1 **Using Multilayer Heterogeneous Networks to Infer Functions of Phosphorylated Sites**

2 Joanne Watson[1, 2], Jean-Marc Schwartz[1] and Chiara Francavilla[2]*,

3 * Corresponding Author: chiara.francavilla@manchester.ac.uk

4 [1]Division of Evolution & Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine &

5 Health, University of Manchester, Manchester, M13 9PT

6 [2]Division of Molecular and Cellular Function, School of Biological Sciences, Faculty of Biology, Medicine &

7 Health, University of Manchester, Manchester, M13 9PT.

8 **Running Title:** Phosphorylation Site-centric Functional Analysis

Abbreviations

RWHN – Random walk on heterogeneous network

PTM – Post-translational modification

PPI – Protein-protein interaction

GO – Gene Ontology

PSP – Phosphosite Plus

EGF -Epidermal Growth Factor

TGF-α – Tumour growth factor α

RWR – Random walk with restart

GOBP – Gene Ontology Biological Processes

FDR – False discovery rate

SILAC – Stable Isotope Labelling of Amino Acids in Cell culture

KDE – Kernel Density Estimation

Abstract

Mass spectrometry-based quantitative phosphoproteomics has become an essential approach in the study of cellular processes such as cell signaling. Commonly used methods to analyze phosphoproteomics datasets depend on generic, gene-centric annotations such as Gene Ontology terms and do not account for the function of a protein in a given phosphorylation state. Thus, analysis of phosphoproteomics data is hampered by a lack of phosphorylated site-specific annotations. Here, we propose a method that combines shotgun phosphoproteomics data, protein-protein interactions and functional annotations from ontologies or pathway databases into a heterogeneous multilayer network. Phosphorylation sites are then associated to potential functions using a random walk on heterogeneous network (RWHN) algorithm. We validated our approach using a dataset modelling the MAPK/ERK pathway and were able to associate differentially regulated sites on the same protein to their previously described functions. Random permutation analysis proved that these associations were not random and were determined by

34 the network topology. We then applied the RWHN algorithm to two previously published datasets; the
35 algorithm was able to reproduce the experimentally validated conclusions from the publications, and
36 associate phosphorylation sites with both new and known functions based on their regulatory patterns.
37 The approach described here provides a robust, phosphorylation site-centric method to analyzing
38 phosphoproteomics data and identifying potential context-specific functions for sites with similar
39 phosphorylation profiles.

40 **1. Introduction**

41 Phosphorylation is the most studied post-translational modification (PTM) due to its central role in
42 cellular regulation. It is thought to be the principal PTM in the human proteome with upwards of 100,000
43 predicted phosphorylation sites [1], as well as an essential mediator of protein-protein interaction (PPI)
44 and protein function. Transient changes occur at specifically regulated phosphorylation sites, of which
45 there may be multiple on each protein. Regulation of phosphorylation is often dependent on
46 perturbations such as the activity of extracellular ligands, drug treatment or physical stimuli in the
47 extracellular environment [2]. By comparing changes in the phosphoproteome of cells in different
48 experimental conditions through mass spectrometry-based phosphoproteomics, phosphorylated sites
49 that are key players in cellular processes and functions can be uncovered in an unbiased, high-throughput
50 manner [3].

51 Functional analysis of phosphoproteomics datasets is typically based on gene-centric enrichment of Gene
52 Ontology (GO) terms or involvement in known pathways [4]. However, this approach disregards
53 information captured by phosphoproteomics data on changes at specific phosphorylated sites, by limiting
54 the analysis to the protein level. The modification state of a protein is inherently coupled to its function;
55 PTMs alter protein activity, as well as the ability to interact with different sets of proteins. Furthermore, if
56 a protein is phosphorylated on multiple sites, each with a different function and regulatory pattern, this
57 information is not revealed by gene-centric analysis [5]. For instance, the well-studied signalling protein
58 MAPK1 has 18 known phosphorylated sites recorded in the database PhosphositePlus (PSP), however
59 only 6 have been annotated to a downstream function[6]. Enrichment analyses that rely on

60 generalizations based on protein-level or gene-centric descriptions exclude details encoded in the
61 phosphorylation signature. Analyses are thus hampered by the lack of phosphorylation site-specific
62 functional annotations.

63 Several network-based methods have been proposed to move towards phosphorylation site-specific
64 analyses of phosphoproteomics data. A significant focus has been on the inference of kinase-substrate
65 networks, such as NetworKIN [7], KSEA [8], and IKAP [9], among others. These methods are useful for
66 reconstructing the architecture of signalling intracellular networks, which can be informative for
67 identifying modules of modified proteins involved in cellular processes, but again remain hampered by
68 the lack of site-specific functional annotation [10]. They may also be biased towards the most studied
69 kinases and the exclusion of non-kinase proteins. Rudolph *et al.* [11] proposed a method to address this
70 issue named PHOTON which identified differentially regulated proteins based on the level of
71 phosphorylation of their binding partners in a high-confidence PPI network, and then used logistic
72 regression to identify the involvement of the phosphorylated proteins in known signalling pathways.
73 Krug *et al.* [12] curated a resource containing literature-derived phosphorylated site-specific signatures
74 to assign functions to sites. However, the use of such a resource is limited by the lack of annotation of
75 these signatures to specific cellular processes.

76 In recent years, heterogeneous or multilayer networks have been used to represent many types of 'omics
77 datasets [13]. These specialized networks are used to describe multiple types of associations with nodes
78 representing different entities. To identify relationships between the different biological layers, random
79 walk algorithms have been applied to these networks. The random walk on heterogeneous network
80 (RWHN) with restart (RWR) method [14] has been particularly popular. Jiang [15] used RWR to prioritize
81 disease candidate genes in a PPI-phenome network; similarly, Soul *et al.* [16] applied it to a PPI-phenome
82 network to identify disease mechanisms. Recent work has extended the method to multiple layers of
83 biological information, for example to infer disease associated m6A RNA methylation via known gene-
84 disease associations [17]. Similar methodology was used to associate phosphorylation sites recorded in
85 the PSP database to diseases via kinase-substrate interactions [18].

86 Here, we propose an algorithm that uses RWHN to associate phosphorylated sites to context-specific
87 function via a heterogeneous multilayer network using shotgun phosphoproteomics data. The network
88 combines three layers of information: phosphorylated sites, protein interactions and GO terms. Clustering
89 of phosphoproteomics data is used to find common features within datasets and is generally followed by
90 enrichment analyses. This is based on the assumption that common patterns of phosphorylation based on
91 temporal changes or in response to a particular stimulus, treatment or environmental context, are a likely
92 indicator of involvement in common functions. We utilize this concept in our algorithm by connecting
93 phosphorylated sites that have been clustered together, and therefore share regulatory patterns, within
94 the multilayer network.

95 To prove the utility of our algorithm, we apply it to a manually-curated validation network and two
96 previously published datasets, describing early signalling events in HeLa cells upon EGF and TGF-α
97 stimulation [19] and phosphorylation-mediated changes in breast cancer cells resistant to lapatinib
98 treatment [20]. We demonstrate that functional annotations can be differentially assigned to
99 phosphorylated sites and this is driven by changes in the context-dependent modification of these sites.
100 Our method is suitable for use with any phosphoproteomics dataset and could be generalized for data
101 describing other PTMs.

102 2. Experimental Procedures

103 A multilayer heterogeneous network was constructed to relate biological function to phosphorylation site
104 via a protein-protein interaction network (Figure 1). Three types of nodes are contained within the
105 network: phosphorylation sites (called 'sites' here on in for brevity), proteins and functional annotations
106 (either Gene Ontology Biological Process (GOBP) terms or KEGG pathways). The edges describe five
107 possible associations which are either bipartite (i.e. site to protein, protein to function) or between the
108 same type of node.

109 2.1 Francavilla et al. Dataset

110 The results of the phosphoproteomics experiment described in Francavilla *et al.* [19] were taken from
111 Supplementary Table S1 of their manuscript. The authors had processed the raw mass spectrometry data
112 using MaxQuant [21], filtered identified phosphorylated sites based on a localization probability of
113 greater than 0.75 and normalized remaining data to the ratio of EGFR at 1 min after stimulation with EGF
114 or TGF-α. Prior to multilayer heterogeneous network construction, the data were further filtered to
115 remove sites with two or more missing ratios in their time series and containing at least one SILAC ratio
116 higher than 2 or lower than 0.5 (as described in the methods of Francavilla *et al.* [19]). Missing data were
117 imputed with random draws from a truncated distribution, as previously described [22], using the
118 *impute.QRLIC* function from the *imputeLCMD* R package.

