OXFORD

Briefings in Bioinformatics, 2025, **26(3)**, bbaf242
https://doi.org/10.1093/bib/bbaf242
Problem Solving Protocol

PhosF3C: a feature fusion architecture with fine-tuned protein language model and conformer for prediction of general phosphorylation site

Yuhuan Liu[1,‡], Xueying Wang[2,3,‡], Haitian Zhong[4], Jixiu Zhai[5], Xiaojuan Gong[2,6], Tianchi Lu[2,*]

[1]Cuiying Honors College, Lanzhou University, 222 South Tianshui Road, 730000 Lanzhou, China
[2]Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong, Hong Kong 999077, China
[3]Department of Computer Science, City University of Hong Kong (Dongguan), No. 8, Kaohsiung Road, Songshan Lake High-Tech Industrial Development Zone, 523808, Dongguan, China
[4]New Laboratory of Pattern Recognition (NLPR), State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences (CASIA), 95 Zhongguancun East Road, 100190, Beijing, China
[5]School of Mathematics and Statistics, Lanzhou University, 222 South Tianshui Road, 730000 Lanzhou, China
[6]Xi'an Jiaotong University, No. 76 West Yanta Road, Xi'an, Shaanxi, 710061, P.R. China

[*]Corresponding author. Tianchi LU, Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong, Hong Kong 999077, China, Tel: +86-13239620274, E-mail: tianchilu4-c@my.cityu.edu.hk
[‡]The authors wish it to be known that, in their opinion, the first two authors (Yuhuan Liu, Xueying Wang) should be regarded as joint first authors.

Abstract

Protein phosphorylation, a key post-translational modification, provides essential insight into protein properties, making its prediction highly significant. Using the emerging capabilities of large language models (LLMs), we apply Low-Rank Adaptation (LoRA) fine-tuning to ESM2, a powerful protein large language model, to efficiently extract features with minimal computational resources, optimizing task-specific text alignment. Additionally, we integrate the conformer architecture with the feature coupling unit to enhance local and global feature exchange, further improving prediction accuracy. Our model achieves state-of-the-art performance, obtaining area under the curve scores of 79.5%, 76.3%, and 71.4% at the S, T, and Y sites of the general data sets. Based on the powerful feature extraction capabilities of LLMs, we conduct a series of analyses on protein representations, including studies on their structure, sequence, and various chemical properties [such as hydrophobicity (GRAVY), surface charge, and isoelectric point]. We propose a test method called linear regression tomography which is a top-down method using representation to explore the model's feature extraction capabilities. Our resources, including data and code, are publicly accessible at https://github.com/SkywalkerLuke/PhosF3C

Keywords: protein phosphorylation; large language model; LoRA; conformer

Introduction

Protein phosphorylation is one of the most crucial post-translational modifications (PTMs). It plays an essential role in regulating various cellular processes. These processes include signal transduction, cell cycle control, metabolism, apoptosis, and protein–protein interactions. Phosphorylation occurs through the addition of a phosphate group to specific amino acids, typically serine, threonine, or tyrosine. This modification can cause structural changes in proteins. It can also modulate enzymatic activity or affect binding affinity and localization within the cell. Phosphorylation is particularly significant in cellular signaling pathways. It acts as a molecular on/off switch to control key processes such as cell division, immune responses, and gene expression [1, 2].

Phosphorylation's role is especially prominent in cancer. Aberrant activation of kinase-driven pathways, such as EGFR and HER2, leads to uncontrolled cell proliferation and tumor growth. This makes kinase inhibitors a vital class of targeted therapies [3, 4]. In neurodegenerative diseases, abnormal phosphorylation of proteins like tau contributes to pathological aggregates. These findings offer insights into therapeutic approaches [5, 6]. In cardiovascular diseases, phosphorylation of regulatory proteins affects heart muscle contractility. Targeting these pathways holds potential for treating heart failure [7, 8]. Given its extensive involvement in disease mechanisms, predicting phosphorylation sites is crucial. This helps identify biomarkers, understand pathological processes, and design effective therapeutic interventions, especially for complex diseases like cancer, neurodegenerative disorders, and heart disease [9].

In recent years, the study of PTMs, particularly phosphorylation, has advanced significantly. This progress is due to the development of high-throughput technologies and machine learning models. These advancements have enabled the identification and mapping of phosphorylation sites at an unprecedented scale. New approaches, such as mass spectrometry and phosphoproteomics, are pivotal in understanding the dynamic nature of phosphorylation. These methods reveal its variations across different cell types and conditions [10–13]. Computational methods have also evolved. Recent models achieve greater accuracy in predicting phosphorylation sites and their roles in various diseases. These

Received: February 14, 2025. **Revised:** April 20, 2025. **Accepted:** May 1, 2025

developments provide deeper insights into how phosphorylation regulates cellular functions and contributes to disease mechanisms [14, 15].

Several models have demonstrated notable performance in protein phosphorylation prediction. Each model has unique strengths and limitations. DeepPSP performs well by utilizing both global and local information for complete protein sequences. However, its effectiveness decreases when analyzing protein fragments, as it struggles without full sequence context [16]. MusiteDeep uses rotational attention to extract dual-dimensional sequence features. Despite this, its reliance on one-hot encoding results in feature loss during embedding, limiting its ability to capture complex sequence information [17]. PtransIPS excels in feature extraction with its dual-embedding approach. It combines sequence and structural information using ProtTrans [18] and EMBER2 [19]. However, it faces significant challenges in computation time and resource requirements when processing large datasets [20].

The rapid evolution of large language models (LLMs) has provided powerful tools for protein phosphorylation site prediction. These models leverage emergent capabilities to automatically extract complex features from protein sequences. They capture subtle relationships between amino acids and phosphorylation sites [21]. LLMs deliver enhanced performance in feature extraction and deep contextual understanding, making them more efficient and accurate than traditional methods. Fine-tuning techniques, such as LoRA, further improve their efficiency. This positions LLMs as a promising approach for advancing phosphorylation prediction [22, 23].

Based on the protein language model ESM2 [24], our model addresses the limitations of existing approaches by achieving highly efficient feature extraction. It improves the prediction of short protein sequence segments while significantly reducing storage and computational resource requirements. This is accomplished through LoRA fine-tuning [22], ensuring task alignment with minimal computational overhead. The model is further enhanced with the conformer architecture [25], which captures both global and local features. Feature coupling unit (FCU) is used to facilitate dynamic interactions between these features, an aspect often overlooked by previous models. Additionally, we applied our model to other protein-related tasks, further demonstrating the high generalizability of this framework. Across various tasks, the model showed strong feature extraction capabilities, proving its effectiveness and adaptability in multiple protein prediction challenges.

