Improving DeepKinZero

STUDENTS / UNIVERSITIES

Kağan Korkmaz/Sabancı University
Selami Doğan Akansu/Sabancı University

SUPERVISOR(S)

Öznur Taştan

Introduction

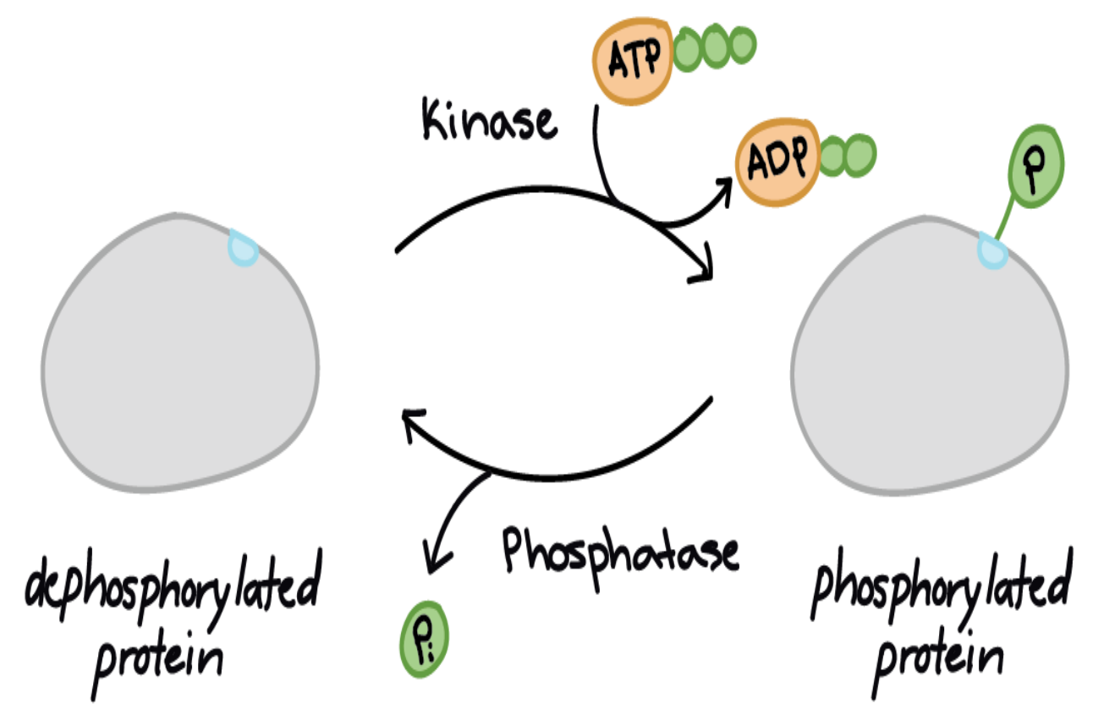


Figure1:Overview of phosphorilation

Protein kinases are a large family of enzymes that catalyze the phosphorylation of other proteins.[1]Phosphorilated proteins do specific functions. (Figure 1).

Aberrant kinase function is associated with cancer, immune system diseases and degenerative diseases.[2]. Protein kinases are major drug targets [3].