119 **2.2 Ruprecht et al. Dataset**

120 Data from the Ruprecht *et al.* [20] study were taken from the Supplementary Data file named 'Filtered and
121 normalized phosphoproteome dataset'. The raw mass-spectrometry data had been processed using
122 MaxQuant, filtered to include only identified phosphorylated sites with a localization probability of
123 greater than 0.75, and normalized. Sites that did not show a significant change (FDR <1%) either between
124 the untreated parental and laptanib-treated resistant (SILAC ratios H/L) or between the lapatinib treated
125 parental and lapatinib-treated resistant conditions (SILAC ratios H/M) experimental conditions were
126 filtered from the data, as described in the methods of Ruprecht *et al.* [20]. Missing data were imputed
127 with the same method used for the Francavilla et al. dataset [22][22][22][22].

128 **2.3 Construction of Multilayer Network**

129 **Phosphorylation Site – Phosphorylation Site Subnetwork**

130 Phosphorylated sites were clustered based on their quantitative values in different experimental
131 conditions using the Fuzzy C-Means [23] or k-means method. Edges were drawn between sites in the
132 same cluster that had a Pearson correlation (R2) between all the data points greater than or equal to 0.99
133 (Figure 1).

134 **Protein - Protein Subnetwork**

135 The protein-protein interaction network was extracted from STRING [24]. All interactions to the proteins
136 included in the phosphoproteomics datasets with an experimental confidence score of greater than 0.4
137 were included (Figure 1).

138 **Function – Function Subnetwork**

139 We used either GOBP terms or KEGG pathways as functional annotators in this work. If GOBP terms were
140 included in the multilayer network, the GOSemSim package from Bioconductor [25] was used to calculate
141 the semantic similarity of enriched GOBP terms. An edge was drawn between terms with semantic
142 similarity greater than 0.7, as calculated using the Wang method. In the case of KEGG pathways, edges
143 were drawn between pathways that had greater than 70% pairwise similarity in their functional
144 annotation profiles, following the method described in Stoney *et al.* [26] (Figure 1)

145 **Phosphorylation Site – Protein Bipartite Edges**

146 Sites and proteins had an edge between them if the residue was found on that protein. Therefore, sites
147 will only have one edge, but proteins will have edges to all the sites found on that protein that were
148 differentially regulated in the data set (Figure 1).

149 **Protein – Function Bipartite Edge**

150 We assumed that closely connected nodes in the protein-protein subnetwork would more likely be
151 involved in similar biological processes. Therefore we computed modules of the protein-protein
152 subnetwork using the Louvain method. Proteins from each module were analyzed for enrichment (FDR <
153 0.05%) of functional annotations (either GOBP terms or KEGG pathways) using the *Enrichr* [27] R
154 interface. This increased the specificity of terms to be included in the network. When GO terms were
155 included, high frequency (annotated to more than 5% of genes) and semantically redundant terms were
156 filtered using the Bioconductor GOSemSim [25] and GO.db packages [28] (Figure 1).

157 **2.4 Random Walk on Heterogeneous Network**

158 The heterogeneous network can be represented as an adjacency matrix as follows:

$$\mathbf{A} = \begin{bmatrix} A_{RR} & A_{RP} & 0 \\ A_{PR} & A_{PP} & A_{PF} \\ 0 & A_{FP} & A_{FF} \end{bmatrix} \quad (1)$$

159 where: A_{RR} are site-site associations, A_{PP} are protein-protein associations, A_{FF} are function-function
160 associations, A_{RP} are site-protein associations with A_{PR} as the transpose and A_{PF} are protein-function
161 associations with A_{FP} as the transpose.

162 As described in previous work [14], a transition matrix (M) was calculated for use in the first stage of the
163 algorithm:

$$M = \begin{bmatrix} M_{RR} & M_{RP} & 0 \\ M_{RP^T} & M_{PP} & M_{PF} \\ 0 & M_{PF^T} & M_{FF} \end{bmatrix} \quad (2)$$

164 The bipartite inter-subgraph transition matrices (M_{RP} and M_{PF}) were calculated as:

$$(M_{RP})_{i,j} = \begin{cases} \lambda\,(A_{RP})_{i,j} \Big/ \sum_j (A_{RP})_{i,j}, if \sum_j (A_{RP})_{i,j} \neq 0 \\ 0, otherwise \end{cases} \quad (3)$$

$$(M_{PF})_{i,j} = \begin{cases} \lambda\,(A_{PF})_{i,j} \Big/ \sum_j (A_{PF})_{i,j}, if \sum_j (A_{PF})_{i,j} \neq 0 \\ 0, otherwise \end{cases} \quad (4)$$

165 where λ is the transition probability (i.e. the likelihood of the walker moving between two layers of the
166 network).

167 The intra-subgraph transition matrices (M_{RR}, M_{PP} and M_{FF}) were calculated as:

$$(M_{RR})_{i,j} = \begin{cases} (A_{RR})_{i,j} \Big/ \sum_j (A_{RR})_{i,j}, if \sum_j (A_{RP})_{i,j} = 0 \\ (1-\lambda)\,(A_{RR})_{i,j} \Big/ \sum_j (A_{RR})_{i,j}, otherwise \end{cases} \tag{5}$$

$$(M_{PP})_{i,j} = \begin{cases} (A_{PP})_{i,j} \Big/ \sum_j (A_{PP})_{i,j}, if \sum_j (A_{PR})_{i,j}, (A_{PF})_{i,j} = 0 \\ (1-\lambda)\,(A_{PP})_{i,j} \Big/ \sum_j (A_{PP})_{i,j}, otherwise \end{cases} \tag{6}$$

$$(M_{FF})_{i,j} = \begin{cases} (A_{FF})_{i,j} \Big/ \sum_j (A_{FF})_{i,j}, if \sum_j (A_{FP})_{i,j} = 0 \\ (1-\lambda)\,(A_{FF})_{i,j} \Big/ \sum_j (A_{FF})_{i,j}, otherwise \end{cases} \tag{7}$$

RWHN is a ranking algorithm; nodes are ranked based on the probabilities of finding the random walker at a given node in the steady state, having started at a given seed node or set of seed nodes. In this work we set the seed nodes to be those sites belonging to a particular cluster. The probability of finding the random walker at each node for each step is calculated based on the iterative equation:

$$\begin{cases} P_{s+1} = (1-r) \times M^T \times P_s + r \times P_0 \\ P_{s+1} - P_s \leq 10^{-6} \end{cases} \tag{8}$$

where r is the restart probability (set to 0.7 as described in Kohler et al. [29]), P_0 is the initial probability vector and P_s is the probability vector at step s. P_0 was calculated such that all seed nodes were given equal probabilities with their sum equal to 1. All other nodes in the site-site subnetwork were assigned an initial probability of 0. Nodes in other subnetworks were assigned equal probabilities with their sum equal to 1 and weighted with the tunable parameters η_P and η_F, that are used to weight the influence of each layer.

178 The output of the algorithm is a ranked list of all the nodes, based on the probability of finding the
179 random walker at each node in the steady state. This list is filtered by retaining functional annotations
180 and removing proteins and sites.

181 We implemented the algorithm in R using packages available from Bioconductor and CRAN and the
182 source code is available at www.github.com/jowatson2011/RWHN_phosphoproteomics. RWHN on the
183 validation network took less than 30 seconds to run however the larger experimental data sets took
184 several hours on a moderately powerful computer (32GB RAM, Intel i7 Processor).

185 **3. Results**

186 **3.1 Overview of Algorithm**

187 To associate phosphorylated sites of unknown function to potential cellular functions we developed an
188 algorithm to apply to shotgun phosphoproteomics data. First, in order to identify context-specific roles of
189 proteins based on the regulation of their phosphorylated sites, a multilayer heterogeneous network is
190 constructed. This represents three layers of biological entities and information: phosphorylated sites,
191 proteins-protein interaction, and biological function (Figure 1). We then apply a ranking algorithm,
192 RWHN, which ranks nodes of each layer based on: (i) distance from the phosphorylated sites of interest,
193 which are assigned as 'seed' nodes and (ii) the topology of the multilayer heterogeneous network [14].
194 Functions that are highly ranked can be considered more correlated with a set of seed nodes.

195 In the multilayer heterogeneous network, edges are drawn between sites based on similarity in the
196 regulatory pattern of their phosphorylation which we determine using k-means or fuzzy C-means
197 clustering. For the protein layer, a PPI network is constructed of the phosphorylated proteins and their
198 interactors using the STRING database [24]. Interactors are included to account for non-phosphorylated
199 proteins, or those below the limit of detection of the experiment [30]. In order to get a comprehensive
200 and specific list of functional annotations, we calculated closely connected groups of proteins in the PPI
201 network using module detection and performed functional enrichment analysis on these modules. Edges
202 between functional annotations terms are drawn based on functional similarity and overlap [25,26].

203 The RWHN algorithm simulates a walker moving from a starting node(s) (called a seed) and then node to
204 node through the multilayer network. Each step is influenced by the probability of transition to another
205 layer (λ), the weighting of the protein and function layers (η_P and η_F) and the probability of restart (that
206 is, teleportation back to the seed node(s), r). The output of RWHN consists in a list of ranks for all the
207 nodes in the network based on the likelihood of the walker reaching that node.

