ZERO- AND FEW-SHOT DARK KINASE–PHOSPHOSITE PREDICTION VIA TASK-AWARE PROTEIN EMBEDDINGS

by
ZEYNEP IŞIK

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfilment of
the requirements for the degree of Master of Science

Sabancı University
July 2025

ZERO- AND FEW-SHOT DARK KINASE–PHOSPHOSITE PREDICTION VIA TASK-AWARE PROTEIN EMBEDDINGS

Approved by:

Assoc. Prof. ÖZNUR TAŞTAN .....
(Thesis Supervisor)

Prof. ARZUCAN ÖZGÜR .....

Asst. Prof. NUR MUSTAFAOĞLU .....

Date of Approval: July 10, 2025

ABSTRACT

ZERO- AND FEW-SHOT DARK KINASE–PHOSPHOSITE PREDICTION VIA TASK-AWARE PROTEIN EMBEDDINGS

ZEYNEP IŞIK

Computer Science and Engineering M.Sc. THESIS, JULY 2025

Thesis Supervisor: Assoc. Prof. ÖZNUR TAŞTAN
Thesis Co-supervisor: Assoc. Prof. RAMAZAN GÖKBERK CİNBİŞ

Keywords: Task Adaptation, Zero-shot Learning, Few-shot Learning, Transformers, Protein Language Models, Kinases, Phosphorylation

Accurately mapping kinases to their substrate phosphosites is fundamental for decoding cellular signaling and understanding disease mechanisms. While high-throughput techniques can identify the phosphosites, finding the kinase that catalyzes the phosphorylation is challenging. Thus, over 95% of experimentally detected human phosphosites lack kinase annotations. It is possible to formulate the kinase-phosphosite association problem as a supervised multi-class classification task; however, a large portion of the human kinases are under-studied (dark kinases) and have few or no phosphosites associated with them, thus dark kinases fall outside the reach of conventional supervised learning methods. In this thesis, we formulate kinase–phosphosite association as zero-shot and few-shot learning tasks: in the zero-shot setting, the model must predict associations for kinases never seen during training; in the few-shot setting, it may leverage only a handful of labeled examples.

We employ transformer-based protein language models (pLMs) to embed both kinase domains and phosphosite peptides, and we systematically explore domain-adaptation strategies—ranging from full fine-tuning and partial layer re-initialization to task-specific pre-training—under severe data constraints. Surprisingly, a de novo–trained ESM-1b model outperforms its fully fine-tuned pretrained counterpart, suggesting that general-purpose pLM embeddings may lack task-specific biochemical context. Our best results are obtained by combining kinase- and phosphosite-aware

pLMs with partial re-initialization of upper transformer layers. On the DARKIN benchmark, this approach delivers state-of-the-art performance in both zero-shot and few-shot kinase prediction, offering a promising direction for illuminating the dark phosphoproteome.

ÖZET

GÖREVE DUYARLI PROTEİN GÖMÜLERİYLE SIFIR- VE AZ-ÖRNEKLİ KARANLIK KİNAZ–FOSFOSİT TAHMİNİ

ZEYNEP IŞIK

Bilgisayar Bilimi ve Mühendisliği Yüksek Lisans TEZİ, TEMMUZ 2025

Tez Danışmanı: Doç. Dr. ÖZNUR TAŞTAN
Tez Eş-Danışmanı: Doç. Dr. RAMAZAN GÖKBERK CİNBİŞ

Anahtar Kelimeler: Göreve Uyarlama, Sıfır-Örnekli Öğrenme, Az-Örnekli Öğrenme, Dönüştürücüler, Protein Dil Modelleri, Kinazlar, Fosforilasyon

Hücre içi sinyal iletimini çözümlemek ve hastalık mekanizmalarını anlamak için kinazların substrat fosfositlerine doğru bir şekilde eşleştirilmesi önemli bir problemdir. Yüksek verimli deneysel teknikler fosfositleri belirleyebilse de, bir fosfositin fosforilasyonunu katalizleyen kinazı bulmak zordur. Bu nedenle, deneysel olarak tespit edilen insan fosfositlerinin %95'ten fazlası kinaz anotasyonundan yoksundur. Kinaz–fosfosit ilişkisinin denetimli çok sınıflı bir sınıflandırma problemi olarak formüle edilmesi mümkündür fakat insan kinazlarının büyük bir kısmı az çalışılmıştır. Karanlık kinazlar adı verilen bu kinazlar için ya çok az fosfosit yeri vardır ya da hiç yoktur; dolayısıyla karanlık kinazlar geleneksel denetimli öğrenme yöntemlerinin kapsamında yer alamaz. Bu tezde, kinaz–fosfosit ilişkilendirmesini sıfır-örnekli (zero-shot) ve az-örnekli (few-shot) öğrenme problemleri olarak ele alıyoruz: sıfır-örnek düzeneğinde model, eğitim sırasında hiç görülmemiş kinazlar için ilişki tahmin etmek zorundadır; az-örnek düzeneğinde ise elindeki sınırlı sayıda etiketli örnekten yararlanabilir.

Dönüştürücü mimarisine dayalı protein dil modelleri (pDM'ler) kullanarak hem kinaz alanlarını hem de fosfosit peptitlerini gömüyor ve sıkı veri kısıtları altında tam ince ayardan katmanların kısmi yeniden başlatılmasına ve görev-odaklı ön eğitime kadar uzanan alan uyum stratejilerini sistematik olarak araştırıyoruz. İlginç bir şekilde, sıfırdan eğitilmiş bir ESM-1b modeli, tamamen ince ayarlanmış

ön eğitimli muadilini geride bırakarak, genel amaçlı pDM gömme yöntemlerinin görev-özgü biyokimyasal bağlamı yakalayamayabileceğini düşündürmektedir. En iyi sonuçlarımızı, kinaz ve fosfosit farkındalıklı pDM'lerin üst dönüştürücü katmanlarının kısmi olarak yeniden başlatılmasıyla elde ettik. DARKIN kıyas setinde, bu yaklaşım sıfır-örnekli ve az-örnekli kinaz tahmininde en iyi performansı sunarak karanlık fosfoproteomu aydınlatmak için umut verici bir yaklaşım olduğunu göstermiştir.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my thesis supervisor, Assoc. Prof. Öznur Taştan for her guidance, support, kindness, and trust in me. Throughout my master's journey, she was always there. She is a remarkable example as a woman and inspires me in various aspects.

I would also like to express my deepest gratitude to my co-advisor, Assoc. Prof. Ramazan Gökberk Cinbiş for his patience, kindness, and especially his generosity in sharing his knowledge and experiences.

I would like to thank my parents and sister. Thanks to my mother, Havva, without her ambition, patience, and motivational talks, this would not be possible. Thanks to my father, Ahmet, he probably would love me if I were a trash can; he is an expert at unconditional love. Thanks to my sister, Rukiye, I would probably have lost the joy of life without her. She always inspires me with her learning process, while I am learning how people and machines learn.

My dear supervisor in Bachelor's, Suzan Üsküdarlı, PhD, and the most lovable ex-boss, Reyyan Yeniterzi, PhD, they helped me attach to academia securely after experiencing hesitations.

I am deeply grateful to my friends: Egemen, Hasan, Mekan, Seçilay, and Semih(x2). They are the best teammates and supporters. Very special thanks go to Emine Ayşe. She always stood with me on good and bad days. She is one of the most just and generous people in the world. Thanks to Şerife, Bahar, Sabriye, and "BOUN-CMPE ve hâlâ" team, I have a great opportunity to find someone to listen to my complaints. Furthermore, I would like to thank Mert, Gökçe, and Merve, who never left me unanswered anytime I needed.

I would like to express my gratitude to the Scientific and Technological Research Council of Turkey (TÜBİTAK) for their funding of the project 122E500. All computational works for this thesis were fully performed at TÜBİTAK ULAKBİM, High Performance and Grid Computing Center. I would like to thank the Faculty of Engineering and Natural Sciences for the conference travel grant, which enabled me to participate in the ICLR 2024 MLGenX Workshop.

Last but not least, I want to thank myself. I could come to the end of this journey and gain wonderful experiences.

to my sister Rukiye
I can't imagine a life without her

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Protein phosphorylation is an important post-translational modification that regulates the functions of proteins (Hunter, 1995). Phosphorylation enables the transfer of phosphate groups from high-energy molecules, such as adenosine triphosphate (ATP), to the amino acids of a target protein. Kinases are the enzymes that catalyze phosphorylation in a target-specific manner (Cohen, 2002). Phosphorylation can lead to a wide range of changes: activating and deactivating the target, altering the interaction between target proteins and other proteins, directing proteins to their place in the cellular localization, and marking the target proteins for degradation (Pawson and Scott, 2005). Through phosphorylation, kinases play crucial roles in cellular signaling pathways.

During phosphorylation, in general, serine, threonine, and tyrosine (S/T/Y) amino acids on the target proteins are phosphorylated (Cohen, 2002). However, there are rare cases in which histidine (H) can be phosphorylated as well (Pesis et al., 1988). The amino acid in the target protein where the phosphate groups are added is called *phosphosite*.

The human kinome comprises over 500 kinases, making them one of the largest gene families in humans (Cohen, 2002). Because kinases are one of the key regulators of the cellular processes, dysregulation in kinase functions is associated with many diseases, including cancer, neuro-degenerative diseases, cardiovascular diseases, and many others (Blume-Jensen and Hunter, 2001; Heineke and Molkentin, 2006; Gaestel et al., 2009; Wang et al., 2012; Müller et al., 2015; Jiang et al., 2025). Drug resistance in cancer is associated with kinases (O'Reilly et al., 2006; Klempner et al., 2013). Additionally, there are many diseases where a mutation in the phosphosite is associated with the disease (Needham et al., 2019). Due to these reasons, kinases have become drug targets of many diseases, and 25% to 33% of designed drugs target kinases (Ferguson and Gray, 2018; Roskoski Jr, 2022). To understand the normal and abnormal signaling in the cell, detecting phosphorylation sites and finding the associated kinases playing a role in the phosphorylation of those regions are necessary.

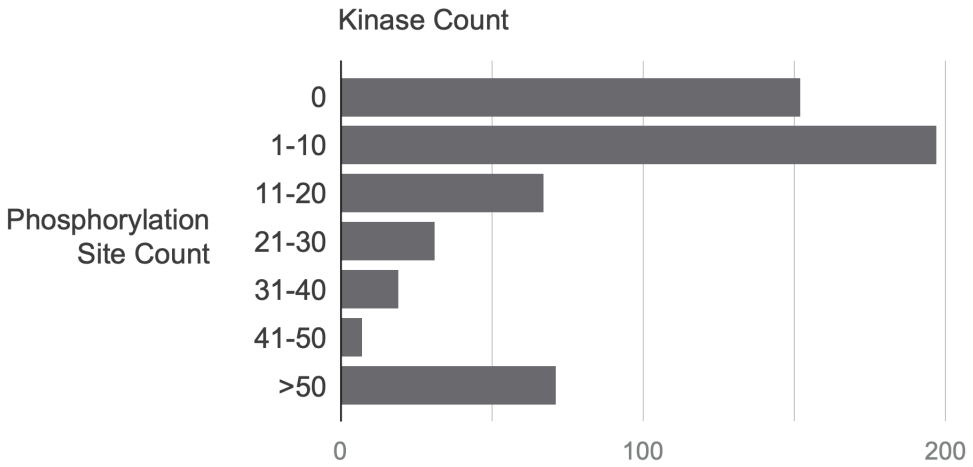


Figure 1.1 The number of phosphosites associated with a kinase reported in the PhosphoSitePlus dataset. For most of the kinases, there are no or few assigned phosphosites. This figure is reproduced from our previous study DARKIN (Sunar et al., 2024).

Developments in mass spectrometry have enabled researchers to experimentally identify phosphorylation sites (Manning et al., 2002). Consequently, today, there are more than 300,000 phosphosites reported in PhosphoSitePlus (Hornbeck et al., 2012) (downloaded on March 18, 2024). After identifying where on a protein is phosphorylated, the next critical question is, "Which kinase phosphorylates this phosphosite?" Answering this question experimentally is challenging due to the transient nature of kinase-target interactions. For this reason, despite kinases being the most studied protein family due to their critical function in the cell, most of the phosphoproteome is in the dark. For 95% of the known phosphosites in human cells, the cognate kinase has yet to be identified; therefore, the biological functions of these phosphosites remain unknown. We will call the phosphosites which have not been associated with any kinase *orphan phosphosites*. Moreover, approximately 150 kinases, which correspond to 25% of kinases, do not have known phosphorylation targets, and approximately 195 kinases have only 1–10 identified target sites, which correspond to around 35% of kinases (Needham et al., 2019). In Figure 1.1, the number of phosphosites per kinase is shown. Needham et al. (2019) named understudied kinases *dark kinases*, and kinases with plenty of known phosphosites *light kinases*. In this thesis, we will use this terminology.

Phosphorylation sites and their cognate kinases are identified through experimental methods, which are not only costly but also time-consuming. For this reason, numerous computational approaches have been proposed to accelerate research on assigning kinases to orphan phosphosites. Several studies that focus on identifying kinase-specific phosphosites and finding the cognate kinase of these phosphosites

have been conducted (Wong et al., 2007; Saunders et al., 2008; Xue et al., 2008; Li et al., 2008; Gao et al., 2010; Zou et al., 2013; Horn et al., 2014; Patrick et al., 2015; Wang et al., 2017a; Song et al., 2017; Wang et al., 2017b; Qin et al., 2017; Ma et al., 2020). Previous computational studies on phosphorylation have been mostly based on position-specific weight matrices (PSWMs) or supervised learning techniques (Altschul et al., 1997; Li et al., 2010; Zou et al., 2013; Horn et al., 2014; Patrick et al., 2015; Song et al., 2017; Wang et al., 2017a,c). However, position-specific matrices cannot capture the intrinsic features of phosphosite and kinase sequences, and these representations fail to capture richer biological context, such as site-specific roles, conformational and functional properties, interactions between neighboring amino acids, and post-translational modifications. Supervised classification techniques are powerful in predicting which kinase is associated with a given phosphosite (Xue et al., 2008; Gao et al., 2010; Zou et al., 2013), yet they require labeled training examples for each kinase. In the case of dark kinases, because there are few or no phosphosites associated, the classical supervised learning setup falls short for making predictions for dark kinases. Hence, we model this problem as zero-shot and few-shot approaches.

Zero-shot learning is a machine learning approach that enables a model to classify data samples whose classes were not observed during training. The main idea of zero-shot learning is to establish a relationship between input samples and class attributes (Larochelle et al., 2008; Palatucci et al., 2009; Lampert et al., 2013; Romera-Paredes and Torr, 2015; Akata et al., 2015a). For the first time in the literature, Deznabi et al. (2020) formulated the dark kinase–phosphosite association task as a zero-shot problem, and proposed a model called DeepKinZero. DeepKinZero transfers the knowledge learned during training from light kinase-phosphosite pairs to the dark phosphoproteome. Thereby, it can provide predictions for phosphosites whose associated kinases are very limited or unknown. In a kinase–phosphosite association prediction task, zero-shot learning can utilize the information about well-studied kinases and phosphosites to guide predictions for dark kinases and orphan phosphosites.

The few-shot formulation has not been explored for kinases yet. This could be valuable for the many kinases where there are few phosphosites available. In the few-shot formulation, a model leverages a small number of known phosphosites to make predictions for under-studied kinases. Because training a model on very few samples carries the overfitting risk, common few-shot strategies first train a model on classes with sufficient data and then adapt the model to classes with only a handful of samples (Koch et al., 2015; Qiao et al., 2018; Qi et al., 2018; Li et al., 2021).

In this study, we approach the kinase-phosphosite prediction task as both a zero-shot and a few-shot learning problem. To capture phosphorylation context, we employ a transformer-based encoder, and we couple it with a bi-linear compatibility model that links kinase and phosphosite embeddings (Vaswani et al., 2017; Sumbul et al., 2018). We adapt the transformer to the task through fine-tuning strategies. Additionally, we retain a single architecture across zero-shot and few-shot regimes, ensuring a fair performance comparison.

We use the protein sequences of the kinase domains and the peptide sequence surrounding the phosphosite. The amino acid sequences must be encoded as numerical vectors before being fed into a machine learning based model. There are several ways to encode proteins, some of which rely on hand-crafted features (Gribskov et al., 1987; Henikoff and Henikoff, 1992; Nanni and Lumini, 2011), while others use dense representations learned by neural networks (Asgari and Mofrad, 2015; Rao et al., 2019; Rives et al., 2021; Meier et al., 2021; Elnaggar et al., 2021; Lin et al., 2023; Ferruz et al., 2022; Brandes et al., 2022; Nijkamp et al., 2023; Hayes et al., 2024; ESM Team, 2024; Ouyang-Zhang et al., 2024; Peng et al., 2025). Recently, protein language models (pLMs) have been trained to capture complex biological features in protein sequences by leveraging techniques from natural language processing, and they have become a powerful source of protein sequence representations for downstream tasks. In our previous study, called DARKIN (Sunar et al., 2024), we compared the representability of various pLMs on a zero-shot kinase-phosphosite prediction task by using Sumbul et al. (2018)'s bi-linear model. Among them, Rives et al. (2021)'s ESM-1b showed the best performance. Hence, in this thesis, we adopt ESM-1b and its task-adapted variants to represent both kinase and phosphosite sequences.

pLMs provide general but task-agnostic representations of protein sequences, and these representations may not encode the subtle biochemical context for accurate kinase assignment in low-data regimes (Unsal et al., 2022). Therefore, many studies have focused on adapting general-purpose pLMs to their specific tasks to make the models task-aware (Zhou et al., 2023; Schmirler et al., 2024; Zhou et al., 2024; Esmaili et al., 2025). Schmirler et al. (2024) fine-tuned five pLMs on eight downstream tasks, and reported that prediction accuracies increased compared to frozen embeddings in almost all of the tasks. Zhou et al. (2024), and Esmaili et al. (2025) fine-tuned existing pLMs directly, and Zhou et al. (2023) firstly obtained a domain-adapted transformer and then fine-tuned the output model for the target task. Recently, layer-selective re-initialization has also emerged as a complementary adaptation strategy (Zaidi et al., 2023). In this strategy, at each training stage, the upper layers of the network are randomly re-initialized, while the lower layers are retained

and remain trainable. This preserves the general features learned by early layers, yet allows the re-initialized upper layers to adapt to task-specific signals.

In this work, to solve the dark kinase-phosphosite association prediction task in the zero- and few-shot settings, we explore several task-aware pLM backbones by employing an existing pLM, ESM-1b. These backbone models are used to represent phosphosite and kinase sequences.

The main contributions of this thesis are as follows:

- We presented and experimented with 11 task-aware pLMs by either fine-tuning or pre-training ESM-1b on various auxiliary tasks related to phosphorylation. These include kinase group prediction, contrastive learning to enforce clear clustering of kinases within their families and within their groups, masked language modeling for kinase- and phosphosite-representation learning, and multi-task learning for optimizing phosphosite-representation learning with phosphorylation prediction. We used these models in the kinase-phosphosite prediction task to explore fine-tuning and pre-training strategies, and to obtain task-aware representations for kinases and phosphosites.
- We proposed a transformer-based approach, which leveraged a transformer as a phosphosite model in zero- and few-shot prediction setups, and applied partial re-initialization and partial fine-tuning strategies to adapt the transformer better to the kinase-phosphosite prediction task.
- We evaluated the prediction performance not only as zero-shot but also as few-shot to show how the prediction model performs when having information about kinases associated with a small number of phosphosites compared to the zero-shot case.

The organization of this thesis is as follows:

- In Chapter 2, we provide the background information on phosphorylation. Then, we present an overview of how proteins are represented as numerical vectors by emphasizing both baseline and pLM-based representations. We subsequently summarize some key deep learning concepts used in this thesis. Finally, we review previous work regarding computational approaches for kinase-phosphosite association prediction.
- In Chapter 3, we detail the experimental setup for zero- and few-shot kinase-phosphosite association prediction and describe the datasets we used. We then detail how we obtained the task-aware pLMs, and in parallel, we explain the datasets we used for every pLM.

- In Chapter 4, we present the zero- and few-shot prediction results. We additionally analyze the effect of distinct re-initialization configurations for two of the adapted best-performing pLMs.
- Lastly, in Chapter 5, we conclude our study by highlighting the contributions of task-aware model representations and the advantages of using transformer-based architecture, and present directions for future work.

