Supplementary methods

Residue co-evolution

In addition to mutual information, we also considered the two other approaches to measure the residue co-evolution.

For a pair of human PTM sites, let A denote all species in the MSA and let C denote “species with co-evolving residue pair” which are defined as those that (1) have the same residue as human at both sites; or (2) have residues different to human at both sites.

1-nHMdist The Hamming distance for a pair of human PTM sites in the MSA is defined as the number of species without co-evolving residue pair $|A \setminus C|$. The 1-normalized Hamming distance (1-nHMdist) is then the proportion of species with co-evolving residue pair, and can be used as a measure of the residue co-evolution:

$$1\text{-}nHMdist = \frac{|C|}{|A|} \quad \text{(S1)}$$

In the example of S15-T18 pair in Figure 2A, the Hamming distance is 2, which is contributed by two species: *O. Latipesand* and *G. Aculeatus*. 1-nHMdist score is therefore 1-2/19=0.895.

nCoMBL The MSA of each human protein can be associated with a phylogenetic tree that describes the evolutionary relationship between the species in the MSA. The topologies and branch lengths of the phylogenetic trees were downloaded from the ‘tree’ dataset of veNOG in the eggNOG v4.0 database (1). For a pair of human PTM sites, the maximum branch length (MBL) between the two species with co-evolving residue pair divided by the MBL of any two species in the MSA can also be used as a measure of residue co-evolution:

$$nCoMBL = \frac{\max_{i \neq j \in C}\{B(i,j)\}}{\max_{i \neq j \in A}\{B(i,j)\}} \quad \text{(S2)}$$

where *B(i, j)* is the branch length between two species *i* and *j* in the phylogenetic tree.

PTM imputation

The PTM databases are still incomplete, especially for non-human species. So the conservation of modification status may be under-estimated due to false negatives in other species. To overcome this issue when evaluating the modification co-evolution, we proposed the following imputation strategy. For a human PTM site *j* in the multiple sequence alignment (MSA), in addition to using an indicator variable to flag the presence of PTM of an amino acid residue in species *i*, we also calculated the probability of PTM for the species *i* that had no modification found in the database but had the same residue as a specie with observed modification:

$$s_{i,j} = \begin{cases} 1 & \cdots \quad M_{i,j} = 1 \\ P_{i,j} & \cdots \quad M_{i,j} = 0 \text{ and } R_{i,j} = 1 \\ 0 & \cdots \quad M_{i,j} = 0 \text{ and } R_{i,j} = 0 \end{cases} \quad \text{(S3)}$$

where $M_{i,j}$ is the indicator of observing PTM at site *j* in species *i*; $R_{i,j}$ is the indicator that residue at site *j* is modified in at least one other species whose residue is the same species *i*.

To derive an formulation of $P_{i,j}$, we relied on the following intuitions: (1) the modification status is more likely conserved if the sequence context of a PTM site is conserved; (2) when compared to the PTM sites observed across the species, proteins with lower number of PTMs have higher chance of missing data. To support the first intuition, we found that among PTM sites shared by human and mouse, 65.4% were fully conserved in the 9 amino acids window centered around the PTM site, and 68.4% for PTM sites shared by human and rat, 84.6% for mouse and rat. The second intuition can be supported by the findings that despite the poor conservation of the modification of single PTM sites, the overall number of PTM sites in proteins is conserved across yeast species (2).

For a protein that can be post-translationally modified, we defined its *potential PTM sites* as the columns in the MSA where PTMs were observed in at least one species, and use *tot* to denote the number of potential PTM sites. In case of two species, mouse and human, to impute unobserved PTM at the potential PTM sites of a protein in mouse, let *obs* be the observed number of PTMs, and *k* be the unknown number of real PTM sites of that protein in mouse. Then for a human PTM site *j* in the protein, the probability of PTM at the site *j* in mouse is then calculated by:

$$P_j = I^j_{human,mouse} \cdot SeqSim^j_{human,mouse} \cdot \sum_{k=obs}^{tot} p(k,tot)\frac{k-obs}{tot-obs} \quad \text{(S4)}$$