Lastly, we conducted a series of studies on protein representations generated by ESM2, applying traditional machine learning methods to analyze specific features. Using techniques such as random forests [26] and multiple linear regression [27], we developed a top-down analysis approach called linear regression tomography (LRT). Inspired by RePE [28, 29], this method identifies the principal directions of properties within the representations. This approach offers a potential pathway for improving model interpretability.

Materials and methods

Dataset

Dataset overview

Our dataset is divided into two parts: general datasets and task-specific datasets.

General datasets

The general dataset used in our study is a comprehensive fusion of data sourced from MusiteDeep [17] and DeepPSP [16]. The dataset integrate phosphorylation site information from key databases including UniPort/SWISSPROT [30], PhosphoSitePlus [31, 32], and Phospho.ELM [33]. This dataset includes general phosphorylation, including kinase-specific data, providing a rich resource for phosphorylation prediction tasks.

In terms of structure, although we incorporate kinase-specific information from several families (CDK, MAPK, CK2, PKA, PKC, AGC, CMGC, CAMK), we opted not to explicitly train separate models for each kinase family, in contrast to methods used in some other studies. This approach helps to maintain model generalization across different types of phosphorylation events while still benefiting from the rich diversity provided by kinase-specific information.

The dataset was then split into training and testing sets. For evaluation, a portion of the dataset was also reserved for independent testing. These methods offer a robust way to measure model performance across various conditions.

In addition to the general datasets, we utilized two specific datasets for phosphorylation site prediction in special tasks: DeepIPS [34] and PhosAF [35].

Table 1. Overview of phosphorylation site data across different datasets

Dataset	MusiteDeep	DeepPSP	PhosAF	DeepIps
S Positive	31 002	132 101	1029	974
S Negative	450 318	587 387	1298	974
S Total	481 320	719 488	2327	1948
T Positive	5473	52 274	245	105
T Negative	287 673	394 233	474	105
T Total	293 146	446 507	719	210
Y Positive	1930	32 213	77	21
Y Negative	135 914	149 500	81	21
Y Total	137 844	181 713	158	42

DeepIPS dataset

The DeepIPS dataset [34] consists of phosphorylation sites from human A549 cells infected with SARS-CoV-2, comprising data from the literature. Protein sequences were truncated into 33-residue segments centered on serine (S), threonine (T), or tyrosine (Y). Positive samples were defined as phosphorylated, while non-phosphorylated segments were treated as negative samples.

PhosAF dataset

The PhoSAF dataset [35] contains phosphorylation sites from human proteins, sourced from UniProt/Swiss-Prot, PhosphoSitePlus, and Phospho.ELM. CD-HIT [36] was used to ensure that the similarity of sequence was less than 40%. Protein fragments centered on serine (S), threonine (T), or tyrosine (Y) were extracted.

The detailed information and sample proportions for each dataset, including the number of positive and negative samples across Serine (S), Threonine (T), and Tyrosine (Y) phosphorylation sites, are presented in Table 1. The overall dataset composition and the positive and negative sample ratios for the merged general dataset, PhosAF, and DeepIps datasets are visually summarized in Supplementary Fig. 2.

Dataset splitting training validating and testing

After merging the general datasets, we randomly split the dataset into a general training set and a general testinng set. Specifically, we isolated 1500 independent (10% of the whole general dataset), full-length sequences to form the general test set, while the

Table 2. Training data overview for serine (S), threonine (T), and tyrosine (Y) phosphorylation sites

Train set	Serine (S)	Threonine (T)	Tyrosine (Y)
Positive	122 074	47 122	29 303
Negative	580 626	395 141	167 830
Total	702 700	442 263	197 133

Table 3. Test data overview for general, PhosAF, and DeepIps datasets

Test set	Merged general	PhosAF	DeepIps
S Positive	13 575	1029	974
S Negative	67 466	1298	974
S Total	81 041	2327	1948
T Positive	5663	245	105
T Negative	45 107	474	105
T Total	50 770	719	210
Y Positive	3366	77	21
Y Negative	18 432	81	21
Y Total	21 798	158	42

remaining sequences were used as the general training set. For the other two datasets, PhosAF [35] and DeepIPS [34], we used only the test portions that were originally defined in their respective datasets.

Next, we randomly split the training set into training and validation sets after separating the testing set. The validation set, which accounts for approximately 10% of the remaining dataset, is used for early stopping during model training to prevent overfitting. In the Supplementary Table 4, we report the mean and variance of the 10-fold cross-validation experiments.

Regarding the training set, due to the class imbalance where negative samples far outnumber positive samples, we applied a strategy to balance the dataset by repeating the entire positive sample set until the number of positive samples exceeds the number of negative samples, and then truncating the positive samples to match the number of negative samples. This ensures an equal number of positive and negative samples for training.

For model training, we only used the general training set, and the train portions of the original PhosAF and DeepIPS datasets were excluded. This strategy allows us to evaluate the model's generalization ability across different test sets, emphasizing the robustness and transferability(different task) of the model developed on the general dataset.

Training set information is presented in Table 2, and test set information is presented in Table 3 and portion in Supplementary Fig. 2.

Method

Method overview

Our approach consists of two main steps designed to enhance protein phosphorylation site prediction through effective feature extraction and integration. The overall framework of our method is illustrated in Fig. 1.

Step 1: Initial fine-tuning with LoRA: We begin by employing LoRA [22] to fine-tune a simple architecture consisting of ESM2 [24] combined with two layers of MLP (multi-layer perceptron). This architecture is specifically used for the protein phosphorylation prediction task. The trained checkpoint is aligned with the identification task, leveraging the self-regressive encoding capabilities of the ESM2 model to capture the abstract features necessary for this task. This enables a rational allocation of weights to the relevant features.

