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

Fengcheng Li1, Ying Zhou1, 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
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 combining 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
Benchmark datasets collected and analyzed in this study

<table>
<thead>
<tr>
<th>Dataset ID</th>
<th>Description</th>
<th>Sample Size</th>
<th>Acquired from</th>
</tr>
</thead>
<tbody>
<tr>
<td>PXD000672</td>
<td>ProteomeXchange</td>
<td>20 samples</td>
<td>PRIDE [44]</td>
</tr>
<tr>
<td>PXD002882</td>
<td>ProteomeXchange</td>
<td>20 samples</td>
<td>PRIDE [44]</td>
</tr>
<tr>
<td>PXD003972</td>
<td>iProx [46]</td>
<td>20 samples</td>
<td>PRIDE [44]</td>
</tr>
<tr>
<td>PXD004880</td>
<td>ProteomeXchange</td>
<td>20 samples</td>
<td>PRIDE [44]</td>
</tr>
<tr>
<td>PXD005144</td>
<td>ProteomeXchange</td>
<td>20 samples</td>
<td>PRIDE [44]</td>
</tr>
<tr>
<td>PXD006129</td>
<td>ProteomeXchange</td>
<td>20 samples</td>
<td>PRIDE [44]</td>
</tr>
<tr>
<td>PXD008840</td>
<td>ProteomeXchange</td>
<td>20 samples</td>
<td>PRIDE [44]</td>
</tr>
</tbody>
</table>

Table 2
FS methods employed and analyzed in this study

<table>
<thead>
<tr>
<th>Method</th>
<th>Theory</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filter</td>
<td>Univariate</td>
<td>fold change analysis, Wilcoxon rank-sum test, etc.</td>
</tr>
<tr>
<td>Filter</td>
<td>Multivariate</td>
<td>correlation-based method, entropy-based filter, etc.</td>
</tr>
<tr>
<td>Wrapper</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Embedded</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Downloaded from https://academic.oup.com/bib/article/23/2/bbac040/6532538 by Zhejiang University user on 29 March 2022
Brief introduction of FS methods employed and analyzed in this study

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

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

Table 2. Seven benchmark proteomics datasets were collected and analyzed in this study. The corresponding dataset was collected from the Proteomics Identification Database (PRIDE) or integrated Proteome resources (iProx).

<table>
<thead>
<tr>
<th>FS method (Abbreviation)</th>
<th>Type</th>
<th>Brief introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFS</td>
<td>Multivariate filter</td>
<td>Evaluate feature subset based on the prediction ability of each feature in and the correlation between them.</td>
</tr>
<tr>
<td>Entropy-based filters</td>
<td>Multivariate filter</td>
<td>Select features based on the contribution of information related to class variables. Compensate for information gain bias.</td>
</tr>
<tr>
<td>FC (ENTROPY)</td>
<td>Univariate filter</td>
<td>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.</td>
</tr>
<tr>
<td>LMEB</td>
<td>Univariate filter</td>
<td>Evaluate the differential abundance of features by drawing a volcano plot, which measures the differentially accumulated features based on fold changes and t statistics.</td>
</tr>
<tr>
<td>PLS-DA</td>
<td>Multivariate filter</td>
<td>Predict variables that maximize differences among predetermined samples. Infer classification of unclassified sample groups based on the calibration set with known class distribution.</td>
</tr>
<tr>
<td>ReliefF (REF)</td>
<td>Multivariate filter</td>
<td>Estimate attributes based on the degree of value differentiation between near instances.</td>
</tr>
<tr>
<td>Significance analysis of microarrays (SAM)</td>
<td>Univariate filter</td>
<td>Score each gene based on the change in gene expression relative to the standard deviation of repeated measurements.</td>
</tr>
<tr>
<td>Univariate t-test (t-test)</td>
<td>Univariate Filter</td>
<td>Rank features based on P-values. Features with a P-value &lt;0.05 are considered to be significant.</td>
</tr>
<tr>
<td>Wilcoxon Rank-sum test (Wilcox)</td>
<td>Univariate Filter</td>
<td>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.</td>
</tr>
</tbody>
</table>

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.
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.