2. BACKGROUND AND LITERATURE

In this chapter, we establish the foundation for the work presented in this thesis. We begin by reviewing essential biological concepts, particularly protein phosphorylation and the role of kinases. Next, we survey the principal methods for encoding protein sequences in silico, dividing them into two categories: conventional (baseline) feature-based encodings and recent embeddings derived from protein language models. We then introduce the deep learning concepts that drive our approaches, including zero- and few-shot learning, contrastive learning, masked language modeling, and the transformer architecture, as well as the concepts of pre-training and fine-tuning. Finally, we present an overview of the existing computational strategies for phosphorylation and kinase-phosphosite association prediction, and the studies on fine-tuning/pre-training the pLMs.

2.1 Phosphorylation and Kinases

Protein kinases are enzymes that catalyze the site-specific attachment of a phosphate group to their substrate proteins (Hunter, 1995; Cohen, 2002). It is a key post-translational modification, where the specific residue at the target protein accepts a phosphate. This residue is called a *phosphosite*. The type of phosphorylated residue is generally one of serine, threonine, or tyrosine (S/T/Y) amino acids. (Cohen, 2002). However, there are rare cases in which histidine (H) is also phosphorylated (Pesis et al., 1988). Phosphorylation can induce conformational changes in the substrate protein, modulating their activity, altering protein–protein interactions, directing subcellular localization, and marking proteins for degradation (Pawson and Scott, 2005). Dysregulated kinase activity might result in a wide range of diseases, such as cancer and Parkinson's disease. Due to their therapeutic significance, kinases constitute a major class of drug targets (Müller et al., 2015; Steger et al., 2016;

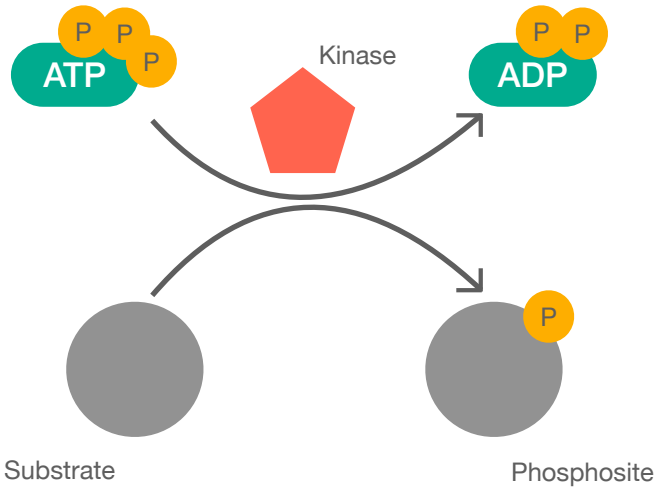


Figure 2.1 Phosphorylation involves the transfer of a phosphate group from a high-energy molecule, such as ATP, to an amino acid residue of the substrate protein. Kinases are the enzymes that catalyze this reaction.

Ferguson and Gray, 2018).

The human kinome comprises more than 500 kinases (the exact number is debatable due to the definition of the kinase domain). The human kinases are organized into 10 groups and 116 families based on similarities in their kinase domain sequences and substrate specificities (Manning et al., 2002). There are classification systems such as the Enzyme Commission numbers (EC). EC numbers provide a numerical categorization scheme for enzymes, based on the chemical reactions they catalyze. EC numbers are available in the ENZYME Database (Bairoch, 2000).

In this study, each phosphosite sequence is represented as a 15-residue peptide sequence where the phosphorylated residue is located in the middle position, and each kinase is represented by its domain sequence as well as its family, group, and EC numbers.

2.2 Protein Representations

Proteins are linear polymers of amino acids, each drawn from a 20-letter alphabet. Formally, a protein of length n is a sequence

$$S = (s_1, s_2, \ldots, s_n),$$

where each s_i is one of the twenty canonical amino acids. A representation method then maps S to a matrix in $\mathbb{R}^{n\times d}$, embedding each residue into a d-dimensional feature space.

Broadly, such encodings fall into two categories:

(i) **Baseline representations:** handcrafted features capturing physicochemical properties (e.g., hydrophobicity, charge), evolutionary substitution likelihoods (e.g., BLOSUM or PAM scores), or simple statistics such as n-gram counts.

(ii) **Representations based on deep learning models:** representations learned by shallow networks or representations derived from protein language models capturing contextual information that reflects complex biochemical and evolutionary patterns.

2.2.1 Baseline Representations

In *one-hot encoding* representation, a protein sequence

$$S = (s_1, s_2, \ldots, s_n)$$

is mapped to a binary matrix

$$X \in \{0,1\}^{n\times 20},$$

where each row x_i is a 20-dimensional vector with a single entry of 1 at the index corresponding to amino acid s_i, and 0s elsewhere. The one-hot encoding does not contain information beyond the amino acid identities.

A *position-specific scoring matrix (PSSM)* is an $n \times 20$ matrix M whose entry $M_{i,a}$ quantifies the log-odds of observing amino acid a at position i within a protein family. Given a multiple sequence alignment of N sequences, let $f_{i,a}$ be the count of amino acid a at column i and q_a its background frequency. Then, for each position i and residue a,

$$M_{i,a} = \log \frac{f_{i,a}/N}{q_a},$$

so that conserved residues receive positive scores, while variable residues receive negative scores (Gribskov et al., 1987).

Another widely used encoding is the *Blocks Substitution Matrix (BLOSUM)*, intro-

duced by Henikoff and Henikoff (1992). A BLOSUM matrix $B \in \mathbb{R}^{20\times 20}$ assigns each amino-acid pair (a, b) a log-odds score

$$B_{a,b} = \log \frac{p_{a,b}}{q_a\, q_b},$$

where $p_{a,b}$ is the observed probability of a aligning with b within conserved blocks of related proteins, and q_a, q_b are their background frequencies. Scores above zero indicate substitutions occurring more often than expected by chance (Henikoff and Henikoff, 1992). The row for a particular amino acid is used as an encoding.

The *Non-linear Fisher (NLF)* representation maps amino-acid sequences into a feature space by applying a non-linear Fisher transform to selected physicochemical property vectors (Nanni and Lumini, 2011). A vector of length 18 represents each amino acid. The position-specific NLF descriptors were computed by the Epitopepredict tool developed by (Farrell, 2021).

2.2.2 Representations Based on Deep Learning Models

ProtVec trains a skip-gram neural language model on overlapping amino-acid tripeptides, producing 100-dimensional embeddings that encapsulate local sequence motifs and their contextual co-occurrence patterns (Asgari and Mofrad, 2015). Although ProtVec and NLF both augment traditional feature sets with richer motif- and physicochemical-level information, neither approach explicitly captures context-dependent relationships within protein sequences.

Protein language models (pLMs) leverage transformer architectures—specifically multi-head self-attention—to learn rich, contextual embeddings of protein sequences. By attending to all residue pairs simultaneously, pLMs capture both local motifs and long-range dependencies, encoding nuanced biochemical and evolutionary signals into dense, high-dimensional vectors. To show how these representations have progressed, we present the main pLM families emphasizing their design choices.

To address the absence of standardized benchmarks and evaluation protocols in semi-supervised learning for proteins, TAPE (Tasks Assessing Protein Embeddings) was introduced by Rao et al. (2019). It comprises five biologically meaningful tasks—secondary structure, residue–residue contact, remote homology, fluorescence, and stability prediction—each with predefined data splits to evaluate and compare model performance across diverse protein-biology challenges.

Elnaggar et al. (2021) introduced ProtTrans, a suite of transformer-based protein language models trained on massive datasets (BFD100, UniRef100, UniRef50) (Bairoch et al., 2005), comprising two auto-regressive and four autoencoder architectures. These models capture conserved motifs, structural classes, and phylogenetic domains, and were benchmarked—against existing literature—on tasks including secondary structure prediction, subcellular localization, and membrane-protein classification.

Meta AI's Evolutionary Scale Modeling (ESM) family has advanced protein sequence representations. ESM-1b (Rives et al., 2021) demonstrated that masked-language pre-training on 250 million sequences captures structural information sufficient for secondary-structure and contact-map prediction via linear projections and unsupervised remote-homology retrieval via cosine similarity. Building on ESM-1b, ESM-1v (Meier et al., 2021) enabled zero-shot mutational scanning—estimating variant effects directly from sequence without retraining a model for every new protein. ESM-2 (Lin et al., 2023), released in model sizes from 8 million to 15 billion parameters, produces embeddings informative enough to predict atomic-level coordinates, which shows that the model can be used as a structure predictor. ESM-IF (Hsu et al., 2022) learned inverse folding from AlphaFold2 structures to predict sequence from structure (Jumper et al., 2021). Very recently, ESM-3 (Hayes et al., 2024) and ESM-C were released. ESM-3 (ESM Team, 2024) is a multi-modal generative language model having knowledge about the sequence, structure, and function of proteins from hundreds of millions of evolutionary years. ESM-C focuses on creating representations by scaling up data and making the training process efficient.

While the ESM family models mostly rely on large-scale training on a vast amount of protein sequences, there are several models that were obtained by injecting additional information, such as structure and post-translational modifications, into pre-trained ESM-2. ISM-2 (Ouyang-Zhang et al., 2024) is one of such models that distills structure tokens, generated via a self-supervised structure encoder, into ESM-2's pretrained weights. SaProt (Su et al., 2023) builds a structure-aware vocabulary by combining sequence tokens with Foldseek-derived 3Di tokens, then fine-tunes on 40 million sequence–structure pairs, boosting performance on protein–protein interaction, metal-binding, and thermostability benchmarks. PTM-Mamba (Peng et al., 2025)(PTM stands for post-translational modification) augments the Mamba architecture with bi-directional gated blocks and explicit PTM tokens (e.g., <N-phosphoserine>), which makes it the first model to jointly embed wild-type and modified residues.

Optimization and scaling are important issues in protein language learning to han-

dle massive datasets efficiently, train models affordably, and reduce inference time. DistilProtBert (Geffen et al., 2022) uses knowledge distillation to halve inference time with minimal loss in accuracy. AMPLIFY (Fournier et al., 2024) questioned the necessity of scaling and showed that, when the training corpus is well-curated in terms of both data quality and quantity and the model architecture is carefully designed, smaller models can outperform larger ones when measured in terms of FLOPs. Ankh (Elnaggar et al., 2023) demonstrated that, with carefully optimized hyperparameters and cautious data curation, a compact model can surpass much larger transformers.

Beyond encoder-only architectures, decoder-only and diffusion-based models have also been developed. ProtGPT-2 (Ferruz et al., 2022) adapts GPT-2 (Radford et al., 2019) to protein sequences by training on 50 million examples, demonstrating the ability to generate novel proteins whose amino-acid residue distributions closely mirror those observed in nature.

Nijkamp et al. (2023) introduced ProGen2, decoder-only protein language models scaled up to 6.4 billion parameters and trained on diverse sequence datasets drawn from multiple omics repositories. ProGen2 models achieved state-of-the-art likelihoods and fitness prediction without any downstream fine-tuning. Recently, DPLM (Wang et al., 2024) brought a new direction to protein modeling with discrete diffusion. It unified generation and representation in a single framework without setting an encoder-decoder architecture.

Finally, in ProteinBERT, Brandes et al. (2022) jointly masked amino acid tokens and their Gene Ontology Annotations (GO Annotations), and trained a single network to recover both of them. Thus, the final embeddings can encode the sequence and functions learned from GO Annotations.

2.3 Fundamental Deep Learning Methods Employed in This Thesis

Zero-shot Learning: Zero-shot learning (ZSL) is a machine learning approach that enables a model to classify data whose classes are not seen during training. Knowledge learned during training is transferred through an auxiliary space such as attribute vectors, an association matrix that links seen and unseen classes (Larochelle et al., 2008; Palatucci et al., 2009; Lampert et al., 2013; Romera-Paredes and Torr, 2015; Akata et al., 2015a). In this thesis, ZSL appears when predicting the cognate

kinase of orphan phosphosites.

Few-shot Learning: Few-shot learning (FSL) is a paradigm that enables models to generalize to new classes given only a very small number of training examples (Koch et al., 2015; Qiao et al., 2018; Qi et al., 2018; Li et al., 2021). Common FSL approaches include metric-based, meta-learning, and non-episodic methods (Li et al., 2021). In this thesis, we extend our ZSL approach into the few-shot setting for dark kinases having a few known phosphosites.

Masked Language Modeling: Masked Language Modeling (MLM) is a self-supervised technique that randomly masks tokens, the smallest units in a sequence, and trains the network to recover them. Thereby, the model is forced to integrate information from tokens and to learn bidirectional, context-aware representations (Devlin et al., 2019). When MLM is applied to *protein sequences*, the model learns both local sequence motifs and long-range evolutionary signals, capturing biochemical context.

Pre-training: Pre-training is a self-supervised stage where the model learns general sequence patterns via any self-supervised objective such as MLM, next-token prediction, or diffusion objectives on a huge amount of protein sequences. The resulting models define an informative prior over sequence space (Ruder et al., 2019). Pre-trained models can be adapted to specific tasks with comparatively small labeled datasets by fine-tuning.

Fine-tuning: Fine-tuning is a form of transfer learning in which the weights of a pre-trained model are updated on a task-specific dataset consisting of a smaller amount of data compared to data used in pre-training (Pan, 2020; Howard and Ruder, 2018). The adaptation might be realized on various objectives and tasks, such as masked-language modeling, binary or multi-class classification. During fine-tuning, early transformer layers might be frozen, all layers might be updated, or some layers might be re-initialized while the rest are fine-tuned. In the end, a task- or context-aware model is obtained (Schmirler et al., 2024).

Contrastive Learning: Contrastive learning is a deep learning technique whose objective is to learn a representation from data by bringing similar data samples closer in the representation space, while pushing the dissimilar instances further apart (Schroff et al., 2015). By maximizing the margin, this method yields representations in which members of the same class form tight clusters and different classes constitute well-separated regions.

Transformer: Transformer architecture, first introduced by Vaswani et al. (2017), is a deep neural network model based entirely on self-attention mechanisms. At its

core, it contains:

- Multi-Head Self-Attention: In each transformer layer, self-attention computes weights that allow every token (e.g., an amino-acid residue in a protein sequence) to attend to all other tokens, enabling the model to capture both short- and long-range dependencies. Multiple attention heads learn complementary patterns, yielding richer representations.
- Positional Encodings: Self-attention does not have prior information of token order, thus positional encodings are added to input embeddings to inject sequence order information.
- Position-Wise Feed-Forward Networks: The output of the attention layers is fed into a fully-connected network, which enables the transformer encoder to extract non-linear features.
- Residual Connections and Layer Normalization: Residual connections facilitate training multi-layered networks by providing shortcut paths which preserve gradient signals during backpropagation, thus they mitigate the problem that gradients become very small. On the other hand, layer normalization maintains normalized feature distributions across layers so that updates on model weights remain consistent.

These components constitute an architecture that enables learning of rich and context-sensitive representations in parallel through multi-head attentions and scaling a deep neural network.

In addition to its architectural power, transformer encoders can be trained on sequential data to learn domain-specific representations by optimizing language modeling objectives. Moreover, pre-trained transformer encoders can be fine-tuned on specific tasks through task-specific components such as classification heads, MLM heads, and decoders, so that encoders can be adapted to distinct tasks. An example of transformer encoder adaptation is Phosformer-ST (Zhou et al., 2024). In that study, the ESM-2 encoder was fine-tuned in a multi-task manner to learn richer kinase representations for kinase-specific phosphorylation prediction.

2.4 Computational Methods on Phosphorylation and

Kinase-Phosphosite Prediction

Computational research on protein phosphorylation has significantly evolved since the early 2000s, with each new generation of methods improving prediction performance and uncovering deeper biological insights. In this section, we review the major computational methods developed for predicting phosphorylation and kinase–phosphosite relationships, including general phosphosite prediction, kinase-substrate assignment, and recent advances in zero-shot learning frameworks.

2.4.1 General Phosphosite Prediction Model

As an early work in phosphosite prediction, NetPhos (Blom et al., 1999) demonstrated that a simple neural network trained on sliding windows of amino acids predicts phosphosite residues (S/T/Y) at rates far above random chance. It achieved a true positive rate ranging from 69% to 96%, depending on the organism. The identification of cognate kinase was out of this study's scope.

PhosphoSVM (Dou et al., 2014) was introduced as a phosphorylation site prediction tool based on support vector machines (SVM). PhosphoSVM used eight distinct sequence-level features, including Shannon Entropy (Shannon, 1948), Relative Entropy (Kullback and Leibler, 1951), Secondary Structure (Garnier et al., 1978), Protein Disorder (Dunker et al., 2000), Solvent Accessible Area (Lee and Richards, 1971), Overlapping Properties (Wu and Brutlag, 1995), Averaged Cumulative Hydrophobicity (Sweet and Eisenberg, 1983), and k-Nearest Neighbor (Cover and Hart, 1967) to predict phosphorylation sites by a single SVM (Cortes and Vapnik, 1995).

After the transformer architecture (Vaswani et al., 2017) was introduced in the field of natural language processing, it has become a powerful tool in computational biology. TransPhos (Wang et al., 2022) was the first to apply a transformer encoder to phosphosite prediction using contextual windows of 33 and 51 residues, thereby capturing long-range dependencies that convolutional (LeCun et al., 1995) and feed-forward networks (Rosenblatt, 1958) overlook. The two windows share an embedding layer and are processed by separate four-layer, four-head transformer encoders; their outputs are refined by parallel 1-D CNN blocks, concatenated, and passed to a softmax classifier. This hybrid attention–convolution design achieved state-of-the-art performance on the Phospho.ELM benchmark.

2.4.2 Kinase-Specific Phosphosite Prediction Models

NetworKIN (Linding et al., 2007) is among the earliest tools to predict the cognate kinase of a given phosphosite. The resulting model was capable of identifying not only if a given phosphosite was phosphorylated or not, but also that it's likely phosphorylated by kinase X, which physically associates with that region of the substrate in cells. NetworKIN's power came from being substrate-aware. To capture the biological context of a substrate, NetworKIN used a network of associations extracted from the STRING database (Mering et al., 2003). NetworKIN is the first effort to bridge motif recognition with real biology.

Around the same time, GPS 2.0 (Xue et al., 2008) introduced a hierarchical, family-aware scoring system such that kinases were grouped by sequence similarity, and each new prediction benefited from transferring strength across well-studied relatives. GPS 2.0 predicted kinase-specific phosphorylation sites for 408 human kinases in hierarchy and showed remarkable accuracy on a large-scale prediction of more than 13,000 mammalian phosphorylation sites.

Building on these foundations, Musite (Gao et al., 2010) aimed to predict both general and kinase-specific phosphosites. They approached the phosphosite prediction problem as a binary classification problem. They trained separate SVMs for each organism, enriched with amino acid frequency around phosphorylation site, k-NN scores, and a protein disorder predictor. k-NN scores were obtained by finding the k most similar samples from both phosphorylated (positive) and non-phosphorylated (negative) samples. The percentage of the positive closest neighbors gives the k-NN score. To test query samples, they took the average of the results of all the trained SVMs.

Another study for predicting kinase-phosphosite association is PKIS (Zou et al., 2013), which employed the composition of monomer spectrum (CMS) encodings and SVMs. It achieved a 73% sensitivity score, which was the highest score of the period among other studies in the same category.

PhosphoPick (Patrick et al., 2014) came up with a new idea that kinase-substrate phosphorylation can be found in the surrounding cellular context in addition to the substrate sequence. To realize this, Patrick et al. (2014) combined cellular context signals into a probabilistic Bayesian Network. They used protein–protein interaction data and cell-cycle profiles. Their model attained a mean AUC of ≈ 0.86 across 59 human kinases, outperforming motif-only baselines.

Traditional phosphorylation predictors rely on hand-engineered features—statistical,

biological, or disorder-based—derived from peptide sequences. With the rise of deep learning (DL), many models now leverage its ability to learn complex, non-linear representations directly from data, eliminating the need for manual feature engineering. DeepPhos (Luo et al., 2019), an earlier DL based study, aimed to uncover richer, non-linear motifs automatically. The authors designed a densely connected CNN architecture that learned high-dimensional representations of protein sequences to use for phosphorylation site prediction. Then, an additional layer-transfer fine-tuning step adapted the phosphorylation prediction model to kinase-specific tasks. For kinase-specific evaluation, the model obtained AUC ≈ 0.92 on the CMGC group and ≈ 0.91 on CAMK, clearly surpassing GPS 2.0.

