DARKIN: A ZERO-SHOT CLASSIFICATION BENCHMARK AND AN EVALUATION OF PROTEIN LANGUAGE MODELS

Emine Ayşe Sunar[1], **Zeynep Işık** [1,†], **Mert Pekey** [1,†], **Ramazan Gokberk Cinbis**[2,*], **Oznur Tastan**[1,*]
[1]Sabanci University, Department of Computer Science and Engineering, İstanbul, Türkiye
[2]Middle East Technical University, Department of Computer Engineering, Ankara, Türkiye
{ayse.sunar,zeynep.isik1,mpekey,otastan}@sabanciuniv.edu
gcinbis@ceng.metu.edu.tr

ABSTRACT

Protein language models (pLMs) aim to capture the complex information embedded within protein sequences and are useful for downstream protein prediction tasks. With many pLMs available, there is a critical need to benchmark their performance across diverse tasks. Here, we introduce a biologically relevant zero-shot prediction benchmark, focusing on dark kinase-phosphosite associations. Kinases are the enzymes responsible for protein phosphorylation, and they play vital roles in cellular signaling. While phosphoproteomics allows large-scale identification of phosphosites, determining the catalyzing kinase remains challenging. We present a zero-shot classification benchmark dataset, DARKIN, for assigning phosphosites to one of the understudied kinases (dark kinases). DARKIN provides training, validation, and test folds that are split based on zero-shot classification, kinase groups, and sequence similarities. Evaluation of pLMs using a novel training-free k-NN-based zero-shot classifier and a bilinear zero-shot classifier reveals superior performance by Esm models, ProtT5-XL, and the recently introduced structure-based SaProt model. We believe this biologically relevant yet challenging benchmark will further facilitate assessing the efficacy of pLMs and aid the exploration of dark kinases.

1 INTRODUCTION

Large language models (LLMs) have demonstrated remarkable capabilities for generative sequence modeling, serving as a mechanism to acquire semantic representations from large datasets. Building on the success of LLMs in natural language processing, protein language models (pLMs) have been developed to capture the complex information embedded within protein sequences (Rao et al., 2019; Elnaggar et al., 2021; Rives et al., 2020; Meier et al., 2021; Lin et al., 2022; Brandes et al., 2022; Ferruz et al., 2022; Geffen et al., 2022; Elnaggar et al., 2023; Su et al., 2023; Zhang & Okumura, 2024). These models present semantic representations of protein sequences to be used in prediction tasks where protein sequences are the input. However, with the growing number of pLMs, benchmarking their performance becomes essential to assess their reliability and applicability across various biological contexts. Previous work has compared the pLMs in their ability to predict proteins' functional properties (Unsal et al., 2022; Schmirler et al., 2023; Zhang & Okumura, 2024), and functional motifs (Savojardo et al., 2023). In this work, we provide a novel biologically relevant zero-shot prediction benchmark on phosphosite-dark kinase associations.

Kinases are the enzymes that catalyze the phosphorylation of other proteins in a target-specific manner (Hunter, 1995). Phosphorylation involves the transfer of a phosphate from adenosine 5'-triphosphate (ATP) to amino acid residues (Cohen, 2002). The amino acid residue that receives the phosphoryl group on the target protein (substrate) is called the *phosphosite*. Phosphorylation is a key regulator of protein function in signal transduction pathways, and their dysfunction is associated with numerous diseases (Blume-Jensen & Hunter, 2001; Müller et al., 2015). For this reason, kinases are major drug targets in diseases such as cancer, infectious diseases, and neurological disorders (Blume-Jensen & Hunter, 2001).

[*]Corresponding Authors.
[†]These authors contributed equally to this work.

Although the advances in phosphoproteomics enable the identification of phosphosites at the proteome level, determining which kinase is responsible for phosphorylating a site remains an experimental challenge. On one hand, more than 95% of reported human phosphosites have no known cognate kinase (Needham et al., 2019). On the other hand, 25% of the kinases are yet to be assigned to a phosphorylation event, and for about 35% of the kinases, there are 1-10 phosphosites reported. Thus, most of the phosphoproteome and the kinome are in the dark (Needham et al., 2019). Associating orphan phosphosites to their cognate kinases is an important task that would aid the understanding of the biological function of these phosphorylation events and help discover new drug targets (Berginski et al., 2021; Deznabi et al., 2020; Needham et al., 2019).

We can summarize the contribution of this work as follows: (i) We propose a benchmark dataset for dark kinase-phosphosite association prediction. (ii) We present a strategy to split the dataset into train, validation, and test splits for zero-shot multi-class prediction tasks, paying attention to the stratification of the kinase-phosphosite pairs by the kinase groups, the number of examples per kinase and the sequence similarity of the kinases. (iii) We introduce a new training-free, k-NN-based zero-shot classification method that allows assessing the pLMs. (iv) We evaluate and compare the protein language models using two different zero-shot classification approaches.

