Supplementary Information for "PTMscape: an open source tool to predict generic post-translational modifications and map modification crosstalk in protein domains and biological processes"

Ginny Li, Christine Vogel, Hyungwon Choi

Supplementary Table 1. Performance evaluation of the linear SVMs across five modification types (acetylation, methylation, ubiquitination, SUMOylation, and phosphorylation) using three different window sizes (11, 15, 25 amino acids). AUC and sdAUC denote area under the curve of the ROC curve and its standard deviation estimated from bootstrap samples. MCC is Matthew's correlation coefficient.

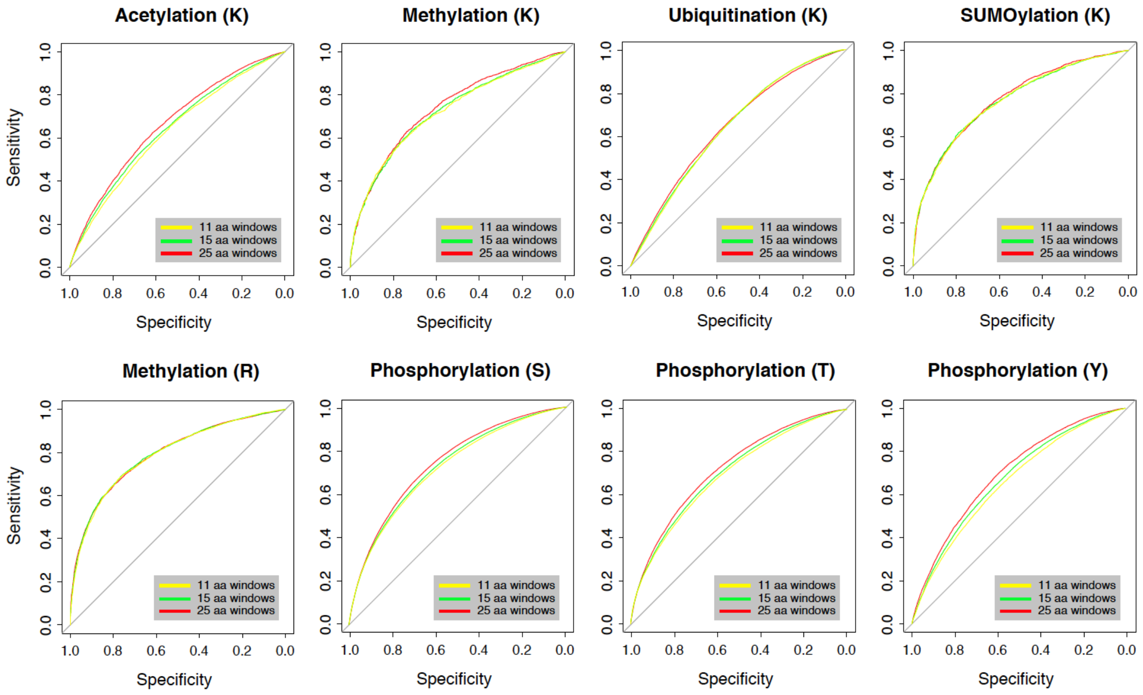
Supplementary Table 2. Prediction performance comparison between PTMscape and ModPred. The number of proteins considered for different modifications vary because each target residue does not appear in all protein sequences.

Supplementary Table 3. Comparison of linear SVM against Artificial Neural Network (ANN) and Random Forest (RF). The three methods are compared in terms of their (i) prediction accuracy in terms of AUC, MCC, sensitivity of detection at score thresholds associated with 99% specificity, and F-measure, and (ii) total computation time. Computation time for RF and ANN was measured on a standard linux computer using the corresponding implementations in R.