GPS 2.0 was improved by Wang et al. (2020), called GPS 5.0, to obtain better prediction performance and wider coverage. Until 2020, no single tool could handle 479 human kinases and provide multi-species support. To improve the performance for predicting kinase-specific phosphorylation sites, they proposed two novel methods: i) position weight determination (PWD), which learned position-specific amino-acid weights via logistic-regression fitting, and ii) scoring matrix optimization (SMO), which refined the 300-pair BLOSUM62 scores by logistic regression to produce a kinase-specific substitution matrix. It demonstrated better performance compared to NetPhos and GPS 2.0.

While GPS 5.0 covered 479 human kinases, users needed a broader taxonomic scope, richer annotations, and higher accuracy, especially for Y kinases. GPS 6.0 (Chen et al., 2023) emerged to meet these needs. GPS 6.0 utilized ten sequence-derived feature sets from GPS 5.0 and iLearnPlus (Chen et al., 2021) and fed them into penalized logistic regression (Hosmer Jr et al., 2013), deep neural network, and Light Gradient Boosting Machine (Ke et al., 2017). Then, via transfer learning, they obtained 577 protein kinase-specific predictors at the group, family, and single protein kinase levels by employing a dataset consisting of 30,043 known site-specific kinase-substrate relations in 7041 proteins. Phosphorylation site prediction accuracy improved over previous versions of GPS.

2.4.3 Kinase Assignment Prediction Models

Ma et al. (2020) presented KSP, whose novelty relied on the fact that most of the predictors merge sequence motifs with protein-protein-interaction (PPI) data, yet KSP used only simple sequence similarity and ignored richer network topology. Ma et al. (2020) noted that capturing neighbor structure in a PPI graph could

boost kinase assignment, especially when motif information was weak. They built a single interaction network as a weighted bipartite graph. Then, they calculated a KSPScore for each kinase–substrate pair such that a four-term similarity score was computed and down-weighted by a degree factor to reduce bias towards well-studied kinases. They also provided an optional overall score by including the frequency and similarity features of sequences. KSP reached ≈ 83 - 86% top-10 accuracy on each fold among 10-fold cross-validation over all pairs.

Ma et al. (2023) introduced a holistic similarity-based prediction approach for understudied kinases. In the study, they drew attention to the issue that greater than 50% of human kinases have fewer than 15 labeled sites. Thus, the supervised methods may not build reliable per-kinase models. The authors built a merged kinase–kinase similarity graph combining the sequence, functional, contextual, and STRING-related similarities. Then, they leveraged positive sites from the most similar family/group neighbors by the top-k rule to train predictive models. Easy negatives were down-sampled to balance training. For 116 human kinases with 5–14 known sites, they achieved a balanced accuracy for kinase group: TK-0.81 , Other-0.78, STE-0.84, CAMK-0.84, TKL-0.85, CMGC-0.82, AGC-0.90, CK1-0.82 and Atypical-0.85.

Recently, Aman et al. (2025) developed a tool, KinAID, which serves as an orthology-based kinase-substrate prediction and analysis tool, not a prediction algorithm. While most of the analysis tools remain human-only, the authors wanted a simple, multi-species utility that assigns kinases, infers kinase-activity shifts, and draws ready-to-publish plots. They collected PWMs for 303 human S/T kinases and 93 human Y kinases from studies conducted by Johnson et al. (2023); Yaron-Barir et al. (2024), respectively. To map one-to-one or ambiguous orthologs in 10 species, they used DIOPT-grouping (Hu et al., 2011) multiple paralogs under a single "family" when necessary. For each input peptide of at least ten residues centered on S/T/Y, they computed the PWM score and kept matches in the top 10% of that kinase's background distribution. The resulting tool can output match tables, kinase-activity z-tests, volcano plots, heat-map clustering of sites by kinase preference, and interactive kinase–kinase networks as downloadable.

2.4.4 Zero-Shot Based Kinase Prediction Models

DeepKinZero (Deznabi et al., 2020) is the first study in the literature framing kinase–phosphosite association as a zero-shot learning problem. DeepKinZero takes

pairs of kinases and 15-residue peptides centered on phosphorylation sites as inputs. Phosphosite sequences are embedded using ProtVec—a 3-mer Word2Vec model trained on Swiss-Prot (Asgari and Mofrad, 2015; Bairoch et al., 2005). Kinases are represented by hierarchical labels (family and group), Enzyme Commission class (Bairoch, 2000), and Kin2Vec embeddings (Asgari and Mofrad, 2015). A bi-linear compatibility model (Sumbul et al., 2018) learns to score kinase-phosphosite pairs; meanwhile, phosphosite sequence representations are refined by an LSTM module (Hochreiter and Schmidhuber, 1997). In the inference phase, the knowledge learned by the bi-linear model is transferred to make predictions for phosphosite sequences among kinases-entirely unseen during training-. DeepKinZero reached 21.52% top-1 accuracy on a test set containing kinases never seen during training. The result is far better than a random guess of 0.89%.

Phosformer (Zhou et al., 2023) asked if a single, sequence-only transformer was capable of learning kinase-family specificity end-to-end and remained interpretable. The authors first pre-trained a 6-layer encoder–decoder transformer on masked-language modeling (MLM) of roughly 300,000 kinase domains and 5,000 substrate proteins. Then they fine-tuned it on kinase-peptide pairs with "whether this phosphosite is phosphorylated by this kinase" binary objective using a novel multi-level negative-sampling; hard negatives (peptides having experimentally proven phosphorylation site but lacking the paired kinase) and easy negatives (peptides having S/T/Y residue yet there is no evidence of being phosphorylated by any kinase). They employed focal loss to handle 1:16 class imbalance ratio. Across 106 test kinases, Phosformer reached 0.86 AUC-ROC score with false-positive rates less than 2%. The authors noted that Phosformer was theoretically capable of zero-shot prediction, yet they would present the details of zero-shot prediction in a forthcoming study.

Later on, the authors extended Phosformer to create Phosformer-ST (Zhou et al., 2024), which is a transformer-based kinase-substrate predictor fine-tuned on a vast, balanced dataset. This dataset, which was curated by the Phosformer-ST study using Johnson et al. (2023)'s kinase atlas as its source, consisted of roughly one million positive/negative peptide–kinase pairs covering 300 S/T kinases. They aimed to reach a powerful zero-shot generalization and explainable outputs. In their approach, an ESM-2 (Lin et al., 2023) backbone was fine-tuned in a multi-task setup: i) masked-language modeling over 295,000 kinase domains from 18,832 organisms, ii) binary classification of phosphorylation between 15-residue peptides and kinase domains. To handle overfitting, they used data augmentation strategies including shifting of domain boundaries, token masking, and negative-sampling. In a zero-shot split where all pairs for 10% of unseen kinases were excluded during training,

Phosformer-ST obtained a 0.886 AUC-ROC score.

In DARKIN (Sunar et al., 2024), our previous study, we presented a dataset and established a benchmark for assessing protein language models on the dark kinase-phosphosite prediction task with zero-shot learning. To evaluate the representability of pLMs on the task, we adapted a k-NN classifier to zero-shot evaluation as a baseline predictor. We further employed Sumbul et al. (2018)'s bi-linear model to learn associations between kinase-phosphosite pairs. The results showed that ESM-1b (Rives et al., 2021), ProtT5-XL (Elnaggar et al., 2021), and SaProt (Su et al., 2023) provided the best representations in this task.

Recently, Esmaili et al. (2025) introduced their study, having the same objective as DeepKinZero, which approaches kinase-phosphosite association as a zero-shot classification problem. They proposed a solution similar to Phosformer and Phosformer-ST did; i.e., to find an answer "whether phosphosite X is phosphorylated by kinase Y". The architecture included ESM-2 (Lin et al., 2023) (having 650 million parameters) as encoder, where full kinase and substrate sequences were fed into, and a four-layer decoder that output a binary interaction score in an auto-regressive manner. While preparing hard-negative examples, they followed this strategy: for every positive pair, negatives were chosen among kinases whose ESM-2 embeddings were located closest in Euclidean space, which forced the network to learn specificity. On the authors' benchmark, the proposed model outperformed both Phosformer and GPS 6.0 in terms of F1-score.

In summary, computational methods for kinase–phosphosite association prediction have significantly evolved over the past two decades, transitioning from simple sequence-based predictors to complex deep-learning frameworks capturing long-range dependencies and kinase specificity.

2.5 Task-Aware Fine-Tuning and Re-initialization Strategies

Protein language models provide rich but task-agnostic embeddings. These representations may not capture the subtle biochemical context to correctly predict kinase-phosphosite association in low-data regimes (Unsal et al., 2022). Recent studies demonstrated that pre-training and/or fine-tuning backbone models to encode task-specific nuances has improved prediction performances (Schmirler et al., 2024; Zhou et al., 2023, 2024; Esmaili et al., 2025).

Schmirler et al. (2024) conducted a comprehensive evaluation of supervised fine-tuning for pLMs. They fine-tuned ESM-2 (different versions with various parameter sizes), ProtT5-XL, and Ankh (base and large versions) on eight benchmark tasks, such as secondary structure, intrinsic disorder, subcellular localization, stability, and mutational-scanning fitness landscape. They compared fine-tuned models with predictors that kept the pLMs frozen. They observed that supervised fine-tuning numerically improved performance for almost every model.

Zhou et al. (2023) (Phosformer) pre-trained a transformer with masked language modeling on peptide sequences from the Uniprot database (Bairoch et al., 2005) to understand the "language of life". Then, it used the pre-trained encoder to fine-tune it on the kinase-phosphosite prediction task. Zhou et al. (2024) (Phosformer-ST) leveraged fine-tuning rather than pre-training. It extended Phosformer's idea with a multi-task learning approach. Phosformer-ST optimized ESM-2's encoder on both masked language modeling objective and the kinase-phosphosite association task. Esmaili et al. (2025) also employed the ESM-2 encoder as a backbone, yet it replaced the classifier in Phosformer-ST with an auto-regressive decoder. Then, it fully fine-tuned ESM-2, and the resulting model was capable of making zero-shot generalization on kinase-specific predictions. All three studies demonstrated improved prediction and, except for Phosformer (which did not report zero-shot results), effective generalization to unseen kinases, which suggests that incorporating task-specific biochemical context improves the understanding of models and predictive ability.

Previous studies have explored different ways of adapting transformer layers for downstream tasks. For instance, ESM-Effect (Glaser and Braegelmann, 2025) compared freezing versus fine-tuning various layers to increase efficiency. It fine-tuned only the last two layers of ESM-2 and matched the accuracy of full-model fine-tuning on four datasets, while training faster and without requiring LoRA (Hu et al., 2022) or other adapters. The results highlighted that unfreezing a slice of the transformer often suffices for downstream tasks. Another technique to adapt deep networks to downstream tasks is layer-wise re-initialization. In the image domain, Zaidi et al. (2023) showed that re-initializing the deeper layers of ResNet (He et al., 2016) models improves accuracy and generalization in data-scarce settings.

3. METHODS

This chapter describes the zero- and few-shot kinase–phosphosite association prediction models developed in this thesis. In Sections 3.1 and 3.2, we introduce the architectures for zero-shot and few-shot prediction and detail the kinase–phosphosite datasets used. Section 3.3 defines the evaluation metric used to quantify model performance. Sections 3.4 and 3.5 present our approaches for enhancing phosphosite and kinase representations, respectively, together with the datasets underpinning each method. Finally, Section 3.6 outlines the experimental design, including zero- and few-shot training protocols, hyperparameter tuning procedures, integration of additional kinase features, and the configuration of transformer layers employed in our prediction models.

3.1 Zero-shot Learning Approach

3.1.1 Problem Definition

We defined the zero-shot kinase-phosphosite prediction task as follows:

Let $\mathcal{X}$ be the space of 15-residue peptide sequences, in which the phosphorylated residue is located in the central position, and let $\mathcal{Y}$ be the set of human kinases. This constitutes the classes in the machine learning task. Kinase-phosphosite association prediction task seeks the kinase $y \in \mathcal{Y}$ that is most likely to catalyze the phosphorylation at the central residue of a given 15-residue peptide window $x \in \mathcal{X}$. Since a phosphosite can be phosphorylated by multiple kinases, we framed the problem as a multi-label classification task.

The classes in the training and test sets do not overlap. We denoted the kinases seen during training as $\mathcal{Y}_{tr} \subset \mathcal{Y}$ and test kinases reserved for testing as $\mathcal{Y}_{te} \subset \mathcal{Y}$, where $\mathcal{Y}_{te}$ includes the zero-shot classes, and is disjoint from $\mathcal{Y}_{tr}$. The training dataset, $D_{tr} = (x_i, y_i), i = 1, \ldots, N_{tr}$, contains training kinase-phosphosite pairs, where $y_i \in \mathcal{Y}_{tr}$. Similarly, the test dataset, $D_{te} = (x_j, y_j), j = 1, \ldots, N_{te}$, contains phosphosite pairings of the test kinases $\mathcal{Y}_{te}$.

3.1.2 Dataset

The dataset used in the zero-shot learning approach was taken from our previous work, DARKIN benchmark (Sunar et al., 2024). This dataset contains kinase-phosphosite pairs and train, validation, and test splits.

The set of human kinases in DARKIN was obtained from the curated resource of Moret et al. (2020). Experimentally validated phosphosites for each kinase were obtained from PhosphoSitePlus as of May 2023 (Hornbeck et al., 2012). To align the zero-shot learning set and prevent information leakage, we clustered all kinase domain sequences at a global sequence-identity threshold of 90%. Each resulting cluster was assigned in its entirety to one of three disjoint partitions—training, validation, or test. The data is partitioned as 80:10:10 split at the kinase level for training, validation, and test. Kinases with fewer than 15 associated phosphosites were retained exclusively in the training partition to ensure that the test data include enough phosphosite examples. This is different than what was done in Deznabi et al. (2020). Thus, the dark and light kinases are switched. The DARKIN's partitioning algorithm was performed with four random seeds to report the splitting sensitivity of the models.

There are a total of 392 human kinases in the original DARKIN splits. We used kinase domain sequences, families, groups, and Enzyme Commission (EC) numbers as kinase features. Family and group memberships were obtained from Manning et al. (2002). EC classifications were retrieved from the ENZYME Database (Bairoch, 2000).

After hyperparameter optimization by using train and validation splits, we retrained zero- and few-shot learning approaches by merging training and validation sets. (For hyperparameter optimization details, please see Section 3.6.1.1)

Table 3.1 summarizes the number of kinase–phosphosite pairs in each DARKIN partition. To ensure a fair comparison between zero-shot and few-shot evaluations, we

held out the same set of test samples in both settings. Specifically, the counts shown for the test split in Table 3.1 show the remaining kinase–phosphosite associations after removing the few-shot samples from the original DARKIN test set.

Table 3.1 The number of kinase-phosphosite pairs in four different DARKIN splits obtained by running DARKIN's partitioning algorithm with four distinct random seeds for zero-shot setup.

	Seed 0	Seed 42	Seed 87	Seed 12345
Training+Validation Data	9855	9941	9921	10043
Test Data	1254	1190	1288	1268
Total	11109	11131	11209	11311
Training+Validation Kinases	352	352	353	354
Test Kinases	40	40	39	38
Total	392	392	392	392

3.1.3 Architecture

In this study, the core of the kinase-phosphosite association prediction model we adopted is the Bi-linear Zero-shot Model (BZSM). BZSM aims to estimate the compatibility between a given pair of phosphosite sequence x and kinase y. Although numerous zero-shot learning approaches have been developed—especially in the context of image classification—, bi-linear compatibility models remain among the most widely adopted (Xian et al., 2017; Akata et al., 2016; Romera-Paredes and Torr, 2015; Frome et al., 2013; Akata et al., 2015b; Kodirov et al., 2017; Sumbul et al., 2018; Deznabi et al., 2020). In this thesis, we adopt the specific bi-linear formulation introduced by Sumbul et al. (2018) and later applied by Deznabi et al. (2020).

The bi-linear compatibility function is defined as:

$$F(x,y) = [\theta(x)^\top \quad 1] W [\phi(y)^\top \quad 1]^\top \tag{3.1}$$

where $\theta(x) \in \mathbb{R}^d$ corresponds to the representation for a 15-residue peptide sequence containing phosphosite, and $\phi(y) \in \mathbb{R}^m$ corresponds to the kinase representation.

BZSM is trained by minimizing the regularized cross-entropy loss:

$$\min_{W} - \sum_{(x,y)\in D_{tr}} \log p(y|x) + \lambda \|W\|^2 \tag{3.2}$$

where the summation is carried out over all kinase-phosphosite pairs in the training set $D_{tr} = (x_i, y_i)$, and $p(y|x)$ is the softmax of F over the light kinases:

$$p(y|x) = \frac{\exp F(x,y)}{\sum_{y' \in Y_{tr}} \exp F(x,y')}. \tag{3.3}$$

The ℓ_2 regularization term $\lambda\|W\|^2$ in Eq. 3.2 is implemented as *weight decay* in practice. At test time, $p(y|x)$ is calculated via softmax over the test kinases.

The BZSM framework was applied in DeepKinZero (Deznabi et al., 2020), where phosphosite representations were refined via an LSTM layer. In contrast, our approach leverages a transformer module to fine-tune the phosphosite embeddings: by employing multi-head self-attention, the transformer selectively attends to key residues within each sequence and captures richer contextual relationships (Vaswani et al., 2017). Figure 3.1 depicts the overall architecture of our zero-shot prediction model.

In this thesis, we used the ESM-1b—a 650 million-parameter transformer model—and its fine-tuned and further pre-trained variants as our phosphosite embedding module. This choice is made based on our previous work, where we evaluated different pLMs' representational power Sunar et al. (2024).

ESM-1b is a 33-layer transformer model with an embedding dimension of 1280 per residue and 20 attention heads per layer. It was pre-trained on the UniRef50 protein sequence database using a masked language modeling (MLM) objective (Rives et al., 2021). Variants of ESM-1b and their specific configurations for zero- and few-shot kinase–phosphosite prediction are described in Sections 3.4, 3.5, and 3.6.2.

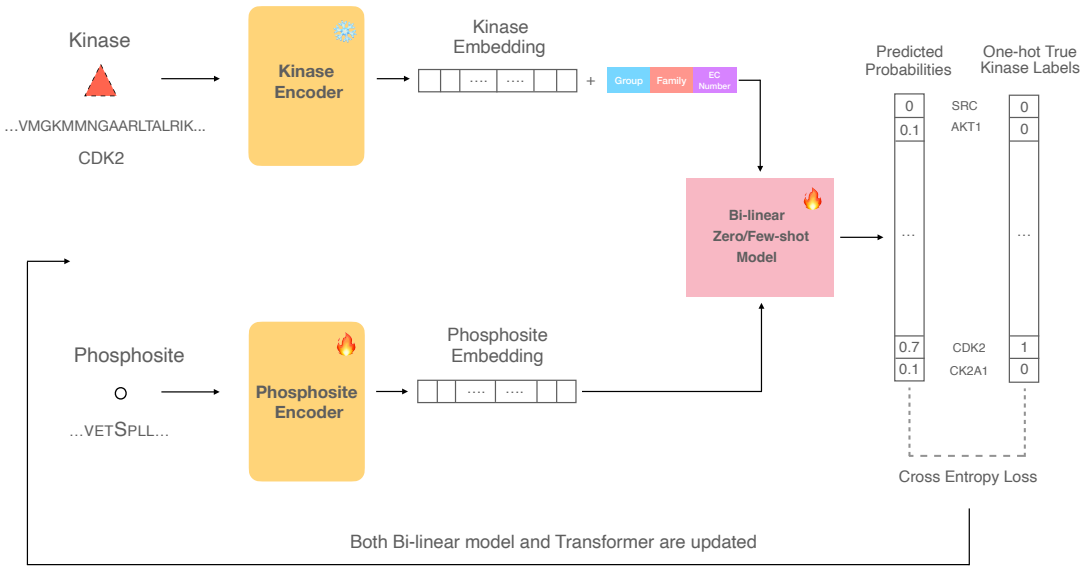


Figure 3.1 Architecture of the Bi-linear Zero-shot Model (BZSM) and its few-shot variant (BFSM). Phosphosite sequences are encoded by a transformer module whose parameters are updated during training, while kinase profiles are embedded via a fixed encoder. A bi-linear layer computes compatibility scores between phosphosite and kinase embeddings, and the model is trained using cross-entropy loss over light kinases (kinases with many known phosphosites). Depending on the evaluation setting, the transformer is either randomly initialized, fully fine-tuned, or partially re-initialized. In inference, the learned bi-linear weights are used to rank phosphosite candidates for dark kinases (under-studied kinases).