Step 2: Feature extraction and integration: Next, we reintegrate the obtained checkpoint back into the ESM2 model, using it as a feature extractor for downstream tasks. We connect the Conformer architecture to this setup. The resulting feature matrix is divided into two parts and subjected to appropriate transformations before being processed through the Conformer's CNN [37] branch and Transformer [38] branch for local and global feature extraction, respectively. Each branch comprises multiple layers of structures. To facilitate effective communication between the branches, we employ FCU for feature fusion, allowing timely information exchange between the two branches. Finally, we derive distinct evaluation parameters from each branch and utilize MLP to obtain predictive assessments for the phosphorylation sites.

LoRA fine-tuning

In conventional deep learning, the weights obtained through gradient descent enable the model's predictions to more closely align with the target labels, thereby reducing the loss function. For illustration, let's consider a multi-layer perceptron (MLP) and denote the forward propagation as model and input as x.

The update rule for the weights can be expressed as:

$$W' = W - \eta \nabla L$$

where W' is the updated weight, W is the original weight, η is the learning rate, and ∇L is the gradient of the loss function.

After the update, we can expand the updated weight matrix into the original weight part and the gradient descent update part. Let the rank of the gradient be denoted as r. Thus, we have:

$$W' = W + \Delta W$$

where

$$\Delta W = -\eta \nabla L$$

Next, substituting this into the forward propagation gives us:

$$\begin{aligned} \text{Output} &= model(W'x) \\ &= model((W + \Delta W)x) \\ &= model(Wx + \Delta Wx) \end{aligned}$$

To analyze the change in predictions, we can derive the difference in predicted values before and after the update:

$$\Delta\text{Output} = model(W'x) - model(Wx)$$

Using the first-order Taylor expansion around W, we find:

$$\Delta\text{Output} \approx \nabla model(Wx) \cdot \Delta Wx$$

However, due to the substantial number of parameters in LLMs, storing and computing these updates can be resource-intensive. Therefore, we aim for a simpler representation of ΔW, proposing that it can be expressed as a product of two lower-rank matrices, A and B, with ranks r_a and r_b, respectively, where both are much smaller than r:

$$\Delta W = A \cdot B$$

Substituting this back into the forward propagation reveals that these can effectively be represented as two subsequent MLP

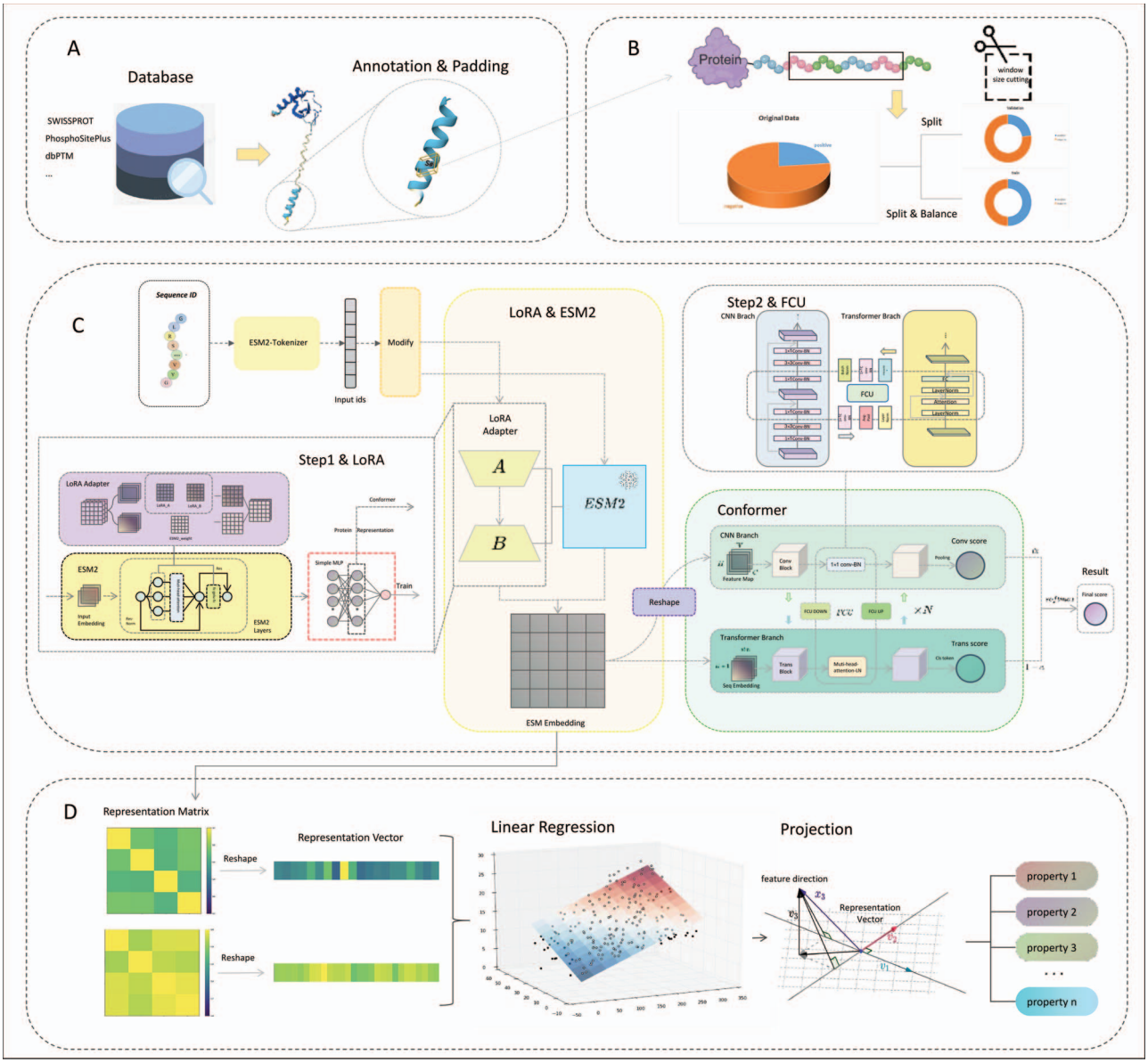


Figure 1. A and B show the data extraction, preprocessing, and generation of training and validation sets. B specifically handles the extraction of window-based phosphorylation site counts from labeled proteins, creating the training and validation sets with balanced data for the training set. C illustrates the model architecture and training process: step 1 involves fine-tuning ESM2 with LoRA for task alignment, and step 2 combines the fine-tuned ESM2 with the downstream Conformer model, where two branches perform separate predictions and FCU facilitates information interaction between them. D demonstrates the LRT process, where the feature matrix is first transformed into feature vectors, linear regression is applied to identify the principal directions of properties, and the projection shows how well the model extracts these properties.

forward propagations:

$$\text{Output} = model((W + A \cdot B)x) = model(Wx + A \cdot (Bx))$$

We can formally view ABx as x passing through two layers of MLP:

$$A \cdot (Bx) = f_1(f_2(x))$$

where f_1 and f_2 represent the two layers of the MLP. We denote these two layers of MLP as the LoRA adapter.

