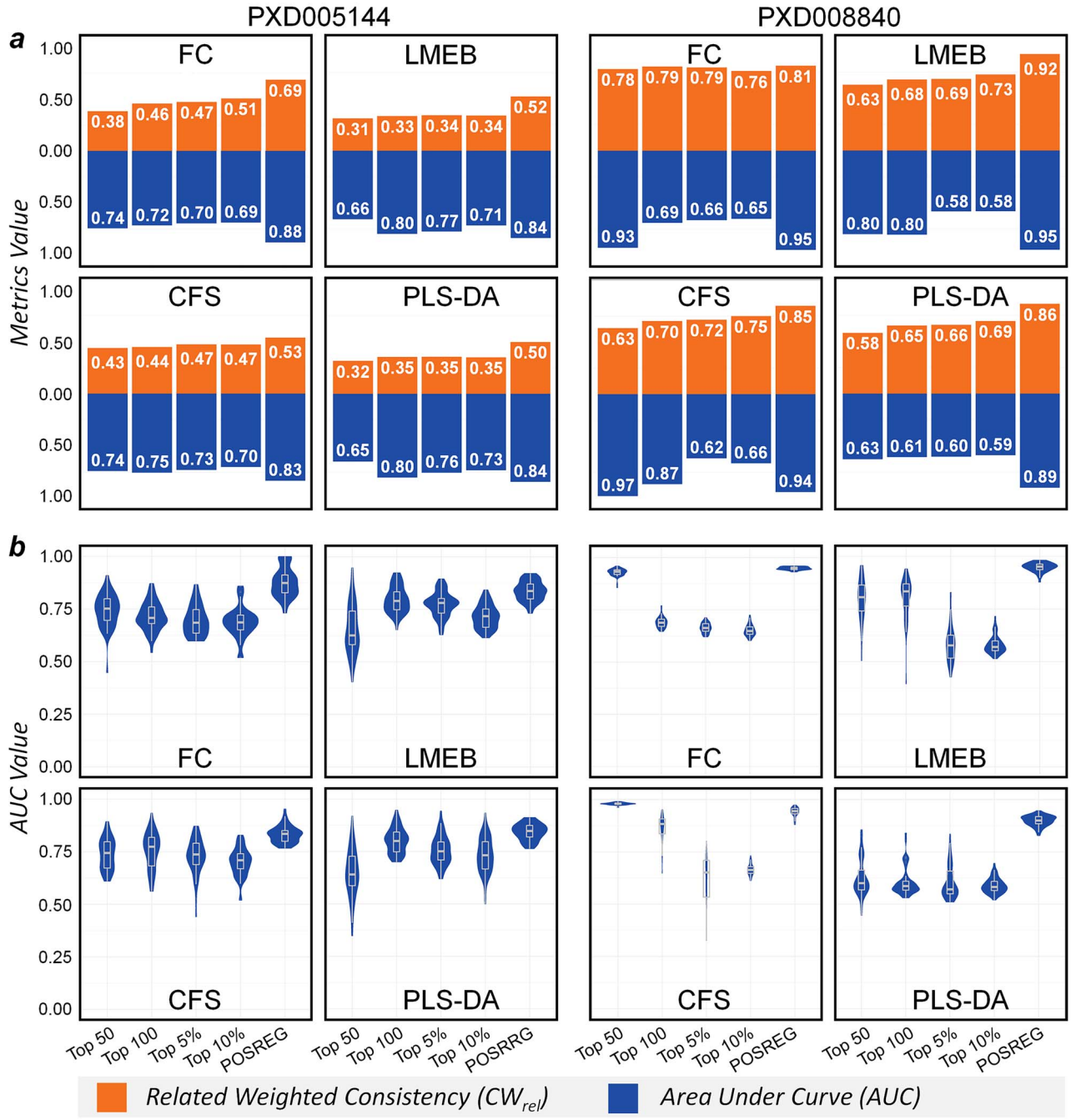

Figure 4. Verifying the capacity of POSREG in improving reproducibility and generalizability of filter FS methods. Two benchmarks PXD005144 and PXD008840 were collected and assessed by both POSREG and four traditional FS workflows using two representative univariate filters (FC, LMEB) and multivariate filters (CFS, PLS-DA), which directly take top 50, top 100, top 5% and top10% of ranked features as the signature. The assessment was repeated 50 times to avoid contingency in the results. (**a**) The mean value of CW_{rel} (orange) and AUC (blue) is drawn as mirrored bar plots. (**b**) The violin plot showing the distribution of the AUC value for 50 repetitions.

was conducted using default FS method Linear Model & Bayes with default parameters on POSREG for 10 times, and the resulting 10 optimal feature lists were collected; (b) for each selected feature, the occurrence in 10 optimal feature lists was calculated; (c) the selected feature was sorted by their occurrences in the 10 optimal feature lists and compared with those reported in PXD005144's publication [112].

