PREDICTING KINASE-SUBSTRATE PHOSPHORYLATION SITE USING AUTOREGRESSIVE TRANSFORMER

Mahdi Pourmirzaei[1,2,*] **Farzaneh Esmaili**[1,*] **Mohammadreza Pourmirzaei**[3]

Mohsen Rezaei[1] **Dong Xu**[1]

[1] University of Missouri, USA
{mpngf, f.esmaili, mrfrg, xudong}@missouri.edu
[2] ProGene
[3] Politecnico di Milano, Italy
mohammadreza.pourmirzaeioliaei@mail.polimi.it
*Equal contribution

ABSTRACT

Protein phosphorylation is a critical post-translational modification regulating cellular processes, yet accurately predicting kinase-specific phosphorylation sites remains challenging due to the vast diversity of kinases and the contextual complexity of substrate recognition. In this work, we introduce a unified model that integrates a pre-trained protein language model with an autoregressive transformer decoder to address three kinase phosphorylation-related tasks directly from substrate and kinase sequences. To overcome the primary limitation of autoregressive decoding in phosphorylation site prediction, we incorporate a self-supervised pre-training stage, removing that limitation and achieving state-of-the-art results in substrate-kinase phosphorylation site prediction, surpassing existing methods by a large margin. Additionally, our framework supports kinase group classification and substrate-kinase interaction prediction as auxiliary tasks, further improving model applicability in real-world biological research. Our results highlight the potential of leveraging generative AI for precise phosphorylation site prediction. The code and model are available at GitHub.

1 INTRODUCTION

Post-translational modifications (PTMs) are crucial processes that regulate cellular functions and biological processes Mann & Jensen (2003). Among these modifications, protein phosphorylation is one of the most significant. It involves the attachment of a phosphate group to specific amino acids—serine (S), threonine (T), or tyrosine (Y)—rendering the target protein phosphorylated Cohen (2002). This process is crucial for signal transduction, cellular communication, and the regulation of various biological activities. Dysregulation of phosphorylation can lead to severe diseases, including cancer, neurodegenerative disorders, and Alzheimer's disease Dashtmian et al. (2024); Kerr et al. (2025). Kinases, a group of specialized enzymes, catalyze this process by transferring phosphate groups to proteins Browning et al. (1985); Kerr et al. (2025). The prediction of kinase-specific phosphorylation sites is challenging due to the vast diversity and structural complexity of kinases, as well as the dynamic nature of protein interactions within cells. Each kinase recognizes specific sequence motifs, but variations in these motifs and context-dependent factors make accurate predictions difficult Cohen (1982).

In recent years, numerous approaches have been developed to predict kinase-specific phosphorylation sites, leveraging computational techniques such as machine learning and deep learning Esmaili et al. (2023). However, most of the existing methods share significant limitations. Many rely on ensemble approaches that use

several, or in some cases hundreds, of models to achieve acceptable accuracy Chen et al. (2023); Ma et al. (2023). While this may improve predictive performance, it introduces inefficiencies in computational resource usage and scalability. Additionally, these methods typically focus on peptides with lengths ranging from 7 to 15 amino acids, which represent only a small fraction of possible protein sequences Huang et al. (2014); Wang et al. (2017); Kirchoff & Gomez (2022). This narrow focus not only fails to cover the full range of protein sequences but also risks losing critical structural information, which is essential for understanding protein-kinase interactions at a more comprehensive level. As a result, the predictive performance of these methods lacks generalizability across diverse protein contexts, limiting their applicability in real-world biological scenarios.

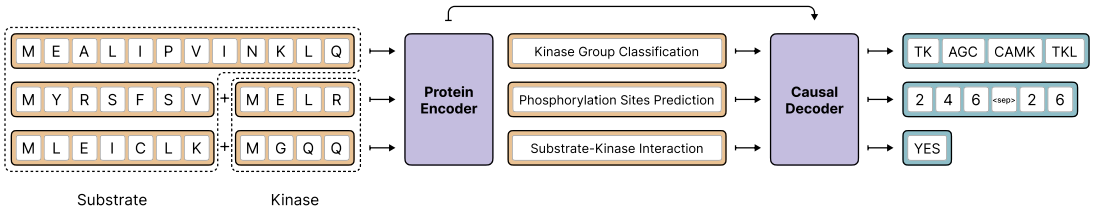

Figure 1: Overview of our approach for kinase-related predictions using substrate and/or kinase sequences. The framework supports three tasks: kinase group classification, protein-kinase interaction prediction, and protein-kinase phosphorylation site prediction. The protein encoder generates sequence embeddings, which are processed by a causal decoder for task-specific predictions.

To bridge these gaps, we reformulate the protein-kinase phosphorylation site prediction problem into three interconnected tasks: kinase group classification, protein-kinase interaction prediction, and protein-kinase phosphorylation site prediction. By drawing inspiration from Prot2Token Pourmirzaei et al. (2024), this study presents a unified model that integrates an autoregressive transformer, called a decoder, with ESM-2 Lin et al. (2022) pre-trained protein language model, to predict those tasks from substrate and kinase sequences in an end-to-end fashion (Figure 1). Specifically, we demonstrate that incorporating self-supervised pre-training tasks to initialize the autoregressive decoder's weights significantly enhances the performance of Prot2Token in predicting protein-kinase phosphorylation sites, surpassing state-of-the-art methods by a large margin. We evaluate our model's performance by addressing each problem as follows:

1. **Kinase group classification.** We first treat kinase group classification as a multi-label learning problem and formulate it as predicting substrate-group interactions using only substrate sequences. Our results show that it is possible to identify kinase groups based solely on substrate sequences.
2. **Protein-kinase interaction.** Moving beyond kinase group-level interactions, we predict direct interactions between individual substrate and kinase sequences. To enhance model robustness and prevent collapse, we generate hard negative samples inspired by data pruning Sorscher et al. (2022) and self-supervised contrastive learning Chen et al. (2020), leading to improved protein-kinase interaction performance.
3. **Protein-kinase phosphorylation site.** Leveraging the results from kinase group classification and protein-kinase interaction prediction, we extend the Prot2Token framework to support protein-kinase phosphorylation site prediction. This is achieved by pre-training the autoregressive decoder with self-supervised learning tasks, significantly refining its ability in identifying phosphorylation sites.