Based on this concept, the LoRA adapter can be applied in any context where MLPs are utilized, such as in the feedforward networks of transformers, or in the attention mechanisms' Q, K, V matrices, as well as in the final output projection layer.

Conformer

CNN branch

The CNN branch follows a feature pyramid structure where the resolution of feature maps decreases with depth while the number of channels increases. The branch is divided into four stages, each consisting of multiple convolution blocks. Each convolution block contains several bottlenecks. A bottleneck includes a 1×1 down-projection convolution, a 3×3 spatial convolution, a 1×1 up-projection convolution, and a residual connection between input and output. Unlike transformers, which project sequence representation patches into vectors in a single step, potentially losing local details, CNNs apply sliding convolution kernels over feature maps, preserving fine local features. This allows the CNN branch to provide detailed local information for the Transformer branch.

Transformer branch

The Transformer branch, inspired by ViT [39], consists of multiple transformer blocks. Each block includes a multi-head self-attention module and an MLP block, both of which are preceded by LayerNorm and include residual connections. A class token is added to the patch embeddings for classification purposes. Since the CNN branch encodes both local features and spatial information, positional embeddings are not required, allowing the model to handle high-resolution representation more efficiently.

FCU

To bridge the gap between the local features from the CNN branch and the global representations from the Transformer branch, we introduce FCU. FCU aligns the dimensionality of the CNN feature maps (C × H × W) with the Transformer patch embeddings ((K + 1) × E, where 1 represents the classification token adding to the head). Here, C denotes the number of channels in the CNN branch, while H and W correspond to the height and width of the representation map in the CNN branch. K represents the length of the protein sequence, and E denotes the embedding dimension of tokens in the Transformer branch. It also uses 1×1 convolutions and down-sampling/up-sampling modules (average pooling and interpolating)to ensure compatibility between the branches. FCU progressively closes the semantic gap between the CNN's local convolutional features and the Transformer's global self-attention mechanisms, enabling effective feature fusion across all layers except the first.

Linear regression tomography

LRT is a method designed to evaluate a model's feature extraction capability for specific properties by isolating feature directions within representation matrices that are strongly correlated with those properties. By constructing linear mapping relationships between target properties and hidden layer representations, LRT disentangles a vector (feature direction) highly relevant to the target properties from high-dimensional representation spaces, enabling preliminary diagnostic analysis of the model's feature extraction mechanisms. As LRT operates solely on intermediate representations generated during forward propagation, it is architecture-agnostic and task-agnostic, making it generalizable across various neural network architectures and task scenarios.

Feature importance via random forests

We start by identifying key features that are important for the phosphorylation site prediction task. Using a Random Forest model (more details are present in Supplementary Fig. 1), we determine the most influential properties of proteins, such as biochemical characteristics like surface charge or hydrophobicity. Once these important features are identified, we obtain their real values (measured experimentally) or predicted values (derived from a specialized prediction model) for a set of proteins.

Representation and feature direction

For each protein, we extract its representation using the ESM2 [24]. Our objective is to find a principal direction within the representation matrix that best correlates with the specific biochemical property of interest. To do this, we apply multivariate linear regression to map the feature values to the corresponding direction in the representation space. Mathematically, this can be formulated as:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \|y - \beta x\|_2$$

where y represents the observed or predicted feature values, X is obtained by flattening the ESM2 representation matrix, β is the vector representing the direction in the representation space that corresponds to the feature. In our experiments, we use the 128 samples with the highest and lowest feature values to ensure a clear separation of extremes.

Projection and feature approximation

After finding the optimal direction β through regression, we project the protein's ESM2 representation onto this direction. The value of the feature for each protein is then approximated as the projection of its representation onto the direction β:

$$\hat{y} = X\beta$$

This projected value $\hat{y}$ serves as an estimate of the protein's biochemical property. By evaluating the quality of feature directions extracted through LRT, we can preliminarily assess the reliability of the model's learned representations. This analysis directly correlates the fidelity of the extracted feature direction with the model's ability to capture task-critical patterns.

Performance evaluation

In our evaluation, we utilize a range of metrics to assess model performance across different aspects of prediction quality:

- **AUC (area under the curve)**:

$$\mathrm{AUC} = \int_0^1 TPR(FPR^{-1}(x))dx$$

- **Accuracy (ACC)**:

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Matthews correlation coefficient (MCC)**:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

- **F1 score**:

$$\mathrm{F1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

- **Recall (REC)**:

$$\mathrm{Recall} = \frac{TP}{TP + FN}$$

- **Precision (PRE)**:

$$\mathrm{Precision} = \frac{TP}{TP + FP}$$

Table 4. Performance index on general dataset categorized by phosphorylation site type

Type	Model	PR–AUC	ROC–AUC	ACC	MCC	F1	Recall	Precision
S	DeepPSP	0.4313	0.7768	0.6703	0.3031	0.4305	0.7441	0.3029
	MusiteDeep2.0	0.4224	0.7737	0.6275	0.2904	0.4167	0.7941	0.2824
	TransPhos	0.4315	0.761	0.6841	0.2926	0.4260	0.7000	0.3062
	PTransIPS	0.2443	0.6687	0.6678	0.2460	0.3918	0.6519	0.2800
	PhosF3C	0.4817	0.7952	0.7281	0.3449	0.4648	0.7050	0.3467
T	DeepPSP	0.3212	0.7542	0.7355	0.2453	0.336	0.5999	0.2333
	MusiteDeep2.0	0.3268	0.7578	0.6625	0.2387	0.3202	0.7127	0.2065
	TransPhos	0.2662	0.7069	0.7537	0.1985	0.3041	0.4824	0.2220
	PTransIPS	0.1592	0.6411	0.6559	0.1782	0.2822	0.6133	0.1832
	PhosF3C	0.3347	0.7630	0.7254	0.2591	0.3435	0.6438	0.2342
Y	DeepPSP	0.2968	0.6987	0.6493	0.2099	0.3571	0.6307	0.2491
	MusiteDeep2.0	0.2940	0.6955	0.6515	0.2024	0.3524	0.6141	0.2471
	TransPhos	0.2296	0.6262	0.6274	0.1429	0.3137	0.5514	0.2192
	PTransIPS	0.1900	0.5989	0.6018	0.1350	0.3058	0.5805	0.2076
	PhosF3C	0.3147	0.7136	0.6334	0.2263	0.3661	0.6857	0.2498