The protein numbers of 10 optimal lists varied from 24 to 34 and their corresponding AUC values were always over 0.95. Figure 6 shows 22 proteins consistently occurred (occurred at least 8 times) in 10 optimal feature lists. Among these 22 consistently occurred proteins, 17 of them were also reported in the original publication of PXD005144 [112], which were colored in blue in Figure 6. Besides, there were also five new proteins only identified by POSREG, which were colored in orange in Figure 6. To determine whether these POSREG's newly identified proteins were relevant to the studied disease, a comprehensive literature review

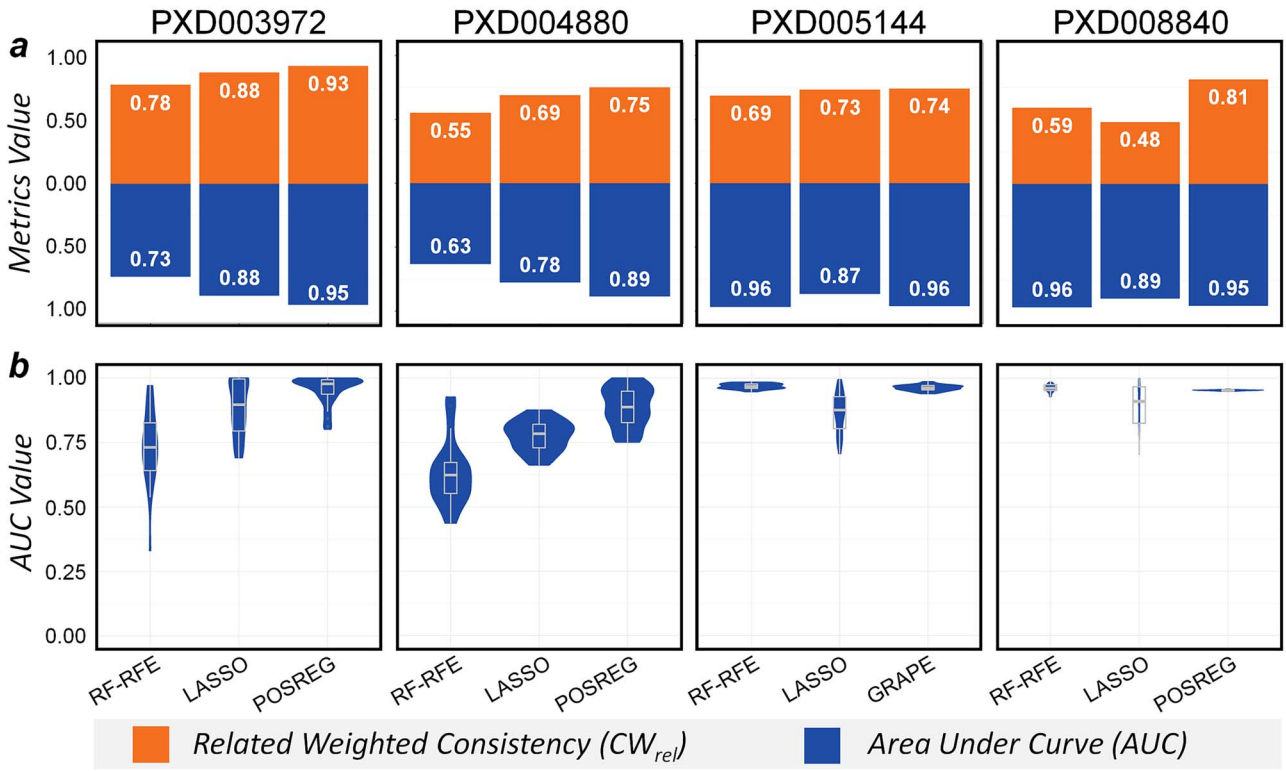

Figure 5. Comparing POSREG with established wrapper/embedded FS techniques. Four benchmark datasets PXD003972, PXD004880, PXD005144 and PXD008840 were used to assess POSREG, RF-RFE (well-established wrapper method) and LASSO method (well-established embedded method), respectively. (**a**) The mean value of CW_{rel} (orange) and AUC (blue) is drawn as mirrored bar plots. (**b**) The violin plot showing the distribution of the AUC value.

was conducted. Vitronectin (UniProt Entry: P04004) was reported as a major driver of the differentiation process in the pancreatic cancer model [119]. Ceruloplasmin (UniProt Entry: P00450) was suggested to be a promising marker for pancreatic patients negative for CA19-9 [120]. Chemokine-like factor superfamily member 1 (UniProt Entry: Q8IZ96) was identified to be prognostic and high expression was unfavorable in pancreatic cancer by HUMAN PROTEIN ATLAS [121]. Epidemiologic evidence indicated that high glucose was linked to an increased risk for pancreatic cancer [122]. High glucose conditions can enhance the cancer progression by upregulating the expression of alpha-mannosidase 2× (MAN2A2, UniProt Entry: P49641) at both mRNA and protein levels in cancer [123]. Although no existing report was showing direct relevance between intraflagellar transport protein 88 homolog (IFT88, UniProt Entry: Q13099) and pancreatic cancer, it had been previously investigated as a tumor suppressor in other cancer such as hepatocellular carcinoma, breast carcinoma and so on [124].

To conclude, POSREG could not only identify proteomic signatures with great generalizability and reproducibility but also provide valuable clues for discovering proteomic features with significant biological meaning. Moreover, the results also implied that there is some relationship between phenotype association with both generalizability and reproducibility, by improving the reproducibility of FS, the generalizability of identified signature would be improved by eliminating the non-predictor features and the phenotype association of selected features would also be improved by reducing the chances of erroneous elimination of predictor features.

