Supplementary Information

Machine Learning Empowers Phosphoproteome Prediction in Cancers

Hongyang Li[1,*], Yuanfang Guan[1,*]

1. Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109, USA

* Corresponding author: hyangl@umich.edu, gyuanfan@umich.edu

Availability of data and material

Breast cancer proteomic and phosphoproteomic data:
https://cptac-data-portal.georgetown.edu/cptac/s/S015
Ovarian cancer proteomic and phosphoproteomic data:
https://cptac-data-portal.georgetown.edu/cptac/s/S020
Breast cancer genomic data:
https://portal.gdc.cancer.gov/legacy-archive/search/f?filters=%7B%22op%22:%22and%22,%22content%22:%5B%7B%22op%22:%22in%22,%22content%22:%7B%22field%22:%22cases.project.program.name%22,%22value%22:%5B%22TCGA%22%5D%7D%7D,%7B%22op%22:%22in%22,%22content%22:%7B%22field%22:%22cases.project.project_id%22,%22value%22:%5B%22TCGA-BRCA%22%5D%7D%7D%5D%7D
Ovarian cancer genomic data:
https://portal.gdc.cancer.gov/legacy-archive/search/f?filters=%7B%22op%22:%22and%22,%22content%22:%5B%7B%22op%22:%22in%22,%22content%22:%7B%22field%22:%22cases.project.program.name%22,%22value%22:%5B%22TCGA%22%5D%7D%7D,%7B%22op%22:%22in%22,%22content%22:%7B%22field%22:%22cases.project.project_id%22,%22value%22:%5B%22TCGA-OV%22%5D%7D%7D%5D%7D
Challenge dataset repository:
https://www.synapse.org/#!Synapse:syn8228304/wiki/448379

Proteomic and phosphoproteomic data generation

In this study, the proteomic and phosphoproteomic data were downloaded from CPTAC data portal. For breast cancer, 77 samples were acquired at the Broad Institute (BI); for ovarian cancer, 69 samples were acquired at Pacific Northwest National Laboratory (PNNL). These data were processed by the common data analysis pipeline from CPTAC, which was described in detail in the original CPTAC publications (Zhang *et al.*, 2016; Mertins *et al.*, 2016). In brief, the tumor sample tissues were first digested into peptide with trypsin. The digested samples were labeled with 4-plex iTRAQ (Ross *et al.*, 2004) and fractionated by basic reversed phase liquid chromatography (LC) to reduce sample complexity. The LC separated samples were subsequently used in the LC-MS/MS system for proteome analysis. For phosphoproteomic data, phosphopeptides were enriched using immobilized metal affinity chromatography (IMAC) before the LC-MS/MS step. MS analysis was performed using the Thermo Scientific LTQ Orbitrap Velos mass spectrometer. Thermo RAW files were processed with DeconMSn (v2.3.0) to determine the m/z values and charge of the precursor ions and saved as CDTA files. Then the CDTA files were processed with DTARefinery (Petyuk *et al.*, 2010) to correct for instrument

calibration errors. The database search engine MSGF+ (Kim and Pevzner, 2014) was used to match the CDTA files against the RefSeq human protein sequence database (release version 37). Subsequently, peptides were identified and assembled into proteins using IDPicker (Ma *et al.*, 2009). The maximum false discovery rate (FDR) was set to 1% at peptide level and a minimum of 3 unique peptides was required to identify a protein. The intensities of iTRAQ reporter ions were extracted using MASIC (Monroe *et al.*, 2008) and the ratio of sample abundance to reference abundance from ion intensities was calculated as the relative protein abundance. The relative abundances were further log2 transformed to obtain the relative expression values. To reduce the sample-specific bias, the protein abundances from the same sample were re-centered to achieve a common median of 0. For the phosphoproteomic data, the phosphorylation site was localized using Ascore algorithm (Beausoleil *et al.*, 2006). The final phosphorylation abundance was calculated by subtracting the log2 relative abundance of the parent protein from the phosphopeptide. This step could remove the confounding effect of changes in expression of parent proteins and reflect the changes in the phosphorylation rates.