3.2 Few-shot Learning Approach

3.2.1 Problem Description

In the few-shot learning approach, $\mathcal{X}$ denotes the space of 15-residue peptide sequences in which phosphorylated amino acids reside in the central position, and $\mathcal{Y}$ denotes the set of all human kinases. The set of kinases is split into a training set of light kinases $\mathcal{Y}_{\text{tr}}$, and a few-shot set of under-studied kinases $\mathcal{Y}_{\text{fs}} \subset \mathcal{Y}$ for which only a handful of labeled examples are available:

$$\mathcal{Y}_{\text{tr}} \subset \mathcal{Y}, \qquad \mathcal{Y}_{\text{fs}} \subset \mathcal{Y}, \qquad \mathcal{Y}_{\text{tr}} \cap \mathcal{Y}_{\text{fs}} = \varnothing.$$

.

For each few-shot kinase $y \in \mathcal{Y}_{\text{fs}}$ we supplied at most K labeled phosphosite sequences (with $K \ll N_{\text{tr}}$), forming the *support set*:

$$S_y = \Big\{(x_{y,1}, y), \ldots, (x_{y,K}, y)\Big\}, \qquad \text{and} \qquad S_{\text{fs}} = \bigcup_{y \in \mathcal{Y}_{\text{fs}}} S_y.$$

Therefore, the complete training dataset is:

$$D_{\text{train}} = D_{\text{tr}} \cup S_{\text{fs}}, \qquad D_{\text{tr}} = \{(x_i, y_i)\}_{i=1}^{N_{\text{tr}}}, \quad y_i \in \mathcal{Y}_{\text{tr}}.$$

At evaluation time we used a *query set*:

$$D_{\text{test}} = \{(x_j, y_j)\}_{j=1}^{N_{\text{te}}}, \qquad y_j \in \mathcal{Y}_{\text{fs}},\ (x_j, y_j) \notin S_{\text{fs}},$$

and the model should predict kinase y in $\mathcal{Y}_{\text{fs}}$ that is likely to catalyze the phosphorylation of a given 15-residue peptide centered on the phosphorylated residue $x \in \mathcal{X}$. The prediction task remains a multi-label classification problem, but under the constraint that only K examples per few-shot kinase are available during training.

3.2.2 Dataset

In our few-shot experiments, we derived the dataset from the DARKIN benchmark by adopting a 5-shot-per-kinase protocol—a setting that is standard in the few-shot learning literature (Snell et al., 2017; Li et al., 2021; Oreshkin et al., 2018; Finn et al., 2017). This choice balances the scarcity of available phosphosites per kinase in the zero-shot test set against the need for enough examples to enable effective model adaptation.

Since some phosphosites in DARKIN are annotated with multiple kinases (i.e., multi-labeled samples), we first decomposed each into separate kinase–phosphosite pairs. Then, for every kinase in the DARKIN test partition, we randomly selected five phosphosites and moved them into the combined training–validation set. This procedure converts the original zero-shot split into a few-shot learning scenario. Table 3.2 summarizes the resulting counts of kinase–phosphosite pairs in each few-shot split.

Table 3.2 The number of kinase-phosphosite pairs in four distinct few-shot DARKIN splits. Since our few-shot prediction model sees all classes, including the few-shot ones, the number of training kinases became 392 in this setup.

	Seed 0	Seed 42	Seed 87	Seed 12345
Train+Validation Data	10055	10141	10116	10233
Test Data	1254	1190	1288	1268
Total	11309	11331	11404	11501
Train+Validation Kinases	392	392	392	392
Test Kinases	40	40	39	38
Total	392	392	392	392

Kinase features were the same as in the zero-shot setup: kinase domain sequence, family, group, and EC number.

3.2.3 Architecture

In the few-shot experiments, we employed the same transformer-based architecture introduced for zero-shot learning (Section 3.1.3), but trained it on the few-shot dataset defined in Section 3.2.2. To avoid ambiguity between the two protocols, the Bi-linear Zero-shot Model (BZSM) shown in Figure 3.1 is referred to as the Bi-linear Few-shot Model (BFSM) when applied in the few-shot setting.

3.3 Evaluation Metric

To assess both zero-shot and few-shot models, we employed macro-averaged average precision (AP). Average precision measures the area under the precision–recall curve across all decision thresholds, thereby capturing the full trade-off between precision and recall. By computing AP separately for each class and then averaging (macro-averaging), we ensured that classes with few examples contribute equally, mitigating the effects of class imbalance. Unlike accuracy, AP does not depend on a single threshold and directly evaluates the model's ranking ability in multi-label or multi-class settings. Formally, if there are C classes and AP_c denotes the area under the precision–recall curve for class c, then the macro-averaged AP is defined as:

$$\text{Macro-AP} = \frac{1}{C}\sum_{c=1}^{C}\text{AP}_c, \tag{3.4}$$

3.4 Approaches to Enhance Phosphosite Representations

Protein language models (pLMs) provide general representations for peptide sequences. However, they perform better on specialized tasks after domain adaptation (Zhou et al., 2024; Schmirler et al., 2024). To check if there is benefit in a task-aware pLM, we trained several phosphosite-aware variants optimized under four distinct objectives:

- Fine-tuning on Phosphorylation Prediction Task – discriminating between phosphorylated and non-phosphorylated sites.
- Fine-tuning on Masked Language Modeling (MLM) Objective – adapting the backbone model to the phosphosite domain with MLM
- Multi-Task Fine-tuning - jointly optimizing the pLM on phosphorylation prediction task and MLM objective.
- Pre-training from scratch on MLM Objective – training a new model de novo on MLM objective.

These efforts yielded six domain-adapted models, each of which has been published on HuggingFace along with full training metadata. The subsections below describe, for each model, the source datasets, training objectives and protocols, and the selected hyperparameters.

3.4.1 Fine-Tuning ESM-1b on Phosphorylation Prediction Task

To tailor ESM-1b to phosphorylation biology, we fine-tuned it on a binary classification task: predicting whether a serine, threonine, or tyrosine (S/T/Y) is phosphorylated or not, given the surrounding residues (a 15-residue window was used). There

are many phosphosites whose cognate kinases are unknown, which we referred to as unlabeled phosphosites. We used these unlabeled phosphosites to fine-tune the ESM-1b model. By optimizing a binary cross-entropy loss and focusing the model's self-attention on local sequence context, we fine-tuned the ESM-1b to recognize phosphorylation sites.

3.4.1.1 Dataset

We assembled a dataset of unlabeled phosphosites—residues experimentally confirmed as phosphorylated but without kinase annotations—from PhosphoSitePlus (March 2024 release) (Hornbeck et al., 2012). Each site is encoded as a 15-residue peptide with the phosphorylated serine, threonine, or tyrosine in the middle. These 15-residue peptides comprise the positive class in our binary phosphorylation prediction task.

Negative samples were constructed following the protocol of Gao et al. (2010). For each phosphosite peptide in PhosphoSitePlus, we retrieved its substrate sequence and identified another residue of the same type (S, T, or Y). We then extracted a 15-residue window (seven residues upstream and downstream) around this non-phosphorylated site and verified that this peptide does not appear in PhosphoSitePlus as a known phosphosite. These sequences were labeled as negative examples for the binary phosphorylation classification task.

In total, we obtained 730,149 15-residue peptides. The overall count is odd because, in some cases, the substrate protein contained only one S, T, or Y residue, making it impossible to generate a corresponding negative example. We named this dataset *BinaryPhPrediction*.

3.4.1.2 Fine-tuning Details

We employed the curated dataset described in Section 3.4.1.1 to fine-tune ESM-1b with an added binary classification head to predict whether a given phosphosite sequence is phosphorylated or not. The model was trained by minimizing the binary cross-entropy loss:

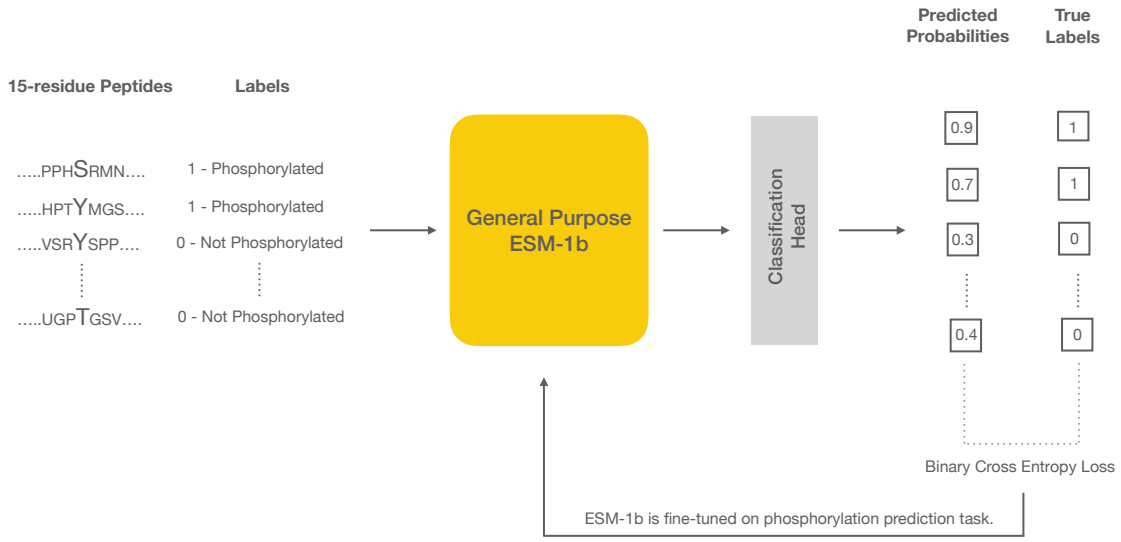


Figure 3.2 Peptide sequences prepared by leveraging PhosphoSitePlus dataset are fed into ESM-1b and a following binary classification layer (classification head) which generates probabilities of whether each sequence is "phosphorylated" (1) or "not phosphorylated" (0). The difference between predictions and true labels is calculated by using Binary Cross-Entropy Loss (See Equation 3.5).

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\Big] \tag{3.5}$$

where $y_i \in \{0,1\}$ is the true label and $\hat{y}_i$ the model's predicted probability for sample i. Figure 3.2 illustrates the fine-tuning scheme.

Training was performed with the AdamW optimizer (learning rate 5×10^{-5}, batch size 512) for three epochs, holding out 10% of the dataset as an internal test set. The model achieved 94% accuracy on this internal test. To evaluate generalization, we tested on the DARKIN split generated with random seed 12345—ensuring no overlap with training sites—and obtained 88% accuracy.

3.4.2 Fine-Tuning of ESM-1b on Masked Language Modeling Objective

To further specialize ESM-1b for phosphorylation biology, we fine-tuned the pre-trained model on the MLM objective using datasets of phosphosite sequences. During MLM training, random residues in each peptide were masked, and ESM-1b was trained to recover the masked residues from their surrounding context. This

targeted adaptation encourages the model to internalize phosphorylation-specific sequence patterns such as position-dependent residue preferences.

3.4.2.1 Dataset

We prepared two distinct datasets to produce two versions of fine-tuned ESM-1b.

3.4.2.1.1 UnlabeledPS: We curated a dataset of unlabeled phosphosite-containing peptide sequences from PhosphoSitePlus. Beginning with substrate protein sequences containing phosphosites, we truncated each sequence to a maximum length of 128 amino acids, ensuring that the phosphosite remained within this window. This truncation was done to accommodate hardware constraints. After processing, the dataset comprised 352,453 peptides. We refer to this collection as *UnlabeledPS*.

3.4.2.1.2 DARKINHomologs: We constructed a second dataset by first collecting the substrate protein sequences that contain the 15-residue phosphosite windows in DARKIN, and by extending each peptide with up to 250 homologous sequences identified via PSI-BLAST (two iterations, $\geq 30\%$ identity, E-value $\leq$ 1e-5) (Kuru et al., 2022). To avoid redundancy, among the fully conserved homologs, we select a single representative sequence. Each homologous sequence was then centered on the annotated phosphosite and truncated to a maximum length of 128 amino acids, ensuring the phosphosite remained within the window. This procedure yielded 702,468 unique peptide sequences, hereafter referred to as *DARKINHomologs*.

3.4.2.2 Fine-tuning Details

We used the two phosphosite-focused datasets introduced in Section 3.4.2.1:

- *UnlabeledPS*: 352,453 truncated sequences from PhosphoSitePlus
- *DARKINHomologs*: 702,468 homolog-augmented sequences via PSI-BLAST

We appended a masked-language-modeling head to ESM-1b's encoder and fine-tuned the entire network on two distinct datasets by minimizing cross-entropy over masked tokens:

$$\mathcal{L}_{\text{MLM}} = -\frac{1}{|M|} \sum_{i \in M} \log p_\theta(x_i \mid x_{\setminus M}), \tag{3.6}$$

where M is the set of masked token indices in each input sequence, x_i the true amino acid, and $x_{\setminus M}$ the sequence with those positions masked with a 0.15 masking ratio. Figure 3.3 visualizes the fine-tuning process.

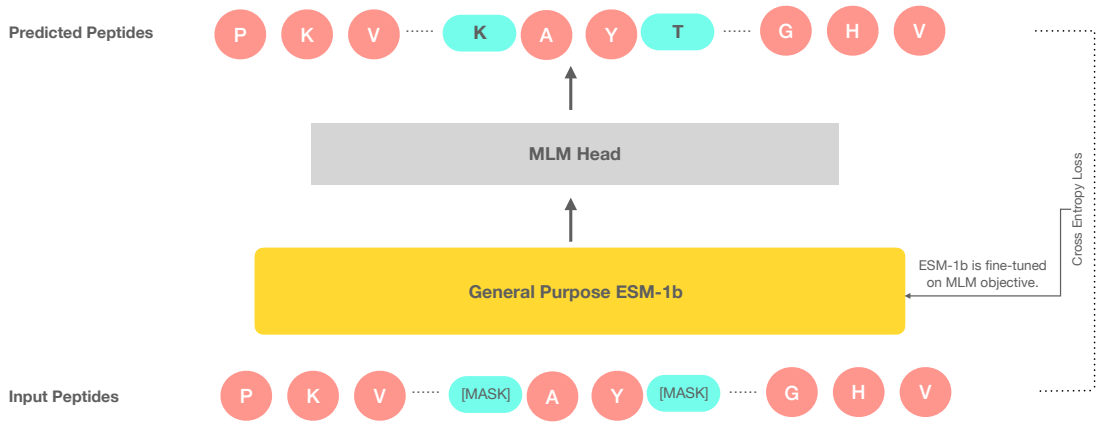


Figure 3.3 Peptide sequences containing masked positions ([MASK]) are fed into ESM-1b having an MLM head on top. The model is fine-tuned to predict masked amino acids via an MLM header. The predictions of the model are compared with the real sequences using the Cross-Entropy Loss function (See Equation 3.6), which minimizes loss to fine-tune ESM-1b.

Both models were fine-tuned for 100 epochs using the AdamW optimizer (learning rate 5×10^{-5}). We employed a per-GPU batch size of 64 with gradient accumulation over four steps, yielding an effective batch size of 256. A random 10% of each dataset was held out for validation, and we tracked perplexity on this split: it decreased from 5.42 to 2.27 on *UnlabeledPS* and from 7.05 to 2.69 on *DARKINHomologs*.

3.4.3 Multi-Task Fine-Tuning of ESM-1b on Masked-Language Modeling and Phosphorylation Prediction Objectives

Another approach to specialize ESM-1b for phosphorylation biology, we performed multi-task fine-tuning of ESM-1b with two concurrent objectives: i) MLM on extended, phosphosite-centered peptides to capture global protein language patterns, and ii) binary classification of phosphorylation status over 15-residue windows to emphasize local phosphosite motifs. We optimized the average of cross-entropy losses

for both branches on a shared encoder, aiming to obtain representations that are both broadly informed and task-aware.

3.4.3.1 Dataset

We used two datasets derived from PhosphoSitePlus; their curation was detailed in Sections 3.4.1.1 and 3.4.2.1 (*BinaryPhPrediction* and *UnlabeledPS*, respectively). For the classification branch, we randomly sampled 175,000 positive and 175,000 negative examples (seed 42) to make both datasets balanced in the MLM and binary classification fine-tuning.

3.4.3.2 Fine-tuning Details

We augmented the ESM-1b encoder with two task-specific heads; an MLM head and a binary classification head. We then fine-tuned ESM-1b jointly on both tasks to optimize a combined loss. The loss function is formulated as:

$$\mathcal{L}_{\text{total}} = \frac{1}{2}\mathcal{L}_{\text{MLM}} + \frac{1}{2}\mathcal{L}_{\text{BCE}}. \tag{3.7}$$

where $\mathcal{L}_{\text{MLM}}$ and $\mathcal{L}_{\text{BCE}}$ are defined in Equations 3.6 and 3.5.

For task scheduling, at each batch step, a Bernoulli trial ($p = 0.5$) selected either the unmasking or the classification task. We fine-tuned the model for three epochs, employing the AdamW optimizer with learning rate 1×10^{-5}, masking probability 0.15, per-GPU batch size 64, and gradient-accumulation as 4. To evaluate masked token prediction and classification performance, we tested on the DARKIN split generated with random seed 12345. The classification head reached an accuracy of 0.77, and in the MLM branch, the perplexity decreased from 11.25 to 5.38.

Fine-tuning process is illustrated in Figure 3.4.

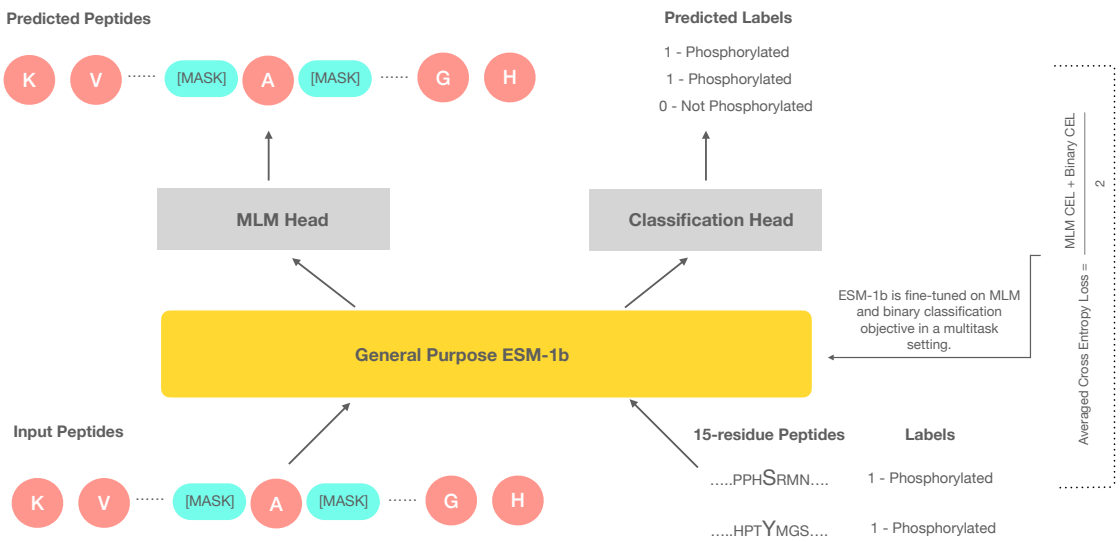


Figure 3.4 15-residue peptides and substrate protein sequences containing masked specific positions ([MASK]) are fed into ESM-1b with a Bernoulli trial $p = 0.5$. ESM-1b is fine-tuned to predict masked amino acids via the MLM head and phosphorylation via the classification head. The predictions of the model are compared with the real labels using a combined loss function (See Equation 3.7) which minimizes loss to fine-tune ESM-1b.

3.4.4 Pre-Training of ESM-1b on Masked Language Modeling Objective

Until now, our fine-tuning strategies have used the original ESM-1b model pre-trained on UniRef50, and thus inherited biases from that broad protein corpus. To explore whether a language model trained exclusively on phosphosite-focused data could yield more specialized embeddings, we pre-trained ESM-1b from scratch—randomly initializing all weights—and optimized the masked language modeling objective using a corpus composed solely of phosphosite-related peptide sequences.