Standard workflow and operating procedure of POSREG

The standard workflow of POSREG can be divided into three steps (Figure 7): (1) reproducibility enhancing by maximizing CW_{rel} and ensemble learning. This step mainly performs multiple FS and then finds the most robust feature size among these generated proteomic signatures and aggregates them into the ensemble feature ranking, which included: (i) data uploading, (ii) data preprocessing (missing value imputation, data filtering, data normalization and data transformation), (iii) multiple FS generating multiple feature ranking (homogeneous, heterogeneous or hybrid), (iv) reproducibility evaluation of multiple feature ranking based on CW_{rel} and (v) ensembling the most robust signatures with highest CW_{rel}. (2) Generalizability assessing using the AUC-based golden section search. The ensemble feature ranking generated in step 1 is further analyzed using the AUC-based golden section search methods proposed in Liu's research [85] to discover the top assemble of features with the highest AUC and assign it as the optimal signature. (3) Phenotype association indicated via functional enrichment analysis. The optimal signature is 'optimal' only in the theoretical perspectives of reproducibility and generalizability, not in practice. Therefore, the final

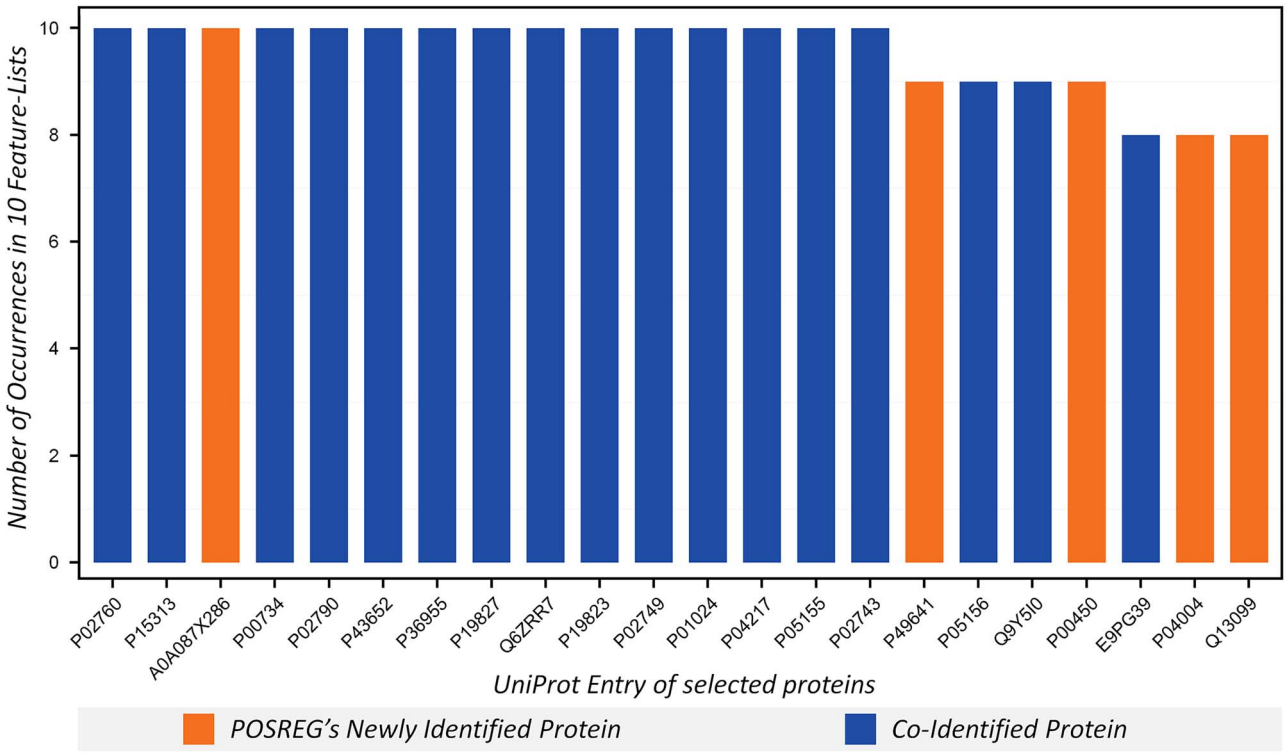

Figure 6. Demonstrating the superiority of POSREG with the case study on PXD005144. The benchmark PXD005144 was analyzed using POSREG 10 times, the resulting 10 optimal feature lists were collected and the number of times each feature occurred in the 10 optimal feature lists was calculated. The number of occurrences of consistently occurred protein (occurred at least 8 times) among 10 optimal feature-lists is drawn as a bar plot. The blue bar represented the co-identified protein which was both identified by POSREG and the original paper, and the orange bar represented POSREG's newly identified protein.

