Accurate, high-coverage assignment of *in vivo* protein kinases to phosphosites from *in vitro* phosphoproteomic specificity data

**Supplemental Materials**

Brandon M. Invergo

1 Translational Research Exchange @ Exeter, University of Exeter, Stocker Road, Exeter EX4 4QJ

*Correspondence: b.invergo@exeter.ac.uk

**Supplemental Methods**

**Data**

*In vitro* kinase substrates were retrieved from ref. [1], excluding mutant kinases and lipid kinases. Autophosphorylation events were omitted from the analysis. In order to stringently control for false positives, kinase substrates were filtered to include only those that passed both the authors’ “SIDIC” and “PTMscore” ($P > 0.75$) criteria [1]. A high-confidence set of human phosphosites was retrieved from ref. [2]. Literature-derived, human phosphosites and kinase-substrate relationships were retrieved from the PhosphoSitePlus database (version 020922) [3] and from ref. [4], and kinase-substrate relationships filtered to remove autophosphorylation.

Kinase physical and indirect interactions were derived from a merged network of interactions from the BioGRID database (version 4.3.194) [5], the BioPlex 293T network (December 2019) [6], and a mass-spectrometry-based assay of kinase interactions [7]. BioGRID interactions were filtered to exclude biochemical reactions (“Experimental System” = “Biochemical Activity”) and genetic interactions. All networks were filtered to exclude self-binding. Protein domains were retrieved from Pfam (version 33.1) [8]. Protein functional association scores were retrieved from the STRING database (version 11.0) [9]. Kinase identities (for IV-KAPhE “substrate is a kinase” feature) and types (S/T or Y) were derived from the *pkinfam.txt* file provided by UniProt and Swiss-Prot (release 2020_02; https://www.uniprot.org/docs/pkinfam). For this purpose, dual-specificity kinases were counted as S/T kinases.
Gene Ontology semantic similarity

Semantic similarity of Gene Ontology (GO) terms between a kinase and a putative substrate protein was calculated using the “GOSemSim” package for R (version 2.18.0) [10, 11]. Semantic similarity was calculated using the “Wang” measure and combined using the “BMA” method. Missing values (due to lack of annotation) were imputed from the full in vivo feature table via multivariate imputation by chained equations using the “mice” package for R (40 maximum iterations; version 3.13.0) [12].

STRING functional association scores

For all STRING component scores considered (“coexpression”, “experimental”, “cooccurrence”, “fusion”, and “neighborhood”), values were transformed to be between 0 and 1 by dividing by 1000. Missing values were set to the prior value of 0.041 used by the STRING developers. STRING uses Ensemble identifiers for proteins, whereas the rest of the analysis used Uniprot identifiers. In some cases, many Ensembl identifiers can map to the same Uniprot identifier. In these cases, the means of the scores assigned to all relevant identifiers were used.

Kinase specificity models

Position Frequency Matrix

Sequence windows were first weighted using a position-based sequence-weighting method [13]. Given a set of $n \geq 20$ substrate sequence windows, $S = S_1, S_2, \ldots, S_i, \ldots, S_{n-1}, S_n$, where $S_i = S_{i1}, S_{i2}, \ldots, S_{i14}, S_{i15}$ and $S_{ij}$ represents the amino acid at position $j$ of sequence $i$, a weight was assigned to each amino acid $a$ at position $j$:

$$w(a, j) = \frac{1}{c_j \sum_{i=1}^{n}(S_{ij} = a)}$$  \hspace{1cm} (1)$$

where $c_j$ is the number of unique amino acids found at position $j$ among the substrates in $S$. Phosphosites near the 5' or 3' tail of the protein may not have 7 residues adjacent to them to form a complete sequence window. At each missing tail position, a 21st amino acid was counted. A weight was calculated for each sequence as the sum of its position-specific residue weights:

$$W(S_i) = \sum_{j=1}^{15} w(S_{ij}, j)$$  \hspace{1cm} (2)$$

Finally, each sequence’s weight was normalized:

$$\hat{W}(S_i) = \frac{W(S_i)}{\sum_{k=1}^{n}W(S_k)}$$  \hspace{1cm} (3)$$

A $20 \times 15$ counts matrix, $r$ was constructed such that entry $r_{aj}$ contains the weighted count of amino acid $a$, ignoring missing tails, at position $j$ across the sequences in $S$:  

