Published in final edited form as:

Pac Symp Biocomput. 2023 ; 28: 73–84.

Prediction of Kinase-Substrate Associations Using The Functional Landscape of Kinases and Phosphorylation Sites

Marzieh Ayati[1,†], **Serhan Yilmaz**[2], **Filipa Blasco Tavares Pereira Lopes**[3,4], **Mark Chance**[3,4,5], **Mehmet Koyuturk**[2,4,5]

[1]Department of Computer Science, University of Texas Rio Grande Valley,Edinburg, TX

[2]Department of Computer and Data Sciences, Case Western Reserve University, Cleveland, OH

[3]Department of Nutrition, Case Western Reserve University, Cleveland, OH

[4]Center for Proteomics and Bioinformatics, Case Western Reserve University, Cleveland, OH

[5]Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH

Abstract

Protein phosphorylation is a key post-translational modification that plays a central role in many cellular processes. With recent advances in biotechnology, thousands of phosphorylated sites can be identified and quantified in a given sample, enabling proteome-wide screening of cellular signaling. However, for most (> 90%) of the phosphorylation sites that are identified in these experiments, the kinase(s) that target these sites are unknown. To broadly utilize available structural, functional, evolutionary, and contextual information in predicting kinase-substrate associations (KSAs), we develop a network-based machine learning framework. Our framework integrates a multitude of data sources to characterize the landscape of functional relationships and associations among phosphosites and kinases. To construct a phosphosite-phosphosite association network, we use sequence similarity, shared biological pathways, co-evolution, co-occurrence, and co-phosphorylation of phosphosites across different biological states. To construct a kinase-kinase association network, we integrate protein-protein interactions, shared biological pathways, and membership in common kinase families. We use node embeddings computed from these heterogeneous networks to train machine learning models for predicting kinase-substrate associations. Our systematic computational experiments using the PhosphositePLUS database shows that the resulting algorithm, NetKSA, outperforms two state-of-the-art algorithms, including KinomeXplorer and LinkPhinder, in overall KSA prediction. By stratifying the ranking of kinases, NetKSA also enables annotation of phosphosites that are targeted by relatively less-studied kinases.

Keywords

Phosphoproteomics; Kinase-substrate association; Network embedding

† marzieh.ayati@utrgv.edu .

1. Introduction

Protein phosphorylation is one of the most important post-translational modifications that play an important role in cellular signaling. Phosphorylation involves phospho-proteins whose activity can be altered by the phosphorylation of their specific sites (a.k.a substrate), kinases that phosphorylate the phospho-proteins at specific sites, and phosphatases that dephosphorylate these proteins. Dysregulation of the kinase-substrate associations are regularly observed in complex diseases, including cancer. Therefore, kinases have emerged as an important class of drug targets for many diseases.[1]

Recent advances in mass spectrometry (MS) based technologies drastically enhanced the accuracy and coverage of phosphosite identification and quantification. However, most identified phosphosites do not have kinase annotations, and large scale and reliable prediction of which kinase can phosphorylate which phosphosites remains challenging. In the last decade, several computational methods are developed to predict kinase-substrate associations (KSAs). The earlier KSA prediction methods focus mainly on sequence motifs recognized by the active sites of kinases..[2–4] Later methods integrate other contextual information such as protein structure and physical interactions to improve the accuracy of prediction methods.[5–8] Recently, we developed CophosK,[9] the first kinase-substrate prediction algorithm that utilizes large-scale mass spectrometry based phospho-proteomic data to incorporate contextual information. While these tools improve the kinase-substrate associations prediction, the knowledge about the substrates of kinases is still unequally distributed, where 87% of phosphosites are assigned to 20% of well-studied kinases.[10]

In parallel, machine learning algorithms that utilize network models gain significant traction in computational biology.[11,12] Inspired by these developments, we here develop a comprehensive framework for integrating broad functional information on kinases and phospho-proteins to build machine learning models for predicting kinase-substrate associations. Our framework uses heterogeneous network models to represent the functional relationships between phosphorylation sites, as well as kinases. Namely, we integrate structural, evolutionary, functional, and contextual information to characterize the landscape of functional relationships and associations among phosphosites and kinases. Since MS-based phosphoproteomic data can present a relatively unbiased view of signaling states, we also incorporate co-occurrence and co-phosphorylation across multiple MS-based phosphoproteomic studies into network construction. After constructing phosphosite association and kinase association networks, we use node embedding algorithms to derive low-dimensional vector representations for phosphosites and kinases, which are in turn used to train machine learning models.

We systematically investigate the predictive performance of reliability of the proposed framework, NetKSA, using established kinase-substrate associations from PhosphositePLUS. Using a cross-validation framework in two problem settings (link prediction and prioritization), we investigate the effect of the network embedding algorithms, the contribution of different types of networks, the value added by network topology, and compare the performance of NetKSA against state-of-the-art algorithms. In order to mitigate the bias toward well-studied kinases in the KSA prediction,[13] we

propose a kinase stratification strategy based on the number of known substrates. Our results show that NetKSA, outperforms state-of-the-art methods in overall prediction performance. Finally, we observe that the performance of NetKSA is robust to the choice of network embedding algorithms, while each type of network contributes valuable information that is complementary to the information provided by other networks.