Ensemble weights selection

The ensemble weights of (A) "proteome", (B) "site-specific", (C) "cross-tissue", (D) "multisite" models were selected by the following steps. We first combined model (B) and (C) using the ratio of 1:1, since they were similar random forest model but trained on different data (within-tissue or cross-tissue data). The temporary result was named as (B+C; 1:1). Then we adjusted this result with model (D), which was the weighted average of all phosphorylation sites from the same protein. The combining ratio between (B+C; 1:1) and (D) was 5:1=10:2=(5+5):2 at this step, since the "multisite" model was similar to a refinement instead of a new machine learning model and we assigned it with a relatively small weight. This temporary results was named as (B+C+D; 5:5:2). Finally, we combined model (A) and (B+C+D; 5:5:2) with the ratio of 1:3=4:12=4:(5+5+2), which was the best result of a grid search only at this step. The resulting final ensemble weights are A+B+C+D = 4:5:5:2. By the above strategy, we generated a robust model of high performance on both the cross-validation dataset and the hold-out test dataset in the DREAM Proteogenomics challenge.

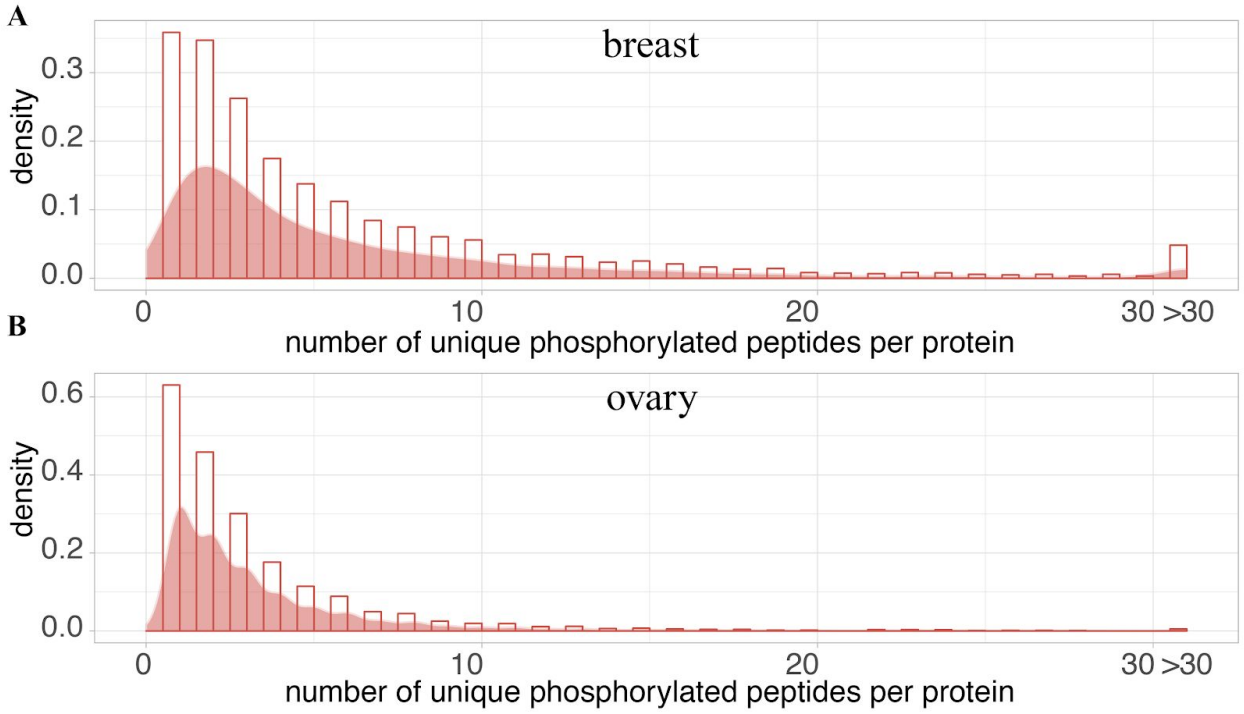

Fig. S1 The number of observed phosphorylation sites per protein in two cancers.
The distributions were shown as histograms and density plots for A. breast and B. ovary. In general, most proteins had less than 10 unique phosphorylated peptides, whereas a small portion of proteins had more than 30 unique ones.