2
\[
\begin{align*}
  r_{a,j} &= n \sum_{i=1}^{n} V(S_{ij}, a) \\
  V(S_{ij}, a) &= \begin{cases} 
    \hat{W}(S_i) & \text{if } S_{ij} = a \\
    0 & \text{otherwise}
  \end{cases}
\end{align*}
\]

To account for the fact that \( S \) necessarily represents a limited sample of possible substrate sequences, a total number of pseudocounts, \( B_j \), were distributed at position \( j \) via a \( 20 \times 15 \) pseudocount matrix \( b \) as described below. The final PFM, \( p \) was then calculated from the counts and pseudocounts such that the likelihood of observing amino acid \( a \) at position \( j \) is:

\[
PFM_{a,j} = p_{a,j} = \frac{b_{a,j} + r_{a,j}}{B_j + \sum_a r_{a,j}}
\]

Following Henikoff and Henikoff [14], a position-specific pseudocount strategy was adopted. For column \( j \) the total number of pseudocounts to be distributed among the amino acids is:

\[
B_j = T_j + m \times c_j
\]

where \( T_j \) is the number of missing tails observed in position \( j \) among the substrates in \( S \), \( c_j \) is defined as in Equation 1, and \( m \) is a tune-able parameter, which was fixed at 1 [15]. Note that positions with lower diversity of observed residues will receive fewer pseudocounts.

Pseudocounts were then distributed according to a \( 20 \times 15 \) expected frequency matrix \( f \). In the case of PFMs evaluated directly for predictive performance or used in the construction of phosphoproteome-backed PSSMs, \( f \) was constructed as a PFM (as above) on the full set of \textit{in vitro} substrates from all kinases, \( S_{BG} \), with no pseudocounts \( (B_j = 0) \). In the case of PFMs used in the construction of proteome-backed PSSMs, \( f \) was constructed such that element \( f_{a,j} = F_a \), where \( F_a \) is the proteomic frequency of residue \( a \). Pseudocount matrix \( b \) was then calculated such that residue \( a \) at position \( j \) receives \( b_{a,j} \) pseudocounts:

\[
b_{a,j} = B_j \times f_{a,j}
\]

Finally, for each position \( j \), a column weight \( \hat{w}_j \) was calculated as the relative entropy (Kullback-Leibler divergence) of the residue frequencies at that position, \( p_j \), versus the background expectation, \( f_j \), rescaled to sum to 15 (the number of positions):

\[
\begin{align*}
  w_j &= \sum_a p_{a,j} \log_2 \left( \frac{p_{a,j}}{f_{a,j}} \right) \\
  \hat{w}_j &= w_j \times \frac{15}{\sum_j w_j}
\end{align*}
\]
To assign an unseen substrate, $S$, to a kinase with PFM $p$ and column weights $\hat{w}$, a score is calculated as:

$$ s_{PFM}(S, p, \hat{w}) = \prod_j \hat{w}_j p_{S,j} $$ \hspace{1cm} (11)

**Position-Specific Scoring Matrix**

20 $\times$ 15 PSSMs were calculated as log-likelihood ratios against the background expectation $f$:

$$ PSSM_{a,j} = \hat{p}_{a,j} = \log_2 \left( \frac{p_{a,j}}{f_{a,j}} \right) $$ \hspace{1cm} (12)

As described above, “proteome-backed” PSSMs were calculated against a background expectation of proteomic residue frequencies, while “phosphoproteome-backed” PSSMs were calculated against a position-specific background expectation derived from all phosphosite sequences observed in the experiment.