2. Materials and Methods

The workflow of the proposed framework for kinase-substrate association prediction is shown in Figure 1. As seen in the figure, we first construct two networks, one to model the functional relationship between phosphorylation sites and the other to model the functional relationship between kinases. Subsequently, for each phosphosite and for each kinase, we compute low-dimensional embeddings using a node embedding algorithm on the respective network. Finally, we use these embedding as feature vectors and kinase-substrate associations obtained from PhosphoSitePLUS as training examples to train models for predicting kinase-substrate associations.

2.1. PhosphoSite Association Network

We define a PhosphoSite Association Network as a network $G_S(V_S, E_S)$ that represents *potential* functional relationships between pairs of phosphosites. In this network, V_S denotes the set of nodes in the network, each of which represents a phosphorylation site. The edge set E_S denotes the set of pairwise functional relationships between phosphosites, where an edge $s_i s_j \in E$ between phosphosites $s_i, s_j \in V$ may represent one of the following relationships:

- **Functional, Evolutionary, and Structural Association.** PTMCode is a database of known and predicted functional associations between phosphorylation and other post-translational modification sites.[14] The associations included in PTMCode are curated from the literature, inferred from residue co-evolution, or are based on the structural distances between phosphosites. We utilize PTMcode as a direct source of functional, evolutionary, and structural associations between phosphorylation sites.
- **Sequence Similarity.** We download the sequences within ±7 residues around each site in the protein sequence from PhosphositePLUS, and perform sequence alignment using BLOSUM62 scoring method. There is an edge between two sites s_i and s_j if their distance is less than 3 standard deviation below average across all pairs of sites.
- **Shared Pathways.** We use PTMsigDB as a reference database of site-specific phosphorylation signatures of kinases, perturbations, and signaling pathways.[15] While PTMSigDB provides data on all post-translational modifications, we here use the subset that corresponds to phosphorylation. There are 2398 phosphosites that are associated with 388 different perturbations and signaling pathways. We represent these associations as a binary network of signaling-pathway associations among phosphosites, in which an edge between two phosphosites

indicates that the phosphorylation of the two sites is involved in the same pathway.

- **Co-Occurrence.** Li et al.[16] show that phosphorylation sites that are modified together tend to participate in similar biological process. Based on this observation, they construct a binary occurrence profile for each phosphosite, where a 1 indicates that the site is identified in a given study, They then assess the co-occurrence of pairs of sites in terms of the mutual information between the respective occurrence profiles. Here, following Li et al.,[16] we use high-throughput MS analyses across 88 different studies from phosphoSitePLUS[17] to assess the co-occurrence of phosphorylation site. These studies include data from 16 human tissue as well as 28 cultural cell lines and 44 disease cells. We include an edge between two sites s_i and s_j if the p-value of their co-occurrence is less than 0.005.
- **Co-Phosphorylation.** Co-phosphorylation (Co-P) refers to correlated phosphorylation of two phosphosites across samples withing a given study.[18] While co-occurrence captures the relationship between pairs of sites that tend to appear in similar contexts at a broader scale, Co-P captures finer-scale correlations between the dynamic ranges of the phosphorylation levels of site pairs. To incorporate Co-P in the site association network, we use data from 9 mass spectrometry-based phosphoproteomic studied that represent a broad range of biological states and provide sufficient number of samples to enable reliable assessment of Co-P.[9] These datasets include data from three breast cancer studies,[19–21] two ovarian cancer studies,[20,22] one colorectal cancer,[23] one lung cancer,[24] one Alzheimer's disease[25] and one retinal pigmented eputhelium data.[26]

Using each pair of sites that are identified in each dataset, we compute as $c_D(i, j)$ the co-P between site i and site j as measured by Biweight-midcorrelation of their phosphorylation profiles in dataset D. We then compute R^2 values for each pair of sites in each dataset by adjusting for the number of samples n_D in dataset D:

$$R^2_D(i,j) = 1 - \frac{n_D - 1}{n_D - 2}\left(1 - c_D(i,j)^2\right) \tag{1}$$

Finally, we integrate these individual co-P scores as follows:

$$c_{integrated}(i,j) = 1 - \prod_{D \in \mathcal{D}_{ij}} \left(1 - R^2_D(i,j)\right) \tag{2}$$

where D_{ij} denotes the set of datasets in which sites i and j are both identified. In the integrated Co-P network, we include an edge between two sites s_i and s_j if the absolute value of their co-phosphorylation is larger than 2 standard deviation of the average across all pairs of sites.