208 To optimize the algorithm, we ran RWHN with each of these parameters (λ, r, η_P and η_F) tested over a
209 range of values (0.2 – 0.8), changing one whilst setting all others to 0.5 (Figure S1). Performance was
210 decidedly stable over the range of η_P and η_F however altering λ and r resulted in a spread of ranks
211 assigned to a term depending on the parameter value. Previous work has suggested an ideal λ and r of 0.7
212 [14,29], whilst η_P and η_F were set to 0.7 and 0.3 respectively to prioritize movement in the protein layer
213 and reduce the number of terms having the same rank regardless of seed.

214 **3.2 Validation Data Case Study**

215 MAPK/ERK signalling has been well studied and the temporal phosphorylation status of component
216 proteins in response to growth factor stimulus are relatively established [31,32](Figure 2A). Based on
217 this general understanding, we compiled a simple validation dataset of phosphorylation dynamics at
218 particular sites of the main signalling proteins within the pathway (Figure 2B). Sites were chosen based
219 on whether their phosphorylation is known to activate or inactivate protein activity, as recorded in the
220 PSP database [6]. We clustered the data into 5 clusters, referred to here as clusters 1-5, using the fuzzy c-
221 means algorithm (Figure 2B). The number of clusters was selected based on the number of temporal
222 trends we could visually identify in the data. The features of the multilayer network that was constructed
223 are summarized in Table 1. Enrichment of GOBP terms was used to form the functional annotation layer
224 of the network.

225 We ran the RWHN algorithm over the multilayer heterogeneous network with seed nodes set to all the
226 sites belonging to one of the clusters; this was then repeated for each cluster (Figure 3A). The highest
227 ranked GOBP term was the same ("Protein autophosphorylation") regardless of the seed nodes; however,
228 it was possible to differentiate between the five clusters based on the ranking of terms below the first one.

229 For example, the inactivating phosphorylation sites of MAPK2K1 and MAPK2K2 (S212 and S216
230 respectively) are clustered separately to the activating sites (MAP2K1_S218, MAP2K1_S222,
231 MAK2K2_S222 and MAP2K2_S226), in clusters 1 and 3 respectively. The ranking of the term "regulation of
232 mitotic cell cycle" is higher for cluster 1 than cluster 3. This is consistent with annotation in PSP, showing
233 that whilst activating MAP2K1/MAP2K2 sites are positive regulators of the cell cycle, MAP2K1 and
234 MAP2K2 are negative regulators of the cellular behaviors downstream of the activating phosphorylation.
235 Another example is the highly specific term "regulation of cysteine-type endopeptidase activity involved in
236 apoptotic process", which is ranked highest when the seed nodes are set to cluster 2. The process is
237 annotated to two of the proteins in this cluster (RPS6KA3 and RAF1) as well as other well-studied signalling
238 kinases found in the PPI layer of the multilayer heterogeneous network (SRC, AKT1, MEK5, BAD and
239 MAP3K5). Higher association with this cluster could indicate a particular role for RAF1-S289 in this
240 process, compared to the other RAF1 phosphorylated sites included in this network.

241 To assess the robustness of the algorithm, we performed a random permutation control. RWHN was run
242 one hundred times with the seeds set as before, random permutations of the subnetworks and bipartite
243 edges maintained. Kernel Density Estimation (KDE) was calculated to assess how often each GO term
244 occurred at each rank in the 100 random permutations for each set of seed nodes. In each case, there were
245 a subset of terms that were more likely to have a rank greater than 50, however none strongly correlated
246 with the actual rankings for each set of seed nodes (Figure 3B). This confirms that the rankings are not
247 random, but primarily determined by the network topology.

248 For a given set of seed nodes, the algorithm is capable of giving high ranks to functions with known
249 associations to those sites. It is also capable of predicting non-random, reasonable functions for sites of
250 unknown function in this context.

251 **3.3 Experimental Case Study: Dissecting EGF- and TGF-α-Induced Dynamic Phosphorylation**

252 We next wanted to test the performance of the algorithm on an already available and complex
253 phosphoproteomics dataset. The phosphoproteomics data retrieved from Francavilla *et al*. [19] describes
254 the effect in HeLa cells of stimulation with EGF or TGF-α over a period of 90 minutes (with time points at

255 1, 8, 40 and 90 minutes). In this study, one of the findings was that the two ligands activating the same
256 cell-surface receptor, regulated phosphorylation dynamics in a different manner, resulting in opposite
257 cellular responses. Phosphorylated sites from Supplementary Table S1 of the original publication were
258 filtered based on regulation by EGF or by TGF-α and divided based on which stimuli the regulation was
259 dependent on. The separated data were clustered into 6 clusters using the fuzzy C-means method, as per
260 the original publication, with each cluster representing a distinct dynamic profile of phosphorylation
261 (Figure S2). A multilayer heterogeneous network was constructed for both sets of data[19][19][19],
262 described in Table 2 and Table 3 respectively.

263 Several terms that were biologically relevant were differentially ranked in the EGF and TGF-α networks
264 (Figure 4A-B). Consistent with the original publication, the term 'regulation of ERK1 and ERK2 cascade'
265 was highly ranked when seeds were set to sites in cluster 3 and 5 (representing EGF transient
266 responders) for the EGF network and the term 'regulation of MAPK cascade' for clusters 4 and 5
267 (representing TGF-α cycling or sustained responders) for the TGF-α network. We used these terms as
268 positive controls, due to the process being well studied and many sites within the cascade having
269 previously documented function we could use to verify the robustness of our approach.

270 EGF cluster 3 contains two EGFR sites described as activating in the PSP database (EGFR_Y1197,
271 EGFR_Y1110), as well as sites belonging to EGFR downstream proteins SHC1 (SHC1_Y427, SHC1_Y349,
272 SHC1_Y350) and PTPN11 (PTPN11_Y584). Cluster 5 contains MAPK1 activating sites (MAPK1_T185,
273 MAPK1_Y187), serving as a positive control for this association, and two sites associated with EGFR
274 internalization in PSP (EGFR_S991, EGFR_S693). Interestingly, cluster 1 (representing EGF cycling
275 responders) contains a tyrosine residue that is also associated with regulation of EGFR internalization
276 (EGFR_Y998) as well as the RAB7 regulatory site experimentally investigated in the original publication
277 (RAB7_Y183) and MAPK3 activating sites (MAPK3_T202, MAPK_Y204). When seed nodes are set to
278 cluster 1 there are higher proportion of terms related to protein localization highly ranked. This
279 potentially indicates nuanced roles between tyrosine and serine/threonine residues on EGFR and
280 MAPK1/MAPK3 activation in receptor internalization. For the TGF-α network, only one site (EGFR-
281 Y1197, activating) was found in one of the two TGF-α clusters associated with MAPK regulation (cluster

282 4). RAB7A_Y183 and RAB11B, important regulators of the receptor recycling investigated in the paper,
283 were also found in this cluster. When seed nodes were set to this cluster, terms associated with cell
284 migration were ranked higher (e.g. `positive regulation of cellular migration`). This was in line with one of
285 the conclusions of the paper, demonstrating a role for recycling and migration upon TGF-α stimulation.
286 When seed nodes were set to sites in cluster 3, the term 'regulation of receptor mediate endocytosis' was
287 ranked high; this cluster contained EGFR sites Y998 and Y1172, both associated to receptor
288 internalization in PSP, and SH3BP4_S279, a known player in this process [33]. Taken together, the
289 functions that were highly ranked were related to the biological functions of the sites set as seed nodes.
290 The algorithm captured the experimentally verified conclusion from the original paper, where EGFR
291 stimulated by TGF-α is associated with receptor internalization/recycling and a strong migratory
292 phenotype, whilst EGFR stimulated with EGF is degraded with fewer potent changes in cell behavior.

293 To consider a single protein example, phosphorylated sites on RAF1 were clustered for the EGF data
294 depending on whether they are activating (S289, S301 in cluster 6, S296 in cluster 2) or inactivating (S259,
295 a 14-3-3 binding site [34], in cluster 5) sites, as annotated in PhosphositePlus (marked in Figure S2). In the
296 EGF network, seeds nodes set as clusters 2 or 6 resulted in 'positive regulation of cellular process' being
297 ranked higher than seed nodes as cluster 5 sites. These clusters also were able to rank 'regulation of PI3K
298 signalling' and 'regulation of protein kinase B signaling' highly, picking up on RAF1 crosstalk with other
299 pathways when active [35,36]. In particular, the cluster containing the sites RAF1_S289 and S301 ranked
300 'regulation of lipid kinase activity higher'. Cluster 5 ranked 'regulation of ERK1 and ERK2 cascade' higher
301 than cluster 2 and 6, however it also counterintuitively ranked 'positive regulation of MAPK cascade' higher.
302 The TGF-α data were also clustered so that RAF_S296 was separated (cluster 3) from the other activating
303 sites (cluster 5), but the negatively regulating site was not quantified in this experimental condition. As in
304 the EGF network, cluster 5 ranked 'inositol lipid-mediated phosphorylation' and 'regulation of lipid kinase
305 activity' higher than cluster 3 did, indicating a stronger correlation between RAF1_S289 and RAF1_S301
306 phosphorylation and lipid kinase pathway crosstalk. These results suggest the algorithm is capable of
307 assigning functional terms to differentially regulated phosphorylation site and validates that it can
308 distinguish between sites of known function.