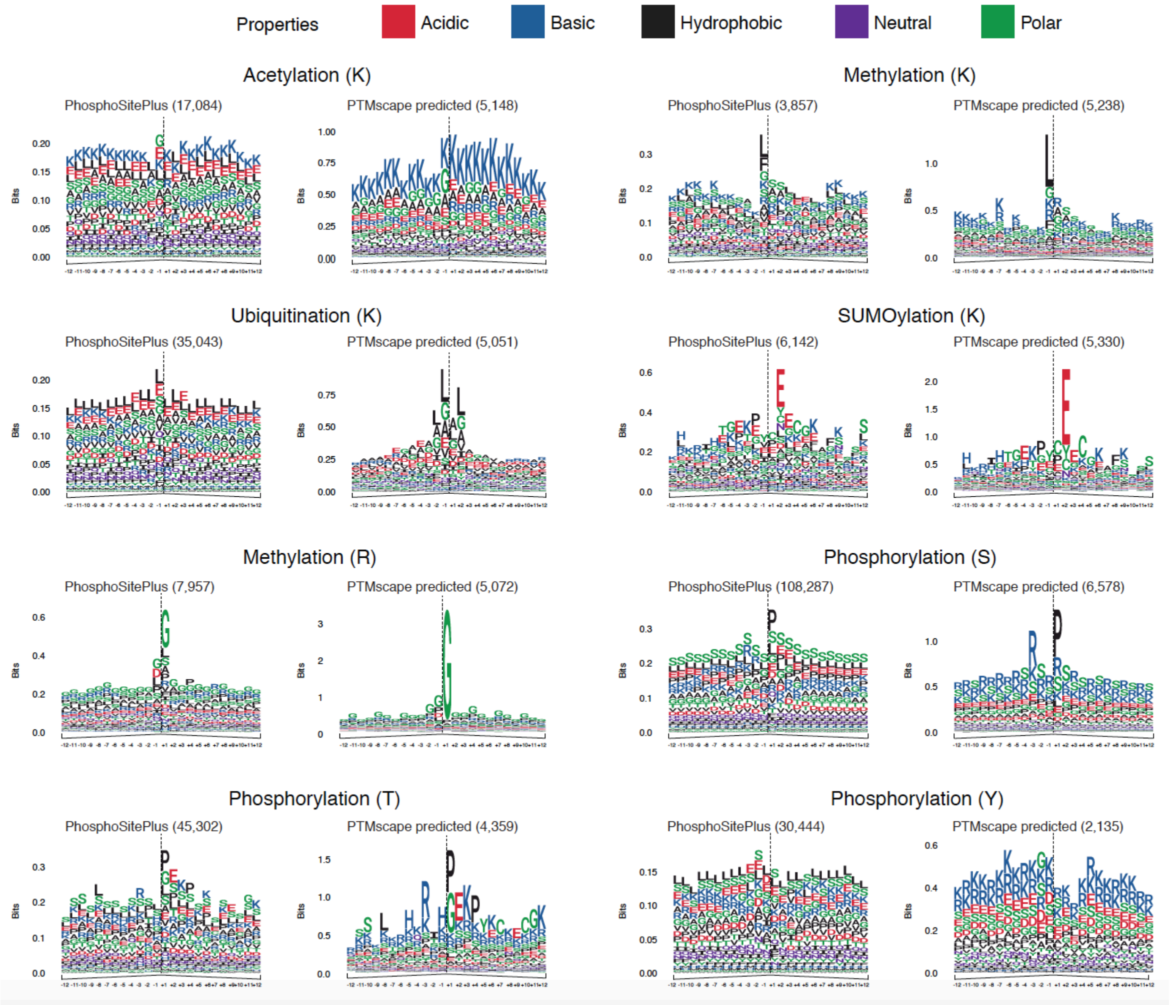
Supplementary Table 4. Protein domains significantly enriched in the combined set of PTM sites (sites in the PSP database and new predictions) per modification type. This output can be directly produced from a post-prediction module in PTMscape.

Supplementary Table 5. Protein domains significantly enriched in the positive and negative crosstalk sites obtained from the combined set of PTM sites (sites in the PSP database and new predictions). This output can also be directly produced from a post-prediction module in PTMscape.