For assigning an unseen substrate, $S$, to a kinase with PSSM $\hat{p}$ and column weights $\hat{w}$, a score is calculated as:

$$ s_{PSSM}(S, \hat{p}, \hat{w}) = \sum_j \hat{w}_j \hat{p}_{S,j} $$ \hspace{1cm} (13)

For multi-label assignment, a min-max normalized PSSM score, $\hat{s}_{PSSM}$, was calculated to take values between 0 and 1 [16]:

$$ a_{\min} = \arg\min_a \hat{p}(a, j) $$ \hspace{1cm} (14)

$$ a_{\max} = \arg\max_a \hat{p}(a, j) $$ \hspace{1cm} (15)

$$ s_{\min} = \sum_j \hat{w}_j \hat{p}(a_{\min}, j) $$ \hspace{1cm} (16)

$$ s_{\max} = \sum_j \hat{w}_j \hat{p}(a_{\max}, j) $$ \hspace{1cm} (17)

$$ \hat{s}_{PSSM} = \frac{s_{PSSM} - s_{\min}}{s_{\max} - s_{\min}} $$ \hspace{1cm} (18)

**Multi-label Naïve Bayes**

We wish to calculate the posterior probability that a phosphosite is phosphorylated by kinase $K$, given its surrounding sequence window $S$. Using Bayes’ Theorem:
\[
P(K|S) = \frac{\mathcal{L}(S|K)P(K)}{P(S)} = \frac{\mathcal{L}(S_1, S_2, \ldots, S_{14}, S_{15}|K)P(K)}{P(S)}
\]

By applying the chain rule and then the Naïve Bayes assumption of conditional independence between the positions in \(S\), this can be simplified as follows:

\[
P(K|S) = \frac{\mathcal{L}(S_1|S_2, \ldots, S_{15}, K)\mathcal{L}(S_2|S_3, \ldots, S_{15}, K) \ldots \mathcal{L}(S_{14}|S_{15}, K)\mathcal{L}(S_{15}|K)P(K)}{P(S)}
\]

\[
= \frac{P(K)\prod_j \mathcal{L}(S_j|K)}{P(S)}
\]

We model the likelihood function, \(\mathcal{L}(S_j|K)\), as a categorical distribution. It is then easy to see from Equation 11 that, with the Naïve Bayes assumption of conditional independence:

\[
\mathcal{L}(S_j|K) = s_{PFM}(S, p_K, \hat{w} = \{1, 1, \ldots, 1, 1\})
\]

By altering the values in \(\hat{w}\), we arrive at a weighted Naïve Bayes formulation:

\[
P(K|S) = \frac{P(K)\prod_j \hat{w}_j \mathcal{L}(S_j|K)}{P(S)} = \frac{P(K)s_{PFM}(S, p_K, \hat{w}_K)}{P(S)}
\]

So, the conditional likelihood function \(\mathcal{L}(S|K)\) is calculated as described for PFMs.

The prior probability, \(P(K)\), was empirically estimated as the proportion of all substrates observed in the experiment (\(S_{BG}\)) that were observed as phosphorylated by kinase \(K\) (\(S_K\)):

\[
P(K) = \frac{|S_K|}{|S_{BG}|}
\]

Typically the marginal probability, \(P(S)\), is ignored in Naïve Bayes assignment as it is just a normalizing constant. However, in multi-label Naïve Bayes, normalization is necessary to compare probabilities between classes (kinases) [17]. By the law of total probability:

\[
P(S) = \mathcal{L}(S|K)P(K) + \mathcal{L}(S|\bar{K})P(\bar{K})
\]
The second term represents the likelihood of sequence $S$ being phosphorylated by any other kinase in the experiment ($\tilde{K}$). This is computed exactly as for kinase $K$, using sequences not phosphorylated by $K$, $S_{BG}$, to build the PFM and calculate the prior probability:

$$S_{BG} = S_{BG} - S_K$$

To perform multi-label assignment of phosphosite sequence $S$, for each kinase $K_i$ in the set of kinases used in the experiment, $K$, posterior probabilities were calculated as

$$P(K_i|S) = \frac{P(K_i)SPFM(S, \hat{p}_K, \hat{w}_K)}{P(K_i)SPFM(S, \hat{p}_K, \hat{w}_K) + P(\tilde{K}_i)SPFM(S, \hat{p}_{\tilde{K}}, \hat{w}_{\tilde{K}})}$$

$$P(\tilde{K}_i|S) = \frac{P(\tilde{K}_i)SPFM(S, \hat{p}_{\tilde{K}}, \hat{w}_{\tilde{K}})}{P(\tilde{K}_i)SPFM(S, \hat{p}_{\tilde{K}}, \hat{w}_{\tilde{K}}) + P(K_i)SPFM(S, \hat{p}_K, \hat{w}_K)}$$

Kinase $K_i$ was assigned to the phosphosite with sequence $S$ if $P(K_i|S) > P(\tilde{K}_i|S)$ [17]. This is equivalent to the condition that $P(K_i|S) > 0.5$, thus $P(K_i|S)$ need not be calculated.