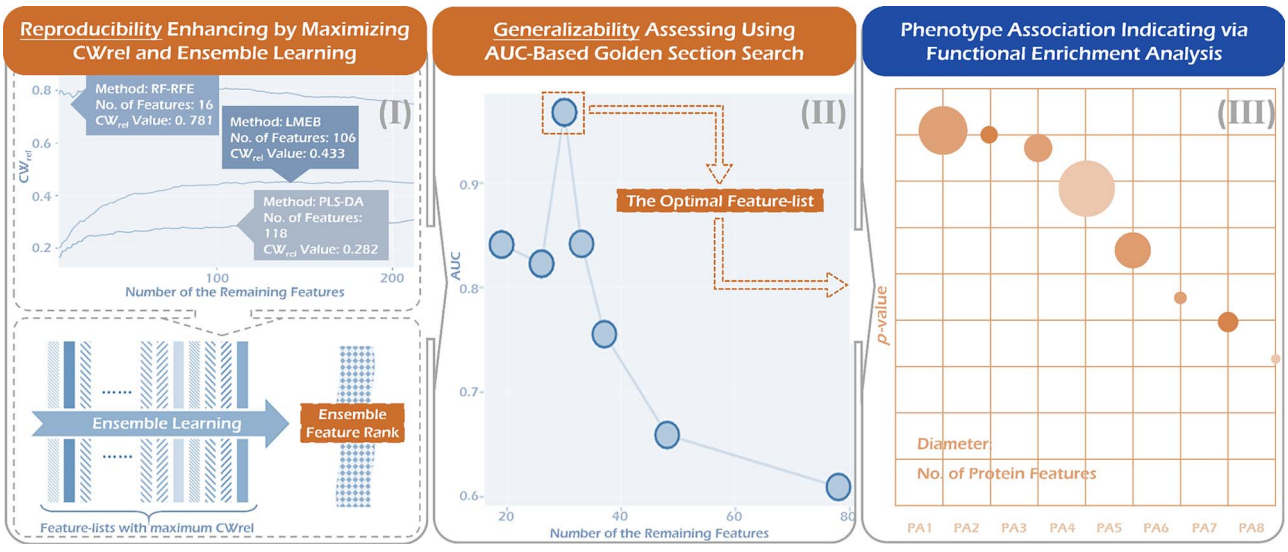

Figure 7. General workflow of POSREG: (I) Reproducibility enhancing by maximizing CW_{rel} and ensemble learning; (II) Generalizability assessing using AUC-based golden section search; (III) Phenotype association indicating via functional enrichment analysis.

step is a GO-based enrichment analysis to indicate the phenotype association level of the optimal signature [125].

Conclusions

POSREG was constructed and validated to enable the generalizable and reproducible discovery of the proteomic signature with phenotype association indication. It is unique for its capacities of identifying proteomic signatures of good reproducibility and generalizability using CW_{rel}, ensemble learning and AUC-based golden-section search. Therefore, POSREG can facilitate current proteomics-based molecular biology researches and has great potentiality for application in proteomic signatures identification and other research requiring FS.

Key Points

- An online tool POSREG was constructed to simultaneously optimize the reproducibility and generalizability of proteomic signature discovery.
- POSREG identified proteomic signatures of good reproducibility by optimizing the CW_{rel} among multiple feature rankings and ensembling the most robust signatures with the highest CW_{rel}.
- POSREG optimized the generalizability of identified signatures by identifying the feature subset with the highest AUC using an AUC-based golden section search strategy.
- POSREG's unique capacities were validated using multiple proteomic benchmarks. It is freely and publicly accessible at: https://idrblab.org/posreg

Supplementary Data

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

Abbreviations

AUC, area under the curve; CW_{rel}, relative weighted consistency; FS, feature selection; FSS, feature subset size; Ensemble-FS, ensemble feature selection; ROC, receiver operating characteristic; DIA, data-independent acquisition; DDA, data-dependent acquisition; LASSO, least absolute shrinkage and selection operator; RF-RFE, random forest - recursive feature elimination

Author Contributions

F.Z. conceived the idea and supervised the work. F.C. and Y.Z. conducted the research. F.C., Y.Z., and Y.J. prepared and analyzed the data. F.Z. and F.C. wrote the manuscript. All authors reviewed and approved the manuscript.

Data availability statament

All data in the manuscript are collected and available in PRIDE database.

Funding

National Natural Science Foundation of China (U1909208 & 81872798); Natural Science Foundation of Zhejiang Province (LR21H300001); Leading Talent of the 'Ten Thousand Plan' - National High-Level Talents Special Support Plan of China; Fundamental Research Fund for Central Universities (2018QNA7023); 'Double Top-Class' University Project (181201*194232101); Key R&D Program of Zhejiang Province (2020C03010). This work was supported by Westlake Laboratory (Westlake Laboratory of Life Sciences and Biomedicine); Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare; Alibaba Cloud; Information Technology Center of Zhejiang University.