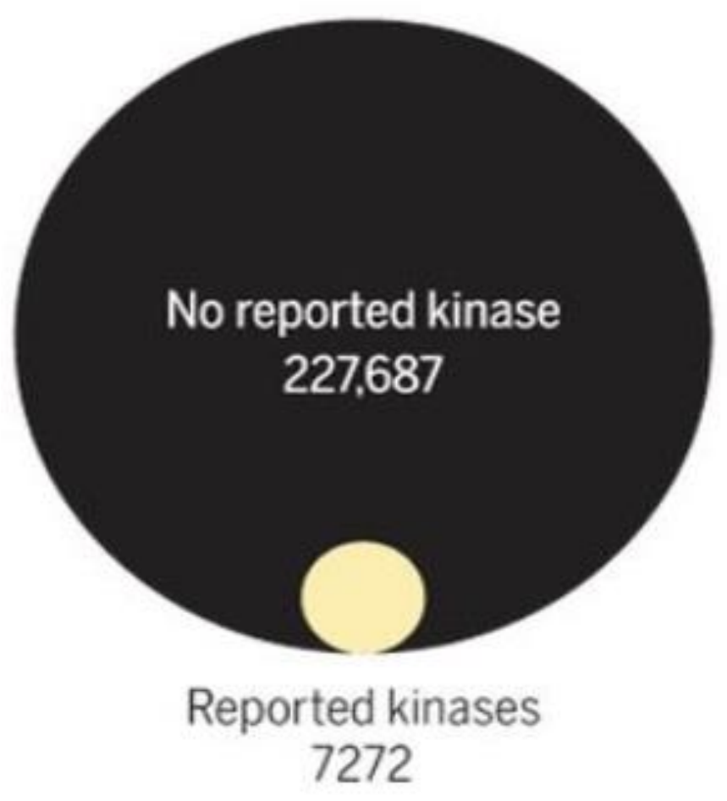


Figure 2. The proportion of human phosphosites with a reported kinase in PhosphoSitePlus Figure from [4].

The advances in enable the identification of phosphosites at the proteome level, most of the phosphoproteome is in the dark: more than 95% of all reported human phosphosites have no known kinase or associated biological function [4] (Figure 2).

A large fraction of the kinome is understudied[4].For most of the kinases there are less than 10 known phosphosites (Figure 3).