Note that the integrated phosphosite association network is a heterogeneous multiplex network, where the nodes are from a common space (phosphorylation sites) and edges

in each network have different semantics. In recent years, many algorithms have been developed for computing embeddings for multiplex networks, which also account for the heterogeneity of the edges.[27–29] However, these algorithms are usually based on the inherent assumption that the overlap between the nodes of the networks is considerably large,[30] which is not the case in our application. For this reason, we here focus on assessing the value of the overall network model, as opposed to the algorithm used for integrating the networks or computing multiplex embeddings. With this motivation, we represent each network as a binary network by applying conservative edge inclusion criteria separately for each network, as described above. Subsequently, we integrate these networks into a single network by including an edge between two sites if there is an edge between them in at least one of the networks.

2.2. Kinase Association Network

We define a Kinase Association Network as a network $G_k(V_k, E_k)$ that represents functional relationship between pairs of kinases. In this network, V_k denotes the set of nodes each of which represents a kinase. The edge set E_k denotes the set of pairwise functional relationships between kinases. There is an edge $k_\ell k_r \in E_k$ between kinases $k_\ell\, k_r \in V_k$ if the two kinases have one of the following relationships:

- **Protein-Protein Interaction (PPI).** If two kinases k_ℓ and k_r physically interact, then there is an edge between k_ℓ and k_r. In our experiments, we use the PPIs that are annotated as ”physical” in the BIOGRID PPI database[31] to infer the PPI edges in the network.
- **Biological Pathways.** If two kinases k_ℓ and k_r are reported to have a role in the same pathway, then there is an edge between k_ℓ and k_r. In our experiments, we use mSigDB, which provides a collection of canonical pathways and experimental signatures.[32]
- **Kinase Families.** If two kinases k_ℓ and k_r belong to the same family according to the Human Kinome database,[33] then there is an edge between them.

2.3. Computing Network Profiles for Sites and Kinases

To obtain a network profile for each phosphosite and each kinase, we use node embedding. Given a graph G, a node embedding is a mapping $f\colon v_i \rightarrow y_i \in R^d$ such that $d \ll |V|$ and the function f preserves some proximity measure defined on graph G.[34] In other words, a node embedding maps each node to a low-dimensional feature vector, aiming to preserve the network proximity between nodes. Many node embedding algorithms have been developed in recent years, and the performance of these algorithms depends on the application, the nature of the learning problem, and the topology of the network. For this reason, in our experiments, we use four different node embedding algorithms[34–37] to comprehensively evaluate the value of the information provided by the networks we utilize, independent of the node embedding algorithm that is used. For each site s_i in G_s, we compute node embedding $x_i \in \mathbb{R}^d$ and for each kinase k_ℓ in G_k we compute node embedding $y_\ell \in \mathbb{R}^d$. We do this separately for each network embedding algorithm, using the default parameters in each algorithm, and using different values of d.

2.4. Predicting Kinase-Substrate Associations

We use the sets of known KSAs obtained from PhosphoSitePLUS (PSP)[17] as a positive reference for training and testing our models. We generate negative training sets of equal size by selecting, uniformly at random, kinase-substrate pairs that are not reported to be associated in PSP. To train the models, we concatenate the network profiles of site-kinase pairs to obtain a $2d$-dimensional feature vector for the pair: $f(s_i, k_\ell) = x_i \| y_\ell = \left(x_i^{(1)}, \ldots, x_i^{(d)}, y_\ell^{(1)}, \ldots, y_\ell^{(d)}\right)$. We consider two variants of KSA prediction:

(I) Link Prediction.—We formulate the KSA prediction problem as a binary classification problem for a given kinase-site pair, i.e., given a list of established kinase-site associations, site-site association and kinase-kinase association networks G_s and G_k, and a kinase-site pair (s_i, k_j), our objective is to assess the likelihood that s_i is a target site for k_j. For this purpose, we train a Random Forest model by using the concatenated embeddings as features. Using 5-fold cross validation, we assess the overall performance of the method using area of the ROC curve (AUC).

(II) Prioritization of Kinases for Phosphosites.—In practice, the kinase-substrate association prediction often manifests itself as a prioritization problem. The scientist discovers a new phosphorylation site that is associated with a certain process and phenotype and would like to know which kinase is responsible for the phosphorylation of that site. This problem is formulated as follows: Given a list of established kinase-site associations, site-site association and kinase-kinase association networks G_s and G_k, and a site s_i, rank kinases based on their likelihood of being associated with s_i. For this task, we use a Random Forest model using concatenated embeddings as well, but we use leave-one-out cross-validation to assess the performance of the resulting models. In this case, we use hit@k accuracy as the performance criterion. Namely, using each site as a test site, we report the fraction of times in which the actual kinase responsible for phosphorylating the site is ranked in the top k for that site, where $k \in \{1, 5, 10, 20\}$.

2.5. Elucidating and Mitigating Bias in KSA Prediction

In order to study the bias in the KSA predictions toward the more well-studied kinases,[13] we stratify the kinases based on the number of their known substrates which are in the phosphosite association network. Letting δ_ℓ denote the number of known substrates of kinase k_ℓ we partition the kinases into three categories: (i) The poor kinases where $\delta_\ell < 5$, (ii) the average kinases, where $5 \leq \delta_\ell < 20$, and (iii) The rich kinases where $\delta_\ell \geq 20$. We then train separate models for each kinase category, by using kinases that belong to a specific category while training the respective model. Subsequently, when prioritizing the kinases for each phosphosite, we rank the kinases within their own category.