References

1. Harel M, Ortenberg R, Varanasi SK, *et al.* Proteomics of melanoma response to immunotherapy reveals mitochondrial dependence. *Cell* 2019;**179**:236–250.e18.
2. Tang J, Fu J, Wang Y, *et al.* ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. *Brief Bioinform* 2020;**21**:621–36.
3. Tang J, Fu J, Wang Y, *et al.* Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains. *Mol Cell Proteomics* 2019;**18**:1683–99.
4. Distler U, Kuharev J, Navarro P, *et al.* Label-free quantification in ion mobility-enhanced data-independent acquisition proteomics. *Nat Protoc* 2016;**11**:795–812.
5. Li YH, Li XX, Hong JJ, *et al.* Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs. *Brief Bioinform* 2020;**21**:649–62.
6. Fu J, Tang J, Wang Y, *et al.* Discovery of the consistently well-performed analysis chain for SWATH-MS based pharmacoproteomic quantification. *Front Pharmacol* 2018;**9**:681.
7. Wang X, Li F, Qiu W, *et al.* SYNBIP: synthetic binding proteins for research, diagnosis and therapy. *Nucleic Acids Res* 2022;**50**:D560–70.
8. Penn-Nicholson A, Hraha T, Thompson EG, *et al.* Discovery and validation of a prognostic proteomic signature for tuberculosis progression: a prospective cohort study. *PLoS Med* 2019;**16**:e1002781.
9. Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. *Nature* 2016;**537**:347–55.
10. Lualdi M, Fasano M. Statistical analysis of proteomics data: a review on feature selection. *J Proteomics* 2019;**198**:18–26.
11. Pes B. Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. *Neural Comput Applic* 2020;**32**:5951–73.
12. Tang J, Mou M, Wang Y, *et al.* MetaFS: performance assessment of biomarker discovery in metaproteomics. *Brief Bioinform* 2021;**22**:bbaa105.
13. Saari P, Eerola T, Lartillot O. Generalizability and simplicity as criteria in feature selection: application to mood classification in music. *IEEE Trans Audio Speech Lang Process* 2011;**19**:1802–12.
14. Zhu F, Li XX, Yang SY, *et al.* Clinical success of drug targets prospectively predicted by in silico study. *Trends Pharmacol Sci* 2018;**39**:229–31.
15. Tyanova S, Albrechtsen R, Kronqvist P, *et al.* Proteomic maps of breast cancer subtypes. *Nat Commun* 2016;**7**:10259.
16. Goh WB, Wong L. Dealing with confounders in omics analysis. *Trends Biotechnol* 2018;**36**:488–98.
17. Pes B, Dessi N, Angioni M. Exploiting the ensemble paradigm for stable feature selection: a case study on high-dimensional genomic data. *Inf Fusion* 2017;**35**:132–47.
18. Alhenawi E, Al-Sayyed R, Hudaib A, *et al.* Feature selection methods on gene expression microarray data for cancer classification: a systematic review. *Comput Biol Med* 2021;**140**:105051.
19. Donnelly DP, Rawlins CM, DeHart CJ, *et al.* Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. *Nat Methods* 2019;**16**:587–94.

20. Christin C, Hoefsloot HC, Smilde AK, *et al.* A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Mol Cell Proteomics* 2013;**12**:263–76.
21. Hong J, Luo Y, Mou M, *et al.* Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery. *Brief Bioinform* 2020;**21**:1825–36.
22. Hong J, Luo Y, Zhang Y, *et al.* Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning. *Brief Bioinform* 2020;**21**:1437–47.
23. Wang L. Feature selection with kernel class separability. *IEEE Trans Pattern Anal Mach Intell* 2008;**30**:1534–46.
24. Ullah M, Han K, Hadi F, *et al.* PScL-HDeep: image-based prediction of protein subcellular location in human tissue using ensemble learning of handcrafted and deep learned features with two-layer feature selection. *Brief Bioinform* 2021;**22**:bbab278.
25. Seijo-Pardo B, Porto-Diaz I, Bolon-Canedo V, *et al.* Ensemble feature selection: homogeneous and heterogeneous approaches. *Knowl Based Syst* 2017;**118**:124–39.
26. Zou Q, Zeng J, Cao L, *et al.* A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 2016;**173**:346–54.
27. Zhang SN, He YF, Li XZ, *et al.* Biolabel-led research pattern positions the effects and mechanisms of Sophorae Tonkinensis radix et rhizome on lung diseases: a novel strategy for computer-aided herbal medicine research based on omics and bioinformatics. *Comput Biol Med* 2021;**136**:104769.
28. Tan MS, Cheah PL, Chin AV, *et al.* A review on omics-based biomarkers discovery for Alzheimer's disease from the bioinformatics perspectives: statistical approach vs machine learning approach. *Comput Biol Med* 2021;**139**:104947.