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, how many features should be added to the training set to achieve the most accurate classifier is a frequently encountered issue. Due to the constrained computation resources, it is impractical to assess the accuracy of every possible combination of proteomic features using the exhaustive method. 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. 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. 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. Under such supposition, the golden-section search algorithm could be adopted to find the optimal signature with the highest AUC.

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. GO resource provides computable knowledge about the function of gene and gene products and is extensively adopted for the analysis of omics-related data. 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.

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 to better visualization of the enrichment result. 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.

**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 but ignored the reproducibility of their proposed signatures. 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. To ensure the stability of identified features and ultimately enhance their practicality, the metrics CWrel was applied to assess the reproducibility in the pipeline of POSREG.

As demonstrated in Materials and Methods and Supplementary Methods S1, CWrel’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. Therefore, it is feasible to use CWrel as a reproducibility assessment metric to find the most robust feature size. To comprehensively validate the feasibility of using CWrel for reproducibility assessment, the benchmark dataset PXD000672 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 CWrel value under particular feature size. These four preceding steps make up one single independent replicated trial which can illustrate the trend of CWrel 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 CWrel.

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\textsubscript{rel}-based FS reproducibility evaluation was assessed by comparing the trends in CW\textsubscript{rel} between different independent replicated trials in benchmark dataset PXD000672 \cite{105}. The x-axis in each sub-figure denoted the proportion of features selected into CW\textsubscript{rel} calculation, and the y-axis denoted the value of CW\textsubscript{rel}. The trend in CW\textsubscript{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.

CW\textsubscript{rel} (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\textsubscript{rel} for different methods tend to trend differently and their maximum points occurred in different coordinates. These results indicate that CW\textsubscript{rel} is a stable and robust metric for FS reproducibility evaluation and the CW\textsubscript{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\textsubscript{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\textsubscript{rel} reached the maximum value with a very small proportion of selected features, whereas the other benchmarks (PXD002882, PXD006129, PXD008840) reached the maximum CW\textsubscript{rel} with a medium proportion of feature. On the other hand, the reproducibility of the
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 CW$_{rel}$-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$_{\text{max}}$($CW_{rel}$), it is not hard to discover from Figures 2 and 3 that PSF$_{\text{max}}$($CW_{rel}$) 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$_{\text{max}}$($CW_{rel}$) 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$_{\text{max}}$($CW_{rel}$), the stability of FS method would decrease because of the difficulty in choosing between the top PSF$_{\text{max}}$($CW_{rel}$) significant features. And vice versa, if the proportion of features need to be selected is more than PSF$_{\text{max}}$($CW_{rel}$), 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$_{\text{max}}$($CW_{rel}$) 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 $PSF_{\text{max}}(CW_{\text{rel}})$ features of multiple feature rankings generated during $CW_{\text{rel}}$ calculation to form an ensemble feature ranking, which further enhanced the FS reproducibility on the basis of maximizing the $CW_{\text{rel}}$.

**Determining the optimal signature by AUC-based golden section search**

The ensemble feature ranking is generated with enhanced reproducibility by maximizing $CW_{\text{rel}}$ and ensemble learning. However, the generalizability of the resulting ensemble feature ranking has not been assessed yet. Moreover, under some circumstances, the overall $CW_{\text{rel}}$ is significantly low [109], so that even if the top $PSF_{\text{max}}(CW_{\text{rel}})$ 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 4a, the reproducibility was assessed by the mean value of $CW_{\text{rel}}$ 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 4b 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 $CW_{\text{rel}}$ 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 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 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 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 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 top 10% 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.

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 diagnostic 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 CWrel 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 CWrel, 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 statement

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 (2018QA7023); ‘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