The premise of this approach is that the kinases in each category should compete with the kinases in the same category as themselves, and scientists should be able to separately investigate the rankings in each category. This will potentially enable discovery and experimental validation of relatively less-studied kinases. We evaluate the performance of the all the methods by considering this stratified analysis, as well as by ranking all kinases. This approach provides insights into the bias associated with each approach, i.e., how much

a method improves its chances of making an accurate prediction by preferring well-studied kinases.

3. Results and Discussion

We use PhosphoSitePLUS as a reference dataset for kinase-substrate associations (KSAs).[17] Considering the phosphosites and kinases in our networks, we use 2083 KSAs from PhosphositePLUS in our computational experiments. To evaluate the performance of the kinase-substrate association prediction method, we limit the site network to the known substrates obtained from PhosphoSitePLUS. We remove the individual nodes that are not connected to any other nodes from both of the networks. The number of sites and edges in the final kinase-kinase and phosphosite-phosphosite association networks and their types are shown in Figure 2(a). The overlaps between different types of association networks are shown in Figure 2(b). The low overlap between different phosphosite-phosphosite association networks suggests that all different types of networks provide information that are potentially complementary with each other.

3.1. Kinase-Substrate Association as Link Prediction

We first use different embedding methods, and 5-fold cross validation to evaluate the performance of NetKSA in predicting KSAs formulated as link prediction. In our computational experiments, we consider different numbers of embedding dimensions and its effect on the performance. We find out that $d = 16$ is optimal for all algorithms considered, thus we perform all remaining experiments using 16 dimensions for the embedding vectors.

The link prediction performance of NetKSA using different embedding algorithms is presented in Figure 3(a). We evaluate the performance for all the KSAs, as well as KSAs that its kinase belongs to different category (i.e. poor, average, rich) separately. In this analysis, there are 103 kinases in the poor category ($\delta < 5$), 64 kinases in the average category ($5 \leq \delta < 20$), and 21 kinases in the rich category ($\delta \geq 20$) (the rest of kinases in the kinase-kinase association network do not have any target sites that are present in the site-site association network). These kinases corresponds to 218 KSAs in poor category, 613 KSAs in the average category and 1252 KSAs in the rich category. The negative set for the training of the model is randomly generated while keeping the proportion of KSA categories. The bar plots show the average across 10 runs. As seen in the figure, the prediction performance highly depends on the the kinase category and the AUC observed by considering all kinases together closely follows the prediction performance for rich kinases. This observation demonstrates the importance of performing stratified analyses to accurately characterize the performance of KSA prediction as a function of what is already known about the kinase and characterize the bias in algorithms.

As seen in Figure 3(a), the prediction performance of NetKSA is robust to the choice of network embedding algorithms. We select DNGR for further analyses due to its slightly better overall performance that is also most balanced across different kinase categories.

To evaluate the value added by the network to the prediction, we randomly permute site association and kinase association networks while preserving the degree distribution and

apply NetKSA by using the permuted networks in place of the actual networks. The results of this analysis are presented in Figure 3(b). As seen in the figure, the prediction performance using original networks is one or more standard deviation(d) above the prediction performance of the method when using permuted networks. This result shows the networks contribute valuable information for KSA prediction. Importantly, randomization of the prediction performance declines more when the phosphosite network is permuted, suggesting that the functional information on the phosphosites provides significant and specific information on the kinase(s) that target(s) the phosphosites.

It is also interesting that the poor kinase category benefits the most in comparison with other categories when the original networks are used. This shows that the information provided by functional associations among sites and kinases reduce the gap between under-studied and well-studied kinases. Note that the models that are based on permuted networks perform better than what would be expected at random, suggesting that these models can learn bias in the benchmarking data to appear as if they are learning what they are designed to learn. However, the performance of the model that is trained on both permuted networks is equal to what would be expected at random for poor kinases, demonstrating that the validation strategy we employ here (stratification of kinases and comparison against permuted networks) provides significant insights into what these models actually learn.

3.2. Contribution of Different Networks on Prediction Performance

In order to evaluate the contribution of different types of networks in capturing the landscape of functional association among phosphosites and kinases, we evaluate the performance of KSA predictions using different networks. For this analysis, we perform KSA prediction using 5-fold cross validation, by adding one network at a time to the integrated network of kinase-kinase and phosphosite-phosphosite associations, while keeping the other network fully integrated. The results of this analysis are shown in Figure 4. As seen in the figure, as we add different types of functional information for the sites and kinases, the prediction performance improves. We also evaluate the KSA coverage as the proportion of existing KSAs for which prediction can be made. The new networks add information about the the individual sites and kinases and connect them to other nodes, and consequently increase the KSA coverage. Finally, we observe that the information contributed by different phosphosite networks is more complementary to each other as compared to the kinase networks, which is not surprising as the overlap between these networks is also considerably low.