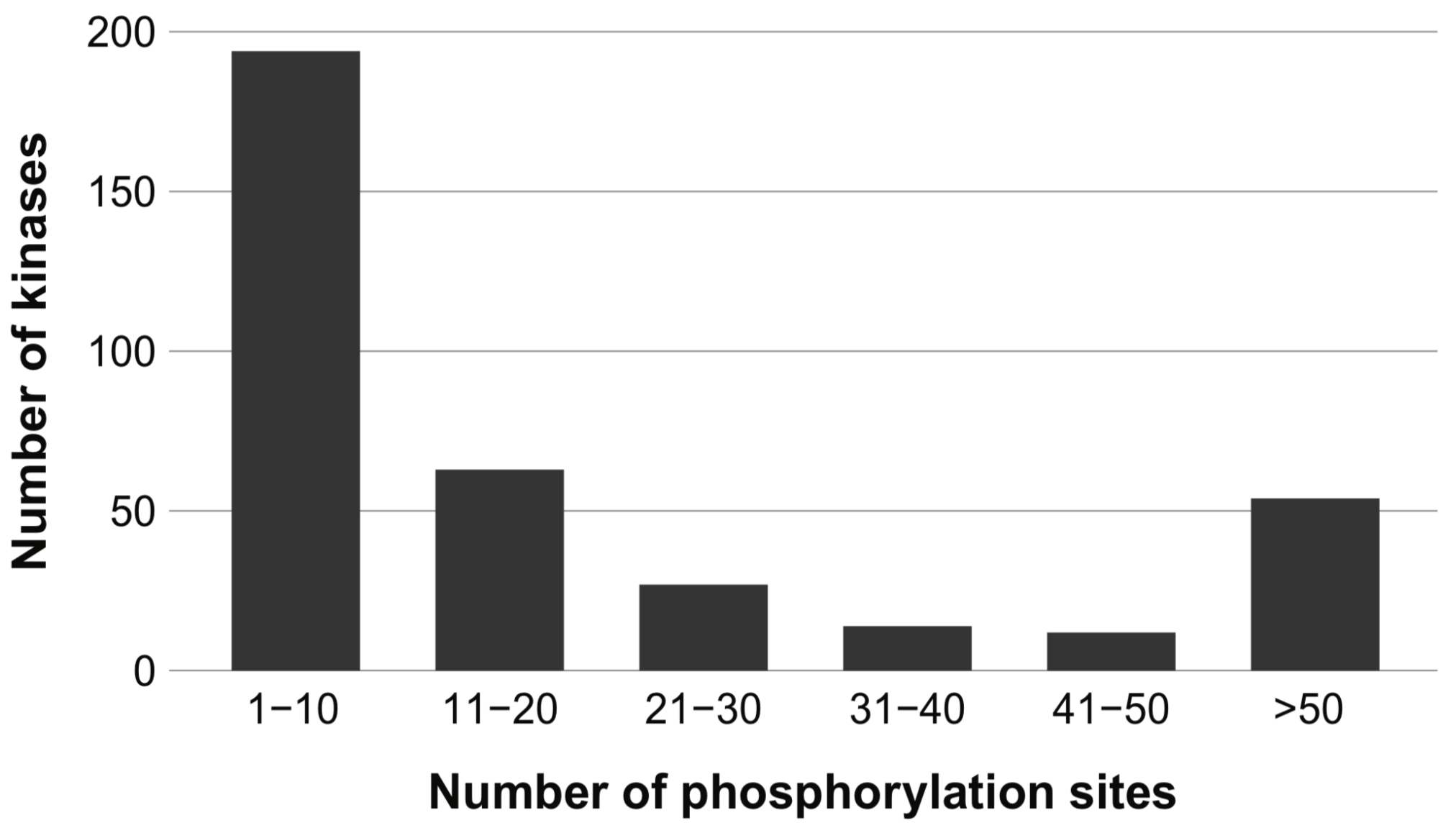


Figure 3: The distribution of the number of experimentally validated target phosphosites for kinases in the human kinome

DeepKinZero is a program that makes predictions for rare kinases, it first learn the association between the phosphosite and kinase embeddings. This idea is shown in Figure 4.

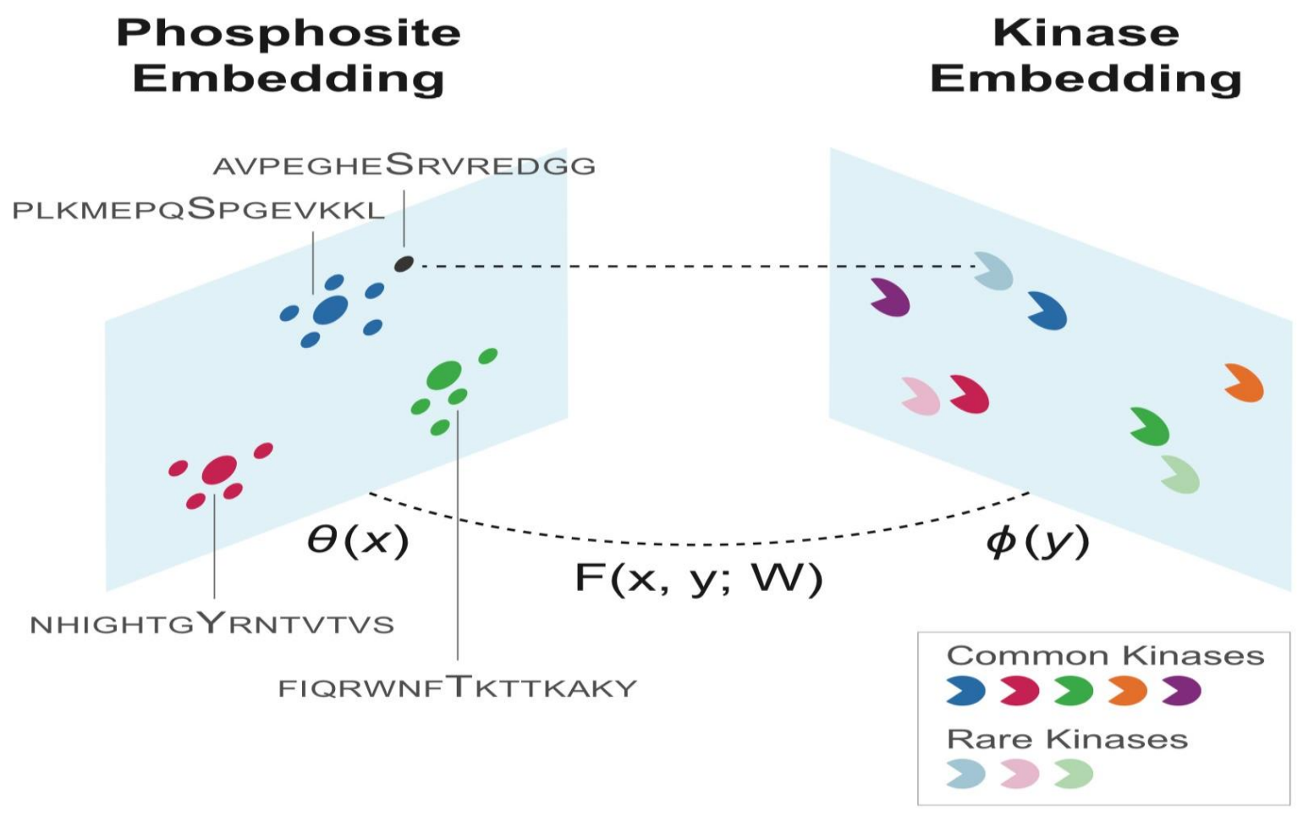


Figure 4: Overview of the application of zero-shot learning to the prediction of kinase-phosphosite associations.