Definitions:

- *TP*: True positive
- *TN*: True negative
- *FP*: False positive
- *FN*: False negative

Result

Independent test of PhosF3C for phosphorylation site identification

To comprehensively evaluate the performance of our models, we compared them with five established phosphorylation site prediction tools, including DeepPSP [16], MusiteDeep [17], TransPhos [40], PtransIPs [20], and PhosF3C. Each model was trained on the general dataset and tested on three independent test datasets, following a consistent training and testing split as described in the *Dataset splitting training and testing* section. It is important to note that all models were trained with same window size 31 (sequence's global information was masked during training) and same training pattern. Predictions were evaluated using a phosphorylation determination threshold set at 0.5, ensuring uniformity across comparisons. This threshold selection avoids bias related to sample imbalance and allows for fair model performance assessment across different datasets.

The results are presented in Table 4 and Fig. 2. For serine (S) and threonine (T) sites, the PhosF3C model achieved the best performance across most metrics, with an AUC of 0.7952 for S and 0.763 for T sites. This surpasses other models like DeepPSP and MusiteDeep by margins of 1.84% and 2.15% for S sites, and 0.88% and 0.52% for T sites, respectively. For tyrosine (Y) sites, although DeepPSP showed competitive performance, the PhosF3C still demonstrated superior AUC (0.7136).

Performance on PhosAF Dataset For the PhosAF dataset [35], PhosF3C achieves the highest overall AUC values for serine (S) and threonine (T) sites, with scores of 0.9155 and 0.8934, respectively. These results are notably higher than the second-best performing models, indicating its superior ability to capture the underlying patterns of these phosphorylation sites. Additionally, it excels in other critical metrics, such as F1 (0.813 for S and 0.7292 for T) and MCC (0.6542 for S and 0.5733 for T), reflecting its balance between precision and recall, as well as its capacity to make accurate predictions. Even for the more challenging tyrosine (Y) site predictions, PhosF3C performs competitively with an AUC of 0.7178 and an MCC of 0.2872.

Performance on DeepIPs dataset On the DeepIPs dataset [34], PhosF3C maintains its strong performance. For serine (S) and threonine (T) sites, it achieves top AUC scores of 0.875 and 0.8584, respectively, which are complemented by solid F1 scores (0.7843 for S and 0.7897 for T). These results indicate that the model not only identifies true positive phosphorylation sites effectively but also minimizes false positives. While the AUC for tyrosine (Y) sites is slightly lower at 0.7302, it remains competitive compared to other models.

More detailed test results are provided in Supplementary Figs 4, 5 and Supplementary Tables 2, 3. This supplementary includes all relevant metrics and further analysis to ensure a comprehensive evaluation of their performance alongside the primary models.

Training with LoRA and FCU improves the performance of PhosF3C

In this section, we present the ablation study to evaluate the effectiveness of key components in our models, specifically focusing on the impact of LoRA [22] and FCU. All experiments were conducted on the general test set.

Our results (Table 5) highlight that the LoRA-enhanced model improves performance by selectively enhancing task-relevant abstract features, thus increasing their contribution to the prediction task. The PhosF3C with LoRA achieves an AUC of 0.7876 for S sites, 0.7606 for T sites, and 0.7011 for Y sites, surpassing the base conformer model (AUC: S: 0.7726, T: 0.7381, Y: 0.6901). The MCC values for LoRA (S: 0.3409, T: 0.2401, Y: 0.2121) are also higher than those of the conformer model (S: 0.2996, T: 0.2016, Y: 0.1976), indicating a stronger correlation with the true labels.

FCU significantly enhances performance beyond LoRA's improvements alone. Specifically, PhosF3C with LoRA achieves higher AUCs (S: 0.7952, T: 0.763, Y: 0.7136) compared to LoRA alone (AUC: S: 0.7876, T: 0.7606, Y: 0.7011). The MCC values also reflect this improvement, with the PhosF3C with FCU and LoRA attaining MCC scores of 0.3449 for S sites, 0.2591 for T sites, and 0.2263 for Y sites, compared to LoRA-only MCCs (S: 0.3409, T: 0.2401, Y: 0.2121).

The PR and ROC curves are shown in Fig. 3, clearly demonstrating that the combined strengths of LoRA and FCU enhance model