29. Bolon-Canedo V, Alonso-Betanzos A. Ensembles for feature selection: a review and future trends. *Inf Fusion* 2019;**52**:1–12.
30. Fu J, Zhang Y, Liu J, *et al.* Pharmacometabonomics: data processing and statistical analysis. *Brief Bioinform* 2021;**22**:bbab138.
31. Chong J, Soufan O, Li C, *et al.* MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res* 2018;**46**:W486–94.
32. Su R, Liu X, Wei L. MinE-RFE: determine the optimal subset from RFE by minimizing the subset-accuracy-defined energy. *Brief Bioinform* 2020;**21**:687–98.
33. Zhu F, Qin C, Tao L, *et al.* Clustered patterns of species origins of nature-derived drugs and clues for future bioprospecting. *Proc Natl Acad Sci U S A* 2011;**108**:12943–8.
34. Li YH, Xu JY, Tao L, *et al.* SVM-Prot 2016: a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity. *PLoS One* 2016;**11**:e0155290.
35. Zhang S, Amahong K, Zhang C, *et al.* RNA-RNA interactions between SARS-CoV-2 and host benefit viral development and evolution during COVID-19 infection. *Brief Bioinform* 2021;**81**:1691–2.
36. Neumann U, Genze N, Heider D. EFS: an ensemble feature selection tool implemented as R-package and web-application. *BioData Min* 2017;**10**:21.
37. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. *Nat Rev Genet* 2018;**19**:129–47.
38. Xue W, Yang F, Wang P, *et al.* What contributes to serotonin-norepinephrine reuptake Inhibitors' dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation. *ACS Chem Neurosci* 2018;**9**:1128–40.
39. Zhang Y, Ying JB, Hong JJ, *et al.* How does chirality determine the selective inhibition of histone deacetylase 6? A lesson from trichostatin a enantiomers based on molecular dynamics. *ACS Chem Neurosci* 2019;**10**:2467–80.
40. Fu T, Zheng G, Tu G, *et al.* Exploring the binding mechanism of metabotropic glutamate receptor 5 negative allosteric modulators in clinical trials by molecular dynamics simulations. *ACS Chem Neurosci* 2018;**9**:1492–502.
41. Xue W, Wang P, Tu G, *et al.* Computational identification of the binding mechanism of a triple reuptake inhibitor amitifadine for the treatment of major depressive disorder. *Phys Chem Chem Phys* 2018;**20**:6606–16.
42. Yin J, Li F, Li Z, *et al.* Feature, function, and information of drug transporter-related databases. *Drug Metab Dispos* 2022;**50**:76–85.
43. Khan T, Khan A, Nasir SN, *et al.* CytomegaloVirusDb: multi-omics knowledge database for cytomegaloviruses. *Comput Biol Med* 2021;**135**:104563.
44. Perez-Riverol Y, Csordas A, Bai JW, *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* 2019;**47**:D442–50.
45. Deutsch EW, Bandeira N, Sharma V, *et al.* The ProteomeXchange consortium in 2020: enabling big data approaches in proteomics. *Nucleic Acids Res* 2020;**48**:D1145–52.
46. Ma J, Chen T, Wu S, *et al.* iProX: an integrated proteome resource. *Nucleic Acids Res* 2019;**47**:D1211–7.
47. Li B, Tang J, Yang Q, *et al.* NOREVA: normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Res* 2017;**45**:W162–70.
48. Yang Q, Wang Y, Zhang Y, *et al.* NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data. *Nucleic Acids Res* 2020;**48**:W436–48.
49. Yang Q, Hong J, Li Y, *et al.* A novel bioinformatics approach to identify the consistently well-performing normalization strategy for current metabolomic studies. *Brief Bioinform* 2020;**21**:2142–52.
50. Wang L, Wang Y, Chang Q. Feature selection methods for big data bioinformatics: a survey from the search perspective. *Methods* 2016;**111**:21–31.
51. Liang S, Ma A, Yang S, *et al.* A review of matched-pairs feature selection methods for gene expression data analysis. *Comput Struct Biotechnol J* 2018;**16**:88–97.
52. Torres R, Judson-Torres RL. Research techniques made simple: feature selection for biomarker discovery. *J Invest Dermatol* 2019;**139**:2068–74.
53. Mahendran N, PM DRV. A deep learning framework with an embedded-based feature selection approach for the early detection of the Alzheimer's disease. *Comput Biol Med* 2022;**141**:105056.
54. Wang Y, Klijn JG, Zhang Y, *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005;**365**:671–9.
55. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci U S A* 2006;**103**:5923–8.
56. Loddo A, Buttau S, Di Ruberto C. Deep learning based pipelines for Alzheimer's disease diagnosis: a comparative study and a novel deep-ensemble method. *Comput Biol Med* 2022;**141**:105032.
57. Li R, Ren C, Zhang X, *et al.* A novel ensemble learning method using multiple objective particle swarm optimization