DeepKinZero use a bi-linear compatibility function *F* to model the mapping between the input and output embeddings. F takes a phosphosite-kinase pair as input and returns a scalar value. The probability that a given site is a target of a given kinase is calculated based on F:

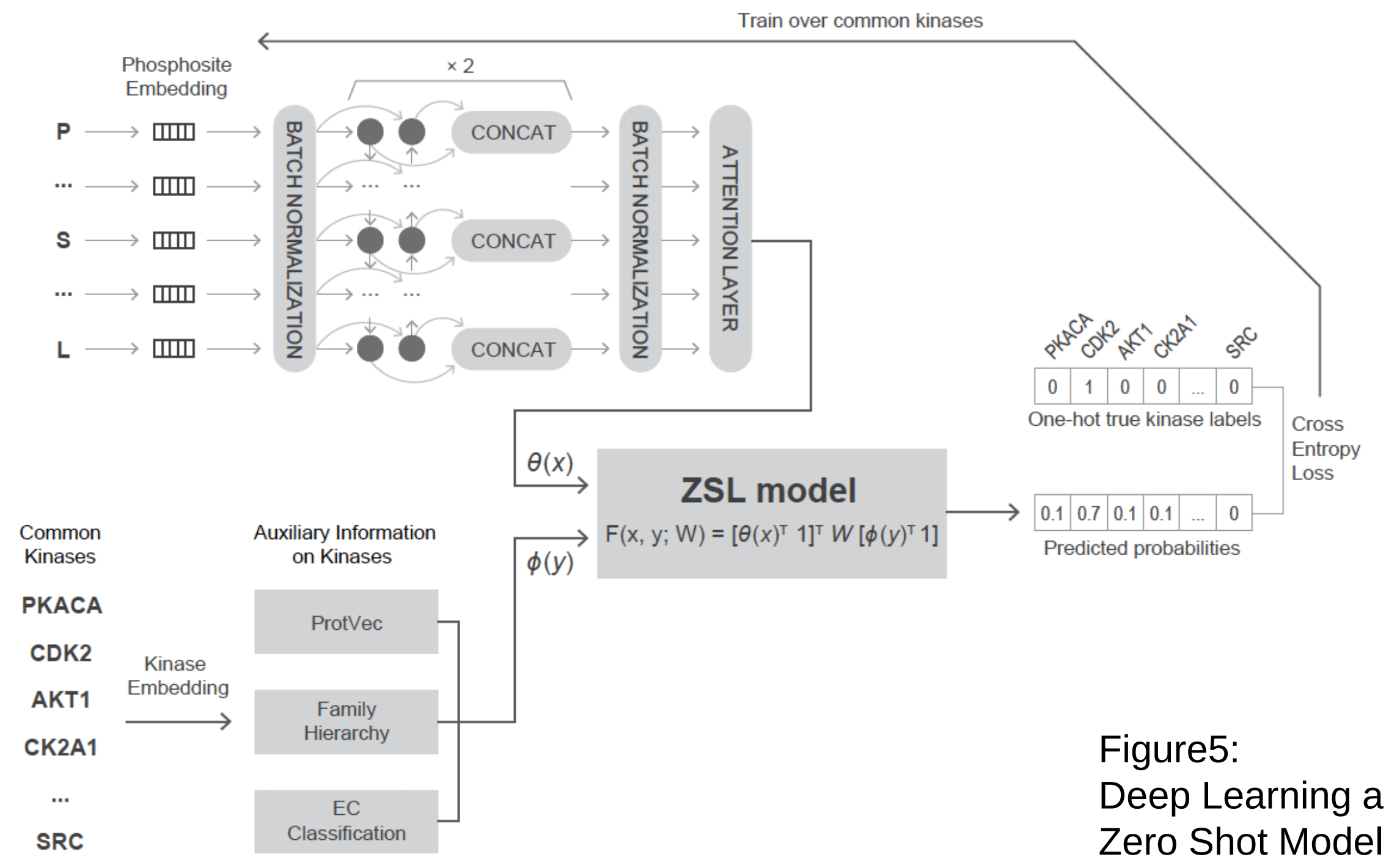


Figure5:
Deep Learning and
Zero Shot Model

To learn phosphosite embeddings, Bi-directional Recurrent Neural Network (BRNN) [5] model is used with an attention mechanism over the training data. Figure 5 illustrates theDeepKinZero model.