3.4.4.1 Dataset

For this task, we employed two datasets described in Section 3.4.2.1, those are: *UnlabeledPS* and *DARKINHomologs* (For details please see Paragraph 3.4.2.1.1 and Paragraph 3.4.2.1.2)

3.4.4.2 Pre-Training Details

We trained two separate ESM-1b models, one on *UnlabeledPS* dataset and the other on *DARKINHomologs* dataset, by loading ESM-1b's configuration (via `AutoConfig.from_pretrained('facebook/esm1b_t33_650M_UR50S')`), ensuring that no pre-trained weights were loaded and all parameters were randomly initialized. Models were then trained for 100 epochs on the MLM objective by optimizing $\mathcal{L}_{\text{MLM}}$ computed over masked tokens (please see Equation 3.6). Figure 3.5 demonstrates the pre-training process.

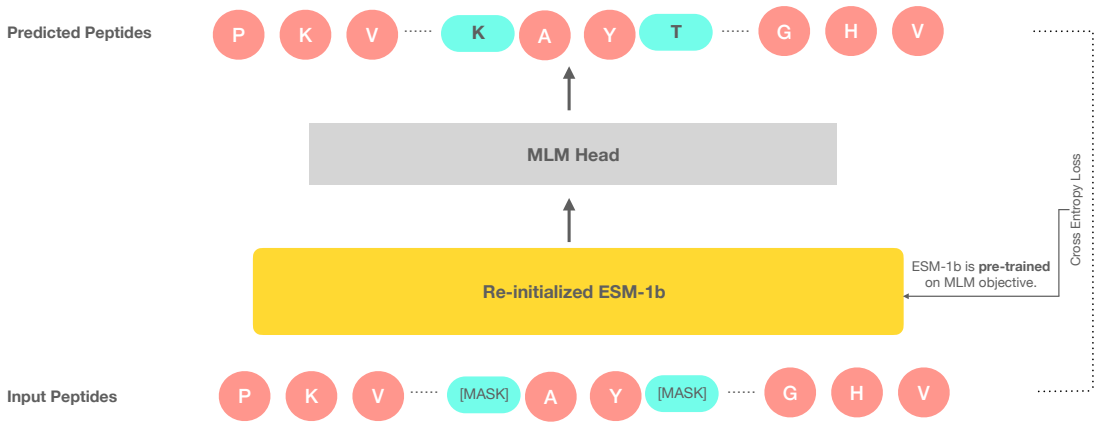


Figure 3.5 Randomly initialized ESM-1b model is pre-trained to predict masked amino acids. The model's predictions are compared with the real sequences using the MLM loss (See Equation 3.6), and the model is updated until the loss is minimized.

We used the AdamW optimizer with a learning rate of 5×10^{-5}, a batch size of 64 (4-step gradient accumulation for an effective batch size of 256), and a masking ratio of 0.15. 10% of each dataset was held out for evaluation, and performance was tracked via perplexity. Perplexity decreased from 13.51 to 2.20 on *UnlabeledPS* and from 12.32 to 1.44 on *DARKINHomologs*.

3.5 Approaches to Enhance Kinase Representations

As noted in Section 3.4, pLMs encode general biological information about a peptide sequence, which may result in overlooking domain-specific knowledge. To integrate kinase-specific information, we developed five variants of kinase-aware models by leveraging ESM-1b, kinase domain sequences, and additional features of kinases.

- Group Classification Fine-tuning – a multi-class classification task that focuses on predicting the group of a given kinase.
- Family-level Contrastive Learning – contrastive fine-tuning to increase the intra-family similarity while maximizing inter-family separation.
- Group-level Contrastive Learning – contrastive fine-tuning designed to boost intra-group cohesion while maximizing inter-group separation.
- Kinase-specific MLM Fine-tuning – adapting ESM-1b on MLM objective
- Kinase-specific MLM Pre-training – training ESM-1b from scratch on MLM objective.

These experiments output five models, and we shared them on HuggingFace. We detailed datasets and training protocols in the following sections.

3.5.1 Fine-Tuning of ESM-1b on Kinase Group Prediction

Kinases within the same phylogenetic group exhibit conserved catalytic mechanisms, activation-loop motifs, and substrate specificities. To encourage ESM-1b's encoder to capture these higher-order patterns, we introduced an auxiliary multi-class classification task: given a kinase's primary sequence, predict its membership among the ten established kinase groups. This objective guides the model to learn representations that reflect both sequence-level features and group-level functional relationships.

3.5.1.1 Dataset

We used the domain sequences of 392 human kinases belonging to ten kinase groups. The kinase group information was obtained from Manning et al. (2002). We named the dataset *KinaseGroups*.

3.5.1.2 Fine-tuning Details

We fine-tuned ESM-1b by appending a classification head that predicted the kinase group from a given kinase sequence. Model parameters were updated with the standard softmax cross-entropy loss computed between the predicted group probabilities and the ground-truth labels. Softmax cross-entropy loss is formulated as:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\log p_\theta\Big(g_i \mid x_i\Big), \tag{3.8}$$

,

where g_i is the true group of given kinase x_i.

Training was performed with the AdamW optimizer (learning rate 5×10^{-5}, batch size 8) over three epochs. On the held-out 15% test split of the kinase dataset, the model achieved an F1-score of 0.82 and an accuracy of 0.92. The network architecture is shown in Figure 3.6.

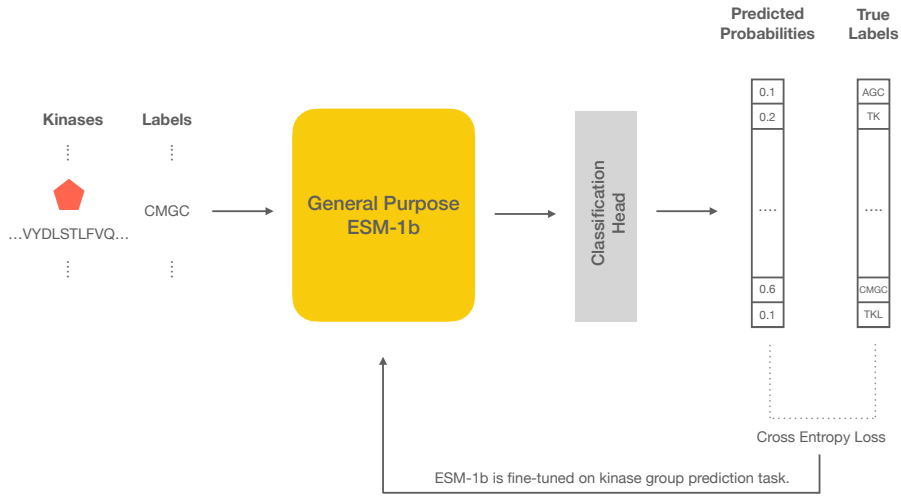


Figure 3.6 General-purpose ESM-1b model receives kinase domain sequences as input. A classification head on top of the encoder predicts the kinase group. Predicted class probabilities are compared with ground truth labels by cross entropy loss, $\mathcal{L}_{\mathrm{CE}}$ (See Equation 3.8). ESM-1b encoder is updated until the loss value is converged.

3.5.2 Contrastive Fine-Tuning of ESM-1b on Family/Group Based Kinase Triplets

To reveal the distinctive features of kinases and to increase the distance between family(or group) clusters, we fine-tuned the ESM-1b on triplets of kinase sequences:

an anchor kinase, a positive kinase from the same family (or group), and a negative kinase drawn from a different family (or group). In this way, we aimed to have the model learn more discriminative kinase representations by pushing the embeddings of different families (or groups) farther apart and clustering the embeddings of the same families (or groups) more tightly through contrastive loss.

3.5.2.1 Dataset

We constructed two kinase triplet datasets. One of them was based on family membership, and the other one was based on group membership.

Each triplet (x_a, x_p, x_n) consisted of the domain sequences of:

- an *anchor* kinase x_a,
- a *positive* kinase x_p that shared the same family (or group) with x_a, and phosphorylated the same residue type (S, T, or Y),
- a *negative* kinase x_n that was from a different family (or group) yet phosphorylated the same type of residue.

Putting a negative kinase into a triplet in which the anchor and positive kinase phosphorylated a different type of amino acid than the negative one would provide us with easy combinations. To prevent this and obtain hard negatives, we set the "same type of residue" criterion. We truncated them into 128-length sequences due to hardware constraints. As a result, we generated 83,000 family-based triplets and 79,000 group-based triplets. We named the dataset containing family-based triplets *FamilyTriplets* and the dataset involving group-based triplets *GroupTriplets*.

3.5.2.2 Fine-tuning Details

By using family-based and group-based triplet datasets, we fine-tuned two ESM-1b models. The triplets (x_a, x_p, x_n) were fed into models to maximize the similarity between x_a and x_p, while maximizing the distance of x_n to them. The learning process was controlled by InfoNCE Loss. This loss function is formulated as:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\Big(\text{sim}(h_{a_i}, h_{p_i})/\tau\Big)}{\exp\Big(\text{sim}(h_{a_i}, h_{p_i})/\tau\Big) + \sum_{j=1}^{N}\exp\Big(\text{sim}(h_{a_i}, h_{n_j})/\tau\Big)} \tag{3.9}$$

,

where $h_a, h_p, h_n \in \mathbb{R}^d$ correspond to the L2-normalized final hidden representations produced by ESM-1b for the anchor, positive, and negative samples respectively, $\text{sim}(\cdot,\cdot)$ corresponds to cosine similarity function and τ corresponds to temperature.

We fine-tuned the models for six epochs by employing the AdamW optimizer. Learning rate was set to 1×10^{-5}, and the batch size was 512. Figure 3.7 shows the fine-tuning process.

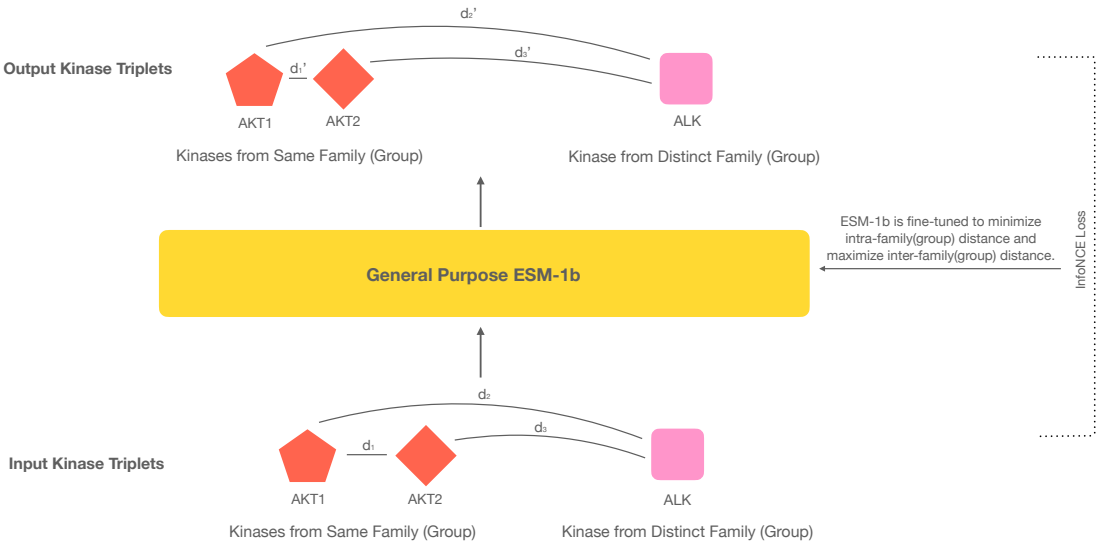


Figure 3.7 ESM-1b, having pre-trained weights, takes kinase triplets as inputs. Two of the kinases are from the same family (or group), and the other is from a distinct family (or group). The model is fine-tuned to maximize inter-family (or group) and minimize intra-family (or family) distance. Input and output similarities are compared by InfoNCE Loss, $\mathcal{L}_{\text{InfoNCE}}$ (See Equation 3.9). Model is updated until the loss converges.

3.5.3 Fine-Tuning of ESM-1b on Masked Language Modeling Objective

To specialize ESM-1b for kinase domains, we fine-tuned its transformer encoder on the masked language modeling (MLM) objective using both kinase domain sequences and their homologs. This targeted MLM training enables the model to internalize

residue-level conservation patterns and activation-loop motifs that underlie kinase function. By incorporating homologous variants, the resulting embeddings capture subtle functional distinctions across the kinome while retaining the broad contextual knowledge acquired during the original pre-training.

3.5.3.1 Dataset

For each kinase domain sequence employed in DARKIN, we obtained 1000 homologous sequences identified via PSI-BLAST (two iterations, $\geq$ 30 % identity, E-value $\leq$ 1e-5) (Kuru et al., 2022). To prevent duplication, identical homologs were consolidated into one representative sequence. Moreover, due to hardware constraints, we truncated the sequences to a maximum of 200 amino acids long, taking into account the inclusion of kinase active sites. In the end, the kinase homologs dataset contains 204,437 sequences. We named this dataset *KinaseHomologs*.

3.5.3.2 Fine-tuning Details

We fine-tuned ESM-1b on the kinase homologs dataset with an MLM objective (masking probability 0.15), using an MLM head to predict masked tokens and updating all encoder parameters by minimizing cross-entropy loss over masked tokens. (Section 3.4.2, Eq. 3.6).

We employed the AdamW optimizer (a learning rate of 5×10^{-5}). The batch size was set to 64 with a gradient accumulation of 4, resulting in an effective batch size of 256. The model was trained for 100 epochs. Fine-tuning process is visualized in Figure 3.8

To monitor fine-tuned model performance, we held 10% of the dataset out as test set. The final model achieved a decrease in perplexity from 1.37 to 1.22 on the evaluation set.

3.5.4 Pre-Training of ESM-1b on Masked Language Modeling Objective

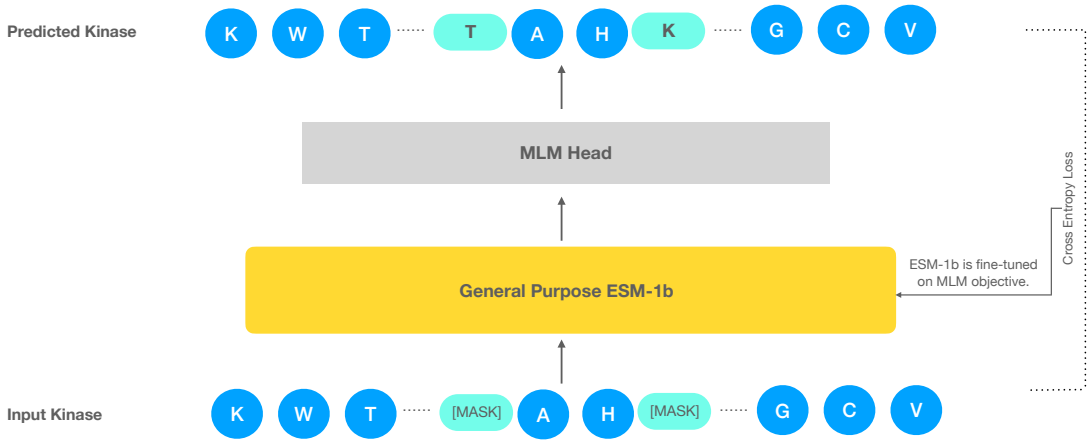


Figure 3.8 ESM-1b's weights are fine-tuned on the MLM objective. Kinase homologous sequences are randomly masked, and the model is trained to predict the masked amino acids by minimizing cross-entropy loss over masked positions.

To obtain representations that reflect kinase-specific biochemistry, we trained a randomly initialized ESM-1b model on the MLM objective using only human kinase domain sequences and their homologs. This domain-focused pre-training was expected to enable the transformer layers to internalize sequence motifs and conserved patterns that were characteristic of kinase catalytic domains.

3.5.4.1 Dataset

We used the *KinaseHomologs* dataset described in Section 3.5.3.1 for pre-training.

3.5.4.2 Pre-Training Details

We initialized the ESM-1b transformer architecture with random weights and pre-trained it from scratch on the MLM objective. Training ran for 100 epochs, optimizing the cross-entropy loss over masked residues (see Equation 3.6), using the same optimizer, learning rate, batch size, and validation protocol detailed in Section 3.5.3.2. The overall pre-training workflow is illustrated in Figure 3.9.

We held out 10% of the data as an internal test set, on which perplexity dropped from 1.77 to 1.35.

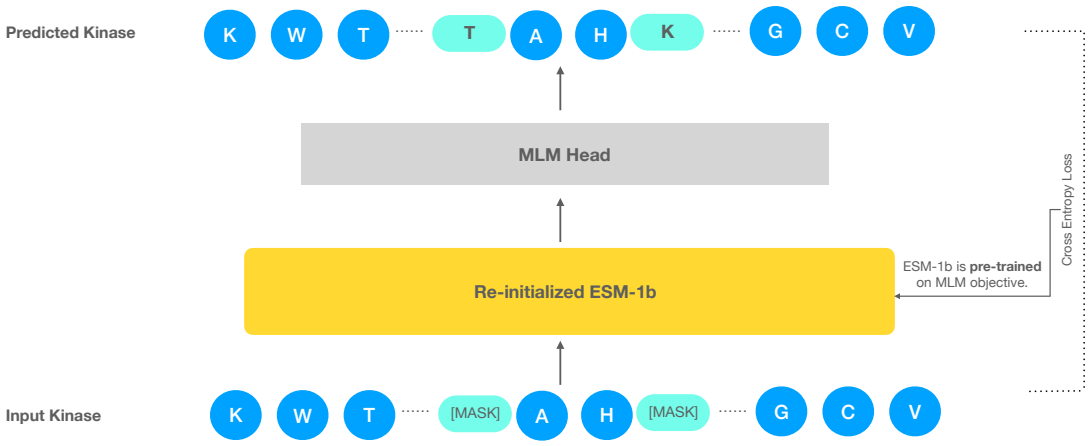


Figure 3.9 ESM-1b is pre-trained from scratch on the MLM objective by using kinase homologous sequences. The model's parameters are randomly initialized and then optimized to reconstruct masked amino acids by minimizing cross-entropy loss.

Before moving on to the next section, we summarized phosphosite- and kinase-aware models in Table 3.3

Table 3.3 We obtained six phosphosite-aware and five kinase-aware models. Models are produced by either fine-tuning a general-purpose ESM-1b or training a randomly initialized ESM-1b from scratch on several tasks and objectives.

Domain	Task/Objective	Fine-tuned/Pre-trained	Dataset
Phosphosite	Phosphorylation Prediction	Fine-tuned	Derived from PhosphoSitePlus, contains 15-residue peptides centered on phosphorylated and not phosphorylated residues
	Masked Language Modeling	Fine-tuned	Derived from PhosphoSitePlus, contains longer peptide sequences covering phosphosites (called UnlabeledPS)
	Masked Language Modeling	Fine-tuned	Contains longer peptide sequences covering DARKIN phosphosites and their homologs (called DARKINHomologs)
	Multi-task Learning	Fine-tuned	UnlabeledPS and dataset used in phosphorylation prediction
	Masked Language Modeling	Pre-trained	UnlabeledPS
	Masked Language Modeling	Pre-trained	DARKINHomologs
Kinase	Kinase Group Prediction	Fine-tuned	Contains human kinase domain sequences employed in the DARKIN dataset and corresponding kinase groups
	Contrastive Learning (Based on Family)	Fine-tuned	Derived from Manning et al. (2002)'s human kinase family classification, contains human kinase sequence triplets
	Contrastive Learning (Based on Group)	Fine-tuned	Derived from Manning et al. (2002)'s human kinase group classification, contains human kinase sequence triplets
	Masked Language Modeling	Fine-tuned	Contains human kinase sequences and their homologs
	Masked Language Modeling	Pre-trained	Contains human kinase sequences and their homologs

3.6 Experimental Setup

The experimental setup was designed to assess the effects of:

- peptide representations obtained from the adapted pLMs
- progressively re-initializing transformer layers
- employing a transformer for representation learning

in our zero- and few-shot kinase-phosphosite prediction approaches.

3.6.1 Zero- and Few-shot Learning Setups

The zero- and few-shot prediction models—whose architectures are detailed in Sections 3.1 and 3.2—each accept two inputs: a 15-residue peptide sequence centered on the phosphosite and the corresponding kinase.

In both the zero-shot and few-shot experiments, kinase embeddings were precomputed using the domain-adapted ESM-1b variants described in Section 3.5 and held fixed throughout training. Consequently, no gradient updates were applied to the kinase representations. Phosphosite sequences were encoded by a transformer initialized with weights from the adapted ESM-1b models (Section 3.4). Unlike kinase embeddings—which remained fixed—both the phosphosite encoder and the bi-linear interaction parameters were updated during training according to the configurations detailed in Section 3.6.2. As a result, task-specific gradient updates continually refined the phosphosite representations throughout both learning paradigms.