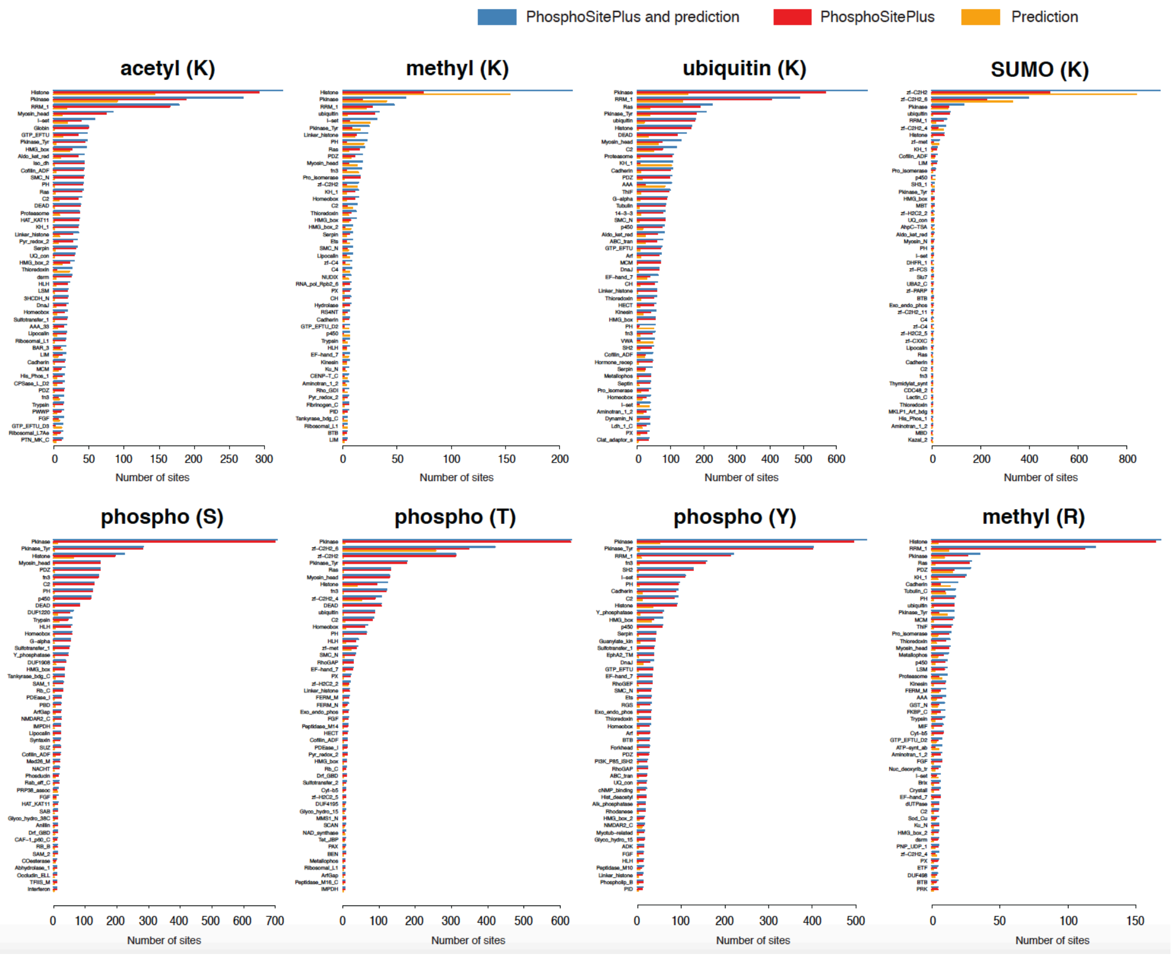
Supplementary Table 6. Amino acid index table after hierarchical clustering-based merging of features to remove correlation. The source and short description of each feature and their clustering information is listed in the "Cluster_composition" tab, and the actual values summarized at the cluster level are listed in the "Cluster_values" tab.