3.3. Prioritization of Kinases for Phosphorylation Sites

To test the effectiveness of our method, we use leave-one-out cross validation. Namely, for each phosphosite, we hide the association between phosphosite and its known kinase (called the target kinase), and we use other reported KSAs to rank the likely kinases for that phosphosite. For this analysis, we use dngr as the embedding method and random forest with 100 classification trees as the score prediction model. For each phosphosite, we rank all kinases based on the calculated score and determine the rank of the target kinase across all kinases. If the target kinase is within the top $k \in \{1, 5, 10, 20\}$, it is considered a the true positive.

We compare our method with two other state-of-the-art methods, KinomeXplorer and LinkPhinder, that also use the network for KSA prediction. KinomeXplorer[5] utilizes the sequences match scoring and network proximity of kinases and substrates to predict KSAs. It is an improved version of NetworKIN[4] and NetPhorest.[38] LinkPhinder[39] is also another predictive model that utilizes the motif characteristics to create a knowledge graph and uses statistical relational learning and node embedding to predict KSAs. The result of this analysis is presented in Figure 5. As seen in the figure, the proposed method with kinase stratification outperform all methods in overall prediction performance, and also average and rich categories. For the poor kinases, the LinkPhinder presents a better result for top 1 and top 5 ranking. We believe integration of different data sources in NetKSA help extracting the relationship among sites and kinases which leads to a better overall performance.

3.3.1. Kinase Stratification—In the kinase prioritization, we rank the kinases in each category (i.e poor, average, rich) separately, and determine if the target kinase is ranked in top k of its category. The premise of this approach is that the kinase that are understudied does not to compete with the well-studies kinases. Using kinase stratification, the hypothesis is that it is more likely that the target kinase wins the competition in ranking compare to the kinases in its own category. We apply this strategy on NetKSAand also KinomeXplorer and LinkPhinder. The result of this analysis is presented in Figure 5. For each bar in the figure, the dark section is the performance without kinase stratification, and the light-color section is the improvement of the performance using the kinase stratification.

4. Conclusion

In this paper, we integrated a multitude of data sources to characterize the landscape of functional relationships and associations among phosphosites and kinases. As a result, we construct two heterogeneous networks presenting functional association among phosphosites and kinases. These networks incorporating static and dynamic data and present an extraordinary value in prediction of kinase-substrate association, and have great potential for analysis of phosphoproteomics data and identification of drug targets. Generalizing the method to include all the identified phosphosites is a challenging task which may point to an interesting research avenue to be addressed by future studies. Moreover, the kinase stratification approach to mitigate the bias toward well-studied kinases provides a great opportunity to researchers to investigate and study kinases in different categories separately.

Acknowledgments

This work was supported by National Institutes of Health grant R01-LM012980 from the National Library of Medicine.

Availability:

The code and data are available at compbio.case.edu/NetKSA/.

References

1. Ferguson FM et al. Nature reviews Drug discovery, 17(5):353–377, 2018. [PubMed: 29545548]
2. Durek P, et al. BMC bioinformatics, 10(1):1–17, 2009. [PubMed: 19118496]