where $I^j_{human,mouse}$ is the indicator variable to indicate that the amino acid residue at site *j* in the mouse protein is the same as human; $SeqSim^j_{human,mouse}$ is the fraction of identical residues between human and mouse in the 9 amino acids windows centered around the site *j*; the second term in the equation S4 is the expectation of $\frac{k-obs}{tot-obs}$, which reflects the prior probability of (*k-obs*) unobserved PTM sites in that protein. In effect, this equation transfers the observed PTM sites from human to mouse by integrating the sequence conservation and priors of unobserved PTM at the potential PTM sites.

More generally, for any one of the potential PTM sites *j* in a protein, if the PTM is unobserved in the species *i*, then its probability can be calculated by:

$$P_{i,j} = \max_{k\in\Omega}\left\{ I^j_{i,k} \cdot SeqSim^j_{i,k} \cdot \sum_{k_i=obs}^{tot} p_i(k_i,tot)\frac{k_i-obs_i}{tot-obs_i} \right\} \quad \text{(S5)}$$

where Ω is the set of species other than *k* for which PTM is observed at site *j*, $I^j_{i,k}$ the indicator variable to indicate that the site *j* is conserved between species *i* and *k*; $SeqSim^j_{i,k}$ is the fraction of identical residues between species *i* and *k* in the 9 amino acids window centered around the site *j*; obs_i is the observed number of PTMs in species *i*; k_i is the unknown number of real PTM sites.

$p_i(k_i,tot)$ is the probability of having k_i real PTM sites in a protein of species *i* which has *tot* number of potential PTM sites. We approximated $p_i(k_i,tot)$ by

$$p_i(k_i,tot) = \int_{\frac{k_i-1}{tot}}^{\frac{k_i}{tot}} f_i(p)dp \quad \text{(S6)}$$

where $f_i(p)$ is the probability density of the fraction of PTM sites (*p*) in a protein in species *i*. Because of the incomplete coverage of PTM sites, using observed PTM sites to estimate $f_i(p)$ will result in under-estimation. So we also considered all potential PTM sites whose $I^j_{i,k}=1$ and $SeqSim^j_{i,k}=1$ as the surrogate of real but unobserved PTM sites for species *i*, and used the union of both the observed and surrogate PTM sites to estimate $f(p)$. Figure S7 shows the estimated probability density of *p* for human, mouse and rat. For human, estimated *f*(*p*) using only observed or both observed and surrogate PTM sites are similar; which is expected since human has the most complete catalog of PTM sites (Figure S7A). For other species, mouse has less observed fraction of PMT sites than human, and rat has the least observed fraction, possibly because PTM has not been well characterized in these species. Nevertheless, adding surrogate PTM sites to the observed ones increases the mean of *f*(*p*) and also results in a more comparable distribution of *f*(*p*) between mouse and rat (Figure S7B,C).

We used both the observed status and imputed probability (Eq S3) to calculate the modification co-evolution using Eq (3) in main text. Compared with only using observed data, the use of imputation could better discriminate the cross-talk and control pairs (AUC=0.644 using actual data and AUC=0.662 using imputed data, Figure 4 and Supplementary Figure S6). However, when combined with other features, using the revised measure of modification co-evolution did not improve the prediction performance of the integrated model (AUC=0.833 using actual data and AUC=0.829 using imputed data, Figure 4 and Supplementary Figure S6). It may be explained by the fact that imputation strategy makes use of the sequence conservation, which was already been accounted for by other features. Therefore, better approaches for imputing modification status may be needed to improve the prediction performance of the integrated model.

Reference

1. Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., Gabaldón, T., Rattei, T., Creevey, C., Kuhn, M., Jensen, L. J., von Mering, C., and Bork, P. (2013) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Research*, gkt1253

2. Beltrao, P., Trinidad, J. C., Fiedler, D., Roguev, A., Lim, W. A., Shokat, K. M., Burlingame, A. L., and Krogan, N. J. (2009) Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. *PLoS biology* 7, e1000134