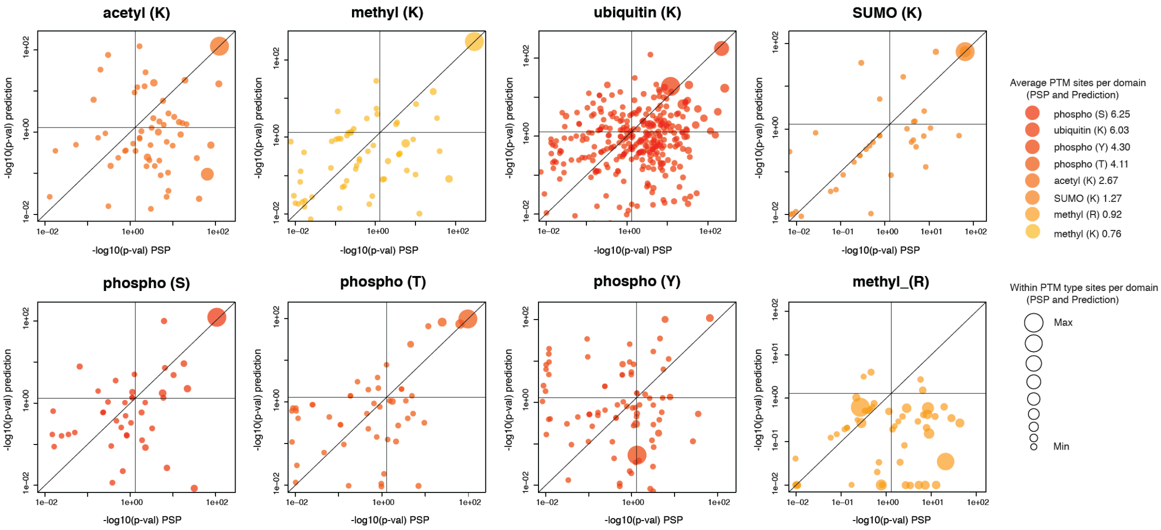
Supplementary Figure 1. Receiver operating-characterstic (ROC) curves for 5 PTMs using three different window sizes in the 10-fold cross-validation. Although long-range windows (25 amino acid long) made difference in some PTMs (e.g. tyrosine phosphorylation), the overall performance was relatively similar across the window sizes.



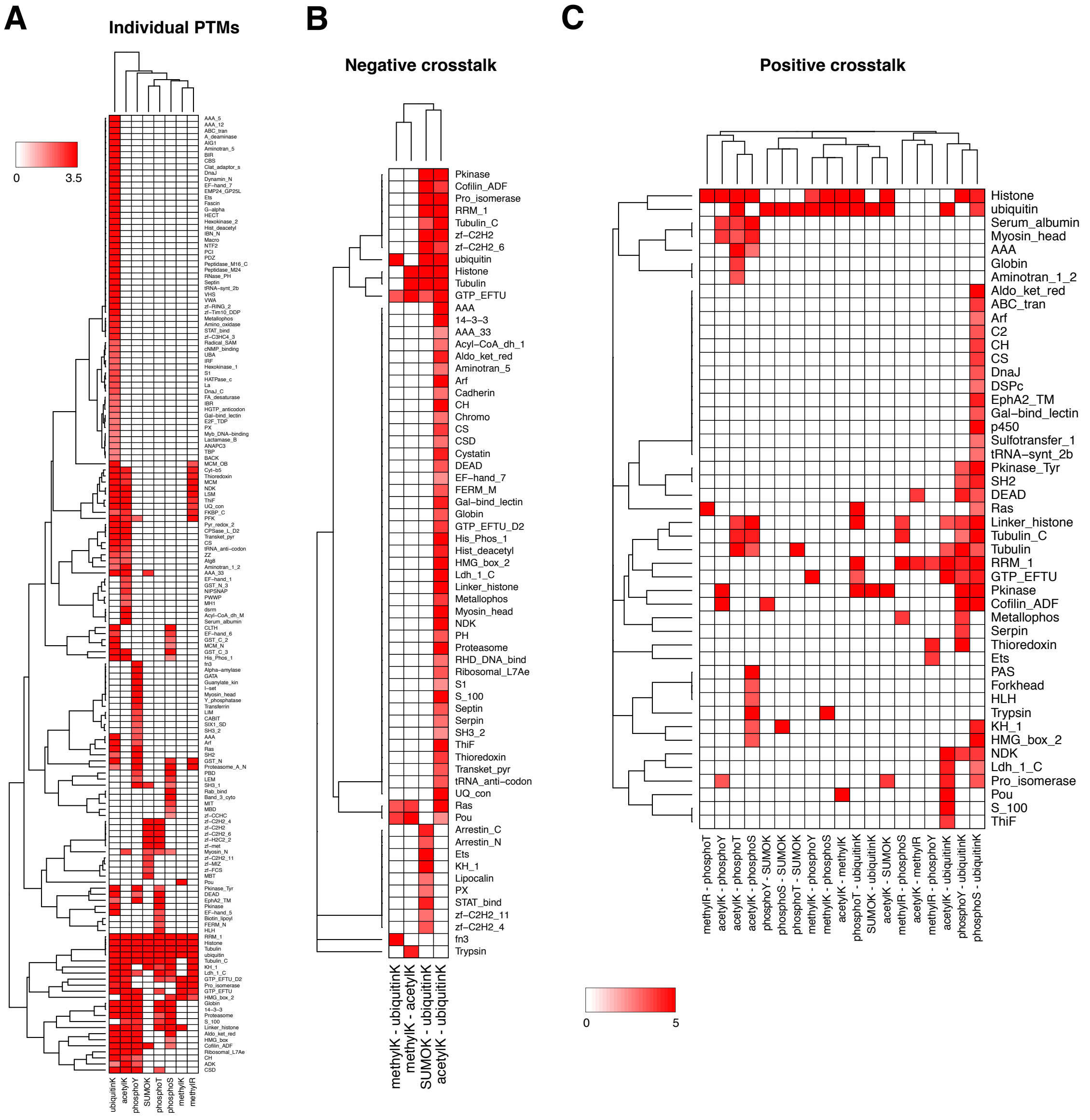
Supplementary Figure 2. Sequence logo plots of the motifs obtained from PSP reported sites (left) and new predictions (right). The y-axis shows bits (information content), whose scale varies depending on multiple factors such as total number of sequences and relative frequency of each amino acid. The counts in the PTMscape predicted windows include the windows that score above the threshold and also appear in the PSP database (as they score high).