309 **3.4 Experimental Case study: Phosphorylation-dependent response to Lapatinib treatment and**
310 **resistance in breast cancer**

311 To verify the applicability of the algorithm with non-temporal shotgun phosphoproteomics data, we
312 applied it to the experimental dataset from Ruprecht *et al.* [20] describing changes in the phosphoproteome
313 upon treatment with the breast cancer drug lapatinib in sensitive ('parental' BT-474) or lapatinib-resistant
314 (BT-474-J4) cells. The paper uncovered and experimentally validated the role of several metabolic enzymes
315 and signaling pathways that were driving lapatinib resistance. In particular, proteins that form the
316 spliceosome, those involved in glycolysis and glycogen catabolism and PI3K/AKT/mTOR pathway
317 members were differentially phosphorylated in lapatinib-resistant cells compared to parental lapatinib-
318 treated cells.

319 We first constructed a multilayer heterogeneous network as described in Experimental Procedures which
320 included 1603 phosphorylated sites that showed significant changes in response to lapatinib treatment in
321 either the parental or resistant cell lines (Table 4). Data were grouped into 5 clusters (referred to as clusters
322 1 – 5) using the k-means method. The number of clusters to use was determined using the elbow plot
323 method [37](Figure S3A), confirming that 5 clusters was sufficient to capture all features found in the data.
324 Indeed, sites sharing similar regulation profiles between the resistant and parental cells were clustered
325 together (Figure S3B). As the original publication used KEGG pathways rather than GO terms, we
326 incorporated these in the functional-annotation layer of our network to investigate the flexibility of our
327 approach. As with the Francavilla *et al.* [19] data, terms that were ranked the same regardless of the seed
328 nodes were excluded from further analysis. There were 26 such pathways out of the total 149 in the
329 network; these included ErbB signaling, MAPK signaling, 'pathways in cancer' and various specific cancer
330 pathways (including breast cancer).

331 The results of RWHN applied to the multilayer heterogeneous network (Figure 5A) show that when the
332 seed nodes are set to sites belonging to clusters 3, 4, or 5, which respectively represent sites that are:
333 rescued in resistance, decreased in resistance, or substantially increased phosphorylation upon lapatinib
334 treatment, the spliceosome pathway was highly ranked. Clusters 3 and 4 also highly ranked the mTOR

335 pathway and metabolism-related pathways (such as 'carbon metabolism in cancer'). Cluster 4 highly
336 ranked 'glycolysis' specifically, along with 'transcriptional misregulation', indicating alternate and nuanced
337 roles for different SRRM2 phosphorylated sites and its association with non-canonical pathways. Metabolic
338 terms were also highly ranked when the seed nodes were set to cluster 1 (representing increased
339 phosphorylation in resistant cells), with terms such as 'glycolysis/gluconeogenesis', 'pentose phosphate
340 pathway' and 'starch and sucrose metabolism' ranked higher. This cluster contains the ALDOA-S39 site
341 highlighted in the original publication and, interestingly, the activating S17 on SRC; this could indicate
342 crosstalk between the ErbB/MAPK and metabolic pathways to be experimentally investigated. Cluster 2
343 (representing decreased phosphorylation upon lapatinib-treatment) contains several of the glycogen
344 catabolism sites highlighted in the paper. There is no KEGG pathway associated with this process, however
345 this cluster highly ranked the mTOR pathway, adding more evidence for the link validated in the paper
346 between mTOR signalling, metabolic rewiring and lapatinib resistance. This cluster also contained several
347 of the key spliceosome sites highlighted by Ruprecht et al. [20] (SRRM2_S1132, SRRM2_S1987,
348 HNRNPU_S59), indicating parallel- or cross-regulation of the 'transcriptional misregulation in cancer'
349 pathway with the metabolic processes. Sites within this cluster could be investigated as therapeutic targets.

350 Much of the investigation by Ruprecht et al. [20] was done purely on the sites found to be regulated in
351 lapatinib-treated resistant cells compared to the untreated parental cells. We therefore constructed a
352 second network using only those sites that were regulated in lapatinib-treated resistant cells (Table 5).
353 These were grouped into four clusters (referred to as clusters 1 – 4) by K-means clustering as before
354 (S3A,C) and KEGG terms were used again in the functional annotation layer. Of the 156 pathways included
355 in the network, 23 were ranked identically regardless of node; these were largely the same as those in
356 common in the first network but did not include 'ErbB signaling'.

357 As described in Figure 5B, two sets of seed nodes (clusters 1 and 4, representing sites downregulated and
358 moderately upregulated) tended to rank metabolic terms more highly (e.g. 'starch and sucrose metabolism',
359 'galactose metabolism', 'pentose phosphate pathway', 'purine metabolism'), with clusters 2 and 3
360 (representing sites slightly and significantly upregulated, respectively) tending to rank other cellular
361 processes and signalling pathways more highly. However, 'mTOR', 'Hippo' and 'cgmp-pkg' signaling were

362 highly ranked in clusters 2, 3 and 4, indicating some cross-regulation between these pathways and the
363 metabolism-associated sites in cluster 4. This is corroborated by the inclusion LDHA_Y10 in cluster 4,
364 discussed by Ruprecht et al. to be an activating site which is known to be phosphorylated by oncogenic
365 kinases such as FGFR and ABL. Cluster 2 contains ABL1_S737; this site has not been recorded in PSP, but
366 could be investigated as a regulator of metabolic rewiring in lapatinib resistance.

367 When sites found in cluster 1 are set as seed nodes, the pathways 'glycolysis/gluconeogenesis' and
368 'glucagon signalling' are ranked high alongside 'spliceosome', adding further support for the relationship
369 between these processes and lapatinib-resistance. Cluster 1 contains ENO1-Y44, which was validated in
370 the original publication as a marker for sustained glycolysis and a target for therapy, PYGB-T59, which is
371 involved in glycogen catabolism, and sites of various spliceosome and ribosomal proteins such as RPS6-
372 S240 and SRRM2-S2123. Furthermore, the high rank of ErbB signalling for this cluster indicates further
373 that it may contain therapeutically relevant sites, as the BT-474 cell-line are HER2 (ERBB2) overexpressing.
374 An avenue of investigation could be the specific roles the sites in this cluster, such as ERBB2-Y1233 and
375 ERBB2 interactors YAP1-S164 and RPS6-S240, may have on signalling and metabolism in the context of
376 lapatinib resistance in HER2-postivive breast cancer.

377 By comparing to the experimentally validated conclusions of Ruprecht et al., we have extracted meaningful
378 associations between specific phosphorylated sites and their functions in driving lapatinib resistance. The
379 algorithm also highlighted several potential sites, such as SRC_S17, that may serve as hubs or intersections
380 of crosstalk between key signalling pathways and the rewired metabolic processes.

381 **4. Discussion**

382 Phosphorylation has an important impact on protein function and thus cellular behavior. Within a network
383 of kinase-substrate interactions, phosphorylation modulates the flow of information and regulation of
384 disparate processes throughout the cell. Understanding the impact of phosphorylation on protein function
385 traditionally required the experimental investigation and manipulation of individual sites, however with
386 the advent of high throughput phosphoproteomics, thousands of novel phosphorylation sites of unknown
387 function have been discovered. As there is limited experimentally validated information available on how

388 regulation at individual phosphorylated sites impacts the downstream cellular output, current methods for
389 analyzing high-throughput phosphoproteomics data rely on general descriptors of phosphorylated protein
390 function. These often do not account for the specific role of a given phosphorylated site. These methods,
391 valuably used in transcriptomics or proteomics analysis, include GO or pathway enrichment analyses or
392 PPI network construction with differentially phosphorylated proteins. Their use in analyzing
393 phosphoproteomics data disregards site-specific or multiple-site regulation, masking the roles of proteins
394 in a particular modification state [10,12]. Here, we developed and tested a method that associates
395 phosphorylated sites to potential function using RWHN. By incorporating the pattern of phosphorylation
396 upon perturbation, we consider more of the information available in phosphoproteomics datasets and
397 establish the phosphorylated sites as the key drivers in the functional analysis. The algorithm is capable of
398 recapturing experimentally proven functions of phosphorylated sites in a non-gene centric manner,
399 improving the accuracy of analysis and extracting more information from phosphoproteomics data.