Throughout this paper, we use the terms substrate sequence and protein sequence interchangeably, treating them as equivalent.

2 Related Work

The prediction of kinase phosphorylation sites has been explored using various machine learning and deep learning approaches to advance our understanding of PTMs and their regulatory roles in cellular

processes Wang et al. (2017); Ma et al. (2023). Computational models leverage sequence-based features, physicochemical properties, structural characteristics, and evolutionary information Esmaili et al. (2023). In contrast, deep learning methods aim for end-to-end learning by directly capturing complex patterns from raw data Yang et al. (2021). Broadly, existing research in this domain falls into two main categories: kinase-specific phosphorylation site prediction and protein-kinase interaction prediction. The former focuses on identifying phosphorylation sites for specific kinases, while the latter aims to determine interactions between kinases and their substrate proteins. A detailed discussion of these categories and related work is provided in Appendix A.1.

3 METHOD

Our approach builds upon a modified Prot2Token architecture in our preliminary work Pourmirzaei et al. (2025) by connecting a causal (autoregressive) Transformer—termed the decoder—to a pre-trained bidirectional Transformer Vaswani (2017), referred to as the encoder. For this encoder, we adopt the ESM-2 650M model and initialize it with the corresponding pre-trained weights. The decoder then leverages the encoder's output through cross-attention. To control the distinct prediction of each task, we use separate tokenizers and embedding tables for the encoder and decoder (Figure 2). The autoregressive transformer factorizes the joint probability of a sequence $x = (x_1, x_2, \ldots, x_t)$ into a product of conditional probabilities, i.e.,

$$p(x) = \prod_{t=1}^{T} p_\theta(x_t \mid x_1, \ldots, x_{t-1})$$

Training proceeds by minimizing the negative log-likelihood of the observed tokens:

$$L(\theta) = -\sum_{t=1}^{T} \log p_\theta(x_t \mid x_1, \ldots, x_{t-1})$$

where θ denotes the model parameters. Employing a causal mask, each token x_t can only attend to the tokens $x_1, \ldots, x_{t-1}$, thereby enforcing the autoregressive property and allowing the model to learn contextual representations of the preceding sequence. We extend this standard autoregressive objective by introducing token-level weights w_t to regulate each token's influence on the loss. Specifically, we set $w_1 = 0$ so that predicting the first token (the prompt) does not affect the loss, while for $t \geq 2$, w_t can be adjusted, granting non-prompt tokens varying degrees of importance. Concretely, the revised training objective is:

$$L(\theta) = -\sum_{t=1}^{T} w_t \log p_\theta(x_t \mid x_1, \ldots, x_{t-1})$$

where each weight $w_t \in [0, \infty)$ is a user-defined parameter for the token x_t. This setup enables flexible fine-tuning by emphasizing specific tokens of interest while removing the prompt token from the loss computation (by assigning it zero weight). Further details on the architecture are provided in Appendix A.2.

Self-supervised Pre-training of the Decoder. Unlike the encoder, which can be initialized with pre-trained weights (e.g., from ESM-2), the decoder typically starts with random weights. However, the original Prot2Token research found that certain tasks like phosphorylation post-translational modification training benefit when they are co-trained with self-supervised tasks. We believe this is because the decoder needs to grasp the label structure (i.e., implicit inductive biases) before making predictions. Suppose the label vocabulary is large (as in phosphorylation) but only a limited number of examples are available. In that case, the model may struggle to learn these biases, which can markedly reduce predictive performance. To mitigate this, we introduce a self-supervised pre-training phase to initialize the decoder weights for initializing substrate-kinase phosphorylation site prediction training. In these self-supervised tasks, amino acid sequences are the input, and the labels indicate the positions of specific amino acids. For instance, for the sequence *"MSGLSNYT"* containing the amino acid 'S', we label the relevant indices—here {2,5} —as the

output. We repeat this process for each of the 20 standard amino acids, creating a set of 20 self-supervised tasks. Crucially, these tasks require no human-generated labels, making them straightforward and cost-free to produce.

Tokenization of labels. We follow the tokenization framework introduced in Prot2Token to convert target labels into discrete tokens. Specifically, we use a non-hierarchical multi-label classification scheme for kinase group prediction, binary classification for protein-kinase interaction, and the same approach from the original work for both protein-kinase phosphorylation site and self-supervised tasks.

Datasets. We utilized the GPS 6.0 Chen et al. (2023) dataset, originally comprising 24,160 phosphorylation sites. After preprocessing and mapping IDs to the UniProt database, we retained 13,401 sequences annotated with kinase information, covering 386 kinases across 12 distinct groups validated against Kinase.com for Homo sapiens species. The final dataset contains Uniprotids, substrate sequences, kinase sequences, group information, and phosphorylation sites. To reduce sequence similarity, CD-HIT Li & Godzik (2006) was applied with a 70% threshold. More details regarding dataset preparation are presented in Appendix A.3.

4 EXPERIMENTS

In this section, we first evaluated the performance of our model on kinase group classification, followed by generating hard negative samples to enhance the protein-kinase interaction task. Building on these results, we performed self-supervised pre-training of the autoregressive decoder and extended our approach to predict protein-kinase phosphorylation sites, benchmarking its performance against state-of-the-art methods.