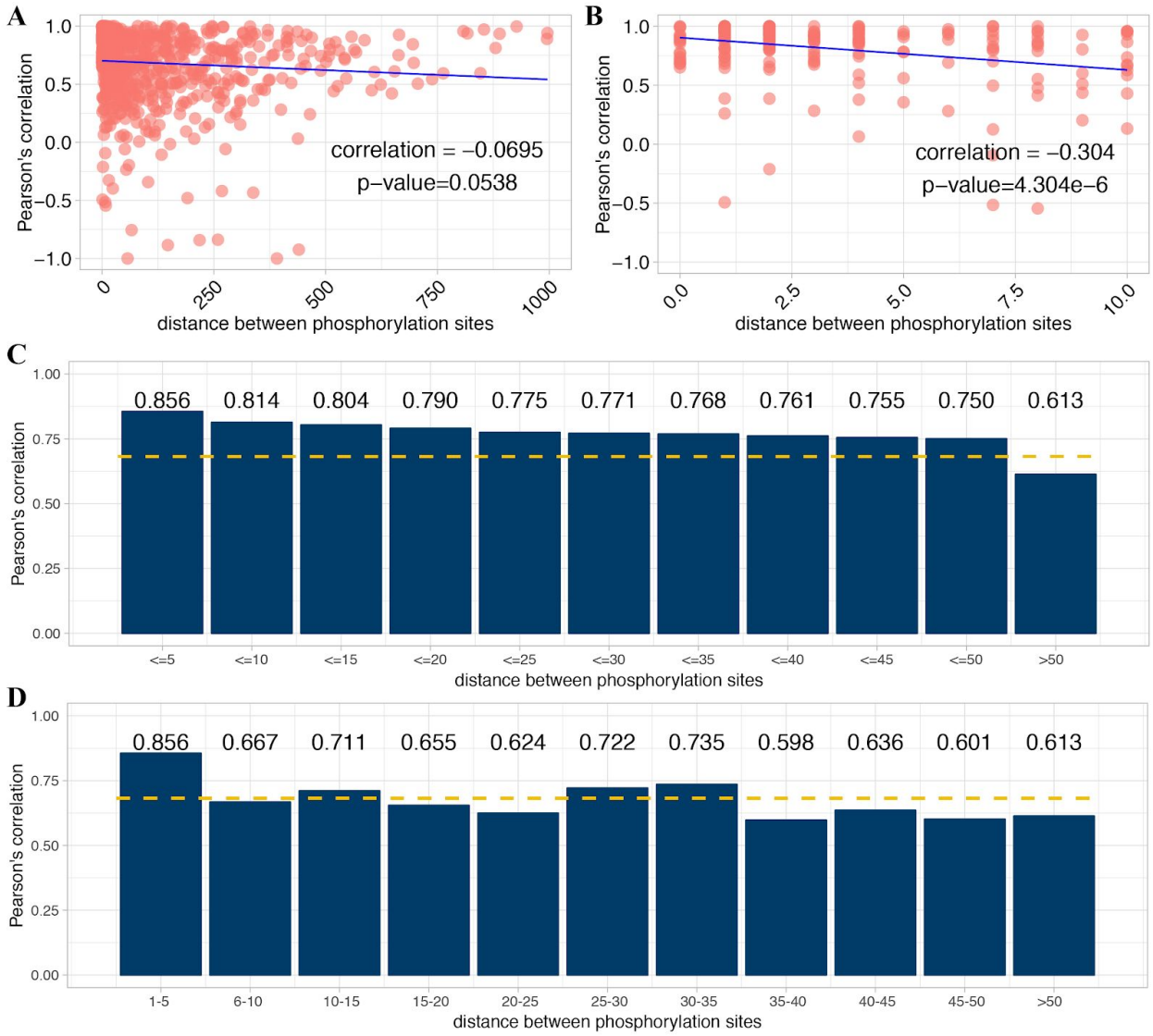

Fig. S2 The cross-sample correlation between multiple phosphorylation sites of the same protein.

A. Overall, the between-site correlation did not significantly correlated with the distance between two phosphorylation sites. **B**. When we only considered phosphorylation sites less than 10 amino acid away from each other, we observed a weak but significant negative correlation of -0.304. **C-D**. In general, phosphorylation sites close to each other had stronger correlations, with the average correlation of 0.682 shown as the yellow dashed line. The value above each bar is the average Peason's correlation for each bin.