400 To prove the utility of our algorithm we applied it to a validation network, which represented a simple
401 model of phosphorylation dynamics in the MAPK/ERK pathway, and to two previously published
402 phosphoproteomics dataset describing temporal cellular signalling events and the impact of resistance to
403 drug treatment in breast cancer, respectively. Our algorithm successfully distinguished between
404 differentially regulated phosphorylated sites from the same protein and associated them to both known
405 and previously uninvestigated functions. For example, when applying this approach to the data from
406 Francavilla *et al.* [19], we can recapture the roles of EGFR sites with functions described in PSP. We also
407 associate EGFR sites to the phenotypes described in the paper. In the case of EGF-driven receptor
408 localization, EGFR_Y998 associated with the RAB7_Y183 site described in the original publication as an
409 indicator of the EGF-EGFR response, and was part of a cluster that ranked the term 'establishment of
410 protein localization to organelle' highly. On the contrary, TGF-α induced receptor recycling and migration
411 were associated here with EGFR sites Y998, Y1172 and Y1197. Francavilla et al. associated these sites with
412 their functions using a combination of gene-centric approaches and experimental validation. Here, we use
413 a single site-centric algorithm to extract these associations directly from the phosphoproteomics dataset,
414 demonstrating the power of our approach in narrowing down candidates for further functional studies. By

415 associating ligand-dependent regulation of EGFR sites, it is possible to disentangle the multifaceted role
416 EGFR plays in regulating cellular signalling networks and downstream cellular behaviors.

417 From the Ruprecht et al. data, we can use the example of SRRM2. This component of the spliceosome is
418 known to be highly phosphorylated, with 675 phosphorylation sites recorded in Phosphosite Plus. At the
419 time of writing none of these sites were are ascribed to a function in the database, despite the importance
420 of phosphorylation and dephosphorylation events in orchestrating splicing events [38]. It is also regularly
421 mutated (>5% of cases in TGCA) in lung, stomach, bladder, endometrial and colorectal cancers [6].
422 Ruprecht et al. uncovered a new role for the spliceosome in modulating lapatinib-resistance, highlighting
423 several sites on spliceosome proteins (including SRRM2_S1132, S1987 and S970) that were differentially
424 modified in resistant and parental cell populations. Previous work has described how SRRM2-depletion in
425 HER2-positive breast and ovarian cancer cells reduced the rate of migration [39]; the spliceosome plays
426 myriad roles in the breast cancer environment, but site-specific analysis is lacking. Using our algorithm, we
427 can begin to uncover the nuanced modification-specific roles of proteins such as SRRM2 in metabolic
428 rewiring. Our analysis showed that SRRM2 sites clustered with sites belonging to proteins involved in
429 glycogen metabolism. When sites in this cluster were set as seed nodes metabolic and transcriptional
430 processes were upregulated along with mTOR signalling. The sites in this cluster could be investigated as
431 points of crosstalk between these different cellular processes. This example highlights the value of our
432 approach in generating hypotheses of modified protein function and signaling pathway crosstalk.

433 Although the associations predicted here between sites and functions cannot be interpreted as causative,
434 phosphorylation sites can act as indirect, and context-dependent, regulators in cellular processes. This is
435 demonstrated by the clustering of SRC_S17, a protein that is typically associated with signalling
436 downstream of cell-surface receptors, with a cluster that ranked metabolic pathways higher than cellular
437 processing or signalling pathways. Critical to this is a robust method of defining clusters, to capture
438 meaningful regulatory patterns at different phosphorylated sites. Here we define clusters using established
439 techniques (e.g. elbow method for K-means) or visually identifying those with differing time-resolved
440 behaviors. By integrating clustering of quantitative phosphoproteomics data and multilayer network

441 construction our algorithm closely associates sites with similar behaviors to predict the influence they have
442 on downstream cellular processes.

443 Multilayer heterogeneous networks are increasingly being used to integrate 'omic data; here, their use
444 allows phosphoproteomics data to be incorporated with PPI networks and functional annotations,
445 overcoming the issue of considering phosphoproteomics data primarily on the protein level. A potential
446 drawback of our approach is the reliance on large semi-curated resources such GO or STRING. For instance,
447 different clusters may have many terms or pathways in common given the involvement of many proteins
448 in the same biological functions and the high proportion of frequently used GO terms [40]. In the work
449 described here, we encounter this issue with the RAF1 inhibitory site S259, which was part of a cluster that
450 resulted in 'positive regulation of MAPK cascade'. We theorize that this may be rectified if the data were
451 clustered into more groups, in order to capture more nuanced phosphorylation patterns. However,
452 enrichment of non-specific terms or false positives remains an issue when analyzing high-throughput
453 'omics data by other commonly used methods too. We have attempted to limit this effect by only including
454 edges in the PPI network with a STRING experimental confidence score of greater than 0.4. This could be
455 taken further by filtering for only high confidence functional annotations or incorporating annotations from
456 multiple sources.

457 An important application of this work could be the opportunity to revise databases like PhosphositePlus,
458 in particular if done in combination with the recently described PTMsigDB phosphorylated-site signature
459 library [12]. There is a pressing need in the phosphoproteomics community to have accurate and context-
460 specific information identified sites, particularly those that are regularly detected in phosphoproteomics
461 experiments. We also anticipate that this method could be used routinely in laboratories generating mass
462 spectrometry-based phosphoproteomics data. Our method is flexible enough to be used with any discovery
463 phosphoproteomics data that describes a change between conditions. This is an improvement on the
464 previously published use of RWHN using multiple sources of phosphoproteomics data to uncover disease-
465 dependent regulation [18]. Moreover, it could easily be generalized to any post-translational modification
466 proteomics dataset, as it incorporates readily available PPI and functional annotation data, as
467 demonstrated here. The fundamental aspect would be maintained with any of these expansions: specific

468 patterns of regulation at modified sites dictate movement through the multilayer heterogeneous network.
469 It can also be noted that although the experimental datasets used in this work are based on SILAC
470 quantification, our method is not dependent on a specific quantitative phosphoproteomics method.

471 We have proposed a site-centric approach to analyze phosphoproteomics data, that provides a robust
472 alternative to gene-centric methods of analysis. We integrated clustered quantitative phosphoproteomics
473 data, a context-specific PPI network and functional annotations into a multilayer, heterogeneous network
474 and use the RWHN method to predict the functions of phosphorylation sites with similar regulatory
475 patterns. Using our algorithm, we extracted experimentally validated associations between
476 phosphorylated sites and their role in cellular processes which could not be captured using the typical
477 gene-centric methods used in previous papers. Moreover, our algorithm has the potential to be used by
478 researchers in predicting novel site-function associations and generating hypotheses to be experimentally
479 validated.

480 **5. Acknowledgements**

481 JW is supported by a Biotechnology and Biological Sciences Research Council (BBSRC) Doctoral Training
482 Programme grant (BB/M011208/1). CF is supported by a Wellcome Trust Sir Henry Dale Fellowship
483 (107636/Z/15/Z), the Biotechnology and Biological Sciences Research Council (BB/R015864/1), and the
484 Medical Rerearch Council (MR/T016043/1). We would like to thank current and former members of the
485 Schwartz and Francavilla labs for their thoughtful and insightful discussion on this work.

486 **6. Data Availability**

487 Data and source code can be found at github.com/JoWatson2011/RWHN_Phosphoproteomics

488 **7. References**

489 1. Sugiyama N, Imamura H, Ishihama Y. Large-scale Discovery of Substrates of the Human Kinome.
490 Sci Rep. 2019;9: 10503. doi:10.1038/s41598-019-46385-4

491 2. Day EK, Sosale NG, Lazzara MJ. Cell signaling regulation by protein phosphorylation: a
492 multivariate, heterogeneous, and context-dependent process. Curr Opin Biotechnol. 2016;40:
493 185–192. doi:10.1016/j.copbio.2016.06.005

494 3. Vyse S, Desmond H, Huang PH. Advances in mass spectrometry based strategies to study receptor
495 tyrosine kinases. IUCrJ. 2017;4: 119–130. doi:10.1107/S2052252516020546

496 4. Munk S, Refsgaard JC, Olsen J V. Systems Analysis for Interpretation of Phosphoproteomics Data.
497 In: von Stechow L, editor. Phospho-Proteomics: Methods and Protocols. New York, NY: Springer
498 New York; 2016. pp. 341–360. doi:10.1007/978-1-4939-3049-4_23

499 5. Sacco F, Perfetto L, Cesareni G. Combining Phosphoproteomics Datasets and Literature
500 Information to Reveal the Functional Connections in a Cell Phosphorylation Network. Proteomics.
501 2018;18: 1700311. doi:10.1002/pmic.201700311

502 6. Hornbeck P V., Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014:
503 Mutations, PTMs and recalibrations. Nucleic Acids Res. 2015;43: D512–D520.
504 doi:10.1093/nar/gku1267

505 7. Linding R, Jensen LJ, Pasculescu A, Olhovsky M, Colwill K, Bork P, et al. NetworKIN: A resource for
506 exploring cellular phosphorylation networks. Nucleic Acids Res. 2008. doi:10.1093/nar/gkm902

507 8. Casado P, Rodriguez-Prados J-C, Cosulich SC, Guichard S, Vanhaesebroeck B, Joel S, et al. Kinase-
508 substrate enrichment analysis provides insights into the heterogeneity of signaling pathway
509 activation in leukemia cells. Sci Signal. 2013;6: rs6. doi:10.1126/scisignal.2003573