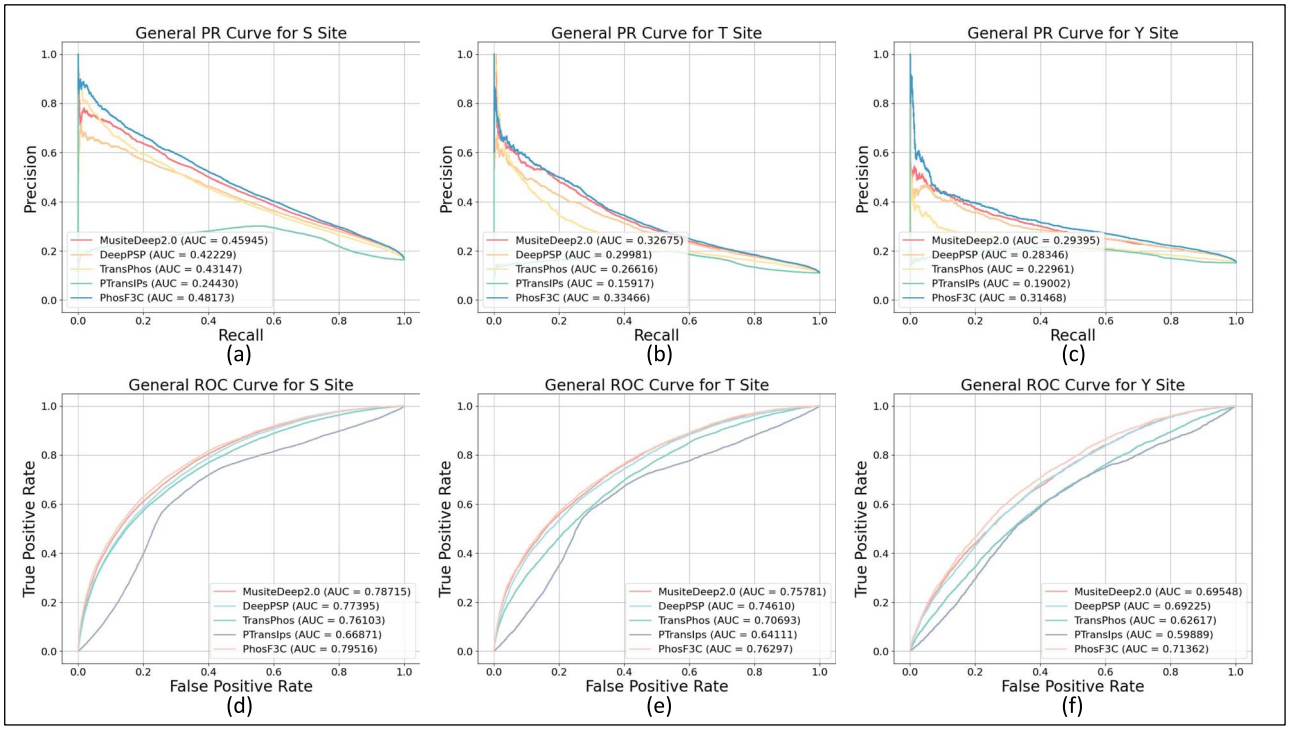


Figure 2. Figures (a–c) present the PR curves for serine (S), threonine (T), and tyrosine (Y) sites on the general test dataset, comparing the performance of the PhosF3C model with other currently well-performing protein phosphorylation prediction models. Figures (d–f) show the corresponding ROC curves for the same sites.

Table 5. Performance index on general dataset categorized by phosphorylation site type

Type	Model	PR–AUC	ROC–AUC	ACC	MCC	F1	Recall	Precision
S	Conformer	0.4285	0.7726	0.6576	0.2996	0.4265	0.7602	0.2964
	LoRA	0.4733	0.7876	0.7435	0.3409	0.4641	0.6628	0.357
	PhosF3C	0.4817	0.7952	0.7281	0.3449	0.4648	0.705	0.3467
T	Conformer	0.2988	0.7381	0.5584	0.2016	0.2855	0.7909	0.1742
	LoRA	0.3385	0.7606	0.6755	0.2401	0.3232	0.6947	0.2106
	PhosF3C	0.3347	0.763	0.7254	0.2591	0.3435	0.6438	0.2342
Y	Conformer	0.29	0.6901	0.5812	0.1976	0.3458	0.7169	0.2279
	LoRA	0.3031	0.7011	0.6096	0.2121	0.356	0.6988	0.2389
	PhosF3C	0.3147	0.7136	0.6334	0.2263	0.3661	0.6857	0.2498

performance. Additionally, Fig. 3 indicates that LoRA effectively aligns ESM2 [24] feature extraction with the task requirements.

Visualizing the feature extraction process of PhosF3C with 2D UMAP and attention map

To investigate the model's capability in distinguishing phosphorylation sites we performed a visualization of features extracted during different stages of training using uniform manifold approximation and projection (UMAP) [41] (as show in Fig. 3). Initially, embeddings from the ESM model showed limited capacity to differentiate between positive and negative phosphorylation sites, as the samples were largely intermingled and lacked discernible boundaries. This observation indicates that the raw ESM embeddings alone were not capable of filtering out the most relevant abstract features for this task. However, after incorporating LoRA [22], we observe a clear improvement in feature alignment, with task-relevant features gradually emerging. This indicates that LoRA plays a critical role in filtering out abstract features and directing the model's focus toward key elements essential for phosphorylation prediction.

In support of our viewpoint, Johnson *et al.* [42] showed that nearly 60% of the human Ser/Thr kinome is characterized by a strong preference for basic residues (arginine and lysine) upstream of the phospho-acceptor site. These positively charged residues create a favorable electrostatic environment that enhances kinase–substrate binding and catalytic efficiency.

To investigate LoRA's impact on feature extraction, we visualized the ESM2 attention maps before and after LoRA fine-tuning (see Fig. 4). Using the full Q9NYP9 sequence, we analyzed serine phosphorylation by extracting adjacent and non-adjacent sites, then plotted the attention weight distribution of a central serine toward its neighboring amino acids, and computed the overall weight distribution across all serine sites. We found that, after fine-tuning, the attention weights for RK residues near phosphorylation sites significantly increased. This observation aligns with the biochemical mechanism described in [42], where basic residues upstream of serine play a crucial role in substrate recognition via electrostatic interactions.

Furthermore, previous studies indicate that certain kinase families, such as the GRK and YANK families, utilize complementary