**Naïve Bayes+**

The Naïve Bayes+ model is an extension of the multi-label Naïve Bayes model described above. In addition to the sequence evidence $S$, additional Bernoulli-distributed features were added – $I_{K,T}$ or $\tilde{I}_{K,T}$: presence or absence of a direct physical interaction between kinase $K$ and substrate protein $T$; $J_{K,T}$ or $\tilde{J}_{K,T}$: presence or absence of an indirect physical interaction between them; $D$ or $\tilde{D}$: the phosphosite is or is not in a predicted domain on $T$; and $E$ or $\tilde{E}$: substrate $T$ does or does not carry a predicted domain that is enriched among $K$’s substrates or physical interaction partners. An indirect interaction was defined as the substrate protein $T$ being “two hops” away from $K$ on the protein interaction network, i.e. they share a neighbor.

Thus, for example, the Naïve Bayes formulation of the posterior probability that kinase $K$ phosphorylates a site with sequence $S$, that is not in a domain $D$, with no direct physical interaction $I_{K,T}$, an indirect interaction $J_{K,T}$, and no enriched domains $E$ would be:

$$P(K|S, \tilde{D}, \tilde{I}_{K,T}, J_{K,T}, \tilde{E}) = \frac{\mathcal{L}(S|K)\mathcal{L}(\tilde{D}|K)\mathcal{L}(\tilde{I}_{K,T}|K)\mathcal{L}(J_{K,T}|K)\mathcal{L}(\tilde{E}|K)P(K)}{P(S)}$$

$$P(S) = \mathcal{L}(S|K)\mathcal{L}(\tilde{D}|K)\mathcal{L}(\tilde{I}_{K,T}|K)\mathcal{L}(J_{K,T}|K)\mathcal{L}(\tilde{E}|K)P(K) +$$

$$\mathcal{L}(S|\tilde{K})\mathcal{L}(\tilde{D}|\tilde{K})\mathcal{L}(\tilde{I}_{K,T}|\tilde{K})\mathcal{L}(J_{K,T}|\tilde{K})\mathcal{L}(\tilde{E}|\tilde{K})P(\tilde{K})$$

For all four additional features, the Bernoulli parameter $p$ was estimated by calculating the fraction of $K$’s substrates that have the feature (or, for $\tilde{K}$, the fraction of all sites that are not substrates of $K$). Domains enriched within the substrate proteins or interaction partners of kinase $K$ were identified via Fisher’s Hypergeometric Test. For each unique domain predicted to be on at least one substrate protein, presence or absence of that domain among proteins were counted contingent upon the proteins being either substrates or physical interaction partners of $K$ or not. $p$-values were adjusted for false discovery rate and tested at a critical value of 0.05.
Approximate logistic relationship between phosphoproteome-backed PSSM score and Naïve Bayes posterior probability

As described above, the definition of the Naïve Bayes posterior probability of assigning kinase $K$ to a phosphosite with sequence window $S$ is:

$$P(K|S) = \frac{\mathcal{L}(S|K)P(K)}{P(S)}$$

(34)

$$\mathcal{L}(S|K) = \mathcal{L}(S_1|K)\mathcal{L}(S_2|K)\ldots\mathcal{L}(S_{14}|K)\mathcal{L}(S_{15}|K)$$

(35)

In terms of the likelihood, the phosphoproteome-backed PSSM score function $s_{PSSM}(S, \hat{p}_K, \hat{w}_K)$ can be re-written as:

$$s_{PSSM}(S, \hat{p}_K, \hat{w}_K) = \log_2 \left( \frac{\mathcal{L}(S|K)}{\mathcal{L}(S)} \right)$$