3. Obenauer JC, et al. Nucleic acids research, 31(13):3635–3641, 2003. [PubMed: 12824383]
4. Linding R, et al. Cell, 129(7):1415–1426, 2007. [PubMed: 17570479]
5. Horn H, et al. Nature methods, 11(6):603–604, 2014. [PubMed: 24874572]
6. Hjerrild M, et al. Journal of proteome research, 3(3):426–433, 2004. [PubMed: 15253423]
7. Song C, et al. Molecular & Cellular Proteomics, 11(10):1070–1083, 2012. [PubMed: 22798277]
8. Hobert EM et al. Journal of the American Chemical Society, 134(9):3976–3978, 2012. [PubMed: 22352870]
9. Ayati M, et al. PLOS computational biology, 2019.
10. Needham EJ, et al. Science signaling, 12(565):eaau8645, 2019. [PubMed: 30670635]
11. Gaudelet T, et al. Briefings in bioinformatics, 22(6):bbab159, 2021. [PubMed: 34013350]
12. Muzio G, et al. Briefings in bioinformatics, 22(2):1515–1530, 2021. [PubMed: 33169146]
13. Deznabi I, et al. Bioinformatics, 36(12):3652–3661, 2020. [PubMed: 32044914]
14. Minguez P, et al. Nucleic acids research, 43(D1):D494–D502, 2014. [PubMed: 25361965]
15. Krug K, et al. Molecular & cellular proteomics, 18(3):576–593, 2019. [PubMed: 30563849]
16. Li Y, et al. PLoS computational biology, 13(5):e1005502, 2017. [PubMed: 28459814]
17. Hornbeck PV, et al. Nucleic acids research, 43(D1):D512–D520, 2014. [PubMed: 25514926]
18. Ayati M, et al. Bioinformatics, 2022.
19. Huang K.-l., et al. Nature communications, 8:14864, 2017.
20. Mertins P, et al. Molecular & cellular proteomics, 2014.
21. Mertins P, et al. Nature, 534(7605):55, 2016. [PubMed: 27251275]
22. Zhang H, et al. Cell, 166(3):755–765, 2016. [PubMed: 27372738]
23. Abe Y, et al. Scientific reports, 7(1):1–12, 2017. [PubMed: 28127051]
24. Wiredja D. PhD thesis, Case Western Reserve University, 2018.
25. Dammer EB, et al. Proteomics, 15(2–3):508–519, 2015. [PubMed: 25332170]
26. Chiang C, et al. Journal of Biological Chemistry, 292(48):19826–19839, 2017. [PubMed: 28978645]
27. Cho H, et al. Cell systems, 3(6):540–548, 2016. [PubMed: 27889536]
28. Valdeolivas A, et al. Bioinformatics, 35(3):497–505, 2019. [PubMed: 30020411]
29. Zeng X, et al. Bioinformatics, 35(24):5191–5198, 2019. [PubMed: 31116390]
30. Li M et al. In International Conference on Complex Networks and Their Applications, pages 39–52. Springer, 2020.
31. Chatr-Aryamontri A, et al. Nucleic acids research, 45(D1):D369–D379, 2017. [PubMed: 27980099]
32. Liberzon A, et al. Bioinformatics, 27(12):1739–1740, 2011. [PubMed: 21546393]
33. Manning G, et al. Science, 298(5600):1912–1934, 2002. [PubMed: 12471243]
34. Grover A et al. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, pages 855–864, 2016.
35. Perozzi B, et al. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 701–710, 2014.
36. Cao S, et al. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 30, 2016.
37. Tsitsulin A, et al. In Proceedings of the 2018 world wide web conference, pages 539–548, 2018.
38. Miller ML, et al. Science signaling, 1(35):ra2–ra2, 2008. [PubMed: 18765831]
39. Nováček V, et al. PLoS computational biology, 16(12):e1007578, 2020. [PubMed: 33270624]

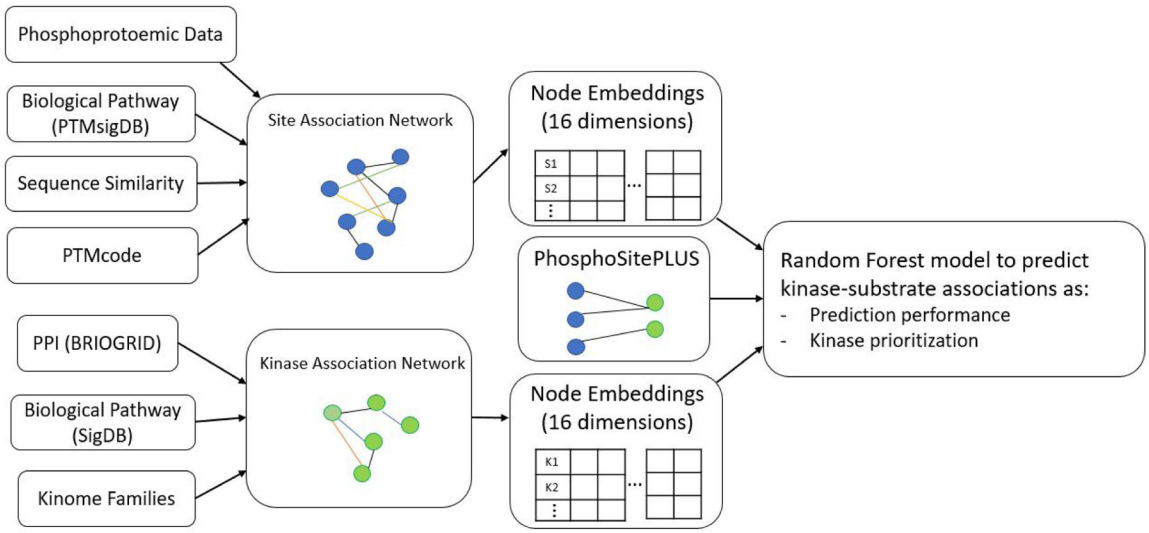

Fig. 1. Workflow of NetKSA.

We first construct two networks to represent the functional relationships and associations among phosphosites and kinases. After construction of networks, we use node embedding algorithms on each network to compute a low-dimensional representation for each node. We then use the kinase-substrate associations (KSAs) obtained from PhosphoSitePLUS to train machine learning models for predicting KSAs.