for subject-independent EEG-based emotion recognition. *Comput Biol Med* 2021;**140**:105080.
58. Yang Q, Li B, Tang J, *et al.* Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data. *Brief Bioinform* 2020;**21**:1058–68.
59. Wang Z, Zou Y, Liu PX. Hybrid dilation and attention residual U-net for medical image segmentation. *Comput Biol Med* 2021;**134**:104449.
60. Cheng LH, Hsu TC, Lin C. Integrating ensemble systems biology feature selection and bimodal deep neural network for breast cancer prognosis prediction. *Sci Rep* 2021;**11**:14914.
61. Zhang M, Zhang L, Zou J, *et al.* Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics* 2009;**25**:1662–8.
62. Geman S, Bienenstock E, Doursat R. Neural networks and the bias variance dilemma. *Neural Comput* 1992;**4**:1–58.
63. Somol P, Novovicova J. Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. *IEEE Trans Pattern Anal Mach Intell* 2010;**32**:1921–39.
64. Fu J, Zhang Y, Wang Y, *et al.* Optimization of metabolomic data processing using NOREVA. *Nat Protoc* 2022;**17**:129–51.
65. Shiri I, Sorouri M, Geramifar P, *et al.* Machine learning-based prognostic modeling using clinical data and quantitative radiomic features from chest CT images in COVID-19 patients. *Comput Biol Med* 2021;**132**:104304.
66. Petkovic M, Slavkov I, Kocev D, *et al.* Biomarker discovery by feature ranking: evaluation on a case study of embryonal tumors. *Comput Biol Med* 2021;**128**:104143.
67. Peeters L, Beirnaert C, Van der Auwera A, *et al.* Revelation of the metabolic pathway of hederacoside C using an innovative data analysis strategy for dynamic multiclass biotransformation experiments. *J Chromatogr A* 2019;**1595**:240–7.
68. Li F, Zhou Y, Zhang X, *et al.* SSizer: determining the sample sufficiency for comparative biological study. *J Mol Biol* 2020;**432**:3411–21.
69. Goh WWB, Wong L. Advanced bioinformatics methods for practical applications in proteomics. *Brief Bioinform* 2019;**20**:347–55.
70. Chen J, Zhang Y, Wei Y, *et al.* Discrimination of the contextual features of top performers in scientific literacy using a machine learning approach. *Res Sci Educ* 2021;**51**:129–58.
71. Yang Q, Li B, Chen S, *et al.* MMEASE: online meta-analysis of metabolomic data by enhanced metabolite annotation, marker selection and enrichment analysis. *J Proteomics* 2021;**232**:104023.
72. Robin X, Turck N, Hainard A, *et al.* PROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;**12**:77.
73. Tabe-Bordbar S, Emad A, Zhao SD, *et al.* A closer look at cross-validation for assessing the accuracy of gene regulatory networks and models. *Sci Rep* 2018;**8**:6620.
74. Ignjatovic V, Geyer PE, Palaniappan KK, *et al.* Mass spectrometry-based plasma proteomics: considerations from sample collection to achieving translational data. *J Proteome Res* 2019;**18**:4085–97.
75. Shi Z, Wen B, Gao Q, *et al.* Feature selection methods for protein biomarker discovery from proteomics or multiomics data. *Mol Cell Proteomics* 2021;**20**:100083.
76. Varoquaux G. Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage* 2018;**180**:68–77.
77. Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification? *Bioinformatics* 2004;**20**:374–80.
78. Vabalas A, Gowen E, Poliakoff E, *et al.* Machine learning algorithm validation with a limited sample size. *PLoS One* 2019;**14**:e0224365.
79. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 2006;**7**:91.
80. Brown MJ, McInnes GT, Papst CC, *et al.* Aliskiren and the calcium channel blocker amlodipine combination as an initial treatment strategy for hypertension control: a randomised, parallel-group trial. *Lancet* 2011;**377**:312–20.
81. Lopez-Rincon A, Martinez-Archundia M, Martinez-Ruiz GU, *et al.* Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection. *BMC Bioinformatics* 2019;**20**:480.
82. Kolde R, Laur S, Adler P, *et al.* Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* 2012;**28**:573–80.
83. Hua J, Xiong Z, Lowey J, *et al.* Optimal number of features as a function of sample size for various classification rules. *Bioinformatics* 2005;**21**:1509–15.
84. Liu Z. Investigation of temperature and feature size effects on deformation of metals by superplastic nanomolding. *Phys Rev Lett* 2019;**122**:016101.
85. Song X, Waitman LR, Hu Y, *et al.* Robust clinical marker identification for diabetic kidney disease with ensemble feature selection. *J Am Med Inform Assoc* 2019;**26**:242–53.
86. Abramowitz MK, Hostetter TH, Melamed ML. The serum anion gap is altered in early kidney disease and associates with mortality. *Kidney Int* 2012;**82**:701–9.
87. Go C. The gene ontology resource: 20 years and still going strong. *Nucleic Acids Res* 2019;**47**:D330–8.
88. Yin J, Li F, Zhou Y, *et al.* INTEDE: interactome of drug-metabolizing enzymes. *Nucleic Acids Res* 2021;**49**:D1233–43.
89. Wang Y, Zhang S, Li F, *et al.* Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res* 2020;**48**:D1031–41.
90. Zhu F, Shi Z, Qin C, *et al.* Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res* 2012;**40**:D1128–36.
91. Yang H, Qin C, Li YH, *et al.* Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res* 2016;**44**:D1069–74.
92. Yu G, Wang LG, Han Y, *et al.* clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 2012;**16**:284–7.
93. Ito K, Murphy D. Application of ggplot2 to pharmacometric graphics. *CPT Pharmacometrics Syst Pharmacol* 2013;**2**:e79.
94. Steenwyk JL, Rokas A. Colorblind-friendly color palettes and ggplot2 graphic system extensions for publication-quality scientific figures. *Microbiol Resour Announc* 2021;**10**:e0087121.
95. Yin J, Sun W, Li F, *et al.* VARIDT 1.0: variability of drug transporter database. *Nucleic Acids Res* 2020;**48**:D1042–50.
96. Fu T, Li F, Zhang Y, *et al.* VARIDT 2.0: structural variability of drug transporter. *Nucleic Acids Res* 2022;**50**:D1417–31.
97. Zhang S, Amahong K, Sun X, *et al.* The miRNA: a small but powerful RNA for COVID-19. *Brief Bioinform* 2021;**22**:1137–49.
98. Gautier L, Cope L, Bolstad BM, *et al.* Affy: analysis of affymetrix genechip data at the probe level. *Bioinformatics* 2004;**20**:307–15.