BZSM and BFSM were trained for 100 epochs by employing hyperparameters, of which the best combinations were found by random search. The hyperparameter tuning process was detailed in the following Section 3.6.1.1.

3.6.1.1 Hyperparameter Tuning

Table 3.4 Hyperparameters and their ranges used to tune BZSM.

Hyperparameter	Values (in a range)
Batch Size	64, 128, 256
Optimizer	Adam, SGD, RMSProp
Learning Rate	$[1 \times 10^{-6}$, 0.1]
Learning Rate Scheduler	ExponentialLR, StepLR, CosineAnnealingLR
Momentum	[0.95 , 0.9999]
Weight Decay	$[1 \times 10^{-5}$, 0.01]

We optimized training hyperparameters via a randomized search over the DARKIN Split1 (seed 12345) training and validation sets. Throughout tuning, both phosphosite and kinase inputs were encoded using the base ESM-1b model. Candidate configurations were sampled from the ranges listed in Table 3.4. Hyperparameter optimization was conducted exclusively on the BZSM, and the optimal settings were then applied to the BFSM.

The hyperparameter configuration that achieved the highest AP on the validation set consisted of a batch size of 64, the SGD optimizer with a learning rate of 0.01, a CosineAnnealingLR scheduler, momentum of 0.97, and a weight decay of 1×10^{-4}.

3.6.1.2 Kinase Additional Features

In Section 3.1, we described the datasets for our zero- and few-shot frameworks and noted that we augmented kinase embeddings with additional features: kinase family, kinase group, and Enzyme Commission (EC) number. Each feature was encoded as a one-hot vector and concatenated to the corresponding kinase domain embedding to enrich the representation.

3.6.2 Configurations of Transformer Module in Zero- and Few-Shot Prediction Setup

To optimize phosphosite-aware models for our zero- and few-shot prediction tasks, we designed a series of experiments across multiple protocols:

- Random initialization: All transformer weights were reset and trained from scratch.

- Full fine-tuning: All transformer weights were updated during downstream training.
- Partial re-initialization: Six cumulative reset experiments were designed. In each experiment $k \in \{1, \ldots, 6\}$, we reset the top k transformer layers to random weights and loaded the remaining layers from the adapted checkpoints:
  - $k = 1$: reset layer 32
  - $k = 2$: reset layers 31–32
  - $k = 3$: reset layers 30–32
  - $k = 4$: reset layers 29–32
  - $k = 5$: reset layers 28–32
  - $k = 6$: reset layers 27–32

For each re-initialized layer, we used Xavier (Glorot) uniform initialization (Glorot and Bengio, 2010), sampling each weight W_{ij} from

$$W_{ij} \sim \mathcal{U}\Big(-\alpha,\, \alpha\Big), \quad \alpha = \sqrt{\frac{6}{\mathrm{fan_{in}} + \mathrm{fan_{out}}}}, \tag{3.10}$$

,

where fan_{in} and fan_{out} are the number of input and output units of the layer. Xavier initialization preserves the variance of activations and back-propagated gradients in a roughly constant manner across layers, which prevents vanishing or exploding gradients.

4. RESULTS

This chapter presents performance results obtained with models trained in zero- and few-shot settings for the dark kinase-phosphosite association task. First, we benchmark phosphosite-aware adaptations of ESM-1b—holding kinase embeddings fixed—to quantify their impact on prediction performance. Next, we fix the two top-performing phosphosite models as backbones and introduce kinase-aware adaptations to evaluate their benefit. Together, these experiments identify the optimal pairing of phosphosite and kinase language models. We compare our best setups with state-of-the-art approaches and present ablation studies on the components of the final model.

Throughout this section, we will use abbreviations to refer to each adapted model. We encoded the name of each model as *Objective(or Task)-Training Method–Dataset*. Abbreviations for each of them are provided in Table 4.1. The adapted models evaluated in this section are listed in Table 4.2.

Table 4.1 Adapted model name components and their abbreviations.

Component	**Abbreviation**	**Description**
Objective / Task	BC	Binary Classification
	MC	Multi-class Classification
	MT	Multi-task Learning
	MLM	Masked Language Modeling
	CL	Contrastive Learning
	GP	General-purpose ESM-1b
Training Method	FT	Fine-tuning on the task dataset
	PT	Pre-training from scratch
Dataset	BinaryPhPrediction	15-residue peptides (positive/negative phosphorylation labels)
	UnlabeledPS	Unlabeled phosphosite-containing peptides
	DARKINHomologs	DARKIN phosphosite peptides plus homologs
	KinaseGroups	392 human kinase domain sequences with group labels
	KinaseHomologs	Kinase domains plus their homologous sequences

Table 4.2 Adapted phosphosite- and kinase-aware ESM-1b models released on Hugging Face.

Domain	Model Name (Abbrv.)	Hugging Face Link
Phosphosite	BC-FT-BinaryPhPrediction	ESM-1b Fine-Tuned Phosphorylation Prediction
	MLM-FT-UnlabeledPS	ESM-1b Fine-tuned on MLM Objective (UnlabeledPS)
	MLM-FT-DARKINHomologs	ESM-1b Fine-tuned on MLM Objective (DARKINHomologs)
	MT-FT-Binary&UnlabeledPS	ESM-1b Fine-tuned on Multi-task Objective
	MLM-PT-UnlabeledPS	ESM-1b Pre-trained on MLM Objective (UnlabeledPS)
	MLM-PT-DARKINHomologs	ESM-1b Pre-trained on MLM Objective (DARKINHomologs)
Kinase	MC-FT-KinaseGroups	ESM-1b Fine-tuned on Kinase Group Prediction Task
	CL-FT-FamilyTriplets	Contrastive Fine-tuning of ESM-1b on Family-wise Triplets
	CL-FT-GroupTriplets	Contrastive Fine-tuning of ESM-1b on Group-wise Triplets
	MLM-FT-KinaseHomologs	ESM-1b Fine-tuned on MLM Objective (Kinase Homologs)
	MLM-PT-KinaseHomologs	ESM-1b Pre-trained on MLM Objective (Kinase Homologs)

4.1 Baseline

We conducted two baseline experiments to establish a reference point so that we could understand the impact of adapted ESM-1b models on zero- and few-shot prediction performance. In these experiments, we obtained the phosphosite and kinase representations from general-purpose (GP) ESM-1b. We ran the experiments by: i) resetting the ESM-1b and training from scratch on phosphosites in parallel with BZSM (or BFSM) training; ii) fully fine-tuning the ESM-1b for phosphosites while simultaneously training BZSM (or BFSM). The results are shown in Table 4.3

Table 4.3 Zero- and few-shot AP scores for baseline kinase–phosphosite models. Kinase embeddings are frozen, while phosphosite sequence embeddings are either i) randomly re-initialized and trained from scratch (denoted "–") or ii) fully fine-tuned.)

Phosphosite Model(ESM-1b)	Model Config	ZSL-AP	FSL-AP
-	All Layers Re-init.	**0.2192**	**0.2208**
GP	Full-FT	0.1808	0.2140

Surprisingly, the ESM-1b initialized with random weights outperformed the general-purpose ESM-1b, which was fine-tuned on our zero- and few-shot classification tasks. This showed that the general representations learned by ESM-1b may not capture the subtle domain-specific features required for our task.

4.2 Impact of Task-Aware Phosphosite Representations

This section evaluates the impact of phosphosite-aware embeddings on zero- and few-shot kinase–phosphosite prediction. Kinase sequences were encoded with frozen, general-purpose (GP) ESM-1b embeddings, while phosphosite embeddings were varied across alternative embeddings. In this way, the contributions of task awareness in the phosphosite representations on performance changes were evaluated.

Table 4.4 Zero- and few-shot kinase–phosphosite prediction AP scores using various phosphosite encodings. "Model Config" denotes whether the phosphosite ESM-1b is initialized randomly (from scratch) or fully fine-tuned. Kinase embeddings remain fixed to GP ESM-1b encodings. The first two rows are baseline models; subsequent rows show task-aware phosphosite variants.

Phosphosite Model (ESM-1b)	Model Config	ZSL-AP	FSL-AP
-	All Layers Re-init.	0.2192	0.2208
GP	Full-FT	0.1808	0.2140
BC-FT-BinaryPhPrediction	Full-FT	0.1967	0.2090
MLM-FT-UnlabeledPS		0.1953	0.2052
MLM-FT-DARKINHomologs		0.2079	0.2157
MT-FT-BinaryPhPrediction&UnlabeledPS		0.2066	0.2136
MLM-PT-UnlabeledPS		**0.2197**	**0.2418**
MLM-PT-DARKINHomologs		**0.2267**	**0.2295**

As Table 4.4 shows, phosphosite-specific pre-training yielded the strongest embeddings for both zero- and few-shot prediction. Both of these outperformed the task-agnostic ESM-1b baseline. We interpreted these results as showing that phosphosite-dedicated pre-training enabled pLMs to produce more informative and richer phosphosite embeddings than fine-tuning.

Considering all phosphosite-aware pLMs together, each outperformed GP ESM-1b under both zero- and few-shot regimes. These findings demonstrated that task-aware phosphosite embedding contributes to improving the prediction of kinase–phosphosite association by capturing nuances related to phosphosites that GP ESM-1b may overlook.

4.3 Impact of Task-Aware Kinase Representations

This subsection evaluates the effect of varying kinase embeddings on zero- and few-shot kinase–phosphosite prediction. For kinases, we substituted frozen representations from each of our kinase-aware pLMs in turn. For phosphosites, we employed the two best-performing phosphosite models from Section 4.2 (MLM-PT-UnlabeledPS and MLM-PT-DARKINHomologs), and fully fine-tuned them in our zero- and few-shot prediction tasks. Comparing the resulting AP scores revealed which kinase-focused model delivers the best performance.

Table 4.5 and Table 4.6 show that the MLM-FT-KinaseHomologs pLM provided the best representations to BZSM and BFSM.

Table 4.5 Impact of kinase-aware embeddings on zero- and few-shot AP scores. The phosphosite model (MLM-PT-UnlabeledPS) is fully fine-tuned, while kinase embeddings are frozen and replaced by six kinase-aware pLM variants.

Phosphosite Model (ESM-1b)	Model Config	Kinase Model (ESM-1b)	ZSL-AP	FSL-AP
-	All Layers Re-init.	GP	0.2192	0.2208
MLM-PT-UnlabeledPS	Full-FT		0.2197	0.2418
MLM-PT-UnlabeledPS	Full-FT	MC-FT-KinaseGroups	0.1317	0.1421
		CL-FT-GroupTriplets	0.2158	0.2367
		CL-FT-FamilyTriplets	**0.2234**	0.2389
		MLM-FT-KinaseHomologs	**0.2359**	**0.2729**
		MLM-PT-KinaseHomologs	0.0797	0.0983

Table 4.6 Impact of kinase-aware embeddings on zero- and few-shot AP scores. The phosphosite model (MLM-PT-DARKINHomologs) is fully fine-tuned, while kinase embeddings are frozen and substituted from six kinase-focused frozen pLM variants.

Phosphosite Model (ESM-1b)	Model Config	Kinase Model (ESM-1b)	ZSL-AP	FSL-AP
-	All Layers Re-init.	GP	0.2192	0.2208
MLM-PT-DARKINHomologs	Full-FT		0.2267	0.2295
MLM-PT-DARKINHomologs	Full-FT	MC-FT-KinaseGroups	0.1326	0.1021
		CL-FT-GroupTriplets	0.2209	0.2237
		CL-FT-FamilyTriplets	**0.2305**	**0.2318**
		MLM-FT-KinaseHomologs	**0.2404**	**0.2488**
		MLM-PT-KinaseHomologs	0.1016	0.1782

While pre-training on the MLM objective resulted in better representations for phosphosites, the same strategy did not work for kinases. The reason might be data diversity: our corpus—392 human kinase domains plus their homologs— may lack the diversity needed for the model to learn discriminative catalytic features. Fine-tuning GP ESM-1b on the kinase homologs set, however, leveraged the broad contextual prior and yielded the stronger kinase representations, raising performance in both evaluation regimes.

Family-level contrastive fine-tuning (CL-FT-FamilyTriplets) clustered kinases within the same family while separating different families and outperformed the

general-purpose baseline, presumably by reducing mis-assignments across families. However, group-level contrastive learning (CL-FT-GroupTriplets) underperformed. The reason may be that merging many biologically distinct families into only ten broad groups forced the model to pull together kinases that differ in catalytic motifs and substrate preferences. This may blur family-specific cues in the embedding space and ultimately lower AP.

Similarly, the multi-class group classifier (MC-FT-KinaseGroups) offered no benefit, indicating that 392 sequences spread across ten classes provide insufficient signal for meaningful adaptation.

4.4 Experiments with Best Combinations of Task-Aware Models on Distinct DARKIN Splits

This section presents the results of experiments that we conducted to test the robustness to zero- and few-shot class distributions of our approach. Until this point, we used one DARKIN split (Split 1 - seed 12345) for experiments. To verify that the gains observed in Split 1 were not valid for a single training-test partition, we repeated the evaluation on three additional DARKIN splits generated with seeds 0, 42, and 87. The results are shown in Table 4.7

For each split, the two task-aware pLMs combinations; MLM-PT-UnlabeledPS & MLM-FT-KinaseHomologs and MLM-PT-DARKINHomologs & MLM-FT-KinaseHomologs outperformed the GP ESM-1b baselines in both zero-and few-shot predictions, which confirmed that the improvements were robust to zero- and few-shot class distributions.

Table 4.7 Zero- and few-shot AP scores on four DARKIN splits. Experiments are conducted by leveraging the two best task-aware model combinations. The phosphosite models are fixed to MLM-PT-UnlabeledPS and MLM-PT-DARKINHomologs, and both are fully fine-tuned. The kinase representations are provided by MLM-FT-KinaseHomologs pLM and frozen. The first two rows for the results of each split provide the baseline.

Split	Phosphosite Model (ESM-1b)	Model Config	Kinase Representation	ZSL-AP	FSL-AP
Split 1	-	All Layers Reinit.	GP	0.2192	0.2208
	GP	Full-FT		0.1808	0.2140
	MLM-PT-UnlabeledPS	Full-FT	MLM-FT-KinaseHomologs	**0.2359**	**0.2729**
	MLM-PT-DARKINHomologs	Full-FT		**0.2404**	**0.2488**
Split 2	-	All Layers Reinit.	GP	0.2352	0.2308
	GP	Full-FT		0.1907	0.2004
	MLM-PT-UnlabeledPS	Full-FT	MLM-FT-KinaseHomologs	**0.2459**	**0.2501**
	MLM-PT-DARKINHomologs	Full-FT		**0.2406**	**0.2518**
Split 3	-	All Layers Reinit.	GP	0.2099	0.2173
	GP	Full-FT		0.1878	0.1886
	MLM-PT-UnlabeledPS	Full-FT	MLM-FT-KinaseHomologs	**0.2198**	**0.2386**
	MLM-PT-DARKINHomologs	Full-FT		**0.2264**	**0.2307**
Split 4	-	All Layers Reinit.	GP	0.2389	0.2294
	GP	Full-FT		0.2231	0.2213
	MLM-PT-UnlabeledPS	Full-FT	MLM-FT-KinaseHomologs	**0.2604**	**0.2542**
	MLM-PT-DARKINHomologs	Full-FT		**0.2523**	**0.2481**

4.5 Ablation Study on Partial Fine-tuning and Partial Re-initialization

This section presents the experiments conducted to understand whether re-initializing only the top transformer layers (from layer 32 to 27) of the phosphosite encoder could enable phosphosite-aware models to adapt better specifically our zero- and few-shot prediction task. We employed the two task-aware pLMs combinations; MLM-PT-UnlabeledPS & MLM-FT-KinaseHomologs and MLM-PT-DARKINHomologs & MLM-FT-KinaseHomologs.

Tables 4.8 and 4.9 show that fully fine-tuning the phosphosite-aware models already outperformed the baseline results. Re-initializing just the top 1–6 layers delivered small additional gains under a few re-initialization setups.

MLM-PT-UnlabeledPS backbone reached the best zero-shot AP of 0.2447 on the experiment conducted by resetting layers [27–32]; the best few-shot AP of 0.2746 in the resetting layers [29–32] setup. MLM-PT-DARKINHomologs backbone obtained 0.2438 zero-shot AP when layers [30–32] were re-initialized; 0.2610 few-shot AP by resetting [27–32] layers.

Figures 4.1 and 4.2 reflect the changes in AP scores.

Table 4.8 Effect of partial layer re-initialization in the phosphosite model MLM-PT-UnlabeledPS. We reset the top transformer layers (e.g. "32-Reinit." means only layer 32 is re-initialized) and the remaining layers are fine-tuned. AP scores are reported for zero-and few-shot evaluation.

Phosphosite Model (ESM-1b)	Model Config	Kinase Model (ESM-1b)	ZSL-AP	FSL-AP
-	All Layers Reinit.	GP	0.2192	0.2208
GP	Full-FT		0.1808	0.2140
-	All Layers Re-init.	MLM-FT-KinaseHomologs	0.2284	0.2292
MLM-PT-UnlabeledPS	Full-FT		0.2359	0.2729
MLM-PT-UnlabeledPS	32-Reinit.	MLM-FT-KinaseHomologs	0.2408	0.2688
	31,32-Reinit.		0.2393	0.2704
	30,31,32-Reinit.		0.2391	0.2736
	29,30,31,32-Reinit.		0.2408	**0.2746**
	28,29,30,31,32-Reinit.		0.2364	0.2700
	27,28,29,30,31,32-Reinit.		**0.2447**	0.2732

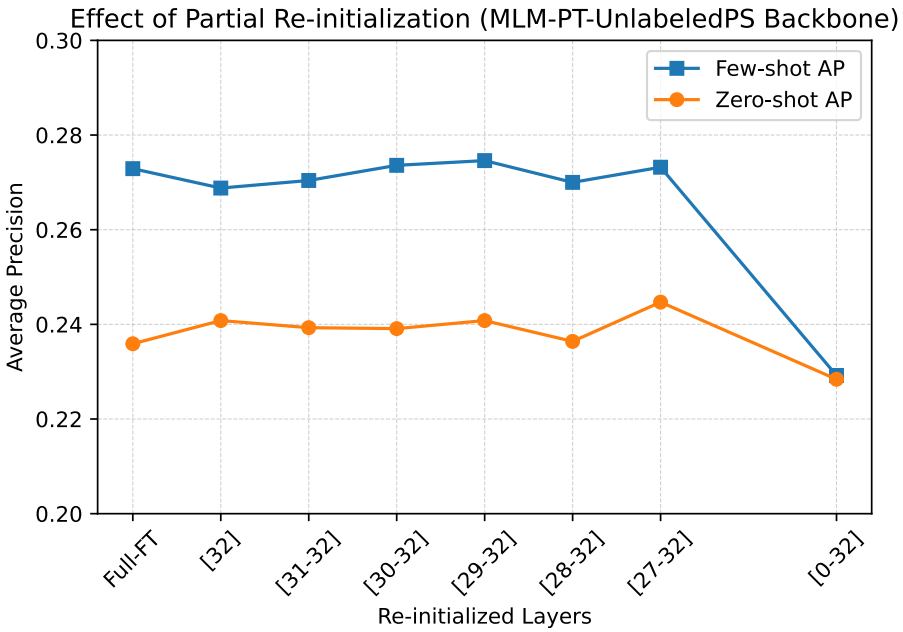


Figure 4.1 Effect of partial re-initialization of top transformer layers using the MLM-PT-UnlabeledPS phosphosite backbone. The x-axis labels indicate which layers (from layer 32 downward) are re-initialized. "[32]" means resetting only layer 32, "[31–32]" means resetting layers 31–32, and so on; "[0–32]" means resetting all layers.

Table 4.9 Effect of partial layer re-initialization in the phosphosite model MLM-PT-DARKINHomologs. We reset the upper layers of the transformer, and the remaining layers are fine-tuned. AP scores are reported for zero-and few-shot evaluation.