(36)

Likelihood $\mathcal{L}(S)$ is the overall likelihood of observing a phosphosite with sequence window $S$, as calculated using the full phosphoproteome background set, $S_{BG}$. Likelihood $\mathcal{L}(S|K)$ is the likelihood of observing a phosphosite with sequence window $S$ among the substrates of kinase $K$, as calculated using a set of known substrates of kinase $K$, $S_K$. Furthermore, we assume:

$$S_K \subset S_{BG}$$

$$|S_K| \ll |S_{BG}|$$

(37)

(38)

We apply a logit transformation to the Naïve Bayes posterior probability $P(K|s)$, using 2 as our logarithm base:

$$\text{logit}(P(K|S)) = \log_2 \left( \frac{P(K|S)}{1 - P(K|S)} \right)$$

$$= \log_2 \left( \frac{P(K|S)}{P(K|\overline{S})} \right)$$

$$= \log_2 \left( \frac{\mathcal{L}(S|K)P(K)}{\mathcal{L}(S|\overline{K})P(\overline{K})} \right)$$

$$= \log_2 \left( \frac{\mathcal{L}(S|K)}{\mathcal{L}(S|\overline{K})} \right) + \log_2 \left( \frac{P(K)}{P(\overline{K})} \right)$$

(39)

Given that $|S_K| \ll |S_{BG}|$ (Equation 38), $\mathcal{L}(S|\overline{K}) \approx \mathcal{L}(S)$. Put simply, the likelihood model constructed from a set of over 14000 phosphosites will not be strongly impacted if one were first to omit a subset of a few hundred sites on average. As a result:
\[
\logit(P(K|S)) \approx \log_2 \left( \frac{\mathcal{L}(S|K)}{\mathcal{L}(S)} \right) + \log_2 \left( \frac{P(K)}{\overline{P(K)}} \right)
\]

\[
\logit(P(K|S)) \approx s_{\text{PSSM}}(S, \hat{p}_K, \hat{w}_K) + \log_2 \left( \frac{P(K)}{\overline{P(K)}} \right) \quad (40)
\]

\[
s_{\text{PSSM}}(S, \hat{p}_K, \hat{w}_K) \approx \logit(P(K|S)) - \log_2 \left( \frac{P(K)}{\overline{P(K)}} \right)
\]

From Equation 38, it also follows that \(|S_K| < |\overline{S}_K|\), so \(P(K) < \overline{P(K)}\) and \(\log_2 \left( \frac{P(K)}{\overline{P(K)}} \right) < 0\). This imposes a dependence of the relationship on \(|S_K|\): a larger \(|S_K|\) causes a lower inflection point in the inverse logistic fit, i.e. the PSSM score \(s_{\text{PSSM}}(S, \hat{p}_K, \hat{w}_K)\) at which \(P(K|S) = 0.5\) and \(\logit(P(K|S)) = 0\).

**Supplemental Figures**

**Fig A:** Using relative entropy as a PFM or PSSM column weight provides greater discrimination of columns than using information content. a) Sequence logo of all S/T sites in a high-confidence human phosphoproteome [2]. b) Information content at each position among S/T sites. c) Relative entropy versus proteomic residue frequencies at each position among S/T sites. Relative entropy provides greater separation between positions. d-f) As in (a-c) for Y sites.
Fig B: **Differing theoretical and empirical score ranges in scoring matrix-based models hinders the selection of a universal score cut-off.** Each vertical bar represents one kinase, colored by S/T (green) or Y kinases (orange). Light bars show theoretical score ranges and dark bars show empirical score ranges for the four sequence-only kinase specificity models. Empirical ranges were identified using the high-confidence human phosphoproteome [2]. PFM scores are shown log_{10}-transformed. Note that PFM and PSSM-based models have very different ranges between kinases. The maximum theoretical range also varies, making universal cutoff definitions difficult. Also, PFMs and PSSMs do not utilize their full range for all kinases when scoring physiologically occurring phosphosites. The Naïve Bayes model exhibits none of these problems.
Fig C: A Venn diagram showing total coverage of the human kinome supported by four kinase-substrate prediction or assignment methods. Values reflect numbers of kinases. GPS 5.0 has the largest coverage, while NetworKIN 3.0’s coverage is significantly lower than the others. IV-KAPhE does not cover any kinases not covered by the other methods.
**Supplemental Tables**

