DeepKinZero: Zero-Shot Learning for Predicting Kinase-Phosphosite Associations Involving Understudied Kinases

Iman Deznabi[1,2], Busra Arabaci[2], Mehmet Koyutürk[3], and Oznur Tastan[4]

[1] College of Information and Computer Science University of Massachusetts Amherst, MA
[2] Departmet of Computer Engineering, Bilkent University, Ankara, Turkey
[3] Center for Proteomics & Bioinformatics, Case Western Reserve University, OH
[4] Faculty of Natural Sciences and Engineering, Sabanci University, Istanbul, Turkey

Abstract. Phosphorylation is a key regulator of protein function in signal transduction pathways. Kinases are the enzymes that catalyze the phosphorylation of other proteins in a target specific manner. Although the advances in phosphoproteomics enable the identification of phosphosites at the proteome level, determining which kinase is responsible for phosphorylating a site remains an experimental challenge. Existing computational methods require several examples of known targets of a kinase to make accurate kinase specific predictions, yet for a large body of kinases, only a few or no target sites are reported. We present DeepKinZero, the first zero-shot learning approach to predict the kinase acting on a phosphosite for kinases with no known phosphosite information. DeepKinZero transfers knowledge from kinases with many known target phosphosites to those kinases with no known sites through a zero-shot learning model. The kinase specific positional amino acid preferences are learned using a bidirectional recurrent neural network. We show that DeepKinZero achieves significant improvement in accuracy for kinases with no known phosphosites in comparison to the baseline model and other methods available. By expanding our knowledge on understudied kinases, DeepKinZero can help to chart the phosphoproteome atlas.

Introduction

Protein kinases are a large family of enzymes that catalyze the phosphorylation of other proteins [1]. Phosphorylation involves the transfer of a phosphoryl group to the side chain of an amino acid residue in the substrate. The amino acid residue that receives the phosphoryl group is called the phosphorylation site, or briefly a *phosphosite*. Since they are the key regulators of protein function in a broad range of cellular activities, aberrant kinase function is implicated in many diseases [2], particularly in cancer [3, 4]. Several pathogenic human mutations also lie on known phosphorylation sites [5]. To this end, understanding the associations between kinases and phosphorylation sites holds the key to understand the signaling mechanisms in the healthy and diseased cells.

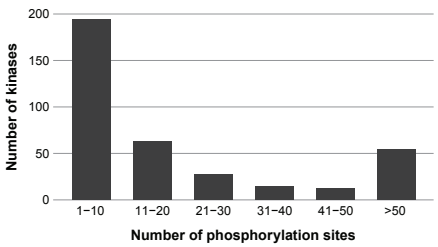

Fig. 1. The histogram of the number of experimentally validated target phosphosites for human kinases in PhosphositePlus database. Note that PhosphositePlus reports phosphosites only for 364 of the 518 human kinases.

With 518 identified kinases in the human genome and the transient nature of kinase-substrate interactions, it is experimentally challenging to determine the cognate kinase of a phosphosite. As underlined by a recent review [5], most of the phosphoproteome is uncharted: more than 95% of reported human phosphosites have no known kinase or associated biological function. There are several computational methods available to predict phosphosites [6–13] and earlier methods reviewed in [14]). Since they also provide kinase specific predictions, they can be used to predict associated kinases of a known phosphosite. These methods either utilize position specific scoring matrices to estimate the position preferences of each kinase or employ supervised machine learning models that use a collection of established kinase-phosphosite associations to model the relationship. However, the application of such tools is limited to kinases for which a substantial number of target phosphosites are available for training. For example,

MusiteDeep [13], uses deep learning to predict binding sites for kinases, and it exclusively focuses on kinase families with at least 100 experimentally verified phosphosites. Existing computational methods require several examples of known phosphosites of a kinase to make accurate predictions, yet for a large body of kinases no or only few target sites are reported (Figure 1). We present DeepKinZero, the first zero-shot learning approach to predict the kinase acting on a phosphosite for kinases with no known phosphosite information.