2 METHODS

2.1 PROBLEM DESCRIPTION

$\mathcal{X}$ denotes the space of phopshosite sequences and $\mathcal{Y}$ denotes the set of all human kinases. The task of kinase-phosphosite association prediction involves identifying the kinase $y \in \mathcal{Y}$ most likely to catalyze the phosphorylation of a given phosphosite sequence $x \in \mathcal{X}$.

Multiple kinases can phosphorylate a phosphosite, we frame the problem as a multilabel classification task. We denote training kinases as $\mathcal{Y}_{tr} \subset \mathcal{Y}$ and test kinases 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 data, $D_{tr} = (x_i, y_i), i = 1, \ldots, N_{tr}$, consists of pairings of train kinases with their associated phosphosites, where $y_i \in \mathcal{Y}_{tr}$. Similarly, the test data contains phosphosite pairings of the test kinases $\mathcal{Y}_{te}$.

2.2 DATASET CURATION AND PROCESSING

Several publicly available human kinase lists are available (Manning et al., 2002; 2023; Eid et al., 2017; UniProt Consortium, 2023a; Moret et al., 2020), and they partially overlap due to ambiguity on kinase domain activities. We resort to an up-to-date list by Moret et al. (2020). This curated kinase contains a dataset of 557 human kinases with at least one kinase domain.

We obtained experimentally validated kinase-phosphosite associations from the PhosphoSitePlus database (downloaded in May 2023). The phosphosites are represented as 15-residue amino acid sequences, with seven neighboring residues of the phosphosite on either side and the phosphosite itself in the middle. Padding was only necessary when the phosphosite was close to the terminal end of the protein to ensure that the phosphosite was in the middle. Kinase-phosphosite associations related to non-human kinases were removed. We did not apply the same restriction to the substrates since the model organism substrates are used to probe the interactions. We removed kinase isoforms and fusion kinases and used the canonical form specified in the UniProt human proteome (Downloaded May 2023).

We retrieved protein sequences from UniProt (UniProt Consortium, 2023b) (Accessed Dec 2023). If the substrate could not be mapped to a Uniprot ID unambiguously, we removed all phosphosite-kinase associations of these substrates. We retrieved the kinase domain sequences using the domain indices from Moret et al. (2020). We obtained the kinase family and group information from Moret et al. (2020). Missing group and family information were imputed according to their similarity with other kinases. We defined a kinase group *Other2* and family *otherFamily* for kinases that could not be imputed into a family or group due to their dissimilarity to the rest of the groups. We downloaded Enzyme Commision (EC) numbers of the kinases (downloaded Jul 2023) (Bairoch, 2000). We obtained protein structure data from the AlphaFold Protein Structure Database via AlphaFoldAPI (DeepMind and European Bioinformatics Institute, 2023) and PDBe (Protein Data Bank in Europe, 2023). For the isoform proteins whose structures are not available in AlphaFold and PDBe, we employed ColabFold to obtain predicted 3D structures (Mirdita et al., 2022).

Table 1: The protein Language Models (pLMs) compared in this study.

PLM	Dataset	Vector Size	Model Size	Representation	Objective	Paper
TAPE	PFAM	768	38M	Sequence	Sequence-based,Structural Feature Prediction	Rao et al., 2019
ProtBERT	BFD100, UniRef100	1024	420M	Sequence	Sequence-based, Structural, Physicochemical Feature Prediction	Elnaggar et al., 2021
ProtALBERT	UniRef100	4096	224M	Sequence		Elnaggar et al., 2021
ProtT5-XL	BFD100, UniRef50	1024	3B	Sequence		Elnaggar et al., 2021
Esm1B	UniRef50	1280	650M	Sequence	Structural, Physicochemical Feature Prediction	Rives et al., 2020
Esm1v	UniRef90	1280	650M	Sequence	Sequence Variant Prediction	Meier et al., 2021
Esm2	UniRef50	1280	650M	Sequence	Structural Feature,Contact Prediction	Lin et al., 2022
ProteinBERT	UniRef90	1562	16M	Sequence	Sequence-based Feature, GO Annotation Prediction	Brandes et al., 2022
ProtGPT2	UniRef50	1280	738M	Subword	Protein Design and Engineering	Ferruz et al., 2022
DistilProtBERT	UniRef50	1024	230M	Sequence	Sequence-based, Structural, Physicochemical Feature Prediction	Geffen et al., 2022
Ankh	UniRef50	1536	1.5B	Sequence	General Purpose Modeling	Elnaggar et al., 2023
SaProt	AlphaFold2, PDB	1280	650M	Sequence, Structure	Structure-Aware Feature, Mutation Effect Prediction	Su et al., 2023

2.3 Evaluated Protein Language Models and Baseline Encodings

We selected pLMs whose models were accessible, reported to perform well in the literature, and are recent. Table 1 presents the pLMs we experimented with and their properties. Processing large dimensions of protein embeddings sets a challenge. For more efficient processing, we computed the column-wise average of the embedding for all pLMs, excluding the padding (PAD) token vectors. Additionally, for pLMs with the classification (CLS) token, we used the vectors corresponding to this token as an embedding summary.