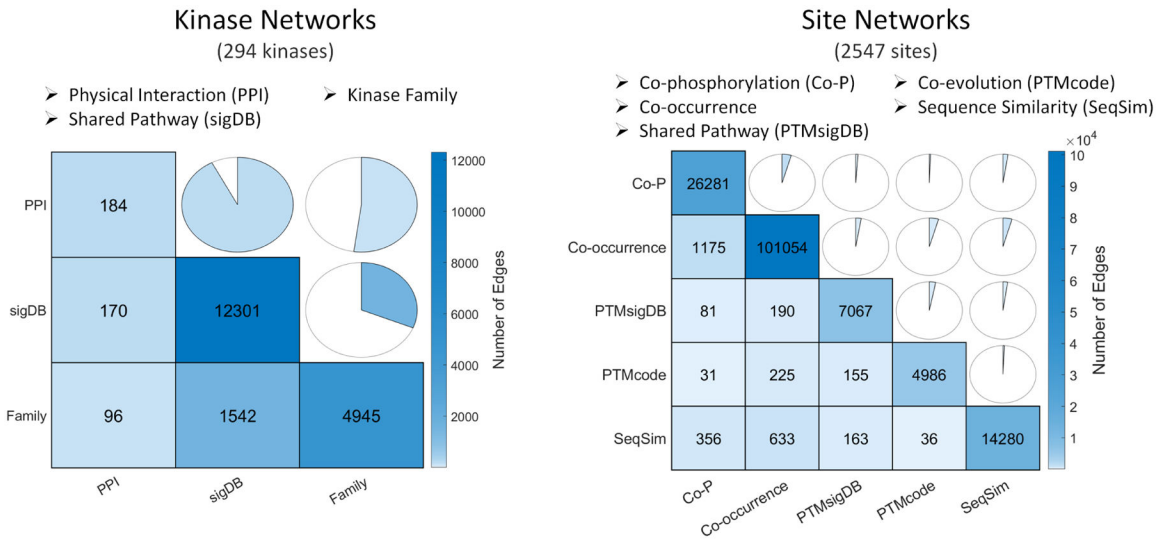

Fig. 2. Kinase-kinase and phosphosite-phosphosite association networks used in this study.

Plots show the edge overlap between different types of networks. Kinase networks are shown on the left, phosphosite networks are shown on the right. The number of edges in each network are given in the diagonals. In each subplot, the pie charts in the top right side indicate the overlap coefficients (size of intersection divided by the smaller of the size of two sets) between any two networks.

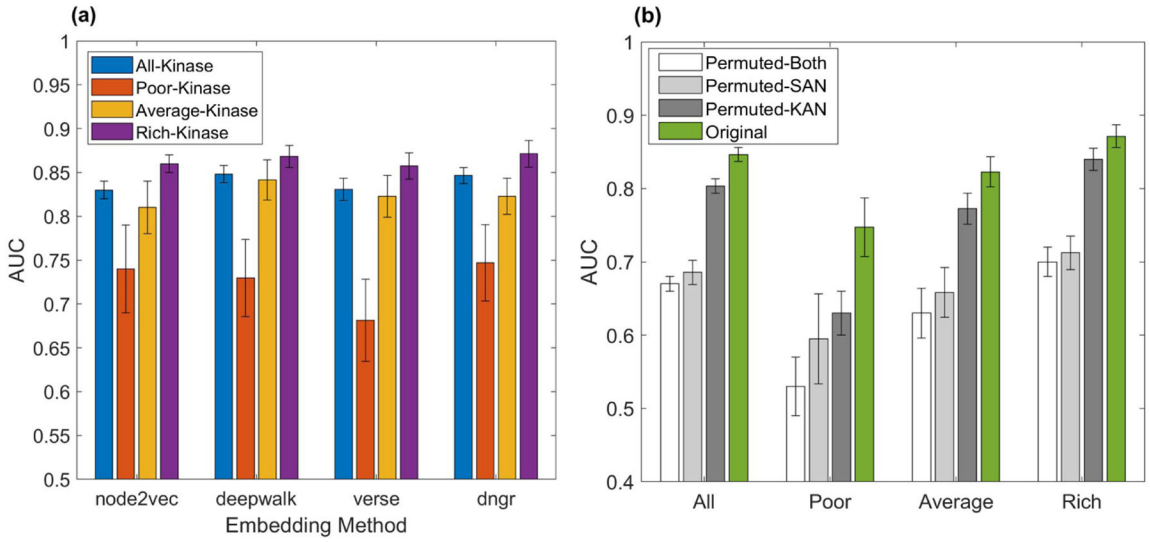

Fig. 3. The contribution of embedding algorithms and functional networks on KSA prediction performance.

(a) The AUC of the predictions of NetKSA using four different node embedding algorithms. For each embedding algorithm, the AUC is shown for all KSAs (blue bar), the KSAs where the kinase belongs to the poor category (red), the average category (gold), and rich category (purple). (b) The prediction performance of NetKSA using DNGR for node embedding using real vs. randomized networks. AUC on the real kinase-kinase and phosphosite-phosphosite association networks (green bar), when only the kinase association network is randomly permuted by preserving node degrees (dark grey), when only the site association network is permuted by preserving node degrees (light grey), when both networks are permuted (white). Each bar shows the average AUC across 10 runs and the error bar shows standard deviation.