510 9. Mischnik M, Sacco F, Cox J, Schneider H-C, Schäfer M, Hendlich M, et al. IKAP: A heuristic
511 framework for inference of kinase activities from Phosphoproteomics data. Bioinformatics.
512 2016;32: 424–431. doi:10.1093/bioinformatics/btv699

513 10. Needham EJ, Parker BL, Burykin T, James DE, Humphrey SJ. Illuminating the dark

514 phosphoproteome. Sci Signal. 2019;12: eaau8645. doi:10.1126/scisignal.aau8645

515 11. Rudolph JD, de Graauw M, van de Water B, Geiger T, Sharan R. Elucidation of Signaling Pathways
516 from Large-Scale Phosphoproteomic Data Using Protein Interaction Networks. Cell Syst. 2016;3:
517 585-593.e3. doi:10.1016/j.cels.2016.11.005

518 12. Krug K, Mertins P, Zhang B, Hornbeck P, Raju R, Ahmad R, et al. A Curated Resource for
519 Phosphosite-specific Signature Analysis. Mol Cell Proteomics. 2019;18: 576–593.
520 doi:10.1074/mcp.TIR118.000943

521 13. Battiston F, Nicosia V, Chavez M, Latora V. Multilayer motif analysis of brain networks. Chaos An
522 Interdiscip J Nonlinear Sci. 2017;27: 047404. doi:10.1063/1.4979282

523 14. Li Y, Patra JC. Genome-wide inferring gene–phenotype relationship by walking on the
524 heterogeneous network. Bioinformatics. 2010;26: 1219–1224.
525 doi:10.1093/bioinformatics/btq108

526 15. Jiang R. Walking on multiple disease-gene networks to prioritize candidate genes. J Mol Cell Biol.
527 2015;7: 214–230. doi:10.1093/jmcb/mjv008

528 16. Soul J, Hardingham TE, Boot-Handford RP, Schwartz J-M. PhenomeExpress: A refined network
529 analysis of expression datasets by inclusion of known disease phenotypes. Sci Rep. 2015;5: 8117.
530 doi:10.1038/srep08117

531 17. Tang Y, Chen K, Wu X, Wei Z, Zhang S-Y, Song B, et al. DRUM: Inference of Disease-Associated m6A
532 RNA Methylation Sites From a Multi-Layer Heterogeneous Network. Front Genet. 2019;10: 266.
533 doi:10.3389/fgene.2019.00266

534 18. Xu X, Wang M. Inferring Disease Associated Phosphorylation Sites via Random Walk on Multi-
535 Layer Heterogeneous Network. IEEE/ACM Trans Comput Biol Bioinforma. 2016;13: 836–844.
536 doi:10.1109/TCBB.2015.2498548

537 19. Francavilla C, Papetti M, Rigbolt KTG, Pedersen A-K, Sigurdsson JO, Cazzamali G, et al. Multilayered
538 proteomics reveals molecular switches dictating ligand-dependent EGFR trafficking. Nat Struct
539 Mol Biol. 2016;23: 608–618. doi:10.1038/nsmb.3218

540 20. Ruprecht B, Zaal EA, Zecha J, Wu W, Berkers CR, Kuster B, et al. Lapatinib resistance in breast
541 cancer cells is accompanied by phosphorylation-mediated reprogramming of glycolysis. Cancer
542 Res. 2017;77: 1842–1853. doi:10.1158/0008-5472.CAN-16-2976

543 21. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range
544 mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008;26: 1367–1372.
545 doi:10.1038/nbt.1511

546 22. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, et al. The Perseus computational
547 platform for comprehensive analysis of (prote)omics data. Nat Methods. 2016;13: 731–740.
548 doi:10.1038/nmeth.3901

549 23. Schwämmle V, Jensen ON. A simple and fast method to determine the parameters for fuzzy c–
550 means cluster analysis. Bioinformatics. 2010;26: 2841–2848. doi:10.1093/bioinformatics/btq534

551 24. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–
552 protein association networks with increased coverage, supporting functional discovery in
553 genome-wide experimental datasets. Nucleic Acids Res. 2018;47: D607–D613.
554 doi:10.1093/nar/gky1131

555 25. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity
556 among GO terms and gene products. Bioinformatics. 2010;26: 976–978.
557 doi:10.1093/bioinformatics/btq064

558 26. Stoney RA, Ames RM, Nenadic G, Robertson DL, Schwartz J-M. Disentangling the multigenic and
559 pleiotropic nature of molecular function. BMC Syst Biol. 2015;9 Suppl 6: S3. doi:10.1186/1752-
560 0509-9-S6-S3

561 27. Kuleshov M V, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a
562 comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44:
563 W90–W97. doi:10.1093/nar/gkw377

564 28. Carlson M. GO.db: A set of annotation maps describing the entire Gene Ontology. R package
565 version 3.8.2. Bioconductor 3.10. 2019.

566 29. Köhler S, Bauer S, Horn D, Robinson PN. Walking the Interactome for Prioritization of Candidate
567 Disease Genes. Am J Hum Genet. 2008;82: 949–958. doi:10.1016/J.AJHG.2008.02.013

568 30. Riley NM, Coon JJ. Phosphoproteomics in the Age of Rapid and Deep Proteome Profiling. Anal
569 Chem. 2016;88: 74–94. doi:10.1021/acs.analchem.5b04123

570 31. Fey D, Croucher DR, Kolch W, Kholodenko BN. Crosstalk and Signaling Switches in Mitogen-
571 Activated Protein Kinase Cascades. Front Physiol. 2012;3. doi:10.3389/fphys.2012.00355

572 32. Arkun Y, Yasemi M. Dynamics and control of the ERK signaling pathway: Sensitivity, bistability,
573 and oscillations. Thattai M, editor. PLoS One. 2018;13: e0195513.
574 doi:10.1371/journal.pone.0195513

575 33. Francavilla C, Rigbolt KTG, Emdal KB, Carraro G, Vernet E, Bekker-Jensen DB, et al. Functional
576 proteomics defines the molecular switch underlying FGF receptor trafficking and cellular outputs.
577 Mol Cell. 2013;51: 707–22. doi:10.1016/j.molcel.2013.08.002

578 34. Chong H, Guan KL. Regulation of Raf through phosphorylation and N terminus-C terminus
579 interaction. J Biol Chem. 2003;278: 36269–36276. doi:10.1074/jbc.M212803200

580 35. Scheid MP, Woodgett JR. Protein kinases: Six degrees of separation? Current Biology. Current
581 Biology Ltd; 2000. pp. R191–R194. doi:10.1016/S0960-9822(00)00349-3

582 36. Moelling K, Schad K, Bosse M, Zimmermann S, Schweneker M. Regulation of Raf-Akt cross-talk. J
583 Biol Chem. 2002;277: 31099–31106. doi:10.1074/jbc.M111974200

584 37. Zambelli AE. A data-driven approach to estimating the number of clusters in hierarchical
585 clustering. F1000Research. 2016;5: 2809. doi:10.12688/f1000research.10103.1

586 38. Naro C, Sette C. Phosphorylation-mediated regulation of alternative splicing in cancer.
587 International Journal of Cell Biology. 2013. doi:10.1155/2013/151839

588 39. Mukherji M, Brill LM, Ficarro SB, Hampton GM, Schultz PG. A phosphoproteomic analysis of the
589 ErbB2 receptor tyrosine kinase signaling pathways. Biochemistry. 2006;45: 15529–15540.
590 doi:10.1021/bi060971c

591 40. Gaudet P, Dessimoz C. Gene Ontology: Pitfalls, Biases, and Remedies. In: Dessimoz C, Škunca N,
592 editors. The Gene Ontology Handbook. New York, NY: Springer New York; 2017. pp. 189–205.
593 doi:10.1007/978-1-4939-3743-1_14

594

595 **8. Tables**

596 Table 1. Multilayer heterogeneous network, constructed from MAPK/ERK validation data. GOBP terms
597 were included as functional annotations.

Subnetwork	Edges	Nodes
Site-site	28	19
Protein-protein	932	313
Function-function	176	71
Site-protein	19	19 sites, 8 proteins
Protein-function	1616	251 proteins, 161 functions

598

599 Table 2. Multilayer heterogeneous network, constructed from Francavilla *et al.* [19] EGF-regulated
600 phosphorylated sites. GOBP terms were included as functional annotations.

Subnetwork	Edges	Nodes
Site-site	4259	729
Protein-protein	17476	4181
Function-function	24	17
Site-protein	729	729 sites, 437 proteins
Protein-function	190	42 proteins, 53 functions

601

602 Table 3. Multilayer heterogeneous network, constructed from Francavilla *et al.* [19] TGFα-regulated
603 phosphorylated sites. GOBP terms were included as functional annotations.