Table A: Protein kinases targeted by the chemical inhibitors used in the quantitative phosphoproteomic experiment of Wilkes et al. [18].

| Condition Name | Chemical Name | Target Kinases (gene symbol) | Target Information Source |
|----------------|---------------|------------------------------|---------------------------|
| AKT<sup>1</sup> | Akt inhibitor VIII | AKT1 AKT2 AKT3 | https://www.selleckchem.com/products/akti-1-2.html |
| AKT<sup>2</sup> | MK-2206 | AKT1 AKT2 AKT3 | https://www.selleckchem.com/products/MK-2206.html |
| CAMK<sup>1</sup> | KN-93 | CAMK2A CAMK2B CAMK2D CAMK2G CAMK1 CAMK1D CAMK1G | https://www.selleckchem.com/products/ kn-93.html |
| CAMK<sup>2</sup> | KN-62 | CAMK2A CAMK2B CAMK2D CAMK2G CAMK1 CAMK1D CAMK1G CAMK4 | https://www.selleckchem.com/products/ kn-62.html |
| EGFR<sup>1</sup> | PD-168393 | EGFR | https://www.selleckchem.com/products/pd168393.html |
| EGFR<sup>2</sup> | PD-153035 | EGFR | https://www.selleckchem.com/products/pd153035.html |
| ERK<sup>1</sup> | ERK inhibitor | MAPK1 MAPK3 | https://www.sigmaaldrich.com/catalog/product/mm/328007 |
| ERK<sup>2</sup> | ERK inhibitor II | MAPK1 MAPK3 | https://www.sigmaaldrich.com/catalog/product/mm/328006m |
| MEK<sup>1</sup> | GSK-1120212 | MAP2K1 MAP2K2 | https://www.selleckchem.com/products/gsk1120212-jtp-74057.html |
| MEK<sup>2</sup> | U0126 | MAP2K1 MAP2K2 | https://www.medchemexpress.com/u-0126.html |
| MTOR<sup>1</sup> | KU-0063794 | MTOR | https://pubchem.ncbi.nlm.nih.gov/compound/Ku-0063794#section=Biomolecular-Interactions-and-Pathways |
| MTOR<sup>2</sup> | Torin-1 | MTOR | https://www.cellsignal.com/products/activators-inhibitors/torin-1/14379 |
| P70S6K<sup>1</sup> | PF-4708671 | RPS6KB1 | https://www.selleckchem.com/products/pf-4708671.html |
| P70S6K<sup>2</sup> | DG2 | RPS6KB1 | https://www.sigmaaldrich.com/catalog/product/mm/559274 |
| PI3K<sup>1</sup> | GDC-0941 | PIK3CA PIK3CB PIK3CD PIK3CG | https://www.stemcell.com/gdc-0941.html |
| PI3K<sup>2</sup> | PI-103 | PIK3CA PIK3CB PIK3CD PIK3CG MTOR | https://www.stemcell.com/gdc-0941.html |
| PKC<sup>1</sup> | Gö-6976 | PRKCA PRKCB PRKCG PRKCD PRKCZ JAK2 FLT3 | https://www.selleckchem.com/products/go6976.html https://pubmed.ncbi.nlm.nih.gov/16956345/ |
| PKC<sup>2</sup> | BIM-1 | PRKCA PRKCB PRKCG | https://www.selleckchem.com/products/gf109203x.html |
| ROCK<sup>1</sup> | H-1152 dihydrochloride | ROCK2 CAMK2A CAMK2B CAMK2D CAMK2G PRKG1 PRKG2 AURKA PRKACA PRKACB PRKACG PRKCA PRKCB PRKCD PRKCE PRKCCH PRKCQ PRKCI PRKCZ | https://www.selleckchem.com/products/h-1152-dihydrochloride.html |
| ROCK<sup>2</sup> | Y-27632 | ROCK1 ROCK2 | https://www.selleckchem.com/products/Y-27632.html |
References

[1] Sugiyama N, Imamura H, Ishihama Y. Large-Scale Discovery of Substrates of the Human Kinome. Scientific Reports. 2019;9(1):10503. doi:10.1038/s41598-019-46385-4.