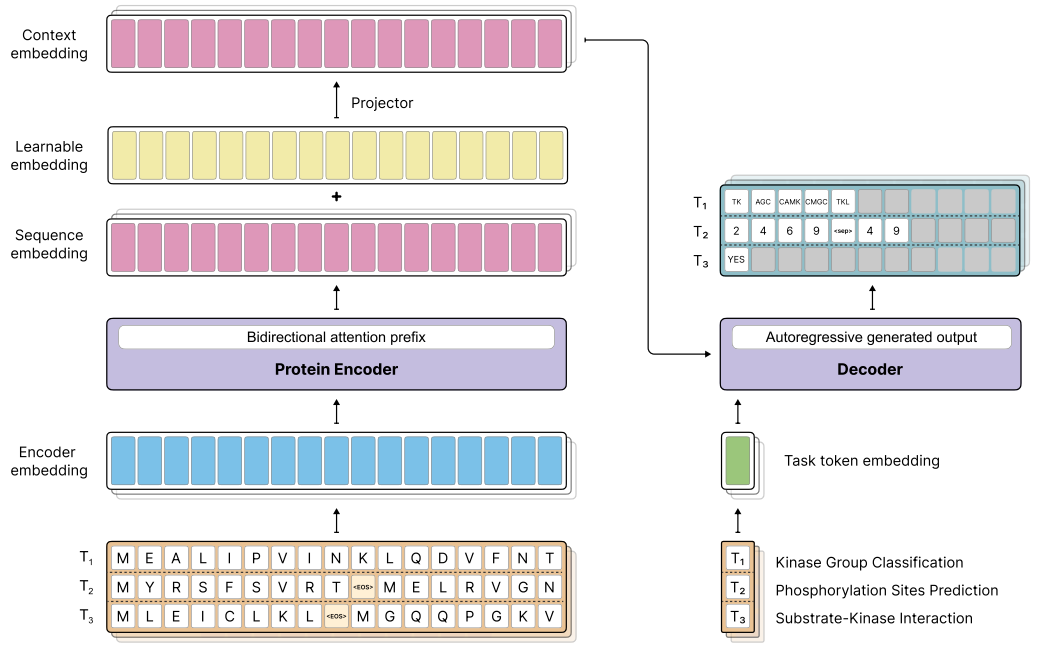

Figure 2: Overview of the proposed model architecture during inference. The bidirectional protein encoder processes substrate and kinase sequences to generate embeddings, which the autoregressive decoder uses to predict output tokens. The figure highlights the integration of the pre-trained ESM-2 model with the autoregressive decoder for kinase group classifications, substrate-kinase interactions, and phosphorylation site prediction.

For all experiments, we initialized the protein encoder of Prot2Token with the pre-trained ESM-2 650M model. We enabled fine-tuning for the weights of the last eight blocks of the encoder, while keeping the embedding parameters of the encoder frozen unless specified otherwise. All other layers in the model were trained during the experiments. We employed the AdamW optimizer Loshchilov (2017) with a weight decay of 0.1, $\beta_1 = 0.9$, $\beta_2 = 0.999$., and epsilon to 1e-16 as default hyperparameters across all training runs. The learning rate followed a cosine annealing schedule with an initial warm-up phase Loshchilov & Hutter (2016), starting from 1e-6 and increasing to 5e-5 over the first 256 steps unless specified otherwise. All training was conducted using the PyTorch 2 framework Ansel et al. (2024) on a single node equipped with 4×A100 80GB Nvidia GPUs.

4.1 KINASE GROUP CLASSIFICATION

In this section, we aimed to predict kinase groups based on substrate sequences. Specifically, we investigated how much information about the related kinase groups the model can infer solely from substrate sequences. To achieve this, we considered our processed training and validation datasets (refer to Appendix A.3), assigning multi-label classification labels by removing "Unknown" group samples from the training set and merging the remaining nine kinase groups associated with each substrate. The model takes a substrate sequence as input and predicts the corresponding kinase groups in alphabetical order. We allow Prot2Token to fine-tune the weights of the last 6 blocks of the protein encoder (ESM-2 650m). The result is presented in Table 1.

Table 1: F1 scores of our method on each kinase group.

Group	AGC	Atypical	CMGC	CAMK	CK1	Other	STE	TK	TKL
F1	0.5551	0.4928	0.634	0.4196	0.5385	0.6047	0.3571	0.6738	0.5

Next, we compared the embedding representations of all unique kinase sequences before and after fine-tuning the protein encoder part on the kinase group classification labels. The fine-tuned model demonstrated better separation of kinase sequences into their respective groups, even though kinase sequences were not explicitly included during training (Table 2 and Figure 3). Additional details are provided in Appendix A.4.

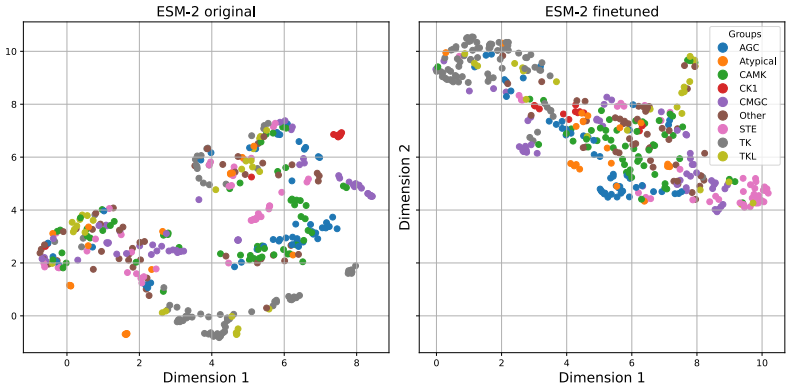

Figure 3: UMAP visualization of unique kinase sequences on the original and fine-tuned checkpoints of ESM-2 650m

Table 2: Unsupervised separability metrics of all unique kinase sequences with respect to the original and fine-tuned version of ESM2 650m model.

Metric	**Original**	**Fine-tuned**
Silhouette Score (cosine)	-0.0393	0.0905
Calinski-Harabasz Index	7.0029	24.0969

Table 3: The details of protein-kinase interaction performance versus other methods on the positive and randomly selected negative samples from the validation set.