Subnetwork	Edges	Nodes
Site-site	2740	683
Protein-protein	17080	4003
Function-function	8	5
Site-protein	683	683 sites, 407 proteins
Protein-function	200	45 proteins, 59 functions

604

605 Table 4. Multilayer heterogeneous network, constructed from Ruprecht *et al.* [20] lapatinib-regulated
606 phosphorylated sites in parental and lapatinib-resistant cell lines. KEGG pathways were included as
607 functional annotations.

Subnetwork	Edges	Nodes
Site-site	84013	1603
Protein-protein	34146	5692
Function-function	11	18
Site-protein	1603	1603 sites, 932 proteins

Protein-function	710	107 proteins, 138 pathways

Table 5. Multilayer heterogeneous network, constructed from Ruprecht *et al.* [20] lapatinib-regulated phosphorylated sites in lapatinib-resistant cell line. KEGG pathways were included as functional annotations.

Subnetwork	Edges	Nodes
Site-site	707785	1920
Protein-protein	39506	6337
Function-function	15	24
Site-protein	1920	1920 sites, 1093 proteins
Protein-function	883	134 proteins, 142 pathways

9. Figures & Figure Legends

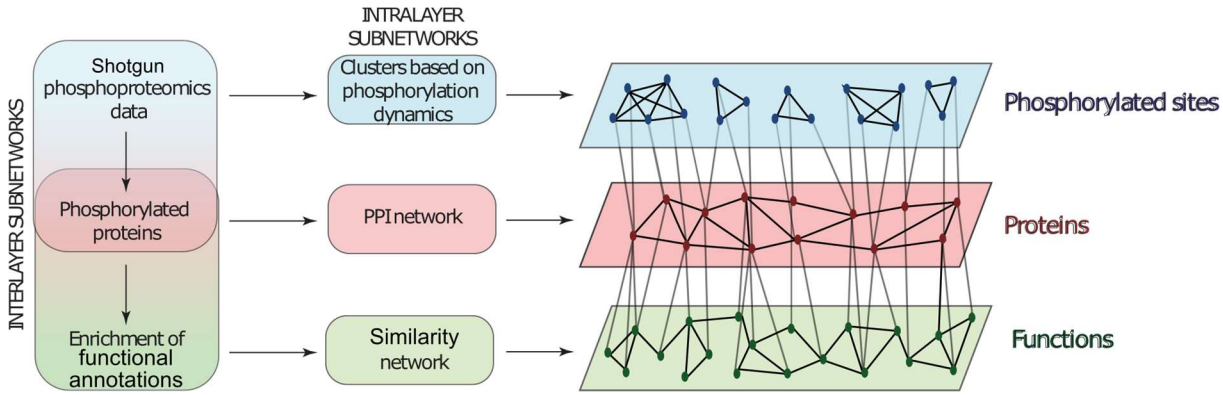
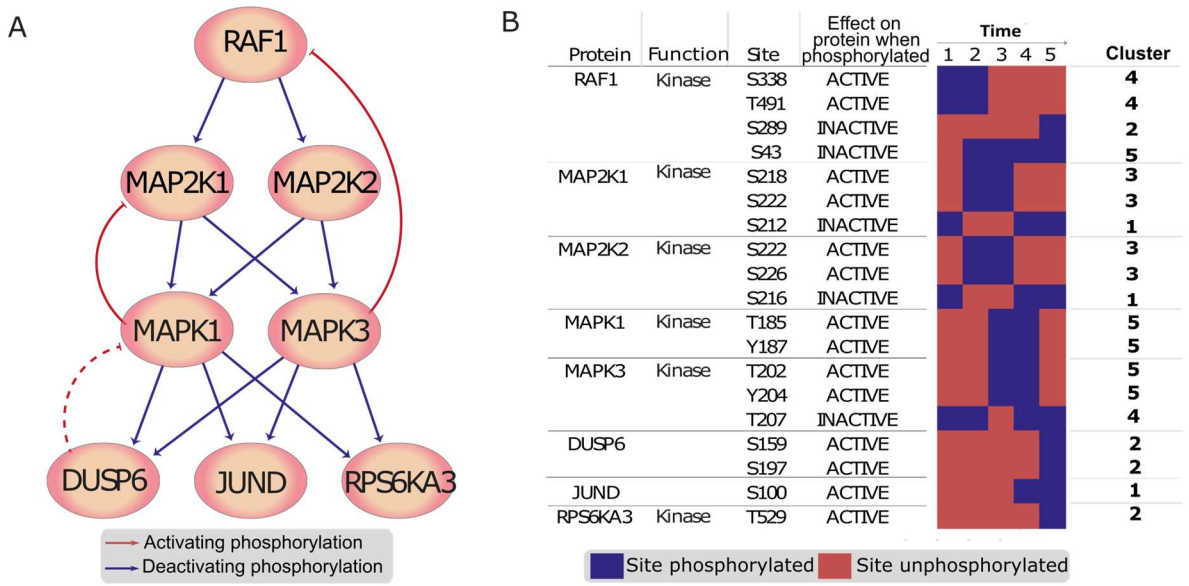


Figure 1. Overview of multilayer heterogeneous network construction from phosphoproteomics data.



616

617 **Figure 2.** Construction of validation data based on the MAPK/ERK pathway. A) Traditional
618 representation of the MAPK/ERK pathway. B) A validation dataset was constructed based on the
619 MAPK/ERK pathway, simulating phosphorylation dynamics over time after pathway activation. The sites
620 were clustered using fuzzy c-means clusters. These data were used to construct a heterogeneous
621 multilayer network.