In addition to the pLMs, we used these baseline representations: (i) **One-hot encoding:** The input sequence is expressed as a binary vector of amino acids. (ii) **BLOSUM62 encoding:** The encoding utilizes the row corresponding to a particular amino acid in the BLOSUM62 matrix, representing the likelihood of substitution of this amino acid with every amino acid. (iii) **NLF encoding:** NLF reflects amino acids' physicochemical properties using a non-linear Fisher's transform (Nanni & Lumini, 2011). Representations are calculated using the Epitopepredict tool (Farrell, 2021). (iv) **ProtVec:** ProtVec is a skip-gram neural network model trained to provide a continuous representation of protein sequences (Asgari & Mofrad, 2015). ProtVec provides 100-dimensional embedding for every 3-gram, and the average embedding is used to represent the sequence.

2.4 Evaluation Methodology

In the zero-shot learning evaluation protocol, it's crucial to ensure class separation during model training and hyperparameter tuning (Xian et al., 2017). Therefore, examples are divided into train, validation, and test sets by their class labels. We establish a setup where a portion of the well-studied kinases (light kinases) are held out as zero-shot classes and are excluded from training. This is to ensure sufficient test data from each kinase for a more robust evaluation. When creating the splits, we follow multiple principles to ensure fair data split evaluations.

- **Number of phosphosites per kinase:** To ensure a robust evaluation in the test and validation sets, we set a minimum threshold for kinase-phosphosite pairs per kinase to avoid relying on very few data points for any kinase class. Thus, we invert the roles of light and dark kinases in evaluation: the test data include well-studied kinases (light kinases), while the training primarily comprises under-studied kinases (dark kinases). However, it's crucial to note that this arrangement is solely for evaluation purposes; the deployed model can predict dark kinases.
- **Stratification based on kinase groups:** Kinases from the same kinase group (Manning et al., 2002) share evolutionary and functional similarities. Thus, we stratify kinases based on their groups, ensuring the representation of kinase groups in the training, validation, and test sets when feasible.
- **Sequence similarity of kinases:** To prevent optimistic performance estimates, kinases with sequence similarity above a specified threshold are grouped and assigned to the same sets (train, validation, or test). Sequence similarity is determined by sequence identity

calculated after pairwise global alignment of the kinase domains (using BLOSUM62, with -11 gap open penalty and -1 gap extension penalty).

Taking all these aspects into consideration, we divide the data set into training (80%), validation (10%), and test (10%) sets. We first identify kinases as train or test kinases according to the number of phosphosites they are associated with. Kinases with fewer than 15 phosphosites are designated as train kinases. Later, kinases with at least 90% sequence identity similarity are grouped together and are randomly defined as train or test kinases altogether. From the remaining kinases, test kinases are randomly selected from each kinase group in a stratified manner. All remaining kinases are defined as train kinases. This process is repeated to determine validation kinases from among train kinases by setting the threshold for kinases in validation to be at least 10 phosphosites per kinase. Finally, the train, validation, and test sets include all train phosphosite-kinase pairs associated with the kinases in that relative set. We split the kinase into these sets in a randomized and reproducible manner, allowing for different DARKIN dataset splits by changing the random seeds.

Due to the imbalance of the multi-class classification problem, we evaluate our methods using the Average Precision (AP) score, which summarizes the precision-recall curve. We compute the macro AP across classes. When multiple kinases can phosphorylate a phosphosite, we accept the predicted kinase as a true positive if it matches any of the true kinases associated with the phosphosite.

2.5 Zero-Shot Classifiers

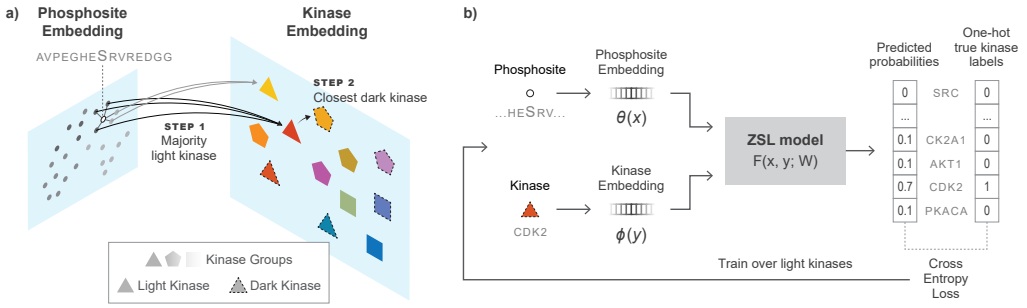

Figure 1: a) k-NN based zero-shot classifier. First, the phosphosite's nearest neighbor phosphosites are determined in the training data. The majority vote is taken among the neighbors' class labels to pick the most likely light kinase. Then, the dark kinase closest to this light kinase is picked. b) The bilinear compatibility function F takes the phosphosite and kinase embedding vectors and is trained to minimize the cross-entropy loss over light kinases.

In our experiments, we employ two zero-shot learning (ZSL) models. The first is a fitting-free method based on an intentionally kept simple adaptation of the k-NN classifier. The second is a well-established bilinear zero-shot compatibility model. The following sections provide further details on these approaches.