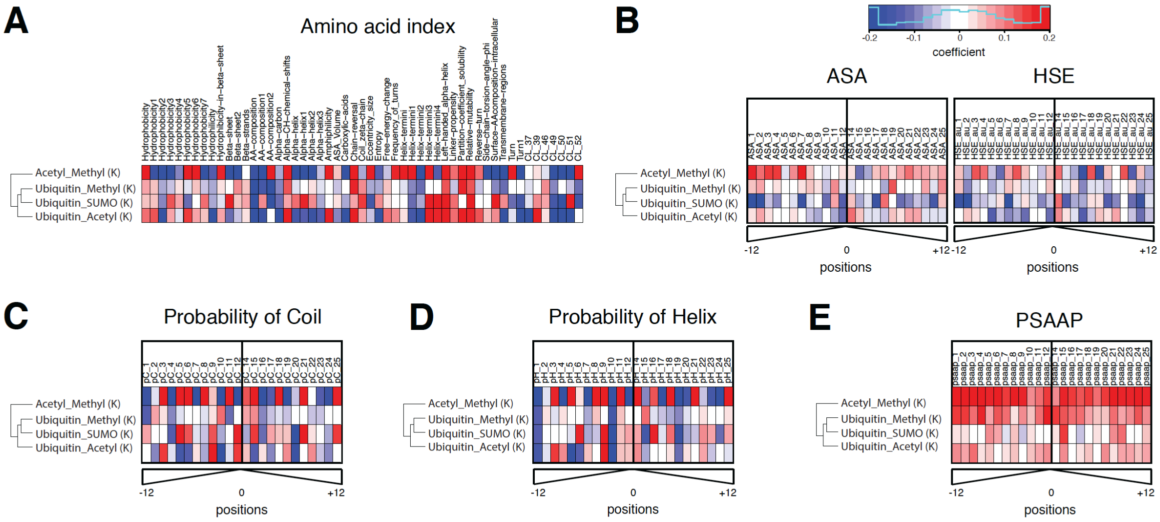
Supplementary Figure 3. The number of PTM sites in Pfam domains (top 50 in each PTM type) counted from the combined set of PTM sites in the experimentally acquired PSP database and the new predictions (blue bar), and separately from the two sets of PTM sites (red bar for PSP, orange bar for predictions). The domains were selected and ordered by the counts of sites from the combined set (blue bar).