622

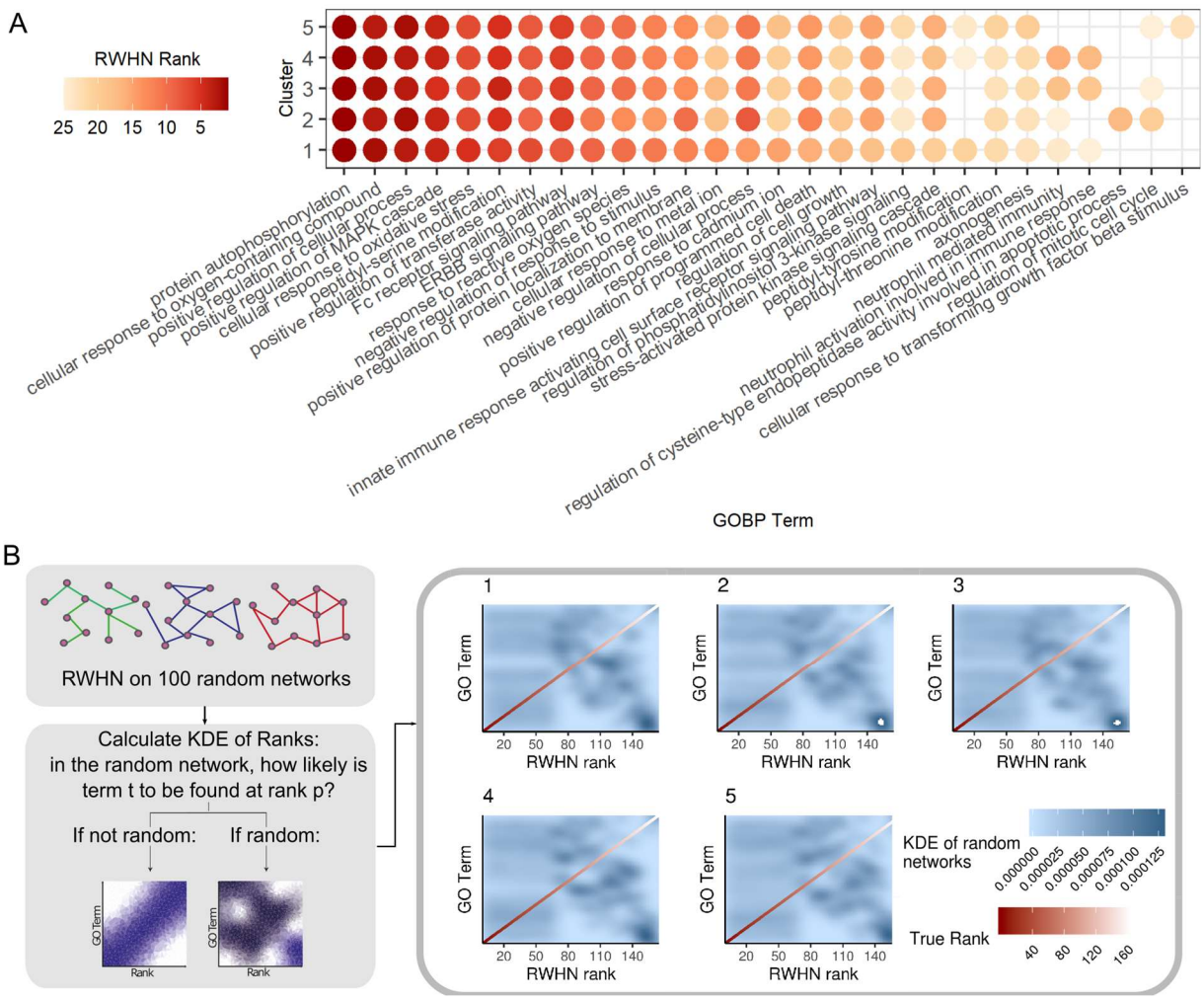
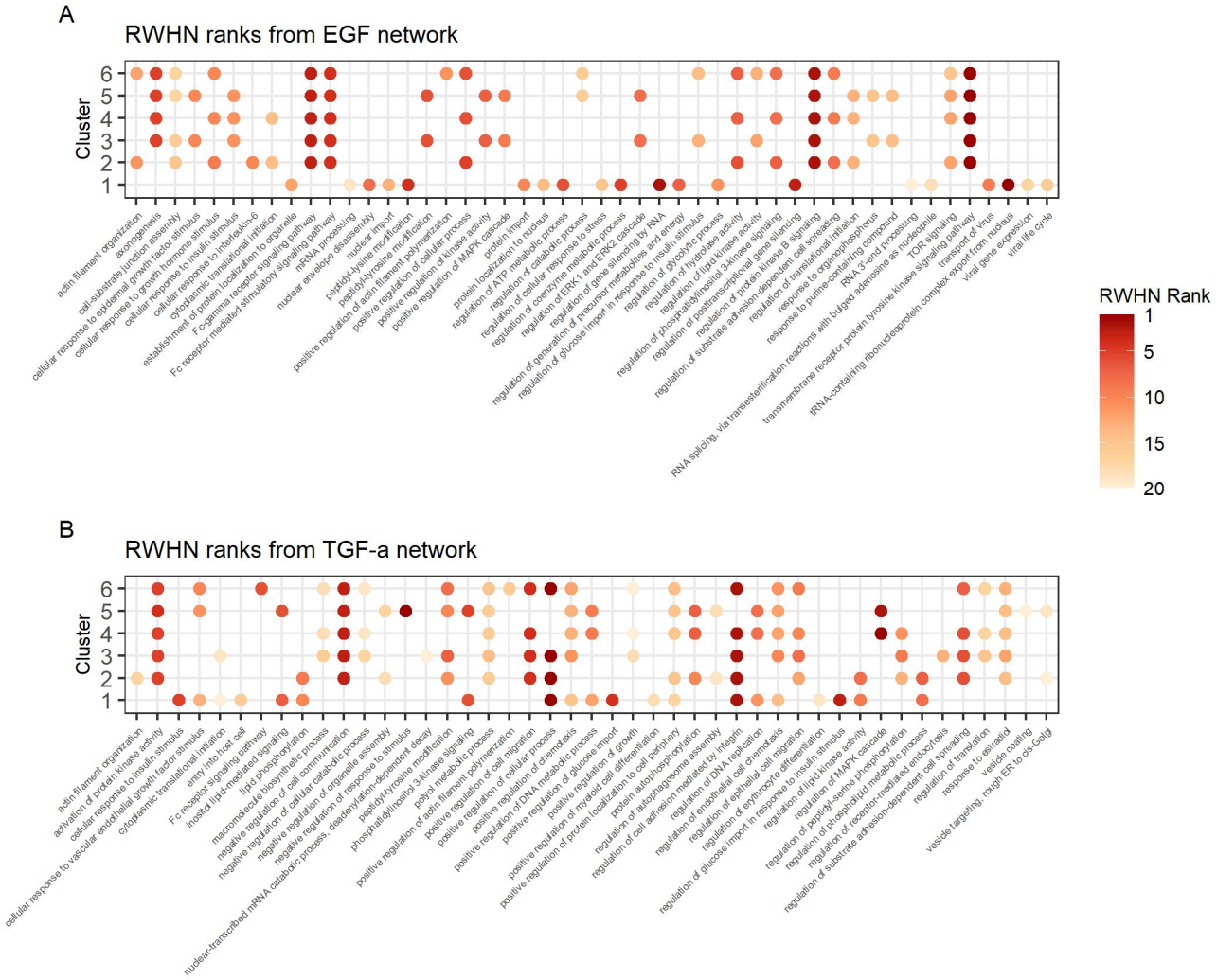


Figure 3. Output of RWHN run on validation data. A) Top ranked terms from each set of seed nodes (equivalent to the clusters, inset) when RWHN was run on the validation network. The top 20 ranked terms from each set of seed nodes is shown. B) The validation network was permuted 100 times and RWHN run over each permutation with seed nodes set to sites found in clusters 1 to 5. KDE was calculated to estimate how frequently each term appeared at each rank in the random networks. If the random networks generated the same result as the actual network we would expect the KDE plots to closely resemble the trend of the true rank.



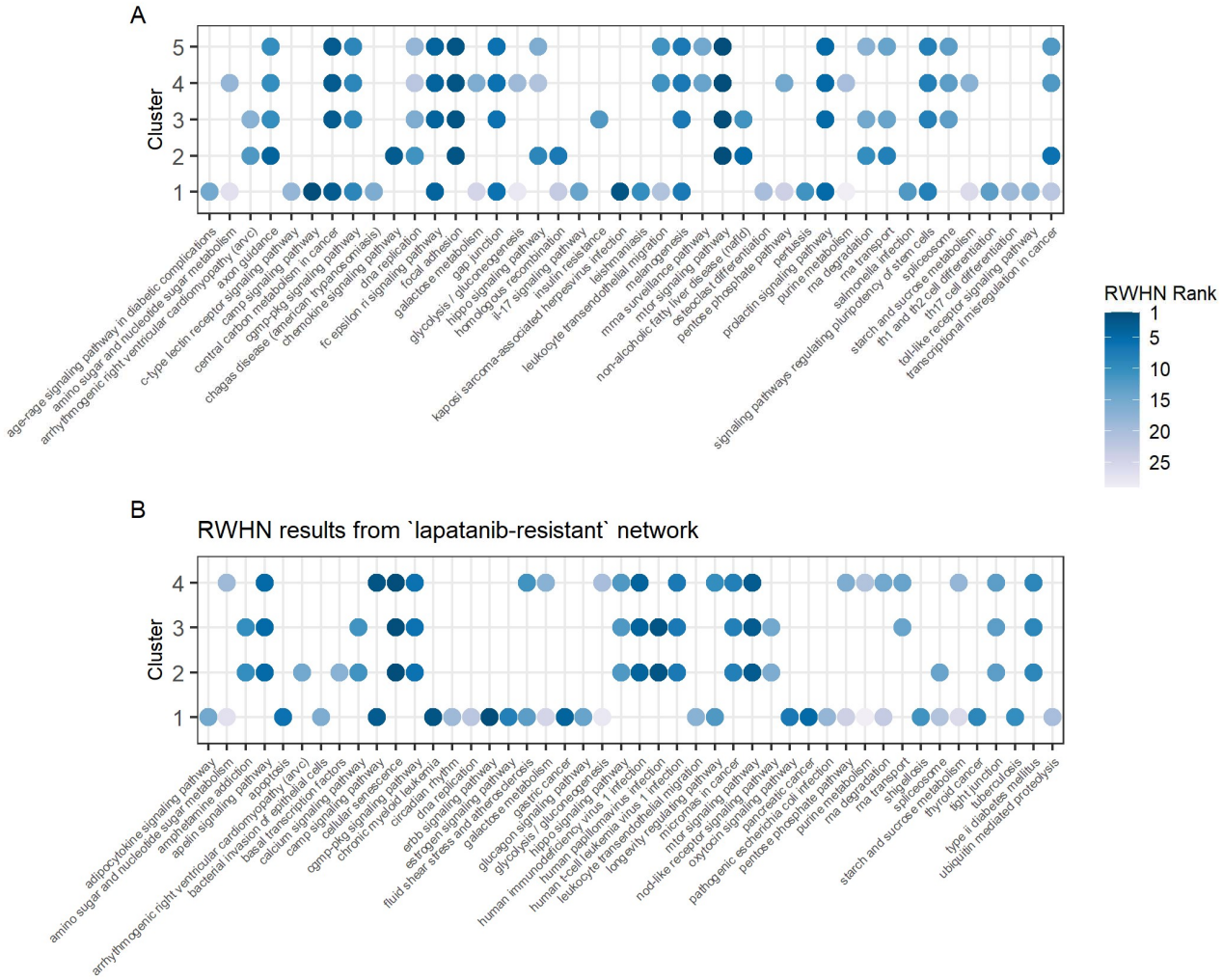
631

632 **Figure 4.** Output of the RWHN algorithm run on the EGF (A) and TGF-α (B) networks constructed from

633 the Francavilla *et al.* [19] dataset. The seed nodes were set to the clustered phosphorylated sites. The top

634 20 ranked terms for each set of seed nodes are shown.

635



636

637 **Figure 5.** Output of the RWHN algorithm run on the total (lapatinib regulated sites in lapatinib-resistant
638 and parental cell lines) (A) and lapatinib-resistant only (B) networks from the Ruprecht *et al.* [20] data
639 with KEGG pathways. The seed nodes were set to the clustered phosphorylated sites. The top 30 ranked
640 pathways for each set of seed nodes are shown.

641