Phosphosite Model (ESM-1b)	Model Config	Kinase Model (ESM-1b)	ZSL-AP	FSL-AP
-	All Layers Reinit.	GP	0.2192	0.2208
GP	Full-FT		0.1808	0.2140
-	All Layers Re-init.	MLM-FT-KinaseHomologs	0.2284	0.2292
MLM-PT-DARKINHomologs	Full-FT		0.2404	0.2488
MLM-PT-DARKINHomologs	32-Reinit.	MLM-FT-KinaseHomologs	0.2405	0.2554
	31,32-Reinit.		0.2366	0.2545
	30,31,32-Reinit.		**0.2438**	0.2564
	29,30,31,32-Reinit.		0.2353	0.2568
	28,29,30,31,32-Reinit.		0.2372	0.2557
	27,28,29,30,31,32-Reinit.		0.2422	**0.2610**

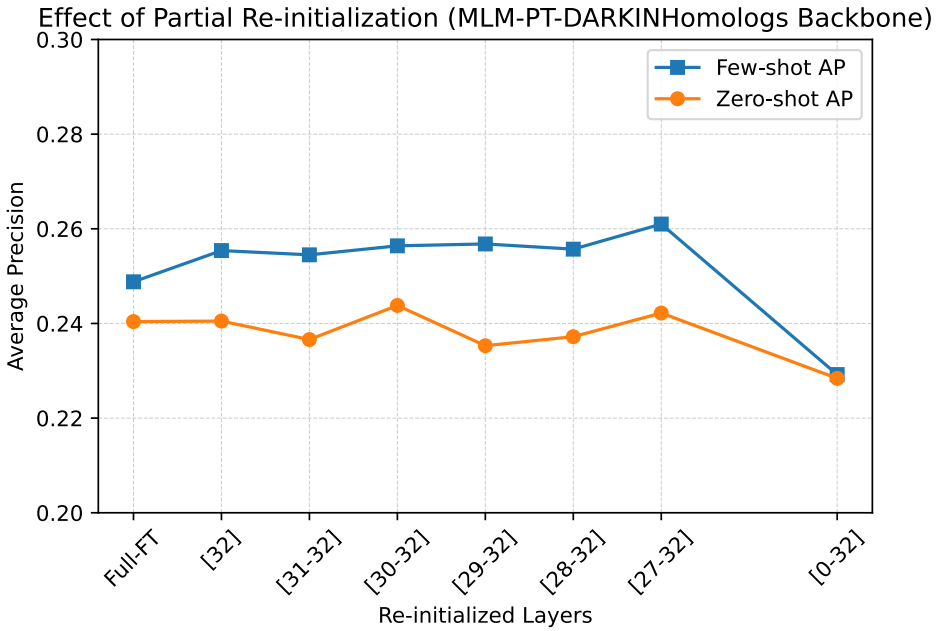


Figure 4.2 Effect of partial layer re-initialization in the phosphosite model MLM-PT-DARKINHomologs. We reset the upper transformer layers and the rest of the layers are fine-tuned. AP scores are reported for zero- and few-shot evaluation.

The effect of re-initialization is modest across the two best-performing phosphosite backbones. Resetting only the final layers may allow phosphosite-aware pLMs to specialize in the high-level features of phosphosites such as kinase-specific recognition motifs, while preserving residue-level bias learned from MLM pre-training.

4.6 Comparison with the Literature

There is only one deep learning based study that handles kinase–phosphosite association prediction as a zero-shot multi-class classification problem, DeepKinZero (DKZ), proposed by our lab (Deznabi et al., 2020). DKZ takes the representations of kinase and 15-residue peptides containing phosphosites as input. During training, it refines the phosphosite representations by employing an LSTM module during the training of BZSM. We also consider the Kinase Library, which focuses on kinase-substrate relationships. This library was developed based on the studies conducted by Johnson et al. (2023); Yaron-Barir et al. (2024). Johnson et al. (2023) and Yaron-Barir et al. (2024) experimentally determined substrate specificities for S/T and Y kinases, respectively. These studies generated Position-Specific Scoring Matrices (PSSMs) reflecting kinase motif preferences. For a given phosphosite sequence, these motif PSSMs allow matching of the phosphosite to its most likely upstream kinase.

To compare our method with the literature, we designed three comparison protocols: i) evaluating task-aware representations on DKZ; ii) comparing our transformer-based pipeline with DKZ; iii) evaluating S/T kinases and Y kinases predictions separately by employing the Kinase Library, DKZ, and our method.

i) Evaluating task-aware representation on DeepKinZero: In DKZ, we used our best-performing task-aware phosphosite and kinase representations (MLM-PT-UnlabeledPS, MLM-PT-DARKINHomologs, and MLM-FT-KinaseHomologs), which resulted in higher AP scores compared to the DKZ experiment leveraging general-purpose ESM-1b representations. Table 4.10 demonstrates the prediction performances.

Table 4.10 Effect of using phosphosite-and kinase-aware representations on DKZ's performance. The DKZ architecture remains unchanged, while the phosphosite and kinase representations are swapped among ESM-1b variants. AP is reported for zero-shot and few-shot evaluation.

Phosphosite Model (ESM-1b)	Kinase Model (ESM-1b)	ZSL-AP	FSL-AP
GP	GP	0.1966	0.2167
MLM-PT-UnlabeledPS	MLM-FT-KinaseHomologs	**0.2015**	**0.2206**
MLM-PT-DARKINHomologs	MLM-FT-KinaseHomologs	**0.2041**	**0.2298**

These results confirmed that phosphosite- and kinase-aware peptide representations provide richer information to the kinase-phosphosite prediction downstream task.

ii) Comparing our transformer-based pipeline with DeepKinZero: Using exactly the same task-aware phosphosite and kinase embeddings (MLM-PT-UnlabeledPS, MLM-PT-DARKINHomologs, and MLM-FT-KinaseHomologs), our

transformer-based prediction method consistently outperformed the original DKZ pipeline. The results are shown in Table 4.11.

Table 4.11 Comparison of our prediction method with DKZ using identical task-aware embeddings (MLM-PT-UnlabeledPS, MLM-PT-DARKINHomologs, and MLM-FT-KinaseHomologs). AP is reported for both zero- and few-shot settings.

Phosphosite Model (ESM-1b)	Model Config	Kinase Model (ESM-1b)	ZSL-AP	FSL-AP
MLM-PT-UnlabeledPS	(DKZ) Full-FT	MLM-FT-KinaseHomologs	0.2015	0.2206
	(Our) Full-FT		**0.2359**	**0.2729**
MLM-PT-DARKINHomologs	(DKZ) Full-FT	MLM-FT-KinaseHomologs	0.2041	0.2298
	(Our) Full-FT		**0.2404**	**0.2488**

These results demonstrated that our transformer–based prediction strategy can benefit from the same embeddings more effectively than an LSTM-based prediction pipeline, which is thanks to the ability of transformers to attend and encode to long-range residue dependencies.

iii) Evaluating S/T kinases and Y kinases predictions separately by employing the Kinase Library, DeepKinZero, and our method: Since the Kinase Library performs the kinase-phosphosite predictions for S/T and Y kinases separately, we partitioned our test set into S/T-only and Y-only subsets. S/T-only subset contains 1046 kinase-phosphosite associations, while the Y-only subset contains 199. We verified that none of the evaluation kinases appear in the original Kinase Library motif catalogue. Then, each subset was used to evaluate: i) Kinase Library, ii) DKZ employing the best-performing task-aware embeddings, and iii) our transformer pipeline leveraging the best-performing task-aware embeddings. The results are reported in Table 4.12. Since we could not provide few-shot samples to the Kinase Library beforehand the evaluation, we did not report the few-shot evaluation for it.

Table 4.12 Comparison of our prediction model with the literature on S/T and Y subsets. AP is reported separately for S/T and Y test sets under zero- and few-shot setups.

Phosphosite Model (ESM-1b)	Model Config	Kinase Model (ESM-1b)	ZSL-AP		FSL-AP	
			S/T	Y	S/T	Y
-	Kinase Library	-	0.1524	0.3294	-	-
MLM-PT-UnlabeledPS	(DKZ) Full-FT	MLM-FT-KinaseHomologs	0.1977	0.3424	0.2182	0.3898
	(Our) Full-FT		**0.2258**	**0.4345**	**0.2578**	**0.4793**
MLM-PT-DARKINHomologs	(DKZ) Full-FT	MLM-FT-KinaseHomologs	0.2005	0.3396	0.2213	0.3853
	(Our) Full-FT		**0.2291**	**0.4222**	**0.2301**	**0.4865**

The results showed that learning-based approaches, DeepKinZero and our method surpassed the motif-based approach Kinase Library on S/T and Y test sets. The motif-based approach depends its decision on phosphosite motif similarity so that

it might not capture the distinctive signals on phosphosite sequences. However, learning-based approaches can optimize the kinase and phosphosite representations and encode broader context, which caused higher prediction performance for S/T and Y kinases.

AP scores on the S/T subset are consistently lower than AP scores on the Y subset. Johnson et al. (2023); Yaron-Barir et al. (2024) supported the intuition that the S/T branch is intrinsically harder because of motif overlap in S/T kinases and sharper specificity in Y kinases. In the study presented by Johnson et al. (2023), more than half of all S/T kinases fall into three broad classes (CAMK, acidophilic, and pro-directed). When many kinases map to the same short motif, the decision boundaries become unsharpened, which causes lower AP scores. In contrast, the profile of Y kinases shows that their intrinsic substrate motifs are more discriminative (Yaron-Barir et al., 2024), which yielded higher AP scores on the Y subset.

In all comparison protocols, our approach reached the highest AP scores. Two main elements supported this success:

- Task-aware peptide embeddings that encode phosphosite- and kinase-specific signals
- Fine-tuning task-aware peptide embeddings on the kinase–phosphosite prediction task via transformer

4.7 Comparison with DARKIN Benchmark

In our previous study, DARKIN, we evaluated zero-shot dark-kinase prediction with frozen pLM embeddings for both kinase and phosphosite sequences. For reference, those baseline results are reproduced in Table 4.13.

The results show the limitation of frozen general-purpose embeddings in low-data settings: for ESM-1b, AP reached 0.1746. When we replaced it with task-aware embeddings obtained through MLM fine-tuning and pre-training, and retrained the phosphosite encoder within our transformer-based pipeline, zero-shot prediction performance rose to 0.2359 for MLM-PT-UnlabeledPS & MLM-FT-KinaseHomologs pLM combination and to 0.2404 for MLM-PT-DARKINHomologs & MLM-FT-KinaseHomologs pLM combination (please see Tables 4.8 and 4.9). Applying the layer-wise re-initialization strategy on the top transformer blocks yielded a further

Table 4.13 The bi-linear zero-shot model performance trained with phosphosite and kinase sequence embeddings-enriched with additional kinase information. The mean macro APs are shown. Of CLS and embedding averaging, only the best-performing model results are listed.

Embedding	Base	+ Family	+ Group	+ EC	+ Family + Group + EC
OneHotEnc	0.0634	0.1107	0.0832	0.0802	0.1098
Blosum62	0.0327	0.0318	0.0310	0.0337	0.0323
NLF	0.0419	0.0391	0.0425	0.0400	0.0426
ProtVec	0.0959	0.1262	0.1129	0.1214	0.1354
ProtBERT (cls)	0.0842	0.1170	0.1077	0.1132	0.1273
ProteinBERT	0.1236	0.1506	0.1215	0.1367	0.1359
ProtT5-XL	0.1552	0.1701	0.1531	0.1674	0.1731
ESM1B (cls)	0.1631	**0.1740**	**0.1688**	**0.1680**	0.1769
ESM1v (cls)	**0.1640**	0.1737	0.1653	0.1652	0.1734
ESM2 (avg)	0.1391	0.1588	0.1453	0.1496	0.1638
DistilProtBERT (cls)	0.1167	0.1360	0.1292	0.1287	0.1441
ProtGPT2	0.1333	0.1476	0.1412	0.1419	0.1557
Ankh-Large	0.0840	0.1417	0.1135	0.1178	0.1594
ProtAlbert (cls)	0.1281	0.1269	0.1276	0.1285	0.1372
SaProt (cls)	0.1292	0.1696	0.1424	0.1434	**0.1800**
TAPE	0.1237	0.1379	0.1333	0.1310	0.1455
ISM2 (cls)	0.1200	0.1275	0.1260	0.1333	0.1374
DPLM (avg)	0.1299	0.1427	0.1318	0.1368	0.1420
AMPLIFY (avg)	0.0896	0.0968	0.0944	0.0969	0.1066
ESM3 (cls)	0.0881	0.1484	0.1220	0.1238	0.1611
ESMC (cls)	0.0866	0.1672	0.1136	0.1401	0.1754
PTM-Mamba (phosphosite)*	0.1218	0.1432	0.1292	0.1346	0.1471

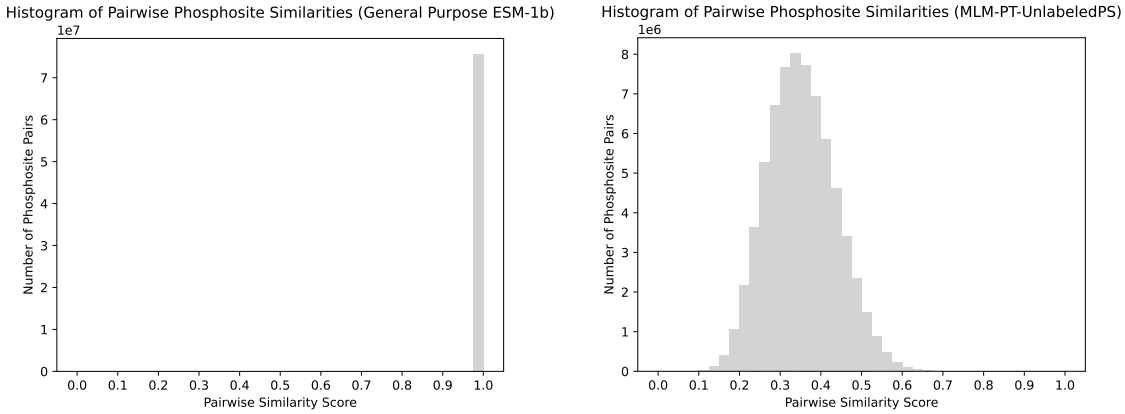
*PTM-Mamba represents kinases with ESM-2 embeddings. Because the architecture has no CLS token and assigns a dedicated special token to the phosphorylated residue, we took the embedding of that residue as the sequence representation.

gain to 0.2447 and 0.2438, respectively. These jumps confirmed: i) task-specific adaptation of pLMs captures biochemical cues that frozen models overlook; ii) selectively re-initializing upper layers can squeeze out additional task-relevant signals even under severe data scarcity.

4.8 Embedding Similarity Analysis

To understand how representation learning impacts the separability of peptide sequences in the embedding space, we computed pairwise cosine-similarity distributions for both phosphosite and kinase embeddings that we used in this thesis. We compared the frozen general-purpose ESM-1b model against the best-performing task-aware models we obtained; MLM-PT-UnlabeledPS and MLM-FT-KinaseHomologs.

For phosphosite sequences, in the general-purpose ESM-1b model, almost all phosphosite pairs cluster in a range of similarity score of [0.9-1.0], indicating that the model failed to capture motif-level diversity (Figure 4.3-a). After task-aware fine-tuning, the distribution widened and shifted leftwards, revealing a greater spread of similarities and thus improved discriminability among sites (Figure 4.3-b).

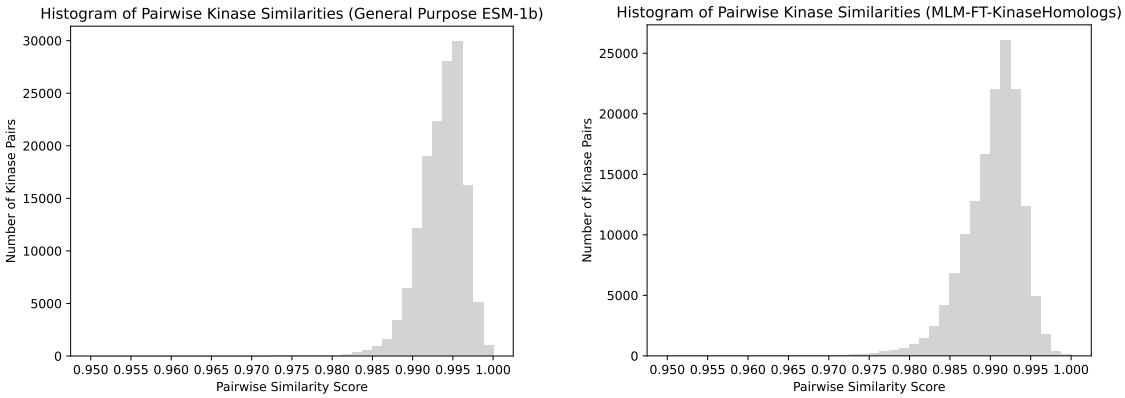


(a) General-purpose ESM-1b, average pairwise similarity is $\mu = 0.996$

(b) MLM-PT-UnlabeledPS, average pairwise similarity is $\mu = 0.353$

Figure 4.3 Pairwise cosine-similarity histograms for phosphosite sequence embeddings before and after task adaptation.

For kinase domain sequences, the embeddings exhibit a narrower change, but the mode shifts slightly to lower similarities with a longer left tail, which is shown in Figure 4.4-a and 4.4-b. This subtle shift suggests that task-aware training introduced finer-grained distinctions between kinases while preserving coarse-grained similarity.



(a) General-purpose ESM-1b, average pairwise similarity is $\mu = 0.994$

(b) MLM-FT-KinaseHomologs, average pairwise similarity is $\mu = 0.990$

Figure 4.4 Pairwise cosine-similarity histograms for kinase domain sequence embeddings before and after task adaptation.

These findings confirmed that the task-aware models encode more informative features, and correlated with the performance gains reported in Sections 4.2 and 4.3,

where the task-aware models yielded the highest AP scores.

4.9 Additional Evaluations on the Best Setup

To provide a complementary view, we evaluated our best-performing setup by employing additional metrics. Thereby, when the prediction model misses the exact kinase, considering the correct family, group, or cluster can still deliver biologically meaningful insights, especially where fine-scale annotations are uncertain. These metrics are:

Top@k Accuracy: As a phosphosite-focused counterpart to the kinase-centric macro-AP, we also report Top@k Accuracy for $k = 1, 3$, and 5. A phosphosite was scored as correct whenever at least one of its ground-truth kinases appears within the first k positions of the model's ranked list. Because a phosphosite might have multiple valid kinases, this measure fits the multi-label setting of our task and indicates how far a user typically needs to scan down the list to encounter a correct prediction.

Phosphosite Average Precision (Ph-AP): Phosphosite AP is calculated in a mirror-image fashion to macro-AP, detailed in Section 3.3. For each phosphosite, we sorted all candidate kinases according to their predicted scores, computed the AP for that ranking, and then took the mean of these AP values across phosphosites. This reflects how well the model orders kinases for a single phosphosite—essentially mimicking the query experience of an end-user.

Attribute-level Metrics (Family-AP, Group-AP, and F-Grain AP): Because kinases are organized into broader functional and evolutionary categories, it is often useful to assess predictions at those coarser levels. Accordingly, we pooled per-phosphosite scores at three hierarchical levels—*Family*, *Group*, and a *Fine-grained Cluster* level (*F-Grain*)—and reported both AP scores for each. After pooling, we applied the same ranking-based formulas used for macro-AP to these attribute labels.

The *Fine-grained Clusters* add an extra level below families and groups, defined by phylogenetic closeness. Starting from the kinase phylogenetic tree of KinBase (2024), we transformed branch lengths into pairwise similarity scores, normalized

the scores, and then clustered the kinases based on the similarity scores so that only highly related kinases shared a cluster. Because each cluster held just a handful of kinases, only very close relatives were grouped.

The results are shown in Table 4.14 and 4.15 for zero-shot and few-shot evaluations, respectively.

Table 4.14 Zero-shot prediction results on our transformer-based prediction model. In the general-purpose (GP) setup, both kinase and phosphosite representations are obtained by task-agnostic ESM-1b. In the "Best" setup, phosphosite representations are obtained from the best-performing phosphosite-aware model, MLM-PT-UnlabeledPS, and kinase representations are obtained from the best-performing kinase-aware model, MLM-FT-KinaseHomologs. The phosphosite encoder is fully fine-tuned, and the kinase encoder is frozen.