Objectives

- Understanding how DeepKinzero work is assential to improve it.
- Understanding Gene Ontology to desribe cellular location and obtaining kinase substrate annotations.
- Understanding and running GO semantic similarity tools leads to calculate similarity between kinases and proteins.
- Running DeepKinzero with the addition feature which is location information of kinases and proteins.
- Hyper parameter tuning makes input appropriate to use in DeepKinZero

Gene Ontology(GO)

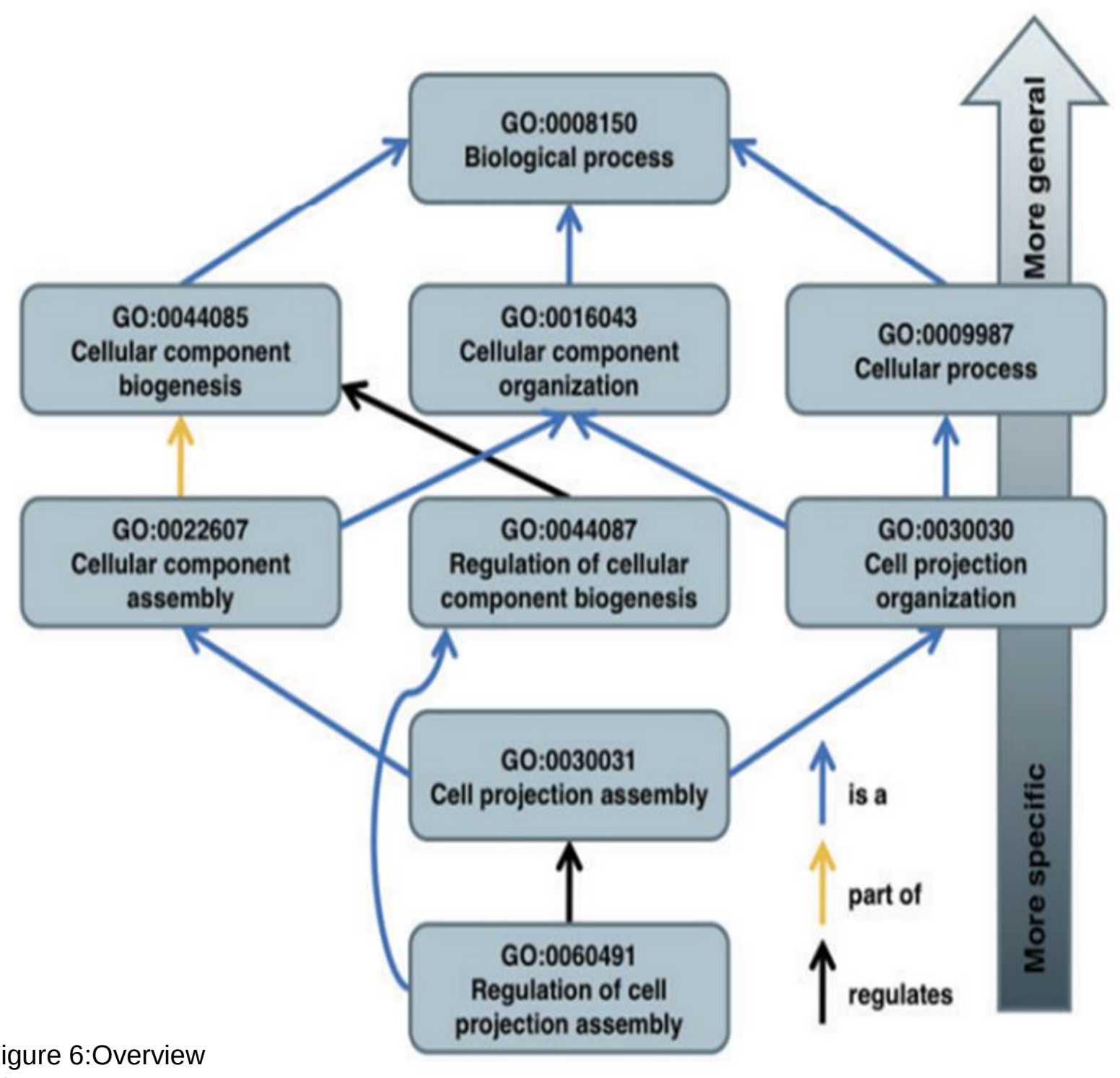
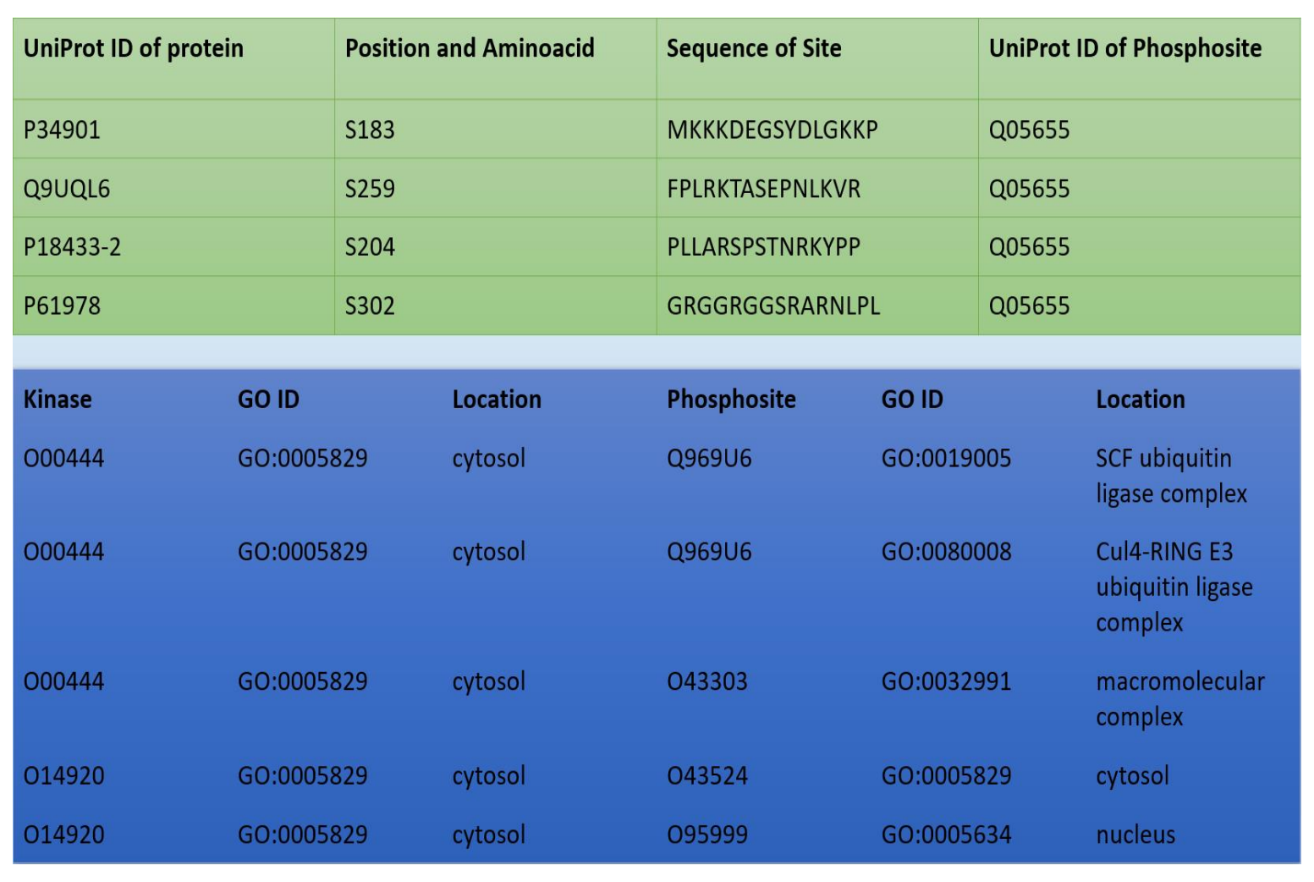