99. Hothorn T, Hornik K, Van de Wiel M, *et al.* A lego system for conditional inference. *Am Stat* 2006;**60**:257–63.
100. Yu G, Wang L, Yan G, *et al.* DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* 2015;**31**:608–9.
101. Rohart F, Gautier B, Singh A, *et al.* mixOmics: an R package for omics feature selection and multiple data integration. *PLoS Comput Biol* 2017;**13**:e1005752.
102. Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* 2013;**29**:1830–1.
103. Stacklies W, Redestig H, Scholz M, *et al.* PCAMethods – a bioconductor package providing PCA methods for incomplete data. *Bioinformatics* 2007;**23**:1164–7.
104. Wang W, Sue AC, Goh WWB. Feature selection in clinical proteomics: with great power comes great reproducibility. *Drug Discov Today* 2017;**22**:912–8.
105. Guo T, Kouvonen P, Koh CC, *et al.* Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. *Nat Med* 2015;**21**:407–13.
106. Boeynaems S, Alberti S, Fawzi NL, *et al.* Protein phase separation: a new phase in cell biology. *Trends Cell Biol* 2018;**28**:420–35.
107. Bron EE, Smits M, Niessen WJ, *et al.* Feature selection based on the SVM weight vector for classification of dementia. *IEEE J Biomed Health Inform* 2015;**19**:1617–26.
108. Gui J, Sun Z, Ji S, *et al.* Feature selection based on structured sparsity: a comprehensive study. *IEEE Trans Neural Netw Learn Syst* 2017;**28**:1490–507.
109. Zhang Y, Zhang HX, Zheng QC. In silico study of membrane lipid composition regulating conformation and hydration of influenza virus B M2 channel. *J Chem Inf Model* 2020;**60**:3603–15.
110. Chen Z, Zhao P, Li F, *et al.* iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 2018;**34**:2499–502.
111. Zeng X, Liu L, Lu L, *et al.* Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* 2018;**34**:2425–32.
112. Saraswat M, Joenvaara S, Seppanen H, *et al.* Comparative proteomic profiling of the serum differentiates pancreatic cancer from chronic pancreatitis. *Cancer Med* 2017;**6**:1738–51.
113. Ge S, Xia X, Ding C, *et al.* A proteomic landscape of diffuse-type gastric cancer. *Nat Commun* 2018;**9**:1012.
114. Seijo-Pardo B, Bolon-Canedo V, Alonso-Betanzos A. On developing an automatic threshold applied to feature selection ensembles. *Inf Fusion* 2019;**45**:227–45.
115. Tang J, Wang Y, Luo Y, *et al.* Computational advances of tumor marker selection and sample classification in cancer proteomics. *Comput Struct Biotechnol J* 2020;**18**:2012–25.
116. Birse K, Arnold KB, Novak RM, *et al.* Molecular signatures of immune activation and epithelial barrier Remodeling are enhanced during the luteal phase of the menstrual cycle: implications for HIV susceptibility. *J Virol* 2015;**89**:8793–805.
117. Caron E, Roncagalli R, Hase T, *et al.* Precise temporal profiling of signaling complexes in primary cells using SWATH mass spectrometry. *Cell Rep* 2017;**18**:3219–26.
118. Sullivan KD, Evans D, Pandey A, *et al.* Trisomy 21 causes changes in the circulating proteome indicative of chronic autoinflammation. *Sci Rep* 2017;**7**:14818.
119. Cabarcas SM, Sun L, Mathews L, *et al.* The differentiation of pancreatic tumor-initiating cells by vitronectin can be blocked by cilengitide. *Pancreas* 2013;**42**:861–70.
120. Brandi J, Dalla Pozza E, Dando I, *et al.* Secretome protein signature of human pancreatic cancer stem-like cells. *J Proteomics* 2016;**136**:1–12.
121. Uhlen M, Fagerberg L, Hallstrom BM, *et al.* Proteomics tissue-based map of the human proteome. *Science* 2015;**347**:1260419.
122. Shikata K, Ninomiya T, Kiyohara Y. Diabetes mellitus and cancer risk: review of the epidemiological evidence. *Cancer Sci* 2013;**104**:9–14.
123. Thonsri U, Wongkham S, Wongkham C, *et al.* High glucose-ROS conditions enhance the progression in cholangiocarcinoma via upregulation of MAN2A2 and CHD8. *Cancer Sci* 2021;**112**:254–64.
124. Lee J, Yi S, Won M, *et al.* Loss-of-function of IFT88 determines metabolic phenotypes in thyroid cancer. *Oncogene* 2018;**37**:4455–74.
125. Yuan N, Chen Y, Xia Y, *et al.* Inflammation-related biomarkers in major psychiatric disorders: a cross-disorder assessment of reproducibility and specificity in 43 meta-analyses. *Transl Psychiatry* 2019;**9**:233.
126. Chuang LY, Yang CH, Wu KC, *et al.* A hybrid feature selection method for DNA microarray data. *Comput Biol Med* 2011;**41**:228–37.
127. Farina D, Kamavuako EN, Wu J, *et al.* Entropy-based optimization of wavelet spatial filters. *IEEE Trans Biomed Eng* 2008;**55**:914–22.
128. Guo L, Lobenhofer EK, Wang C, *et al.* Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat Biotechnol* 2006;**24**:1162–9.
129. van Ooijen MP, Jong VL, Eijkemans MJC, *et al.* Identification of differentially expressed peptides in high-throughput proteomics data. *Brief Bioinform* 2018;**19**:971–81.
130. Bartel J, Krumsiek J, Theis FJ. Statistical methods for the analysis of high-throughput metabolomics data. *Comput Struct Biotechnol J* 2013;**4**:e201301009.
131. Urbanowicz RJ, Meeker M, La Cava W, *et al.* Relief-based feature selection: introduction and review. *J Biomed Inform* 2018;**85**:189–203.
132. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 2001;**98**:5116–21.
133. Wilcoxon F. Individual comparisons of grouped data by ranking methods. *J Econ Entomol* 1946;**39**:269.