Setup	AP	Top@1 Acc	Top@3 Acc	Top@5 Acc	Phosphosite AP	Family AP	Group AP	F.Grain AP
GP	0.1808	0.1356	0.3588	0.5386	0.3025	0.2334	0.4377	0.2008
Best	0.2359	0.1601	0.3588	0.4614	0.3118	0.3113	0.4728	0.2712

Table 4.15 Few-shot performance of our transformer-based model under two evaluation setups. General-purpose (GP) setup uses task-agnostic ESM-1b embeddings for both kinases and phosphosites. The best setup uses the best-performing task-aware embedding combination (MLM-PT-UnlabeledPS and MLM-FT-KinaseHomologs). In both setups, phosphosite encoders are fully fine-tuned and kinase encoders are frozen.

Setup	AP	Top@1 Acc	Top@3 Acc	Top@5 Acc	Phosphosite AP	Family AP	Group AP	F-Grain AP
GP	0.2140	0.1979	0.4117	0.5513	0.3573	0.2867	0.4382	0.2341
Best	0.2729	0.2129	0.4298	0.5536	0.3669	0.3582	0.5117	0.3124

In zero- and few-shot experiments, the results show that Top@1 improves while Top@3 remains unchanged or improves, and Top@5 declines or slightly improves, indicating that early ranks become more precise (Top@1) while fewer alternative true kinases remain inside the top five (Top@5). Phosphosite-centric AP (Ph-AP) increases modestly, consistent with the limited kinase label space per site; however, the large gains at family/group/cluster levels show that even when the exact kinase is missed, the model is more often "locally correct" within the evolutionary hierarchy, providing still-actionable biological guidance.

Overall, the task-aware representations deliver both higher early precision and richer hierarchical signal beyond what is obtainable from the frozen general-purpose encoder.

5. CONCLUSION & FUTURE WORK

In this thesis, we proposed a novel approach for kinase-phosphosite association prediction under zero-shot and few-shot settings. We leveraged several task-aware pLMs. Our methodology addresses the key limitations in existing computational approaches, particularly those related to effectively capturing kinase-specific and phosphosite-specific contextual features, and improves predictions under low-data constraints.

Firstly, we demonstrated the utility of specialized pre-training and fine-tuning strategies in producing task-aware representations. Our experiments demonstrated that adapting the general-purpose ESM-1b model through the MLM objective on phosphosite- and kinase-specific datasets yielded superior embedding quality. This context-awareness significantly improved prediction performance for zero-shot and few-shot kinase-phosphosite predictions.

Secondly, our work revealed the advantages of using transformer-based architectures over LSTM-based and motif-based models. The inherent capacity of transformers to model long-range dependencies and contextual cues within protein sequences enabled a significant improvement in predictive performance. Furthermore, we introduced a layer-selective re-initialization strategy, which further enhanced the adaptability of transformers to task-specific nuances. Through this strategy, transformers showed modest but promising improvements.

Comparisons with the literature show the effectiveness of our approach. Specifically, our task-aware transformer models achieved higher predictive performance than experiment-based approaches (Kinase Library) and the LSTM-based model (DeepKinZero).

These advancements offer promising directions for future research:

- Further improvements might be achieved for protein representations by incorporating structural annotations directly into pLMs, extending beyond sequence-based context.

- Distinct few-shot learning techniques, such as meta-learning, prototype learning, might further enhance the prediction performance for kinases with very limited associated phosphosites.
- Future work may also focus on the integration of our computational predictions with experimental validations, which may establish the utility and accuracy of our approach in biological contexts.

Overall, this thesis demonstrates the significant potential of task-aware peptide representations to enhance kinase-phosphosite prediction, laying a foundation for further innovations in elucidating the dark phosphoproteome.

BIBLIOGRAPHY

Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. (2015a). Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438.

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.

Akata, Z., Reed, S., Walter, D., Lee, H., and Schiele, B. (2015b). Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2927–2936.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402.

Aman, J. M., Zhu, A. W., Wühr, M., Shvartsman, S. Y., and Singh, M. (2025). Kinaid: an orthology-based kinase-substrate prediction and analysis tool for phosphoproteomics. *Bioinformatics*, page btaf300.

Asgari, E. and Mofrad, M. R. K. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLOS ONE*, 10(11):e0141287.

Bairoch, A. (2000). The enzyme database in 2000. *Nucleic acids research*, 28(1):304–305.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L.-S. L. (2005). The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 33(suppl_1):D154–D159.

Blom, N., Gammeltoft, S., and Brunak, S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of molecular biology*, 294(5):1351–1362.

Blume-Jensen, P. and Hunter, T. (2001). Oncogenic kinase signalling. *Nature*, 411(6835):355.

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. (2022). ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110.

Chen, M., Zhang, W., Gou, Y., Xu, D., Wei, Y., Liu, D., Han, C., Huang, X., Li, C., Ning, W., et al. (2023). Gps 6.0: an updated server for prediction of kinase-specific phosphorylation sites in proteins. *Nucleic acids research*, 51(W1):W243–W250.

Chen, Z., Zhao, P., Li, C., Li, F., Xiang, D., Chen, Y.-Z., Akutsu, T., Daly, R. J., Webb, G. I., Zhao, Q., et al. (2021). ilearnplus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic acids research*, 49(10):e60–e60.

Cohen, P. (2002). The origins of protein phosphorylation. *Nature Cell Biology*, 4(5):E127–E130.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20:273–297.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Deznabi, I., Arabaci, B., Koyutürk, M., and Tastan, O. (2020). Deepkinzero: zero-shot learning for predicting kinase–phosphosite associations involving understudied kinases. *Bioinformatics*, 36(12):3652–3661.

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.

Dunker, A. K., Romero, P., Obradovic, Z., Garner, E. C., and Brown, C. J. (2000). Intrinsic protein disorder in complete genomes. *Genome informatics*, 11:161–171.

Elnaggar, A., Essam, H., Salah-Eldin, W., Moustafa, W., Elkerdawy, M., Rochereau, C., and Rost, B. (2023). Ankh: Optimized protein language model unlocks general-purpose modelling.

Elnaggar et al., A. (2021). Prottrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

ESM Team (2024). Esm cambrian: Revealing the mysteries of proteins with unsupervised learning.

Esmaili, F., Qin, Y., Wang, D., and Xu, D. (2025). Kinase-substrate prediction using an autoregressive model. *Computational and Structural Biotechnology Journal*, 27:1103–1111.

Farrell, D. (2021). epitopepredict: A tool for integrated mhc binding prediction. *bioRxiv*.

Ferguson, F. M. and Gray, N. S. (2018). Kinase inhibitors: the road ahead. *Nature reviews Drug discovery*, 17(5):353–377.

Ferruz, N., Schmidt, S., and Höcker, B. (2022). Protgpt2 is a deep unsupervised language model for protein design. *Nature Communications*, 13.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *CoRR*, abs/1703.03400.

Fournier, Q., Vernon, R. M., van der Sloot, A., Schulz, B., Chandar, S., and Langmead, C. J. (2024). Protein language models: Is scaling necessary? *bioRxiv*.

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*, pages 2121–2129.

Gaestel, M., Kotlyarov, A., and Kracht, M. (2009). Targeting innate immunity protein kinase signalling in inflammation. *Nature Reviews Drug Discovery*, 8(6):480.

Gao, J., Thelen, J. J., Dunker, A. K., and Xu, D. (2010). Musite: a tool for global prediction of general and kinase-specific phosphorylation sites. *Molecular & Cellular Proteomics*, pages mcp–M110.

Garnier, J., Osguthorpe, D. J., and Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of molecular biology*, 120(1):97–120.

Geffen, Y., Ofran, Y., and Unger, R. (2022). Distilprotbert: a distilled protein language model used to distinguish between real proteins and their randomly shuffled counterparts. *Bioinformatics*, 38(Supplement_2):ii95–ii98.

Glaser, M. and Braegelmann, J. (2025). Esm-effect: An effective and efficient fine-tuning framework towards accurate prediction of mutation's functional effect. *bioRxiv*, pages 2025–02.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences*, 84(13):4355–4358.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., Badkundri, R., Shafkat, I., Gong, J., Derry, A., Molina, R. S., Thomas, N., Khan, Y., Mishra, C., Kim, C., Bartie, L. J., Nemeth, M., Hsu, P. D., Sercu, T., Candido, S., and Rives, A. (2024). Simulating 500 million years of evolution with a language model. *bioRxiv*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Heineke, J. and Molkentin, J. D. (2006). Regulation of cardiac hypertrophy by intracellular signalling pathways. *Nature reviews Molecular cell biology*, 7(8):589–600.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

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.

Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2012). Phosphositeplus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic acids research*, 40(D1):D261–D270.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. (2022). Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pages 8946–8970. PMLR.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Hu, Y., Flockhart, I., Vinayagam, A., Bergwitz, C., Berger, B., Perrimon, N., and Mohr, S. E. (2011). An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC bioinformatics*, 12(1):357.

Hunter, T. (1995). Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell*, 80(2):225–236.

Jiang, W., Jaehnig, E. J., Liao, Y., Shi, Z., Yaron-Barir, T. M., Johnson, J. L., Cantley, L. C., and Zhang, B. (2025). Deciphering the dark cancer phosphoproteome using machine-learned co-regulation of phosphosites. *Nature Communications*, 16(1):2766.

Johnson, J. L., Yaron, T. M., Huntsman, E. M., Kerelsky, A., Song, J., Regev, A., Lin, T.-Y., Liberatore, K., Cizin, D. M., Cohen, B. M., et al. (2023). An atlas of substrate specificities for the human serine/threonine kinome. *Nature*, 613(7945):759–766.

Jumper, J. M., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D. A., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

KinBase (2024). http://kinase.com/human/kinome/phylogeny.html. [Online; accessed 1-November-2024].

Klempner, S. J., Myers, A. P., and Cantley, L. C. (2013). What a tangled web we weave: emerging resistance mechanisms to inhibition of the phosphoinositide 3-kinase pathway. *Cancer discovery*, 3(12):1345–1354.

Koch, G., Zemel, R., Salakhutdinov, R., et al. (2015). Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, pages 1–30. Lille.

Kodirov, E., Xiang, T., and Gong, S. (2017). Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3174–3183.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Kuru, N., Dereli, O., Akkoyun, E., Bircan, A., Tastan, O., and Adebali, O. (2022). Phact: Phylogeny-aware computing of tolerance for missense mutations. *Molecular Biology and Evolution*, 39(6):msac114.

Lampert, C. H., Nickisch, H., and Harmeling, S. (2013). Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):453–465.

Larochelle, H., Erhan, D., and Bengio, Y. (2008). Zero-data learning of new tasks. In *AAAI*, volume 1, page 3.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

Lee, B. and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology*, 55(3):379–IN4.

Li, T., Du, P., and Xu, N. (2010). Identifying human kinase-specific protein phosphorylation sites by integrating heterogeneous information from various sources. *PloS one*, 5(11):e15411.

Li, T., Li, F., and Zhang, X. (2008). Prediction of kinase-specific phosphorylation sites with sequence features by a log-odds ratio approach. *Proteins: Structure, Function, and Bioinformatics*, 70(2):404–414.

Li, W., Dong, C., Tian, P., Qin, T., Yang, X., Wang, Z., Huo, J., Shi, Y., Wang, L., Gao, Y., and Luo, J. (2021). Libfewshot: A comprehensive library for few-shot learning. *CoRR*, abs/2109.04898.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

Linding, R., Jensen, L. J., Ostheimer, G. J., van Vugt, M. A., Jørgensen, C., Miron, I. M., Diella, F., Colwill, K., Taylor, L., Elder, K., et al. (2007). Systematic discovery of in vivo phosphorylation networks. *Cell*, 129(7):1415–1426.

Luo, F., Wang, M., Liu, Y., Zhao, X.-M., and Li, A. (2019). Deepphos: prediction of protein phosphorylation sites with deep learning. *Bioinformatics*, 35(16):2766–2773.

Ma, H., Li, G., and Su, Z. (2020). Ksp: An integrated method for predicting catalyzing kinases of phosphorylation sites in proteins. *BMC genomics*, 21:1–10.

Ma, R., Li, S., Parisi, L., Li, W., Huang, H.-D., and Lee, T.-Y. (2023). Holistic similarity-based prediction of phosphorylation sites for understudied kinases. *Briefings in Bioinformatics*, 24(2):bbac624.

Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303.

Mering, C. v., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). String: a database of predicted functional associations between proteins. *Nucleic acids research*, 31(1):258–261.

Moret, N., Liu, C., Gyori, B. M., Bachman, J. A., Steppi, A., Hug, C., Taujale, R., Huang, L.-C., Berginski, M. E., Gomez, S. M., et al. (2020). A resource for exploring the understudied human kinome for research and therapeutic opportunities. *BioRxiv*, pages 2020–04.

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):818–821.

Nanni, L. and Lumini, A. (2011). A new encoding technique for peptide classification. *Expert Systems with Applications*, 38(4):3185–3191.

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.

Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N., and Madani, A. (2023). Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978.

O'Reilly, K. E., Rojo, F., She, Q.-B., Solit, D., Mills, G. B., Smith, D., Lane, H., Hofmann, F., Hicklin, D. J., Ludwig, D. L., et al. (2006). mtor inhibition induces upstream receptor tyrosine kinase signaling and activates akt. *Cancer research*, 66(3):1500–1508.

Oreshkin, B. N., López, P. R., and Lacoste, A. (2018). TADAM: task dependent adaptive metric for improved few-shot learning. *CoRR*, abs/1805.10123.

Ouyang-Zhang, J., Gong, C., Zhao, Y., Krähenbühl, P., Klivans, A. R., and Diaz, D. J. (2024). Distilling structural representations into protein sequence models. *bioRxiv*.

Palatucci, M., Pomerleau, D., Hinton, G. E., and Mitchell, T. M. (2009). Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, pages 1410–1418.

Pan, S. J. (2020). Transfer learning. *Learning*, 21:1–2.

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.

Patrick, R., Lê Cao, K.-A., Kobe, B., and Bodén, M. (2015). Phosphopick: modelling cellular context to map kinase-substrate phosphorylation events. *Bioinformatics*, 31(3):382–389.

Pawson, T. and Scott, J. D. (2005). Protein phosphorylation in signaling–50 years and counting. *Trends in biochemical sciences*, 30(6):286–290.

Peng, F. Z., Wang, C., Chen, T., Schussheim, B., Vincoff, S., and Chatterjee, P. (2025). Ptm-mamba: a ptm-aware protein language model with bidirectional gated mamba blocks. *Nature Methods*.

Pesis, K. H., Wei, Y., Lewis, M., and Matthews, H. R. (1988). Phosphohistidine is found in basic nuclear proteins of physarum polycephalum. *FEBS letters*, 239(1):151–154.

Qi, H., Brown, M., and Lowe, D. G. (2018). Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5822–5830.

Qiao, S., Liu, C., Shen, W., and Yuille, A. L. (2018). Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7229–7238.

Qin, G.-M., Li, R.-Y., and Zhao, X.-M. (2017). Phosd: inferring kinase–substrate interactions based on protein domains. *Bioinformatics*, 33(8):1197–1204.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., and Song, Y. S. (2019). Evaluating protein transfer learning with tape.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118.

Romera-Paredes, B. and Torr, P. (2015). An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Roskoski Jr, R. (2022). Properties of fda-approved small molecule protein kinase inhibitors: A 2022 update. *Pharmacological research*, 175:106037.

Ruder, S., Peters, M. E., Swayamdipta, S., and Wolf, T. (2019). Transfer learning in natural language processing. In Sarkar, A. and Strube, M., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Saunders, N. F., Brinkworth, R. I., Huber, T., Kemp, B. E., and Kobe, B. (2008). Predikin and predikindb: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites. *BMC bioinformatics*, 9(1):245.

Schmirler, R., Heinzinger, M., and Rost, B. (2024). Fine-tuning protein language models boosts predictions across diverse tasks. *Nature Communications*, 15(1):7407.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Snell, J., Swersky, K., and Zemel, R. S. (2017). Prototypical networks for few-shot learning. *CoRR*, abs/1703.05175.

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.

Steger, M., Tonelli, F., Ito, G., Davies, P., Trost, M., Vetter, M., Wachter, S., Lorentzen, E., Duddy, G., Wilson, S., et al. (2016). Phosphoproteomics reveals that parkinson's disease kinase lrrk2 regulates a subset of rab gtpases. *elife*, 5:e12813.

Su, J., Han, C., Zhou, Y., Shan, J., Zhou, X., and Yuan, F. (2023). Saprot: Protein language modeling with structure-aware vocabulary. *bioRxiv*.

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.

Sunar, E. A., Işık, Z., Pekey, M., Cinbis, R. G., and Tastan, O. (2024). Darkin: A zero-shot classification benchmark and an evaluation of protein language models. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*.

Sweet, R. M. and Eisenberg, D. (1983). Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *Journal of molecular biology*, 171(4):479–488.

Unsal, S., Atas, H., Albayrak, M., Turhan, K., Acar, A. C., and Doğan, T. (2022). Learning functional properties of proteins with language models. *Nature Machine Intelligence*, 4(3):227–245.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Wang, C., Xu, H., Lin, S., Deng, W., Zhou, J., Zhang, Y., Shi, Y., Peng, D., and Xue, Y. (2020). Gps 5.0: an update on the prediction of kinase-specific phosphorylation sites in proteins. *Genomics, proteomics & bioinformatics*, 18(1):72–80.

Wang, D., Zeng, S., Xu, C., Qiu, W., Liang, Y., Joshi, T., and Xu, D. (2017a). Musitedeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics*, 33(24):3909–3916.

Wang, M., Wang, T., Wang, B., Liu, Y., and Li, A. (2017b). A novel phosphorylation site-kinase network-based method for the accurate prediction of kinase-substrate relationships. *BioMed research international*, 2017(1):1826496.

Wang, M., Wang, T., Wang, B., Liu, Y., and Li, A. (2017c). A novel phosphorylation site-kinase network-based method for the accurate prediction of kinase-substrate relationships. *BioMed research international*, 2017.

Wang, X., Zhang, Z., Zhang, C., Meng, X., Shi, X., and Qu, P. (2022). Transphos: A deep-learning model for general phosphorylation site prediction based on transformer-encoder architecture. *International Journal of Molecular Sciences*, 23(8):4263.

Wang, X., Zheng, Z., Ye, F., Xue, D., Huang, S., and Gu, Q. (2024). Diffusion language models are versatile protein learners. In *International Conference on Machine Learning*.

Wang, Y., Shi, M., Chung, K. A., Zabetian, C. P., Leverenz, J. B., Berg, D., Srulijes, K., Trojanowski, J. Q., Lee, V. M.-Y., Siderowf, A. D., et al. (2012). Phosphorylated α-synuclein in parkinson's disease. *Science translational medicine*, 4(121):121ra20–121ra20.

Wong, Y.-H., Lee, T.-Y., Liang, H.-K., Huang, C.-M., Wang, T.-Y., Yang, Y.-H., Chu, C.-H., Huang, H.-D., Ko, M.-T., and Hwang, J.-K. (2007). Kinasephos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic acids research*, 35(suppl_2):W588–W594.

Wu, T. D. and Brutlag, D. L. (1995). Identification of protein motifs using conserved amino acid properties and partitioning techniques. In *ISMB*, pages 402–410.

Xian, Y., Schiele, B., and Akata, Z. (2017). Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4582–4591.

Xue, Y., Ren, J., Gao, X., Jin, C., Wen, L., and Yao, X. (2008). Gps 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Molecular & cellular proteomics*, 7(9):1598–1608.

Yaron-Barir, T. M., Joughin, B. A., Huntsman, E. M., Kerelsky, A., Cizin, D. M., Cohen, B. M., Regev, A., Song, J., Vasan, N., Lin, T.-Y., et al. (2024). The intrinsic substrate specificity of the human tyrosine kinome. *Nature*, 629(8014):1174–1181.

Zaidi, S., Berariu, T., Kim, H., Bornschein, J., Clopath, C., Teh, Y. W., and Pascanu, R. (2023). When does re-initialization work? In *Proceedings on*, pages 12–26. PMLR.

Zhou, Z., Yeung, W., Gravel, N., Salcedo, M., Soleymani, S., Li, S., and Kannan, N. (2023). Phosformer: an explainable transformer model for protein kinase-specific phosphorylation predictions. *Bioinformatics*, 39(2):btad046.

Zhou, Z., Yeung, W., Soleymani, S., Gravel, N., Salcedo, M., Li, S., and Kannan, N. (2024). Using explainable machine learning to uncover the kinase–substrate interaction landscape. *Bioinformatics*, 40(2):btae033.

Zou, L., Wang, M., Shen, Y., Liao, J., Li, A., and Wang, M. (2013). Pkis: computational identification of protein kinases for experimentally discovered protein phosphorylation sites. *BMC bioinformatics*, 14(1):247.