Zero-Shot Learning Model

Zero-shot learning aims at solving classification problems wherein the available training data does not contain examples of the desired classes [15]. The key to making predictions for classes with no training data (referred to as *unseen* or *zero-shot* classes) is to have side information which can be used to relate the classes. Based on these relations among classes, it becomes possible to transfer the knowledge obtained from classes that have positive training samples (referred to as *seen* class) [15] to the previously unseen classes. In this problem, we do not observe any phosphosites that are associated with a rare kinase (unseen class) in training, the zero-shot learning framework enables us to recognize a target site of this kinase by transferring knowledge from common kinases to the rare kinases. This can be achieved by establishing a relationship between the kinases using relevant auxiliary information, such as functional, sequence and structural characteristics of kinases. Figure 2 illustrates this idea.

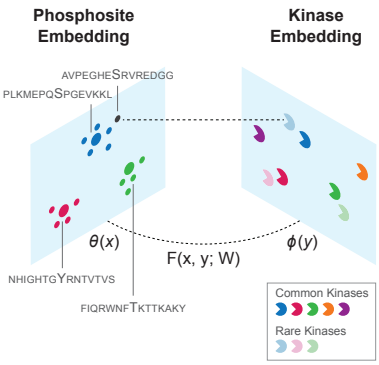

Fig. 2. Overview of the application of zero-shot learning to the prediction of kinase-phosphosite associations.

DeepKinZero takes the sequence of 15 residues centered on the phosphosite as input. We denote the associated kinase with class label y. The problem is formalized as a multi-class classification problem with many classes, where each input phosphosite sequence is associated with a kinase. For each kinase $y \in Y$, a "kinase embedding" vector $\phi(y) \in \mathbb{R}^m$ is computed based on information available on kinases. Following the work in structured output prediction [16] and prior work in zero-shot learning [15,17–22], we use a compatibility function $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ to model the mapping between the input and output embeddings. In this model, F takes a phosphosite - kinase pair (x_i, y_j) as input and returns a scalar value which is proportional to the confidence of associating the site, x_i, with kinase y_i. The probability that a given site is a target of a given kinase is calculated logistically from the compatibility function F:

$$p(y|x) = \frac{\exp(F(x,y))}{\sum_{y' \in Y_{te}} \exp(F(x,y'))} \tag{1}$$

As in [22], we use the following bi-linear compatibility function for input x and y :

$$F(x,y,W) = [\theta(x)^\top \quad 1] W [\phi(y)^\top \quad 1]^\top. \tag{2}$$

Kinase and Phosphosite Embeddings: We use four different data sources to represent kinases i) kinase hierarchy information as obtained from kinase.com, ii) Enzyme Commission(EC) classification of kinases, iii) ProtVec representation of kinase domain sequences, and iv) participating in common KEGG pathway. We expect "similar" kinases to be close according to the Euclidean metric in the embedded space. Similarly, for each phosphosite $x \in \mathcal{X}$, we compute phosphosite embedding vector, $\theta(x) \in \mathbb{R}^d$, that represents the phosphosite sequence in a d-dimensional space. To learn this embedding, we use two layers of Bidirectional Recurrent Neural Networks(BRNN) followed

by a dot attention layer over phosphosite embeddings. To avoid overfitting, we employ drop-out regularization. We also applied batch normalization to the output of LSTM cells to normalize the embeddings passed onto the ZSL model. The overall architecture is provided in (Figure 3).

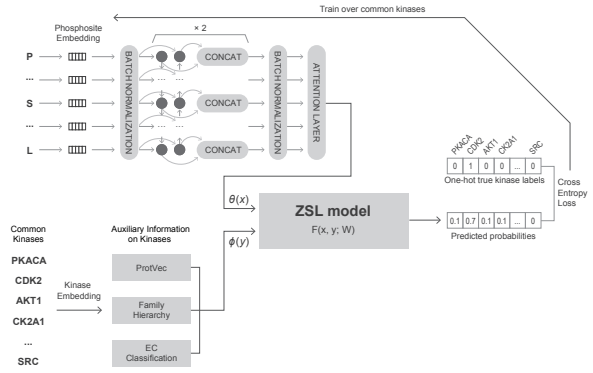

Fig. 3. The DeepKinzero model.