2.5.1 The Zero-Shot k-NN Classifier

To benchmark the zero-shot dark kinase prediction performance, we devise a simple baseline method by adapting the principles of the k-NN algorithm for supervised classification to our zero-shot classification task. For a given test phosphosite, we first locate the k most similar training phosphosites in the phosphosite representation space. Subsequently, we identify the most common *light kinase* among the kinases associated with the nearest neighbor phosphosites. In cases where there is no majority, we utilize the kinase of the nearest neighbor. In the subsequent step, we predict the most resembling the majority kinase in the kinase representation space as the zero-shot *dark kinase*. Kinase similarity is assessed using the cosine similarity of the kinase embedding vectors. This procedure is depicted in Figure 1a. Our motivation in devising this method is to be able to evaluate the pLMs as *directly* as possible, in the sense that the approach does not involve numerical optimization, and the only hyper-parameter is k. This simplicity provides an additional view of the relative strengths of the pLMs, avoiding model selection effects to a great extent.

2.5.2 THE BILINEAR ZERO-SHOT LEARNING MODEL

The second zero-shot learning method that we use is a bilinear compatibility model. While a variety of other zero-shot learning methods, particularly in image classification, have been proposed over the years, variants based on bilinear compatibility models are arguably among the most established ones (Xian et al., 2017; Akata et al., 2016; Romera-Paredes & Torr, 2015; Frome et al., 2013; Akata et al., 2015; Kodirov et al., 2017; Sumbul et al., 2018; Deznabi et al., 2020). Therefore, they are particularly suitable for our pLM evaluation purposes.

The bilinear zero-shot model (BZSM) aims to estimate the compatibility between a given pair of phosphosite x and kinase y (illustrated in Figure 1b). In our work, we use the formulation variant proposed and used in Sumbul et al. (2018); Deznabi et al. (2020), which defines the compatibility function $F(x, y) = [\theta(x)^\top \quad 1]W[\phi(y)^\top \quad 1]^\top$ where $\theta(x) \in \mathbb{R}^d$ is the phosphosite representation, and $y \in \mathbb{R}^m$ is the kinase representation. The augmentation of both representations with separate bias dimensions increases the expressivity of the model (Sumbul et al., 2018), which can more clearly be observed when the definition is expanded:

$$F(x, y) = \theta(x)^\top W \phi(y) + \theta(x)^\top W_{\cdot,m} + W_{d,\cdot}\phi(y) + W_{d+1,m+1}. \tag{1}$$

In this formulation, the first term estimates pairwise compatibility. The second term acts analogous to a $\log p(x)$ prior, formulated via a linear estimator conditioned on $\theta(x)$. Similarly, the third term is the $\log p(y)$ prior, expressed as a linear function of $\phi(y)$. And finally, the last term is simply a trainable scalar. The model is trained by minimizing the regularized cross-entropy loss:

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

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

$$p(y|x) = \frac{\exp F(x, y)}{\sum_{y' \in Y_{tr}} \exp F(x, y')}. \tag{3}$$

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

Table 2: Mean macro AP of 3-NN and the BZSM using only pLM embeddings. For pLMs with CLS and average token, the best performing one is shown.

Embedding	AP (3-NN)	AP (BZSM)
OneHotEnc	0.0897	0.0634 $\pm$0.0034
Blosum62	0.0897	0.0327 $\pm$0.0008
NLF	0.0902	0.0419 $\pm$0.0030
ProtVec	0.0808	0.0959 $\pm$0.0010
Esm1B (cls)	0.1119	0.1631 $\pm$0.0011
Esm1v (cls)	0.1121	**0.1640** $\pm$0.0028
Esm2 (avg)	0.0957	0.1391 $\pm$0.0057
Ankh-Large	0.1106	0.0840 $\pm$0.0012
DistilProtBERT (avg)	0.0811	0.1269 $\pm$0.0084
ProtBERT (avg)	0.0540	0.1044 $\pm$0.0015
ProtAlbert (cls)	0.0915	0.1281 $\pm$0.0049
ProteinBERT	0.1168	0.1236 $\pm$0.0023
ProtGPT2	0.1054	0.1333 $\pm$0.0020
ProtT5-XL	0.1172	0.1552 $\pm$0.0011
SaProt (avg)	0.0973	0.1466 $\pm$0.0026
TAPE	**0.1200**	0.1237 $\pm$0.0018

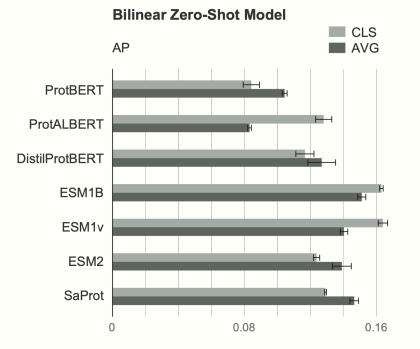

Figure 2: Performance comparison of BZSM trained with CLS and average embedding vector for all pLMs.