GO:0008150
Biological process
GO:0044085
Cellular component biogenesis
GO:0016043
Cellular component organization
GO:0009987
Cellular process
GO:0022607
Cellular component assembly
GO:0044087
Regulation of cellular component biogenesis
GO:0030030
Cell projection organization
GO:0030031
Cell projection assembly
GO:0060491
Regulation of cell projection assembly
is a
part of
regulates
More general
More specific

Figure 6:Overview of GO

The Gene Ontology (GO) **knowledgebase** is the world's largest source of information on the functions of genes. GO is a directed acyclic graph with different realationship types(Figure 6). We used GO for calculating similarity between proteins by transforming uniprot IDs(protein ID) to Gene IDs. We used Wang and Bma method with 'Celllular Component' ontology to find the relationship between Kinase and Substrate peers in terms of their location data in the Gene Ontology.

Similarity Calculation



UniProt ID of protein	Position and Aminoacid	Sequence of Site	UniProt ID of Phosphosite
P34901	S183	MKKKDEGSYDLGKKP	Q05655
Q9UQL6	S259	FPLRKTASEPNLKVR	Q05655
P18433-2	S204	PLLARSPSTNRKYPP	Q05655
P61978	S302	GRGGRGGSRARNLPL	Q05655

Kinase	GO ID	Location	Phosphosite	GO ID	Location
O00444	GO:0005829	cytosol	Q969U6	GO:0019005	SCF ubiquitin ligase complex
O00444	GO:0005829	cytosol	Q969U6	GO:0080008	Cul4-RING E3 ubiquitin ligase complex
O00444	GO:0005829	cytosol	O43303	GO:0032991	macromolecular complex
O14920	GO:0005829	cytosol	O43524	GO:0005829	cytosol
O14920	GO:0005829	cytosol	O95999	GO:0005634	nucleus

Figure 7: Kinase-Phosphosite peers and location IDs.

First, we calculated related GO IDs of Uniprot protein IDs by using Python. We realized that for each substrate and kinase there might be multiple related GO terms, then this data is merged with location ID of proteins. Expectation was similartiy between kinases and substrates which have similar info of cellular componenets will be high.

After that point similarity is calculated by using GoSemSim package in R. Wang method with combining of BMA method was used.

WANG METHOD

This method determines the semantic similarity of two GO terms based on both the locations of these terms in the GO graph and their relations with their ancestor terms

BMA METHOD

The BMA method, used the Best-Match Average strategy, calculates the average of all maximum similarities on each row and column in GO ID matrix.