Model Training: We train the model end-to-end by connecting the BRNN model to ZSL model (Figure 3) by minimizing cross-entropy loss using Adam optimizer [23]. We employ drop-out regularization [24] and batch normalization in LSTM cells [25] to normalize the embeddings passed onto the ZSL model. The attention weights are initialized randomly from a normal distribution with a mean of 0 and standard deviation of 0.05. The learning rate and the number of iterations are optimized on validation data (see below for an explanation of the validation data). To reduce the variance of the model, we ensemble 10 models each of which trained with different initializations of the model parameters. The final class probabilities are obtained by averaging output probabilities over the ensemble.

Data Sources and Evaluation: We train and evaluate our models on the experimentally validated kinase-phosphosite associations obtained from the PhosphoSitePlus database [26]. After removing isoform and fusion kinases, following the evaluation protocol suggested in [17], we split the data into training, validation and test sets based on the number of sites that are associated with each kinase. Kinases with more than five sites are considered as training classes. The BRNN model and zero-shot learning models are trained on this set, which contains 12,901 phosphorylation sites associated with a total of 214 kinases. The validation set includes the kinase-phosphosite associations of kinases for which there are exactly five phosphorylation sites. This validation set includes 80 phosphorylation sites interacting with 17 different kinases. The remaining kinases with less than five positively labeled examples constitute the test or zero-shot classes. The test data includes 237 phosphorylation sites that belong to 112 classes.

Results

Performance: Figure 4 summarizes the results of using different phosphosite sequence embeddings. With respect to hit@1 and hit@3 metrics, the model trained with a BRNN coupled with ProtVec vectors performs the best, where the true kinase is predicted as the top kinase for more than 20% of the sites, and it is among the top 3 for more than 30% of the sites. With respect to hit@5 metric, the input representations have less effect on the prediction performance, where amino acid properties with BRNN delivers the highest hit@5 accuracy with the true kinase being among the top 5 for more than 40% of the sites. Additionally, we observe that the use of BRNN model improves the performance. The model without BRNN embeddings that uses One-Hot sequence embedding as input only returns the true kinase as the top prediction in 10.55% of the test cases. On the other hand, the model with BRNN and ProtVec site embeddings predict the right class with 21.52% accuracy.

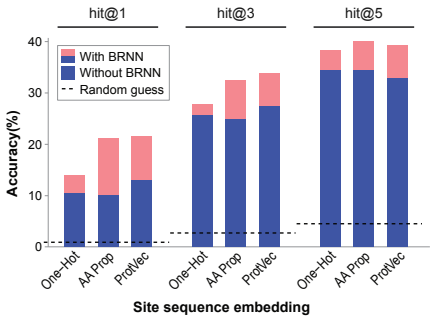

Fig. 4. Performance comparison of the models with the site sequence embeddings and with and without using a BRNN.

Note that random guess will achieve only 0.89% accuracy since there are 112 test classes. We also evaluated different combinations of kinase embedding (data not shown). The kinase hierarchy of kinases contributes the most to the accuracy of the model, achieving 17.72% accuracy when used as the sole auxiliary information on kinases.

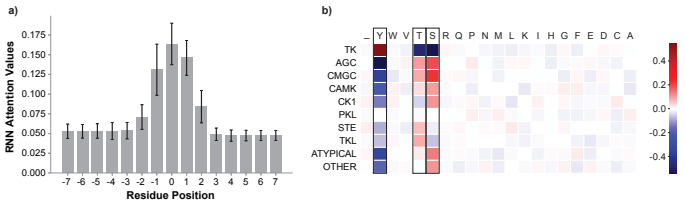

Fig. 5. (Best viewed in color) a) Average attention weights of the residue positions calculated over the ensemble BRNN model. Residue position 0 is the phosphosite position. b) Average zero-shot learning weights for each amino acid type at the phosphosite.