3 RESULTS

Hyperparameter tuning. We use macro AP on the validation set for model selection in all cases. For the k-NN based ZSL, we choose k from $\{3, 5, 7\}$. For the bilinear ZSL, we search hyperparameters among random combinations of learning rate ($0.000001...0.1$), optimizer (Adam, SGD, RMSprop), learning rate schedule (Exponential, Step, CosineAnnealing), momentum ($0.95...0.9999$) and the weight decay ($0.00001...0.01$). Finally, to measure the effect of initialization, unless otherwise stated, we train BZSM models three times and report the mean and standard deviation of the macro AP values.

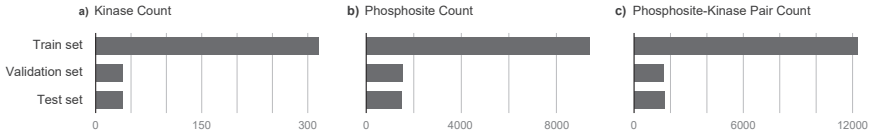

Figure 3: a) The number of kinases b) The number of unique phosphosites c) The number of kinase-phosphosite pairs in each train, validation, and test folds of the default DARKIN split dataset.

Table 3: The BZSM performance trained with sequence embedding and other kinase information. The mean macro APs are shown. The best-performing results of CLS and embedding averaging are shown.

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

DARKIN benchmark statistics. We present four DARKIN splits* to ensure consistency in our experiments and to make consistent decisions on the data, particularly addressing the instability associated with ZSL setups. The experiments are conducted in Darkin split 1 unless otherwise specified. Counts of kinase, distinct phosphosite, and phosphosite-kinase associations for split 1 are shown in 3. For other statistics, refer to the Appendix A.1.

Comparison of Protein Language Models. We initially assess the effectiveness of pLM-based embeddings using both k-NN and BZSM methods. Table 2 presents macro AP scores obtained through the k-NN and BZSM when different pLM embeddings (detailed in Table 1) are used to represent the 15-mer around the phosphosite sequence and the kinase domain sequence. When employing pLM embeddings, we utilized embeddings sourced from the same pLM for the phosphosite and kinase. To establish baseline performance, we also present results obtained with three sequence encoding methods: one-hot encoding, BLOSUM62, and NLF encoding (Section 2.3). In both models, we observe that most of the pLM representations are above the baseline encodings, indicating that they indeed capture the relevant characteristics of the protein sequences better.

The TAPE embeddings perform the best among the k-NN models (0.12 AP score); the Esm models, ProtT5-XL, are close to TAPE's results (Table 2). In the BZSM models, though, the TAPE embeddings fall behind Esm1B and Esm1v. The superior performance of TAPE in the k-NN could be due to it being a lower dimensional vector (see Table 1). In BZSM, when employing the CLS token, both Esm1B and Esm1v exceed 0.16 macro AP. ProtT5-XL is the third close runner-up, and SaProt (cls) also performs well.

CLS token embedding versus averaging. Several pLMs provide a CLS token whose embedding is commonly used as the sequence summary (Devlin et al., 2018). However, it is not clear whether the CLS token or the average of all token embeddings provides a better summary of this task. The performance differences between these two alternatives are shown in Figure 2, which shows that (i) the results can depend on this detail and (ii) the right option varies across the pLMs.

*https://github.com/tastanlab/darkin

Table 4: Comparison of the two best-pLMs, Esm1B and SaProton four random DARKIN splits. The mean macro AP scores and their standard deviations are shown for the BZSM.

	Split 1	**Split 2**	**Split 3**	**Split 4**
Esm1B (cls)	0.1769 $\pm$0.0022	0.1536 $\pm$0.0020	0.1531 $\pm$0.0018	0.1652 $\pm$0.0020
SaProt (cls)	**0.1800** $\pm$0.0015	**0.1599** $\pm$0.0029	**0.1627** $\pm$0.0021	**0.1690** $\pm$0.0050

Incorporating additional kinase information. We augment the kinase sequence embedding vectors with additional information regarding kinase family hierarchy and EC classification. The one-hot encoded vectors encode additional information and are appended to the sequence embedding vectors. Here, we experiment only with the BZSM as it outperforms the k-NN. Including each additional information individually enhances the performance of all models (Table 3), especially the inclusion of the kinase family information improves the results. The models based on Esm1B, Esm1v, and SaProt, using the CLS token embeddings, benefit the most and emerge as the best performers in this augmented case. These findings underscore additional information in these kinase categorizations that cannot be captured solely with sequence information.

Comparing the best-performing pLMs on different DARKIN splits. As Esm1B and SaProt emerge as the two top-performing pLMs when paired with the BZSM model (Table 3), we further evaluated their performance on three additional random splits of the DARKIN dataset to facilitate a more comprehensive comparison between these two pLMs. While both models demonstrate competitiveness, SaProt consistently outperforms Esm1B slightly on these four different splits (Table 4). The performance of SaProt underscores the added value of structural information.