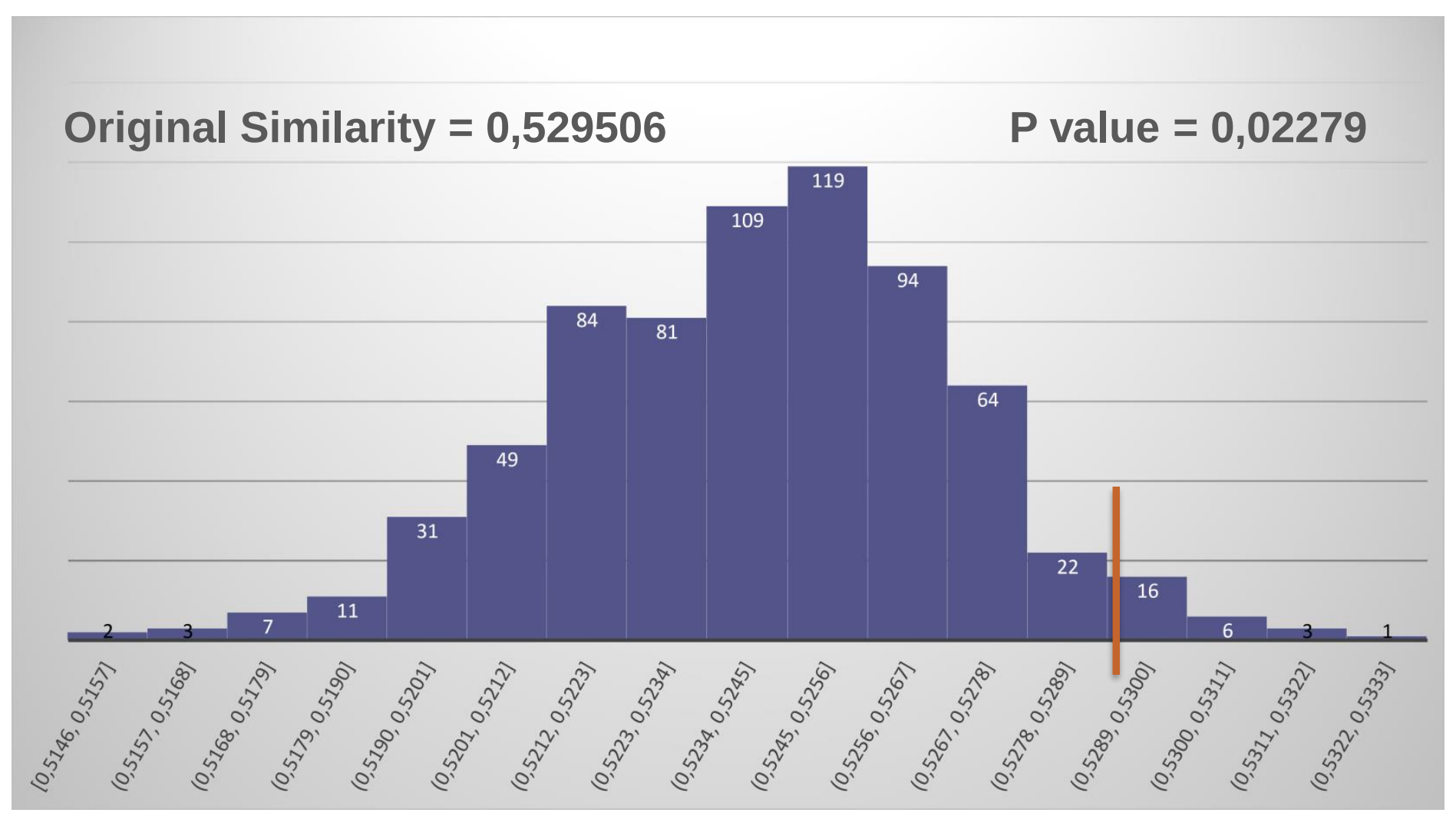


Figure 8:Results of shuffled data

After similarity calculation of 4354 Kinase and Substrate peers the average value of similarity resulted by 0,529506(between 0-1). To test the reliability of the cellular component factor in order to add a feature to DeepKinZero we shuffled the real Kinase-Substrate peers by permutating one column 700 different times .

Figure 8 displays the average GO semantic similarity calculated for the randomized cases. The avarage similairities are so close but interestingly P value shows data data is very reliable to adding as feature to improve DeepKinZero.

With these resulted data, now we are implementing new feature to DeepKinZero by using TensorFlow library in Python.

Conclusion

We observed that similarity of kinase and substrate peers based on cellular component feature is a very significant feature for improving DeepKinzero. However, similarities so close but still it is an usable feature. We are trying to improve new algorithms to calculate similarities in a more distinct way.

Future Work

After the improve similarity calculation and implementing of new cellular component feature, we will run DeepKinZero and observe is the new feature increase prediciton power of DeepKinZero. Then to improve the impact of new feature we will make hyper parameter tuning.

References

1. Hunter, T. Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. Cell 80, 225–236 (1995).
2. Blume-Jensen, P. & Hunter, T. Oncogenic kinase signaling. Nature 411, 355 (2001).
3. Klaeger, S. et al. The target landscape of clinical kinase drugs. Science 358, 4368 (2017).
4. Needham, E. J., Parker, B. L., Burykin, T., James, D. E. & Humphrey, S. J. Illuminating the dark phosphoproteome. Sci. Signal. 12, 8645 (2019).
5. Schuster, M. & Paliwal, K. K. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing 45, 2673–2681 (1997).
6. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. Jan 2019;47(D1):D330-D338.
7. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S (2010). "GOSemSim: an R package for measuring semantic similarity among GO terms and gene products." *Bioinformatics*, **26**(7), 976-978.
8. Deznabi, Iman, et al. "DeepKinZero: Zero-Shot Learning for Predicting Kinase-Phosphosite Associations Involving Understudied Kinases." 2019, doi:10.1101/670638.