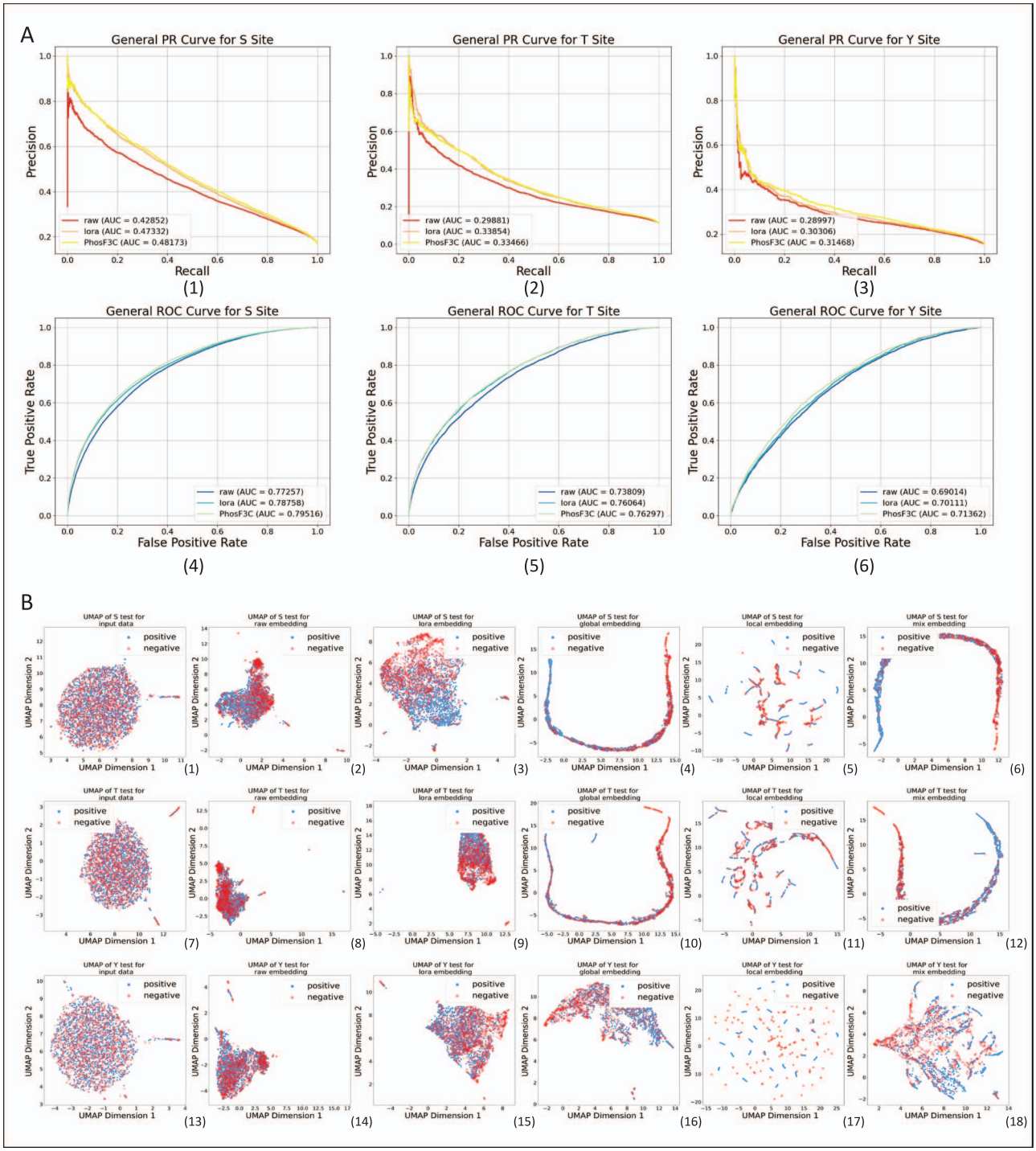


Figure 3. Ablation study for phosphorylation site prediction, showing PR and ROC curves for different model variations: ESM2 (raw checkpoint) + Conformer, ESM2 (finetuned by LoRA) + 2 MLP layers, and ESM2 (finetuned by LoRA) + Conformer. A shows the PR and ROC curves for serine (S), threonine (T), and tyrosine (Y) sites, respectively;B shows the UMAP visualization in a 2D representation during the training process, reflecting the distribution of protein features at different stages in low-dimensional space, with negative and positive samples annotated to highlight classification performance. The stages include: raw (pre-fine-tuned ESM2 embedding), lora (post-LoRA fine-tuning), global (features processed by the transformer branch), local (features processed by the CNN branch), and mix (a combination of both feature processes).

basic residues to enhance substrate specificity [42]. The increased attention to RK residues suggests that LoRA helps ESM2 [24] learn these biologically relevant features. Overall, our results provide strong evidence that LoRA enhances ESM2's ability to capture key electrostatic properties and kinase-specific interaction patterns essential for accurate phosphorylation prediction.

As we progress further, conformer's FCU exhibits its strength in facilitating the interaction between local and global information. This dual extraction from the CNN and Transformer branches results in a distinct separation of positive and negative samples, with both branches contributing complementary perspectives to improve classification accuracy.