4 CONCLUSION

Focused on the task of assigning dark kinases to phosphosites, DARKIN offers a novel benchmark for evaluating PLMs. We assess the pLM's semantic representation capabilities by using them as input to two zero-shot classifiers. Our results demonstrate the superior performance of the Esm models, the ProtT5-XL, and the SaProt models. We hope this novel benchmark will facilitate comprehensive evaluations of pLMs and dark kinase prediction models, contributing to protein biology research.

ACKNOWLEDGMENTS

This study was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grant Number 122E500. The authors thank TUBITAK for their support. The numerical calculations were partially performed at TÜBİTAK TRUBA resources.

REFERENCES

Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2927–2936, 2015.

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7): 1425–1438, 2016.

Ehsaneddin Asgari and Mohammad R. K. Mofrad. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLOS ONE*, 10(11):e0141287, November 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0141287. URL `http://dx.doi.org/10.1371/journal.pone.0141287`.

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

Matthew E Berginski, Nienke Moret, Changchang Liu, Dennis Goldfarb, Peter K Sorger, and Shawn M Gomez. The dark kinase knowledgebase: an online compendium of knowledge and experimental results of understudied kinases. *Nucleic acids research*, 49(D1):D529–D535, 2021.

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

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 02 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac020. URL https://doi.org/10.1093/bioinformatics/btac020.

Philip Cohen. The origins of protein phosphorylation. *Nature cell biology*, 4(5):E127–E130, 2002.

DeepMind and European Bioinformatics Institute. AlphaFold API. https://alphafold.ebi.ac.uk/api-docs, 2023. Accessed: 2023-10-24.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Iman Deznabi, Busra Arabaci, Mehmet Koyutürk, and Oznur Tastan. Deepkinzero: zero-shot learning for predicting kinase–phosphosite associations involving understudied kinases. *Bioinformatics*, 36(12):3652–3661, 2020.

Salah Eid, Stefan Turk, Andrea Volkamer, Friedrich Rippmann, and Simone Fulle. Kinmap: a web-based tool for interactive navigation through human kinome data. *BMC Bioinformatics*, 18:16, 2017.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Wang Yu, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. 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*, pp. 1–1, 2021. doi: 10.1109/TPAMI.2021.3095381.

Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. Ankh: Optimized protein language model unlocks general-purpose modelling, 2023.

Damien Farrell. epitopepredict: A tool for integrated mhc binding prediction. *bioRxiv*, 2021. doi: 10.1101/2021.02.05.429892. URL https://www.biorxiv.org/content/early/2021/02/07/2021.02.05.429892.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. A deep unsupervised language model for protein design. *bioRxiv*, 2022. doi: 10.1101/2022.03.09.483666. URL https://www.biorxiv.org/content/early/2022/03/12/2022.03.09.483666.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pp. 2121–2129, 2013.

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

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

Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. *arXiv preprint arXiv:1704.08345*, 2017.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*, 2022. doi: 10.1101/2022.07.20.500902. URL https://www.biorxiv.org/content/early/2022/10/31/2022.07.20.500902.

Gerard Manning, David B Whyte, Ricardo Martinez, Tony Hunter, and Sucha Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, 2002.

Gerard Manning, David B Whyte, Ricardo Martinez, Tony Hunter, and Sucha Sudarsanam. Kinome tables, 2023. URL http://kinase.com/human/kinome/tables/. Accessed on 2023-12-14.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 2021. doi: 10.1101/2021.07.09.450648. URL https://www.biorxiv.org/content/early/2021/11/17/2021.07.09.450648.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.

Nienke Moret, Changchang Liu, Benjamin M Gyori, John A Bachman, Albert Steppi, Clemens Hug, Rahil Taujale, Liang-Chin Huang, Matthew E Berginski, Shawn M Gomez, et al. A resource for exploring the understudied human kinome for research and therapeutic opportunities. *BioRxiv*, pp. 2020–04, 2020.

Susanne Müller, Apirat Chaikuad, Nathanael S Gray, and Stefan Knapp. The ins and outs of selective kinase inhibitor development. *Nature chemical biology*, 11(11):818, 2015.

Loris Nanni and Alessandra Lumini. A new encoding technique for peptide classification. *Expert Systems with Applications*, 38(4):3185–3191, 2011. doi: https://doi.org/10.1016/j.eswa.2010.09.005. URL https://www.sciencedirect.com/science/article/pii/S0957417410009097.

Elise J Needham, Benjamin L Parker, Timur Burykin, David E James, and Sean J Humphrey. Illuminating the dark phosphoproteome. *Sci. Signal.*, 12(565):eaau8645, 2019.

Protein Data Bank in Europe. Pdbe. https://www.ebi.ac.uk/pdbe/pdbe-kb/, 2023. Accessed: 2023-11-09.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with tape. 2019.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv*, 2020. doi: 10.1101/622803. URL https://www.biorxiv.org/content/early/2020/12/15/622803.

Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pp. 2152–2161, 2015.

Castrense Savojardo, Pier Luigi Martelli, and Rita Casadio. Finding functional motifs in protein sequences with deep learning and natural language models. *Current Opinion in Structural Biology*, 81:102641, 2023. ISSN 0959-440X. doi: https://doi.org/10.1016/j.sbi.2023.102641. URL https://www.sciencedirect.com/science/article/pii/S0959440X2300115X.