Comparison with Other Methods: In the literature, there are no models that we can directly compare our method against as they do not make predictions on the understudied kinases. However, there are two methods [27, 28] that aim at a different but a related problem. They predict the phosphosites for kinases with no known sites, which is the reverse scenario of our problem; we predict the kinase of a given phosphosite. Predikin [27] operates with a set of rules governing the amino acids around the phosphosites that are derived from 3D structures of kinases bound to their substrates. Because the Predikin server was not available, we were not able to carry out a comparison with this method.

The method proposed by [28] is based on the idea that, as compared to a random set of proteins, interaction partners of a kinase are more likely to be phosphorylated by that kinase. The method finds enriched motifs in the interaction partner sequences and use these motifs to predict protein sequences that a kinase can bind to. Our method predicts the kinase of a given phosphosite, whereas Wagih et al. predicts the phosphosite of a kinase. Thus, the two methods are not directly comparable but still, we conduct the following comparison. For the 112 zero-shot kinases, we predict the motifs by Wagih et al. model. If we consider the top motif returned, the method correctly matches 11 of the phosphosites of the 112 kinases, leading to 9.8% hit@1 accuracy. If we consider the top 5 motifs returned for each kinase, the correct phosphosite sequence matches 26 of phosphosites of the 112 kinase motifs leading to 23% hit@5 accuracy. These numbers are significantly lower than what DeepKinZero can achieve (21.52% and 40.08%).

Validation on an External Data: We also evaluated DeepKinZero on an external test data we had retrieved from PhosphoELM database [29]. We removed all the kinases and their associated phosphosites that were in our training and validation set from PhosphositePlus dataset. DeepKinZero trained on PhopsphositePlus and evaluated on this PhosphoELM dataset achieves hit@1 accuracy of 33.96%, hit@3 accuracy of 52.83%, 62.26% hit@5 accuracy and 77.36% hit@10 accuracy. Although the dataset is small, it provides confidence that the model generalizes to other datasets.

Inspecting the Model Weights: We further analyze the learned weights in the model to gain further insight into the model. First, we inspect BRNN attention weights. Figure 5a shows the average attention assigned to each position in the input sequence by the BRNN model. The center residue emerges as the most important residue, thus the model correctly learns to assign more weight to the center, where the phosphosite is located at. The immediate neighbors and the residues within 2 positions are the next most important residues. To investigate the weights assigned to each amino acid type at the phosphosite embedding, we calculate the average weights assigned to different amino-acid types for each group of kinases at the phosphosite. As clearly seen in Figure 5b S, Y and T correctly receive the largest weights. Moreover, the weights assigned to a different type of amino acids in each group align well with existing knowledge of kinase groups. For example, the TK family, which exclusively works on thyrosine residue (Y), puts a very large positive weight on tyrosine while other families do not.