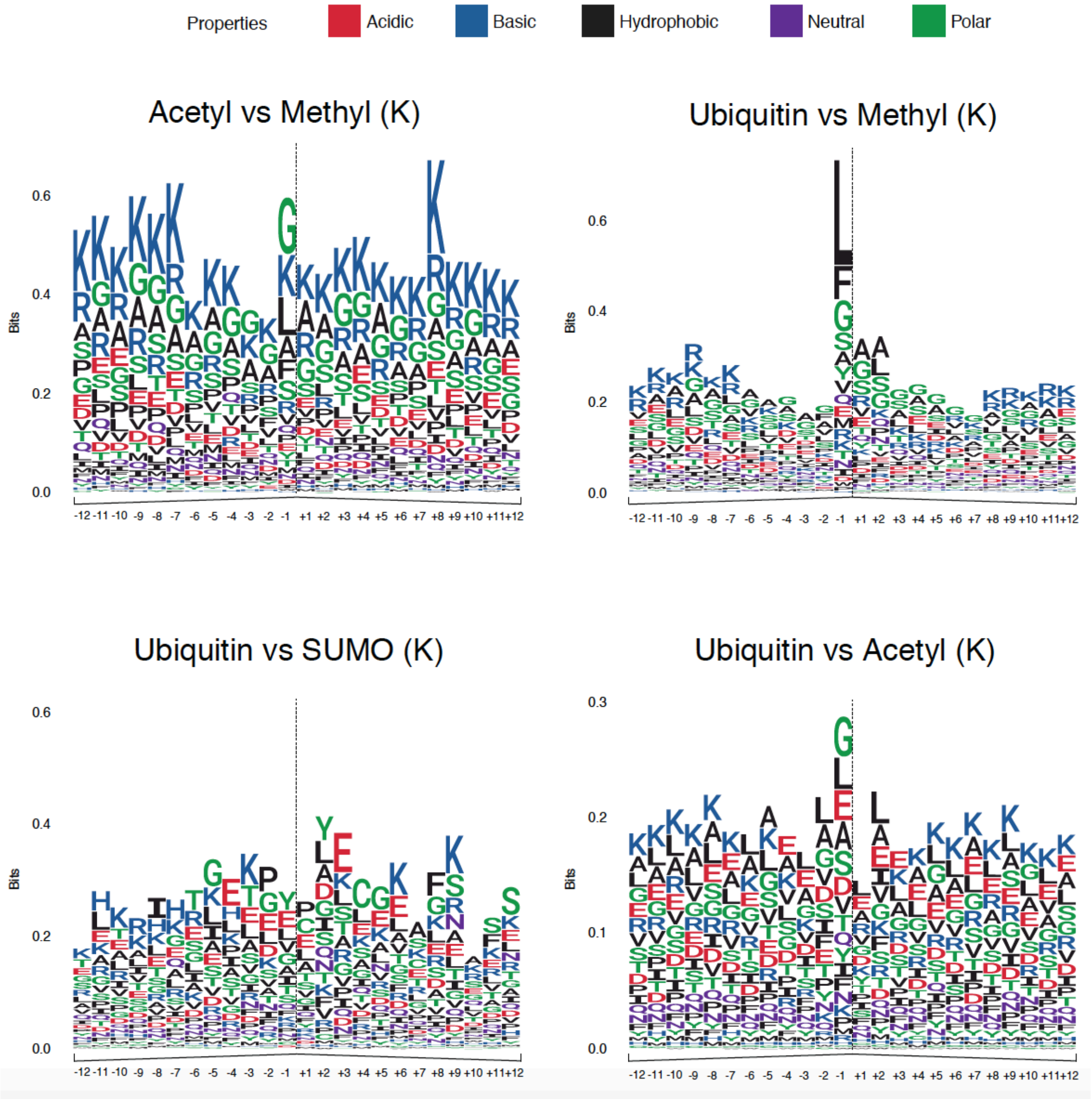
Supplementary Figure 4. Comparison of statistical significance scores for the enrichment of the Pfam domains between the PTM sites in the (experimentally acquired) PSP database (x-axis) and the new predictions (y-axis). The axes are minus the logarithm of p-values (base 10) from hypergeometric tests. The size of circle reflects the number of sites occurring in each domain.