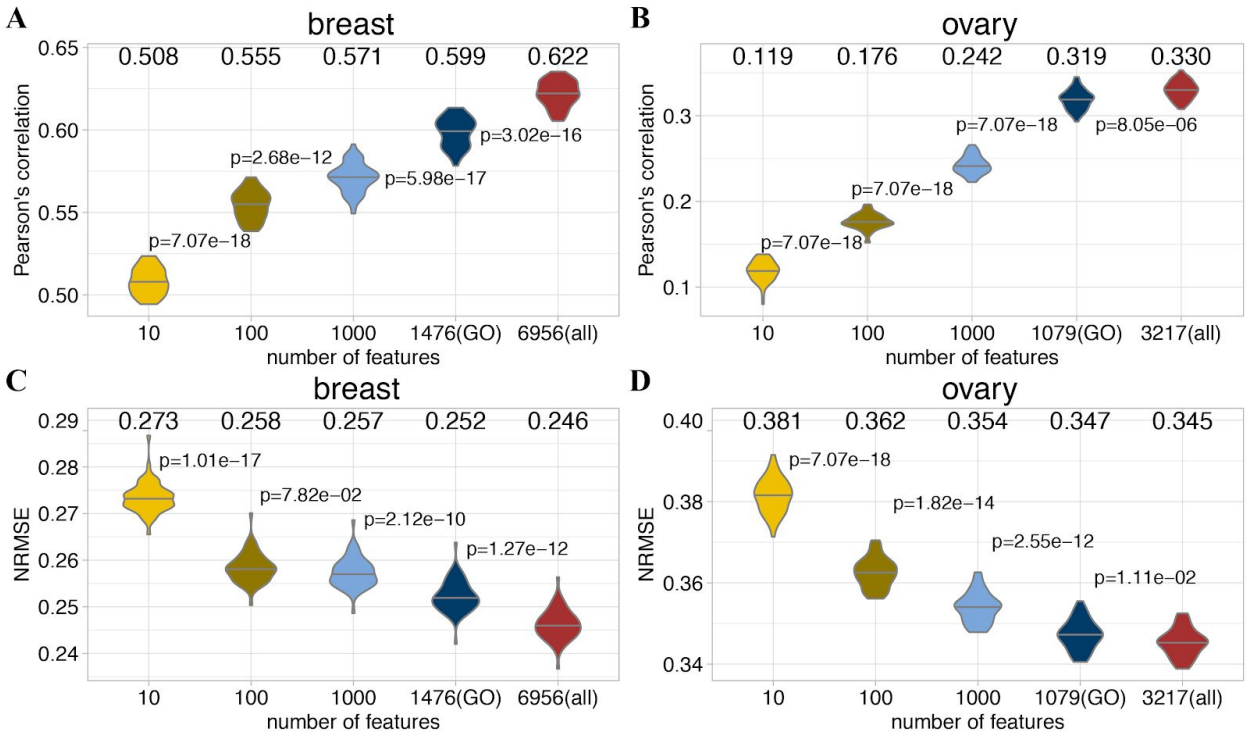

Fig. S3 The prediction performance of models using different number of proteins as features.

From left to right, the performance was compared in violin plots of models using the top "10", "100", "1000" expressed proteins, phosphorylation-related proteins defined by GO terms, and all available proteins. The Pearson's correlations between these approximations and experimental observations were calculated in **A.** breast and **B.** ovary. The Wilcoxon signed-rank tests were performed for every pair of adjacent models, and the p-values were shown between two violin plots. Similarly, the NRMSE were also calculated in **C.** breast and **D.** ovary.