Robert Schmirler, Michael Heinzinger, and Burkhard Rost. Fine-tuning protein language models boosts predictions across diverse tasks. *bioRxiv*, 2023. doi: 10.1101/2023.12.13.571462. URL https://www.biorxiv.org/content/early/2023/12/14/2023.12.13.571462.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. *bioRxiv*, 2023. doi: 10.1101/2023.10.01.560349. URL https://www.biorxiv.org/content/early/2023/10/02/2023.10.01.560349.1.

Gencer Sumbul, Ramazan Gokberk Cinbis, and Selim Aksoy. Fine-grained object recognition and zero-shot learning in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):770–779, 2018.

UniProt Consortium. https://www.uniprot.org/uniprotkb?query=%28taxonomy_id%3A9606%29+AND+%28ft_domain%3Akinase%29+AND+%28organism_id%3A9606%29&facets=reviewed%3Atrue, 2023a. [Online; accessed 14-December-2023].

UniProt Consortium. Uniprot api. https://www.uniprot.org/help/api, 2023b. Accessed: 2023-12-29.

Serbulent Unsal, Heval Atas, Muammer Albayrak, Kemal Turhan, Aybar C Acar, and Tunca Doğan. Learning functional properties of proteins with language models. *Nature Machine Intelligence*, 4 (3):227–245, 2022.

Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4582–4591, 2017.

Yiming Zhang and Manabu Okumura. Prothyena: A fast and efficient foundation protein language model at single amino acid resolution. *bioRxiv*, 2024. doi: 10.1101/2024.01.18.576206. URL https://www.biorxiv.org/content/early/2024/01/22/2024.01.18.576206.

A Appendix

A.1 DARKIN Dataset Statistics

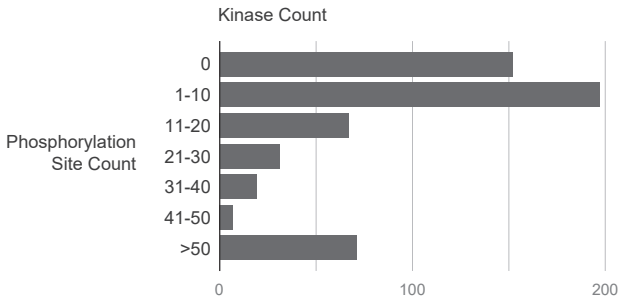

Figure 4: The histogram of the number of phosphosites associated with a kinase in the PhosphositePlus dataset. For most of the kinases, no or few phosphosites are assigned.

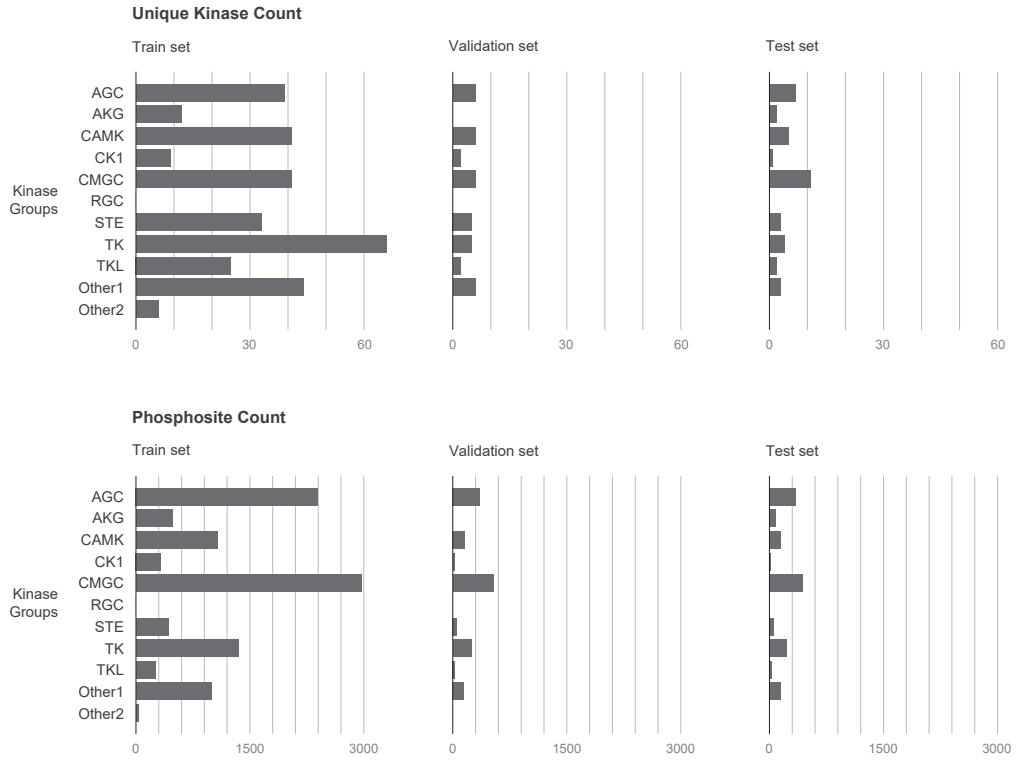

Figure 5: Splits are stratified based on kinase groups. (Upper panel) The number of kinases in each kinase group, in train, validation, and test folds. (Lower panel) The number of unique phosphosites that are associated with a kinase that falls into the kinase group in train, validation, and test folds.