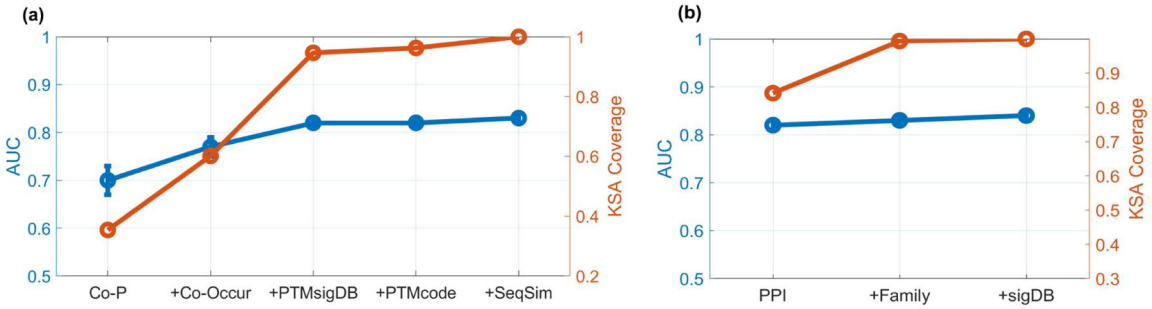

Fig. 4. Contribution of different types of networks on the prediction of KSAs.

The cumulative effect of each (a) phosphosite-phosphosite association network and (b) kinase-kinase association network on the AUC of predictions (left y axis; blue), and the coverage of kinase-substrate associations (right y axis; red) - the fraction of KSAs for which both the kinase and the site are present in the integrated network so that a prediction can be made.

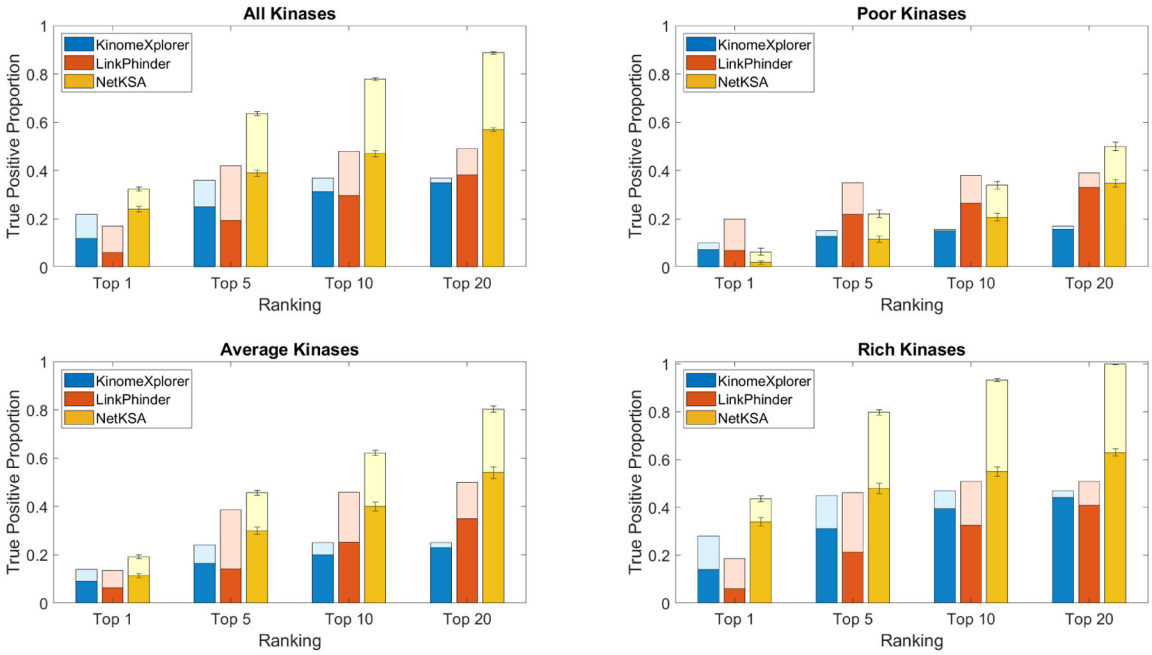

Fig. 5. Performance of NetKSA, KinomeXplorer and LinkPhinder in prioritizing kinases for a given phosphosite.

For each phosphosite, we perform leave-one-out cross validation by hiding the association between the phosphosite and one of its associated kinases (target kinase) to rank the likely kinases for the phosphosite using KinomeXplorer(blue), LinkPhinder(red), and proposed method using constructed networks (gold). We report the fraction of phosphosites for which the target kinase is ranked in the top 1, top 5, top 10 and top 20 predicted kinases by each method. For each bar, the dark section presents the result when all the kinases are ranked together, and the light section presents the improvement of performance when the target kinase is ranked within its category (with stratification). Each panel presents the performance on each category of kinases: poor ($\delta < 5$), average($5 \leq \delta < 20$), and rich ($\delta \geq 20$) kinases (as indicated in each panel).