[2] Ochoa D, Jarnuczak AF, Viéitez C, Gehre M, Soucheray M, Mateus A, et al. The Functional Landscape of the Human Phosphoproteome. Nat Biotechnol. 2020;38(3):365–373. doi:10.1038/s41587-019-0344-3.

[3] Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014: Mutations, PTMs and Recalibrations. Nucleic Acids Research. 2015;43(Database–issue):D512–20. doi:10.1093/nar/gku1267.

[4] Bachman JA, Gyori BM, Sorger PK. Assembling a Phosphoproteomic Knowledge Base Using ProtMapper to Normalize Phosphosite Information from Databases and Text Mining. bioRxiv. 2019; p. 822668. doi:10.1101/822668.

[5] Oughtred R, Rust J, Chang C, Breitkreutz BJ, Stark C, Willems A, et al. The BioGRID Database: A Comprehensive Biomedical Resource of Curated Protein, Genetic, and Chemical Interactions. Protein Sci. 2021;30(1):187–200. doi:10.1002/pro.3978.

[6] Huttlin EL, Bruckner RJ, Navarrete-Perea J, Cannon JR, Baltier K, Gebreab F, et al. Dual Proteome-Scale Networks Reveal Cell-Specific Remodeling of the Human Interactome. Cell. 2021;184(11):3022–3040.e28. doi:10.1016/j.cell.2021.04.011.

[7] Buljan M, Ciuffa R, van Drogen A, Vichalkovski A, Mehnert M, Rosenberger G, et al. Kinase Interaction Network Expands Functional and Disease Roles of Human Kinases. Mol Cell. 2020;79(3):504–520.e9. doi:10.1016/j.molcel.2020.07.001.

[8] Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam Protein Families Database: Towards a More Sustainable Future. Nucleic Acids Research. 2016;44(D1):D279–85. doi:10.1093/nar/gkv1344.

[9] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets. Nucleic Acids Res. 2019;47(D1):D607–D613. doi:10.1093/nar/gky1131.

[10] Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: An R Package for Measuring Semantic Similarity among GO Terms and Gene Products. Bioinformatics. 2010;26(7):976–978. doi:10.1093/bioinformatics/btq064.

[11] Yu G. Gene Ontology Semantic Similarity Analysis Using GOSemSim. In: Kidder BL, editor. Stem Cell Transcriptional Networks: Methods and Protocols. Methods in Molecular Biology. New York, NY: Springer US; 2020. p. 207–215.

[12] van Buuren S, Groothuis-Oudshoorn K. Mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;45(1):1–67. doi:10.18637/jss.v045.i03.

[13] Henikoff S, Henikoff JG. Position-Based Sequence Weights. Journal of Molecular Biology. 1994;243(4):574–578.
[14] Henikoff JG, Henikoff S. Using Substitution Probabilities to Improve Position-Specific Scoring Matrices. Comput Appl Biosci. 1996;12(2):135–143.

[15] Invergo BM, Petursson B, Akhtar N, Bradley D, Giudice G, Hijazi M, et al. Prediction of Signed Protein Kinase Regulatory Circuits. Cell Syst. 2020;10(5):384–396.e9. doi:10.1016/j.cels.2020.04.005.

[16] Wagih O, Reimand J, Bader GD. MIMP: Predicting the Impact of Mutations on Kinase-Substrate Phosphorylation. Nat Methods. 2015;12(6):531–533. doi:10.1038/nmeth.3396.

[17] Zhang ML, Peña JM, Robles V. Feature Selection for Multi-Label Naive Bayes Classification. Information Sciences. 2009;179(19):3218–3229. doi:10.1016/j.ins.2009.06.010.

[18] Wilkes EH, Terfve C, Gribben JG, Saez-Rodriguez J, Cutillas PR. Empirical Inference of Circuitry and Plasticity in a Kinase Signaling Network. Proc Natl Acad Sci USA. 2015;112(25):7719–7724. doi:10.1073/pnas.1423344112.