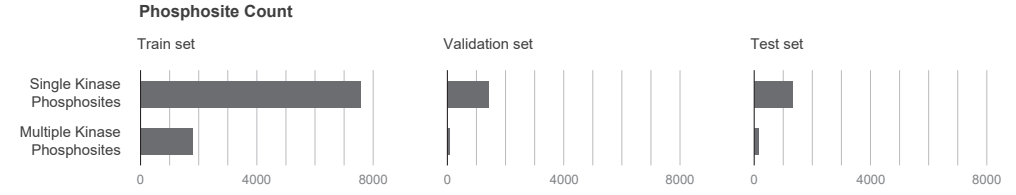

Figure 6: The number of multi-label cases for train, validation, and test. In training we binarize the multi-labels, in the testing, we count the prediction as true positive if the predicted kinase is in the set. The number of multi-label cases is small in the test set.

Table 5: All AP scores of all pLMs on the 3-NN and the BZSM Models with kinase additional information.

	kNN (3-NN)					BZSM				
Embedding	**AP**	**+ Family**	**+ Group**	**+ EC**	**+ Family + Group + EC**	**AP**	**+ Family**	**+ Group**	**+ EC**	**+ Family + Group + EC**
OneHotEnc	0.0897	0.0901	0.0868	0.0864	0.0917	0.0634	0.1107	0.0832	0.0802	0.1098
Blosum62	0.0897	0.0897	0.0901	0.0901	0.0897	0.0327	0.0318	0.0310	0.0337	0.0323
NLF	0.0902	0.0903	0.0907	0.091	0.0913	0.0419	0.0391	0.0425	0.0400	0.0426
ProtVec	0.0808	0.0963	0.1123	0.0993	**0.1230**	0.0959	0.1262	0.1129	0.1214	0.1354
ProtBERT (avg)	0.0540	0.0871	0.0767	0.0722	0.0985	0.1044	0.1053	0.0982	0.1017	0.1018
ProtBERT (cls)	0.0855	0.0867	0.0930	0.0904	0.0968	0.0842	0.1170	0.1077	0.1132	0.1273
ProteinBERT	0.1168	0.1182	0.1182	0.1180	0.1227	0.1236	0.1506	0.1215	0.1367	0.1359
ProtT5-XL	0.1172	0.1164	0.1170	0.1172	0.1240	0.1552	0.1701	0.1531	0.1674	0.1731
Esm1B (avg)	0.1122	0.1107	0.1112	0.1105	0.1132	0.1512	0.1523	0.1553	0.1546	0.1554
Esm1B (cls)	0.1119	0.1101	0.1110	0.1110	0.1134	**0.1631**	**0.1740**	**0.1688**	**0.1680**	0.1769
Esm1v (avg)	0.1114	0.1114	0.1110	0.1112	0.1111	0.1404	0.1494	0.1426	0.1395	0.1430
Esm1v (cls)	0.1121	0.1125	0.1116	0.1131	0.1129	0.1640	0.1737	0.1653	0.1652	0.1734
Esm2 (avg)	0.0957	0.0981	0.0977	0.1015	0.1022	0.1391	0.1588	0.1453	0.1496	0.1638
Esm2 (cls)	0.0982	0.1024	0.1021	0.1024	0.1095	0.1238	0.1443	0.1358	0.1433	0.1518
DistilProtBERT (avg)	0.0811	0.0961	0.0977	0.0967	0.1085	0.1269	0.1380	0.1265	0.1356	0.1385
DistilProtBERT (cls)	0.1021	0.1046	0.1084	0.1101	0.1139	0.1167	0.1360	0.1292	0.1287	0.1441
ProtGPT2	0.1054	0.1061	0.1068	0.1094	0.1111	0.1333	0.1476	0.1412	0.1419	0.1557
Ankh-Large	0.1106	0.1077	0.1119	0.1055	0.1137	0.0840	0.1417	0.1135	0.1178	0.1594
ProtAlbert (avg)	0.1004	0.1055	0.1072	0.1041	0.1124	0.0831	0.0964	0.0854	0.0909	0.0845
ProtAlbert (cls)	0.0915	0.0993	0.0958	0.0982	0.1026	0.1281	0.1269	0.1276	0.1285	0.1372
SaProt (avg)	0.0973	0.1109	0.1098	0.1066	0.1209	0.1466	0.1684	0.1529	0.1490	0.1667
SaProt (cls)	0.1056	0.1100	0.1146	0.1107	0.1204	0.1292	0.1696	0.1424	0.1434	**0.1800**
TAPE	**0.1200**	**0.1217**	**0.1211**	**0.1249**	**0.1230**	0.1237	0.1379	0.1333	0.1310	0.1455