References

1. Hunter, T. (1995) Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell,* **80**(2), 225–236.
2. Gaestel, M., Kotlyarov, A., and Kracht, M. (2009) Targeting innate immunity protein kinase signalling in inflammation. *Nature Reviews Drug Discovery,* **8**(6), 480.
3. Blume-Jensen, P. and Hunter, T. (2001) Oncogenic kinase signalling. *Nature,* **411**(6835), 355.
4. Müller, S., Chaikuad, A., Gray, N. S., and Knapp, S. (2015) The ins and outs of selective kinase inhibitor development. *Nature chemical biology,* **11**(11), 818.
5. Needham, E. J., Parker, B. L., Burykin, T., James, D. E., and Humphrey, S. J. (2019) Illuminating the dark phosphoproteome. *Sci. Signal.,* **12**(565), eaau8645.
6. Horn, H., Schoof, E. M., Kim, J., Robin, X., Miller, M. L., Diella, F., Palma, A., Cesareni, G., Jensen, L. J., and Linding, R. (2014) KinomeXplorer: an integrated platform for kinome biology studies. *Nature methods,* **11**(6), 603.
7. Dou, Y., Yao, B., and Zhang, C. (2014) PhosphoSVM: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. *Amino acids,* **46**(6), 1459–1469.
8. Patrick, R., Lê Cao, K.-A., Kobe, B., and Bodén, M. (2014) PhosphoPICK: modelling cellular context to map kinase-substrate phosphorylation events. *Bioinformatics,* **31**(3), 382–389.
9. Ismail, H. D., Jones, A., Kim, J. H., Newman, R. H., and Kc, D. B. (2016) RF-Phos: a novel general Phosphorylation site prediction tool based on random Forest. *BioMed research international,* **2016**.
10. Wang, M., Wang, T., Wang, B., Liu, Y., and Li, A. (2017) A Novel Phosphorylation Site-Kinase Network-Based Method for the Accurate Prediction of Kinase-Substrate Relationships. *BioMed research international,* **2017**.
11. Song, J., Wang, H., Wang, J., Leier, A., Marquez-Lago, T., Yang, B., Zhang, Z., Akutsu, T., Webb, G. I., and Daly, R. J. (2017) PhosphoPredict: A bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. *Scientific Reports,* **7**(1), 6862.
12. Qin, G.-M., Li, R.-Y., and Zhao, X.-M. (2016) PhosD: inferring kinase–substrate interactions based on protein domains. *Bioinformatics,* **33**(8), 1197–1204.
13. Wang, D., Zeng, S., Xu, C., Qiu, W., Liang, Y., Joshi, T., and Xu, D. (2017) MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics,* **33**(24), 3909–3916.
14. Trost, B. and Kusalik, A. (2011) Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics,* **27**(21), 2927–2935.
15. Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. (2016) Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence,* **38**(7), 1425–1438.
16. Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. (2005) Large margin methods for structured and interdependent output variables. *Journal of machine learning research,* **6**(Sep), 1453–1484.
17. Xian, Y., Schiele, B., and Akata, Z. (2017) Zero-shot learning-the good, the bad and the ugly. *arXiv preprint arXiv:1703.04394,*.
18. Romera-Paredes, B. and Torr, P. (2015) An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning* pp. 2152–2161.
19. Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al. (2013) Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems* pp. 2121–2129.
20. Akata, Z., Reed, S., Walter, D., Lee, H., and Schiele, B. (2015) Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp. 2927–2936.
21. Kodirov, E., Xiang, T., and Gong, S. (2017) Semantic autoencoder for zero-shot learning. *arXiv preprint arXiv:1704.08345,*.
22. Sumbul, G., Cinbis, R. G., and Aksoy, S. (2018) Fine-Grained Object Recognition and Zero-Shot Learning in Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing,* **56**(2), 770–779.
23. Kingma, D. P. and Ba, J. (2014) Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980,*.
24. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014) Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research,* **15**(1), 1929–1958.
25. Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016) Layer normalization. *arXiv preprint arXiv:1607.06450,*.
26. Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E. (2014) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic acids research,* **43**(D1), D512–D520.
27. Ellis, J. J. and Kobe, B. (2011) Predicting protein kinase specificity: Predikin update and performance in the DREAM4 challenge. *PloS one,* **6**(7), e21169.
28. Wagih, O., Sugiyama, N., Ishihama, Y., and Beltrao, P. (2016) Uncovering phosphorylation-based specificities through functional interaction networks. *Molecular & Cellular Proteomics,* **15**(1), 236–245.
29. Diella, F., Gould, C. M., Chica, C., Via, A., and Gibson, T. J. (2007) Phospho. ELM: a database of phosphorylation sites—update 2008. *Nucleic acids research,* **36**(suppl_1), D240–D244.