Metric	**Ours**	**GPS 6.0** Chen et al. (2023)	**PhosphormerST** Zhou et al. (2024)
Accuracy (%)	**72**	58	49
F1	**0.71**	0.61	0.603

4.2 PROTEIN-KINASE INTERACTION PREDICTION

In this section, we refined our approach by shifting from kinase group-level interactions to individual protein-kinase pairs. Specifically, we formulated the problem as determining whether a given protein substrate interacts with a specific kinase sequence through phosphorylation. To achieve this, we constructed both positive and hard negative samples using our training dataset. The results are presented in Table 3. The detailed methodology for crafting negative labels is described in Appendix A.4.

4.3 SELF-SUPERVISED PRE-TRAINING

In the initial stage, we randomly sampled 4 million protein sequences from the UniRef50 database Suzek et al. (2015) for training and 4,000 sequences for validation. From these, we artificially generated 80 million training samples and 20,000 validation samples by treating each amino acid type within a protein as an individual sample. Subsequently, we further randomly sampled 1 million training and 1,000 validation samples to construct the final datasets of this part. For training, we used an input sequence length of 1,280, a weight decay of 0.01, and a batch size of 192 samples, equivalent to 73,728 tokens. The warm-up phase consisted of 512 steps. During training, we froze all the encoder weights while allowing all other parameters to be updated. After 16 epochs, the model achieved a validation perplexity of 2.31, indicating that it could almost perfectly reconstruct protein sequences from the encoder's embeddings.

4.4 PROTEIN-KINASE PHOSPHORYLATION SITE PREDICTION

Building on our ability to predict protein-kinase interactions, we extended our approach to precise phosphorylation site prediction. To achieve this, we selected all protein-kinase sequence pairs along with their corresponding phosphorylation sites and jointly trained them alongside 20 self-supervised tasks. This fine-tuning phase utilized the latest checkpoint from the self-supervised pre-training stage as its initial checkpoint. For this phase, we reduced the number of self-supervised tasks to a total of 20,000 samples. Additionally, substrate sequences exceeding 1,280 amino acids in length were excluded during training and evaluation. The results are shown in Table 4 and 5. Note that there is a strong possibility of data contamination between our validation set and rare group test sets, with the GPS 6.0 training set. As a result, GPS 6.0 may achieve artificially high performance on those sets due to memorization, while its real-world performance on unseen substrates could be even lower. More details are presented in A.4.

Table 4: Comparative results of our method against leading tools (KinasePhos3 and GPS 6.0) across the validation, GPS test, and rare groups test sets.

Method	Validation Set			GPS Test Set			Rare Groups Test Set		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
KinasePhos3 Ma et al. (2023)	0.9773	0.0383	0.0747	0.0215	0.9856	0.0421	0.6800	0.2073	0.3178
GPS 6.0 Chen et al. (2023)	0.2323	0.4549	0.3076	0.1564	**0.5054**	0.2398	**1.0000**	**0.3333**	**0.5000**
Ours	**0.8050**	**0.8200**	**0.8124**	**0.3673**	0.3103	**0.3364**	0.6364	0.3182	0.4242

Table 5: Validation set performance results for each kinase group, presented as F1 scores.

Method	AGC	Atypical	CMGC	CAMK	CK1	Other	STE	TK	TKL
KinasePhos3	0.0566	0.0967	0.0658	0.0443	0.2828	0.0711	0.1489	0.2312	0.2516
GPS 6.0	0.2751	0.5138	0.2940	0.2852	0.0000	0.0000	0.0000	0.4050	0.4638
Our method	**0.8039**	**0.7949**	**0.8108**	**0.6960**	**0.9286**	**0.8080**	**0.6512**	**0.8661**	**0.7619**

4.5 DISCUSSION

Our results first demonstrate that finding the right kinase groups can be effectively retrieved from only substrate sequences. This finding highlights that substrate sequences contain intrinsic information indicative of kinase group associations, even without explicit kinase sequence data. In practical applications, this capability could narrow down potential kinase groups that interact with a given substrate, reducing the number of experimental assays required to confirm phosphorylation events. Building on this, we formulated protein-kinase interaction prediction by constructing a dataset that includes both positive interactions and

hard negative samples. We strategically selected hard negatives to ensure the model learns to distinguish true interactions from closely related but non-interacting pairs. By incorporating kinase sequences alongside substrate sequences and leveraging these challenging negative samples, our approach outperformed existing methods, leading to more precise predictions of individual substrate-kinase interactions. This design significantly refines the search space, improving the efficiency of candidate selection in phosphorylation studies. Finally, we extended the Prot2Token framework for protein-kinase phosphorylation site prediction by introducing a self-supervised pre-training stage for the decoder. Our approach yielded state-of-the-art performance, significantly surpassing prior methods.

REFERENCES

Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Covariance pooling for facial expression recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 367–374, 2018.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, et al. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2, pp. 929–947, 2024.

Michael D Browning, Richard Huganir, and Paul Greengard. Protein phosphorylation and neuronal function. Journal of neurochemistry, 45(1):11–23, 1985.

Miaomiao Chen, Weizhi Zhang, Yujie Gou, Danyang Xu, Yuxiang Wei, Dan Liu, Cheng Han, Xinhe Huang, Chengzhi Li, Wanshan Ning, et al. Gps 6.0: an updated server for prediction of kinase-specific phosphorylation sites in proteins. Nucleic acids research, 51(W1):W243–W250, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pp. 1597–1607. PMLR, 2020.

Philip Cohen. The role of protein phosphorylation in neural and hormonal control of cellular activity. Nature, 296(5858):613–620, 1982.

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

UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. Nucleic acids research, 47(D1): D506–D515, 2019.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. arXiv preprint arXiv:2307.08691, 2023.

Anna Roshani Dashtmian, Fereshteh B Darvishi, and William David Arnold. Chronological and biological aging in amyotrophic lateral sclerosis and the potential of senolytic therapies. Cells, 13(11):928, 2024.