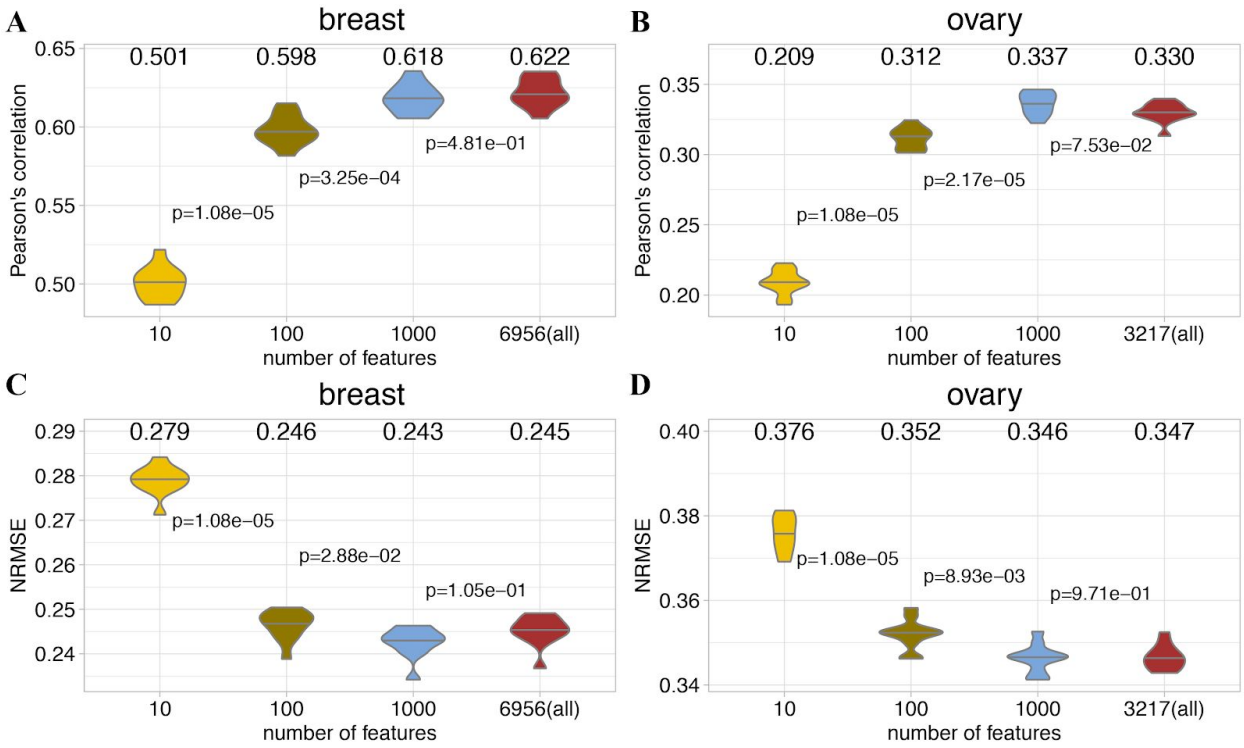

Fig. S4 The prediction performance of models using top proteins with the highest variances as features.

From left to right, the performance was compared in violin plots of models using the top "10", "100", "1000" proteins with the highest variances across samples, and all available proteins. The Pearson's correlations between these approximations and experimental observations were calculated in **A.** breast and **B.** ovary. The Wilcoxon signed-rank tests were performed for every pair of adjacent models, and the p-values were shown between two violin plots. Similarly, the NRMSE were also calculated in **C.** breast and **D.** ovary. Of note, the performances between the "1000" and "all" models did not have significant differences (p-values > 0.05).

Fig. S5 The cross-dataset prediction performance.

The "final test breast" and "final test ovary" datasets are the hold-out test datasets used in the final evaluation during the DREAM Proteogenomics Challenge. Of note, these two scores were not cross-validation results so that they were shown as two lines instead of violin plots. For the "cross-tissue breast" and "cross-tissue ovary" datasets, we trained our model on one dataset and evaluated it against the other dataset in a five-fold cross validation fashion.

References

Beausoleil,S.A. *et al.* (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.*, **24**, 1285–1292.

Kim,S. and Pevzner,P.A. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.*, **5**, 5277.

Ma,Z.-Q. *et al.* (2009) IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J. Proteome Res.*, **8**, 3872–3881.

Mertins,P. *et al.* (2016) Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, **534**, 55–62.

Monroe,M.E. *et al.* (2008) MASIC: A software program for fast quantitation and flexible visualization of chromatographic profiles from detected LC–MS(/MS) features. *Computational Biology and Chemistry*, **32**, 215–217.

Petyuk,V.A. *et al.* (2010) DtaRefinery, a software tool for elimination of systematic errors from parent ion mass measurements in tandem mass spectra data sets. *Mol. Cell. Proteomics*, **9**, 486–496.

Ross,P.L. *et al.* (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics*, **3**, 1154–1169.

Zhang,H. *et al.* (2016) Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell*, **166**, 755–765.