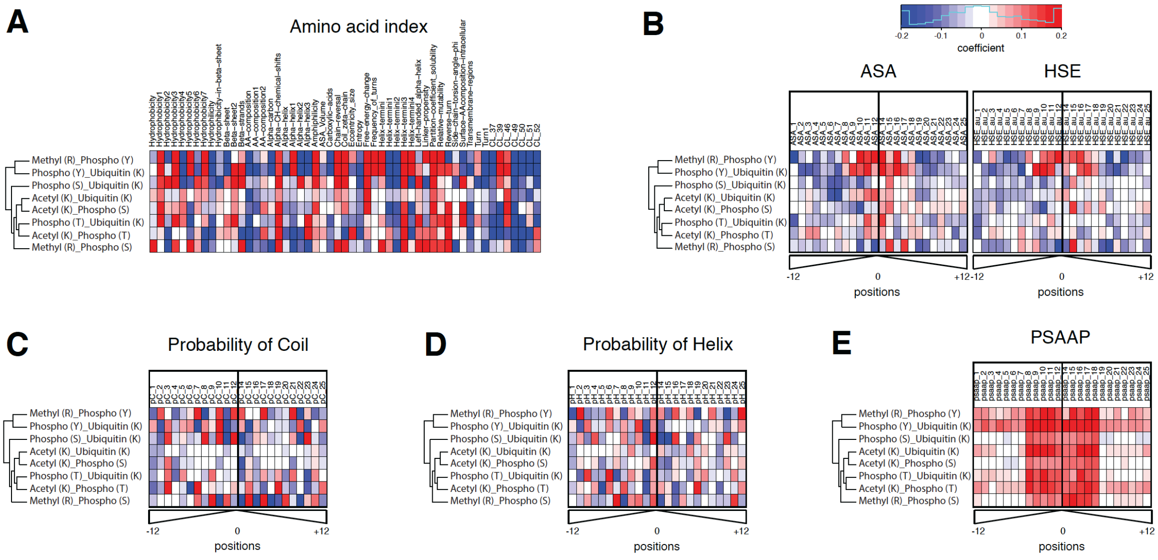
Supplementary Figure 5. Protein domains enriched in (**A**) individual PTM sites, (**B**) negative crosstalk sites, and (**C**) positive crosstalk sites. The color of the heatmaps was –log10 of *q*-values computed from chi-squared tests (**A**) and Fisher exact tests (**B, C**).



Supplementary Figure 6. The feature weight coefficients obtained from the linear SVM analysis of negative crosstalk between four pairs of lysine modifications in heatmaps. Following the same style of presentation of **Figure 2**, the heatmaps were organized into six different sets of features, including (**A**) amino acid indexes, (**B**) accessible surface area (ASA) and half-sphere exposure (HSE), (**C**) probability of coil and (**D**) probability of helix, and (**E**) Position-specific amino acid propensity (PSAAP) information obtained from known and additionally predicted sites. The names and description of amino acid index clusters can be found in **Supplementary Table 6**. The hierarchical clustering of five PTM types was performed using all variables.



Supplementary Figure 7. Sequence logo plots of the motifs obtained from the negative crosstalk sites. The y-axis shows bits (information content), whose scale varies depending on multiple factors such as total number of sequences and relative frequency of each amino acid.



Supplementary Figure 8. The feature weight coefficients obtained from the linear SVM analysis of positive crosstalk between eight pairs of PTMs in heatmaps. Following the same style of presentation of **Figure 2**, the heatmaps were organized into six different sets of features, including (**A**) amino acid indexes, (**B**) accessible surface area (ASA) and half-sphere exposure (HSE), (**C**) probability of coil and (**D**) probability of helix, and (**E**) Position-specific amino acid propensity (PSAAP) information obtained from known and additionally predicted sites. The names and description of amino acid index clusters can be found in **Supplementary Table 6**. The hierarchical clustering of five PTM types was performed using all variables.