Farzaneh Esmaili, Mahdi Pourmirzaei, Shahin Ramazi, Seyedehsamaneh Shojaeilangari, and Elham Yavari. A review of machine learning and algorithmic methods for protein phosphorylation site prediction. Genomics, Proteomics & Bioinformatics, 21(6):1266–1285, 2023.

Yaping Guo, Wanshan Ning, Peiran Jiang, Shaofeng Lin, Chenwei Wang, Xiaodan Tan, Lan Yao, Di Peng, and Yu Xue. Gps-pbs: a deep learning framework to predict phosphorylation sites that specifically interact with phosphoprotein-binding domains. Cells, 9(5):1266, 2020.

Kai-Yao Huang, Hsin-Yi Wu, Yi-Ju Chen, Cheng-Tsung Lu, Min-Gang Su, Yun-Chung Hsieh, Chih-Ming Tsai, Kuo-I Lin, Hsien-Da Huang, Tzong-Yi Lee, et al. Regphos 2.0: an updated resource to explore protein kinase–substrate phosphorylation networks in mammals. Database, 2014:bau034, 2014.

Nathan R Kerr, Anna R Dashtmian, Fereshteh B Darvishi, Charles D Brennan, Sindhuja N Ayyagari, Peter J Moore, Jose A Viteri, Meifang Wang, Mark M Rich, Brian C Clark, et al. 5-ht2c agonism as a neurotherapeutic for sarcopenia: preclinical proof of concept. GeroScience, pp. 1–18, 2025.

Kathryn E Kirchoff and Shawn M Gomez. Ember: multi-label prediction of kinase-substrate phosphorylation events through deep learning. Bioinformatics, 38(8):2119–2126, 2022.

Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22(13):1658–1659, 2006.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. BioRxiv, 2022:500902, 2022.

Zexian Liu, Fang Yuan, Jian Ren, Jun Cao, Yanhong Zhou, Qing Yang, and Yu Xue. Gps-arm: computational analysis of the apc/c recognition motif by predicting d-boxes and ken-boxes. PloS one, 7(3):e34370, 2012.

I Loshchilov. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983, 2016.

Fenglin Luo, Minghui Wang, Yu Liu, Xing-Ming Zhao, and Ao Li. Deepphos: prediction of protein phosphorylation sites with deep learning. Bioinformatics, 35(16):2766–2773, 2019.

Renfei Ma, Shangfu Li, Wenshuo Li, Lantian Yao, Hsien-Da Huang, and Tzong-Yi Lee. Kinasephos 3.0: redesign and expansion of the prediction on kinase-specific phosphorylation sites. Genomics, proteomics & bioinformatics, 21(1):228–241, 2023.

Matthias Mann and Ole N Jensen. Proteomic analysis of post-translational modifications. Nature biotechnology, 21(3):255–261, 2003.

Mahdi Pourmirzaei, Farzaneh Esmaili, Mohammadreza Pourmirzaei, Duolin Wang, and Dong Xu. Prot2token: A multi-task framework for protein language processing using autoregressive language modeling. bioRxiv, pp. 2024–05, 2024.

Mahdi Pourmirzaei, Ye Han, Farzaneh Esmaili, Mohammadreza Pourmirzaei, Salhuldin Alqarghuli, Kai Chen, and Dong Xu. Extending prot2token: Aligning protein language models for unified and diverse protein prediction tasks. bioRxiv, 2025. doi: 10.1101/2025.03.03.641065.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. Advances in Neural Information Processing Systems, 35:19523–19536, 2022.

Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics, 31(6):926–932, 2015.

A Vaswani. Attention is all you need. Advances in Neural Information Processing Systems, 2017.

Chenwei Wang, Haodong Xu, Shaofeng Lin, Wankun Deng, Jiaqi Zhou, Ying Zhang, Ying Shi, Di Peng, and Yu Xue. Gps 5.0: an update on the prediction of kinase-specific phosphorylation sites in proteins. Genomics, Proteomics and Bioinformatics, 18(1):72–80, 2020.

Duolin Wang, Shuai Zeng, Chunhui Xu, Wangren Qiu, Yanchun Liang, Trupti Joshi, and Dong Xu. Musitedeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. Bioinformatics, 33(24):3909–3916, 2017.

Yu Xue, Zexian Liu, Jun Cao, Qian Ma, Xinjiao Gao, Qingqi Wang, Changjiang Jin, Yanhong Zhou, Longping Wen, and Jian Ren. Gps 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. Protein Engineering, Design & Selection, 24(3):255–260, 2011.

Hangyuan Yang, Minghui Wang, Xia Liu, Xing-Ming Zhao, and Ao Li. Phosidn: an integrated deep neural network for improving protein phosphorylation site prediction by combining sequence and protein–protein interaction information. Bioinformatics, 37(24):4668–4676, 2021.

Feng-Feng Zhou, Yu Xue, Guo-Liang Chen, and Xuebiao Yao. Gps: a novel group-based phosphorylation predicting and scoring method. Biochemical and biophysical research communications, 325(4):1443–1448, 2004.

Zhongliang Zhou, Wayland Yeung, Nathan Gravel, Mariah Salcedo, Saber Soleymani, Sheng Li, and Natarajan Kannan. Phosformer: an explainable transformer model for protein kinase-specific phosphorylation predictions. Bioinformatics, 39(2):btad046, 2023.

Zhongliang Zhou, Wayland Yeung, Saber Soleymani, Nathan Gravel, Mariah Salcedo, Sheng Li, and Natarajan Kannan. Using explainable machine learning to uncover the kinase–substrate interaction landscape. Bioinformatics, 40(2):btae033, 2024.

A Appendix

A.1 Related work

We have categorized approaches in this domain into two main areas: those focused on kinase-specific phosphorylation site prediction and those centered on kinase-protein interaction prediction.

A.1.1 Kinase-specific phosphorylation sites

In the first category, studies focus on predicting phosphorylation sites specific to particular kinases, kinase families or kinase groups. These methods often leverage sequence-based features, structural properties, and evolutionary information to identify phosphorylation sites associated with a specific kinase. In this category family of algorithms, GPS Zhou et al. (2004); Xue et al. (2011); Liu et al. (2012); Guo et al. (2020); Wang et al. (2020); Chen et al. (2023) uses several models integrated to predict different groups and number of p-sites for each peptide with size of 15 amino acids. KinasePhos3.0 Ma et al. (2023) used 771 predictive models, developed at different levels, including kinase group, family, and individual kinase levels (SVM and XGBoost) with SHAP-based features to predict kinase-specific phosphorylation sites from 15-residue sequence windows surrounding phosphorylation sites, to predict number of positions and different kinase levels, MusiteDeep Wang et al. (2017) used convolutional neural network, DeepPhos Luo et al. (2019) used densely connected convolutional neural network with different window sizes of 25, 33, 51 peptides, PhosIDN Yang et al. (2021) integrates sequence features extracted through a self-attention-enhanced CNN and protein-protein interaction embeddings processed via a deep neural network, combining them with a bilinear module to predict general and kinase-specific phosphorylation sites.

A.1.2 Protein-kinase interaction

The second category addresses the broader challenge of protein-kinase interaction prediction, where the goal is to identify interactions between kinases and their substrate proteins. This task often involves integrating sequence embeddings, structural data, and sometimes even contextual biological information to predict interaction patterns. Phosphormer-ST Zhou et al. (2024) and Phosphormer Zhou et al. (2023) both tried to predict substrate-kinase interaction. Phosphormer used transformer-based architecture and Phosphormer-ST finetuned on ESM2-650M Acharya et al. (2018) parameters. It uses a shared encoder to generate embeddings for both kinase and peptide sequences.

A.2 ARCHITECTURE

The proposed architecture integrates an encoder-decoder framework, where each component plays a distinct role in processing and generating sequences. The encoder, denoted as $f_{\text{enc}}(\cdot)$, processes the input sequence of amino acid tokens $x = (x_1, x_2, \ldots, x_N)$ to produce contextual embeddings h_{enc}:

$$h_{\text{enc}} = f_{\text{enc}}(x)$$

where $h_{\text{enc}} \in \mathbb{R}^{N \times d_{\text{enc}}}$, N is the sequence length, and d_{enc} is the encoder's hidden dimension. To enhance the contextual embeddings from the encoder, a learnable embedding layer $g_{\text{pos}}(\cdot)$ is introduced, which adds positional information to the encoder's output. The final encoder representation h_{aug} is obtained by summing the positional embeddings p with the encoder output:

$$h_{\text{aug}} = h_{\text{enc}} + g_{\text{pos}}(p)$$

where $p \in \mathbb{R}^{N \times d_{\text{enc}}}$is the learnable positional embedding. Since the decoder operates at a potentially different hidden dimension d_{dec} a linear projection layer $W_{\text{proj}} \in \mathbb{R}^{d_{\text{enc}} \times d_{\text{dec}}}$is applied to align the encoder's output to the decoder's dimensionality:

$$h_{\text{proj}} = h_{\text{aug}} W_{\text{proj}}$$

where $h_{\text{proj}} \in \mathbb{R}^{N \times d_{\text{dec}}}$. The decoder, $f_{\text{dec}}(\cdot)$, receives two inputs: the projected encoder output h_{proj} and a sequence of task tokens t, where $t = (T_1, T_2, \ldots, T_m)$. Task tokens t are mapped to their respective embedding via a learnable embedding layer $g_{\text{task}}(\cdot)$:

$$e_{\text{task}} = g_{\text{task}}(t)$$

where $e_{\text{task}} \in \mathbb{R}^{m \times d_{\text{dec}}}$. Finally, the decoder attends to both the encoder's projected output using cross-attention and, the task token embeddings, producing the sequence of predictions $y = (y_1, y_2, \ldots, y_{N-m})$. The decoder is formulated as:

$$y = f_{\text{dec}}(h_{\text{proj}}, e_{\text{task}})$$

where the decoder predicts all non-task tokens conditioned on the projected encoder output and task token embeddings. The encoder portion of the model follows the exact architecture of the ESM-2 650m model. Its output is first combined with a learnable embedding, then projected from 1280 to 640 dimensions using a learnable linear projector. The decoder is a standard causal (autoregressive) Transformer with a hidden size of 640, a feed-forward dimension of 2560, GeLU activations and 16 attention heads, which iterated over 16 blocks. Additionally, FlashAttention2 Dao (2023) is employed to improve both training speed and memory efficiency.

A.3 DATASETS

The whole dataset was gathered from GPS, where its information on kinase families and groups is based on the kinome of *Homo sapiens*. All kinase sequences were extracted from Uniprot DB Consortium (2019) and Kinase.com. To create a diverse and non-redundant dataset, we applied CD-HIT clustering with a 70% sequence similarity threshold, grouping similar protein substrate sequences and retaining representative sequences from each cluster by removing similar substrate sequences. Representative positive pairs were chosen using the following criteria:

- **Cross-cluster selection:** Substrate-kinase pairs spanning different clusters were retained to preserve diversity across the dataset.
- **Within-cluster selection:** For substrates within the same cluster, only one unique kinase pair was kept, minimizing redundancy while ensuring distinct associations.

Each substrate is treated as an individual sample for model input, with one or more group labels and corresponding phosphorylation sites assigned per sample. Since substrates can be associated with multiple kinase groups and contain multiple phosphorylation sites, the problem is naturally formulated as a multilabel classification task. After preprocessing, the final dataset consisted of 5,385 unique substrates for training

and 969 unique substrates for validation. To ensure rigorous evaluation, we defined three distinct test sets, carefully designed to prevent any data contamination between the test, training, and validation sets:

Rare-Group Test Set. This set includes 14 samples from two rare kinase groups, *'RGC'* and *'PKL'*, which have a limited number of available samples. These groups were completely excluded from the training set to assess the model's ability to generalize to underrepresented kinase groups. This dataset is specifically used for evaluating on phosphorylation site prediction.

GPS-Test Set. To have a direct comparison with existing methods such as GPS 6.0, we adopted the test set used in the GPS study. This dataset contains 146 samples of substrate-kinase pairs, including phosphorylation site (p-site) and kinase group annotations. All samples belong to the *'CMGC'* kinase group.

Validation Set. This set was created using a random split strategy, ensuring a balanced distribution of kinase groups across both the test and training sets. Additionally, substrates in this set were selected to have minimal sequence similarity to each other, providing a robust measure of the model's generalization performance. Table 6 presents the number of samples in each set, while Table 7 details the distribution of samples across kinase groups in each dataset.

While serine (S), threonine (T), and tyrosine (Y) are the primary residues studied in phosphorylation research and commonly targeted by prediction tools, other amino acids—such as histidine (H) and aspartate (D)—are also known to undergo phosphorylation, particularly in prokaryotic systems and specific eukaryotic contexts. In our datasets, we identified phosphorylation sites on additional amino acids. We excluded these sites from our positive sites and finalize each set. Table 8 presents the distribution of phosphorylation sites exclusively on S, T, and Y residues.

Table 6: Dataset statistics, including the number of samples, phosphorylation sites (p-sites), and kinase groups for the training, validation, GPS test, and rare group test sets, along with overall dataset totals.

Dataset	Number of samples	Number of p-sites	Number of groups
All samples	6,514	13,401	12
Training set	5,385	10,621	10
Validation set	969	2,455	9
GPS-test	146	300	1
Rare-Group	14	25	2

Table 7: Distribution of samples across kinase groups for the training, validation, GPS test, and rare group test sets.

Group	Training set	Validation set	GPS-test	Rare-Group
AGC	1,446	231	-	-
Atypical	270	58	-	-
CAMK	653	96	-	-
CK1	100	27	-	-
CMGC	1,466	264	146	-
Other	491	99	-	-
STE	211	34	-	-
TK	677	149	-	-
TKL	68	111	-	-
RGC	-	-	-	2
PKL	-	-	-	12
UNK	4	-	-	-

A.4 Experiments

A.4.1 Kinase Groups Classification

We analyzed the sequence embeddings of unique kinase sequences extracted from all samples in the GPS 6.0 dataset. For this analysis, kinase sequences from the *"RGC"*, *"PKL"*, and *"UNK"* groups were ex-

cluded. These sequences were processed using the pre-trained ESM-2 650M model to generate token-wise embeddings, with a maximum sequence length of 2048. After extracting the model's outputs, we removed the beginning-of-sequence (BOS) and end-of-sequence (EOS) tokens and applied average pooling to obtain fixed-length representations of dimension 1280, matching the model's embedding size.

Table 8: Distribution of positive sites across kinase groups for the training, validation, GPS test, and rare group test sets.

Dataset	**Serine (S)**	**Threonine (T)**	**Tyrosine (Y)**
Training set	338,209	208,927	99,645
Validation set	65,872	39,845	18,054
Rare groups	958	632	317
GPS Test	9,913	5,998	2,449

Next, we employed t-SNE and UMAP for dimensionality reduction, enabling visualization of the embeddings in a two-dimensional space according to their group assignments. Given that the group labels for the kinase sequences were known, these labels were used to show the clusters visually on t-SNE and UMAP graphs with different colors. Also, we calculated unsupervised clustering metrics, such as the silhouette score and the Calinski-Harabasz index. We repeated the entire process for the fine-tuned ESM-650M checkpoint that is trained on the training set of kinase group classification labels. Similar to the first part, we performed dimensionality reduction visualizations and computed clustering metrics to evaluate the differences of both visual and unsupervised clustering metrics. The results are illustrated in Figures 3 and 4 and Table 2.

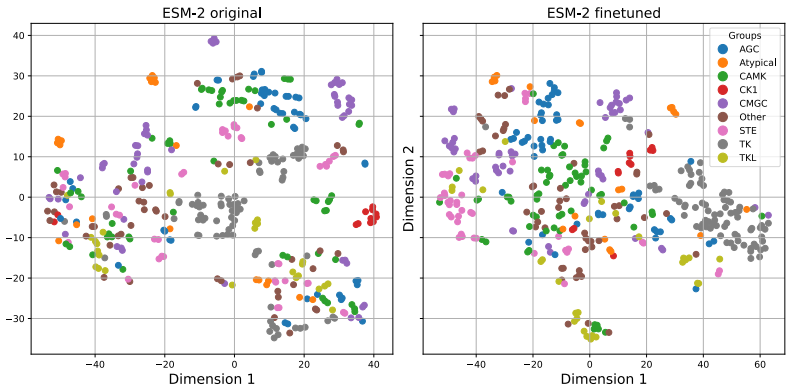

Figure 4: t-SNE visualization of unique kinase sequences on the original and fine-tuned checkpoints of ESM-2 650m.

A.4.2 Protein-Kinase Interaction

The positive data pairs are unique protein-kinase pairs with information of full protein sequence and full kinase sequence. The novelty of our method lies in the strategic selection of hard negative samples. Inspired from data pruning Sorscher et al. (2022) method, hard negative samples preparation relies on choosing kinase sequences with embeddings that exhibit minimal Euclidean distances to other kinases.

To construct the negative pairs, we iterated through all unique substrates in the dataset, processing each substrate individually. For each substrate, we first identified its associated positive kinases from the positive pairs. Using pre-computed embeddings, we calculated the Euclidian distance between each unique kinase embedding in the dataset to form the distance map. From these distances, we selected the top k closest kinases, excluding the positive kinase itself, to form k hard negative pairs for the substrate (Figure 5). This process was repeated for every substrate, ensuring that each one was paired with both its positive kinases and the most challenging negative kinases based on all kinase embeddings. By focusing on embedding distances

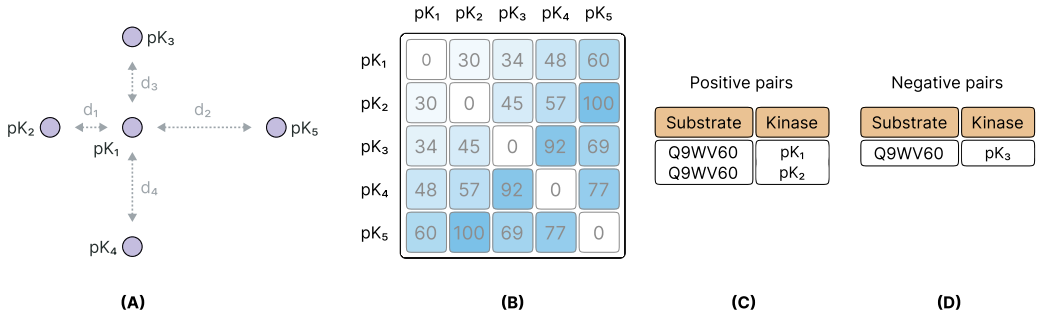

Figure 5: (A) Visualization of kinase embedding distances: The diagram depicts a substrate (center, pK_1 and its surrounding kinases (pK_2, pK_3, pK_4), and pK_5, illustrating the Euclidean distances (d_1, d_2, d_3, d_4) between their embeddings in the learned space. (B) Distance matrix of kinase embeddings: A heatmap representing pairwise Euclidean distances between kinase embeddings. Lighter shades correspond to shorter distances, indicating higher similarity, while darker shades represent greater dissimilarity. (C) Positive sample pairs: Examples of positive protein-kinase pairs (e.g., substrate Q9WV60 with kinases pk_1 and pk_2, representing known functional phosphorylation interactions. (D) Negative sample pairs: Examples of hard negative protein-kinase pairs (e.g., substrate Q9WV60 with kinase pk_3, selected based on minimal Euclidean distances in the embedding space to encourage robust discriminative learning.

to identify closely related negatives, this approach ensured a challenging dataset that effectively trained the model to distinguish between positive and negative pairs. We evaluated the impact of training protein-kinase interaction with hard negative samples versus random negative samples, focusing exclusively on the positive samples in the validation set. The results are summarized in Table 9.

During the training, we set a maximum context length of 1,280 tokens for the combined substrate-kinase sequences, truncating them when necessary to fit within this limit.

We evaluated GPS 6.0 Chen et al. (2023) and PhosphormerST Zhou et al. (2024) using our own validation set. For GPS 6.0, substrate sequences were provided as input, and in contrast to PhosphormerST, GPS 6.0 does not support kinase domain sequence as the input. To add the information of kinases during the prediction of GPS 6.0, we have selected each group and its specific kinase name in the webserver. The method generated a score and cutoff value for each phosphorylation site prediction. An interaction between the substrate and kinase was valid if the model predicted at least one phosphorylation site in the substrate sequence. On the other hand, PhosphormerST takes peptides and kinase domains as inputs and generates a prediction score for each sample. We created peptides with a length of 15 and included kinase domain information as the other part of the input. Interactions were considered valid if the score exceeded 0.5.

Table 9: Comparison of different negative sampling strategies for protein-ligand interaction training, evaluated on the positive samples of the validation set.

Negative sample strategy	Random	Hard
F1 score	0.771	0.8325

A.5 Protein-Kinase Phosphorylation Site Prediction

In contrast to other common approaches, a key objective of this paper was to input the complete substrate and kinase sequences into the model, enabling phosphorylation site predictions based on comprehensive sequence information. During the experiment, we set the total sequence length (including both substrate and kinase sequences) to 2,048 tokens for the encoder, truncating kinase sequences as needed to fit within this limit. The batch size was configured to accommodate 98,304 tokens per iteration. It is important to highlight that, although the self-supervised tasks could have been entirely excluded from the fine-tuning stage, retaining a subset of these samples led to a noticeable improvement in the model's performance on protein-kinase phosphorylation site prediction. We also found that without the self-supervised checkpoint,

the model's performance in phosphorylation site prediction dropped sharply, reaching an F1 score of less than 0.1 on the validation set, highlighting the necessity of pre-training for maintaining predictive accuracy.

We compared our results with two phosphorylation prediction tools, GPS 6.0 and KinasePhos3 Ma et al. (2023). Predictions with scores above 0.7 were considered true positive, and the predicted phosphorylation sites were compared to experimentally validated sites. To generate the results, we selected each specific kinase name individually on both their tools.

A.6 Availability of Resources

To support further research and development in the field, we make our codes, trained models, datasets, and a Python package publicly available for use by researchers and the broader scientific community. These resources can be accessed at this GitHub repository.

A.7 Broader Impact

By improving the accuracy of protein-kinase phosphorylation site prediction, our approach deepens our understanding of cellular signaling and its role in disease mechanisms while accelerating drug discovery, particularly for rare or less-characterized kinase families. Accurate phosphorylation site prediction is essential because phosphorylation plays a central role in diseases like cancer, neurodegenerative disorders, and metabolic syndromes. Enhanced prediction capabilities support biomarker discovery and precision medicine by guiding targeted experiments and enabling the design of more selective kinase inhibitors with fewer off-target effects. This not only improves therapeutic efficacy but also helps in minimizing side effects, thereby advancing patient care.

Beyond its direct clinical implications, our work bridges AI and biomedical research to benefit both the scientific community and society at large. By making our code, models, and datasets publicly available, we encourage transparency, reproducibility, and collaboration. This open approach paves the way for future research in protein engineering, synthetic biology, and other areas where understanding phosphorylation dynamics is key.

We also recognize potential risks and limitations. While our model significantly improves prediction accuracy, it is designed to complement—not replace—experimental validation. Over-reliance on computational predictions without proper validation could misdirect research efforts. Therefore, we emphasize careful interpretation, transparent reporting of uncertainties, and the need for rigorous experimental follow-up.