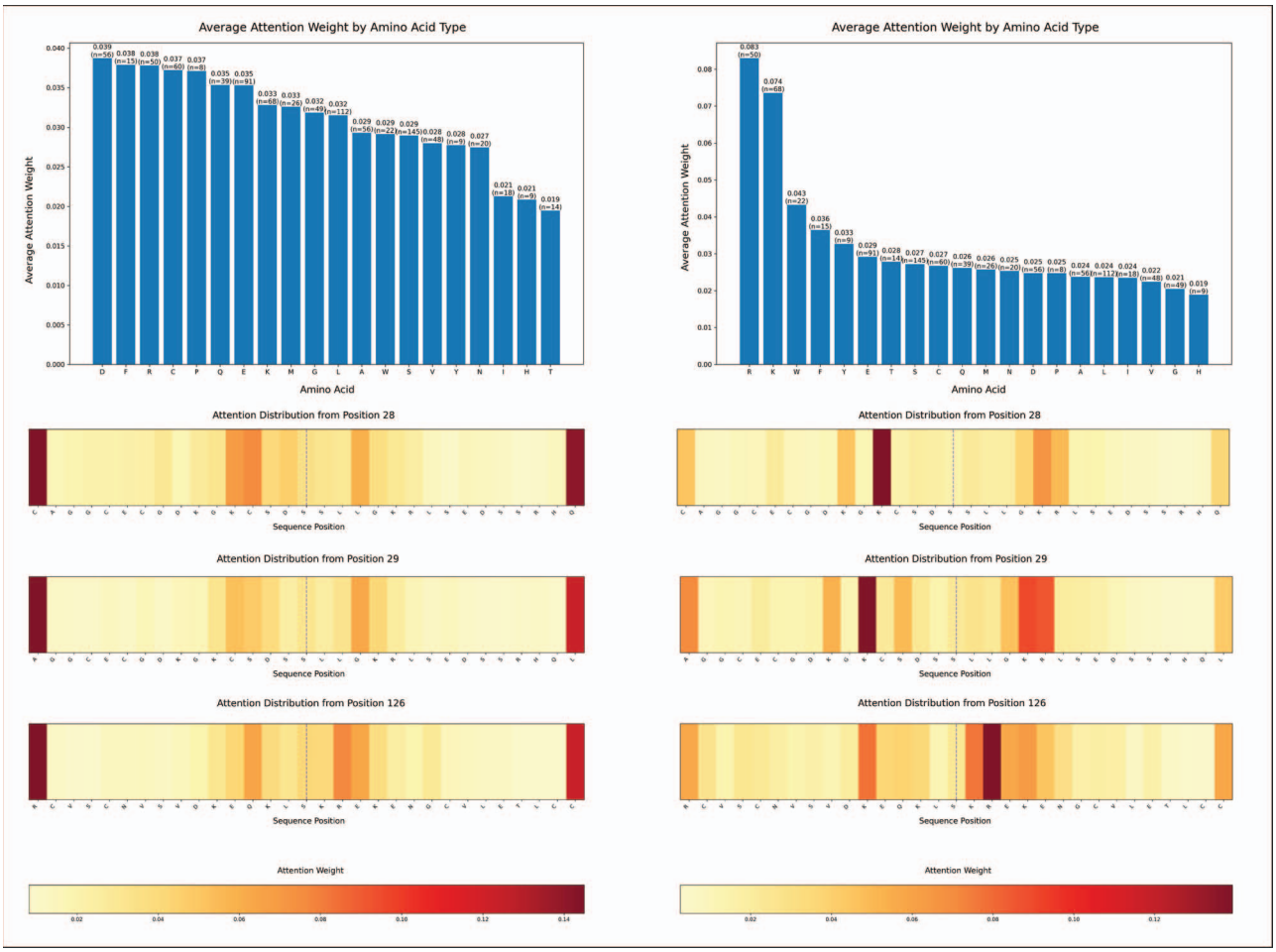


Figure 4. Attention maps of ESM2 for serine phosphorylation prediction on the full Q9NYP9 sequence before (left) and after (right) LoRA fine-tuning. The maps display the distribution of attention weights from a central serine (S) residue toward its neighboring amino acids. Notably, after LoRA fine-tuning, there is a marked increase in weights assigned to adjacent basic residues (arginine and lysine), highlighting the model's enhanced focus on electrostatic features that are critical for accurate phosphorylation prediction.

Applying LRT in feature extracting analysis

In our experiments, LRT was applied to investigate how varying window sizes—which directly influence the dimensions of the representation matrices—affect a model's ability to capture partial features. By analyzing the relationship between window size configurations and the quality of extracted feature direction, we evaluated the impact of this hyperparameter on feature extraction performance.

First, we predicted various physicochemical properties of the proteins using Biopython [43] over 65 window sizes. For each protein sequence, the values of these properties were distributed, and their importance for phosphorylation site discrimination was evaluated using a random forest classifier. We also calculate the property's information entropy [44] to ensure the properties contain a wealth of information that can be analyzed:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

where $H(X)$ represents the entropy of a property, x_i are the distinct property values, and $p(x_i)$ denotes the probability of occurrence of each value.

Based on this analysis, we selected the top three most important properties-h-Hydrophobicity (GRAVY), isoelectric point, and surface charge—as shown in Supplementary Fig. 3. These properties were standardized, with the following values for the three property:

- **Hydrophobicity (GRAVY)**:
  - Importance: S:0.743, T:0.752, Y:0.740
  - Entropy: S:0.376, T:0.348, Y:0.348
- **Isoelectric point**:
  - Importance: S:0.744, T:0.752, Y:0.741
  - Entropy: S:0.380, T:0.349, Y:0.349
- **Surface charge**:
  - Importance: S:0.744, T:0.753, Y:0.742
  - Entropy: S:0.377, T:0.348, Y:0.348

The standardization formula is:

$$z = \frac{x - \mu}{\sigma}$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

We then took 128 boundary values from the distribution of each property and paired them with the corresponding protein segments. The representations for these protein segments were extracted from the LoRA fine-tuned ESM2 model using embedding windows of sizes 31, 21, and 11, respectively. To ensure a consistent input length for the embedding process and eliminate potential overfitting caused by varying latent dimensions during LRT, all windows were symmetrically padded to a fixed length of 31 centered on the original target residue before being fed into ESM2. Then, we transform the embedding matrix into a vector representation and apply multivariate linear regression, using real-valued labels, to identify property directions within these embeddings.

Subsequently, all other protein samples were projected onto the identified property directions, and we derived their distributions. Finally, we compared these distributions to those predicted by Biopython (Supplementary Fig. 3) and computed the distribution differences. Based on these comparisons, we assess the quality of the extracted feature directions by evaluating how well they align with known biochemical property distributions.

The results highlight the effectiveness of LRT in analyzing the model's ability to extract meaningful features. Our findings indicate that as the window size increases, the feature directions obtained through LRT exhibit improved quality. This is evidenced by the decreasing discrepancy between the property distributions projected from these feature directions and the true distributions. Based on this trade-off analysis between feature fidelity and computational costs (GPU memory consumption and training time), we ultimately selected a window size of 31 for PhosF3C—a configuration that optimally balances high-quality feature extraction with manageable resource demands, ensuring both efficiency and usability. This exploration provides critical insights into the relationship between architectural hyperparameters and model capability, offering practical guidance for balancing efficiency and accuracy in real-world applications. Notably, this decision demonstrates how LRT-driven diagnostics can directly inform architecture design under concrete computational constraints.

Analyzing the reasons for suboptimal predictive capabilities

To investigate the reasons behind the model's insufficient predictive capabilities, we hypothesized that similarities among phosphorylation sites might contribute to the model's difficulties in distinguishing between them. To validate this conjecture, we conducted an experiment on the general test dataset.

We computed the representation matrices of two proteins fine-tuned with ESM2 [24] and performed matrix subtraction. We then calculated the Frobenius norm [45] $||A||_F$ of the contrast matrix as a measure of the difference between the two proteins:

$$||A||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$$

where $||A||_F$ is the Frobenius norm of matrix A, m is the number of rows in the matrix, n is the number of columns in the matrix, a_{ij} represents the element in the ith row and jth column of matrix A.

Since the representation matrices capture rich abstract features, not only sequence similarity but also chemical properties and structural similarities would be accounted for. We selected the two proteins with the low norms, identified by their IDs and positions, and analyzed their sequences and structures (see Fig. 5). This confirms the validity of measuring differences in this manner.

Next, we divided the 10% general test dataset into two parts based on their Frobenius norms: a high-norm group and a low-norm group. The differences in norms between any negative and positive samples in the high-norm group were set to be greater than or equal to a dynamically determined threshold (approximately allowing 0.4% of the original sample size in each group). The low-norm group was handled similarly.

The specific data obtained are as follows:

- **Serine (S) sites:**
  - High-norm group: $F > 56$ (positive: 886, negative: 171)
  - Low-norm group: $F < 39$ (positive: 192, negative: 502)
- **Threonine (T) sites:**
  - High-norm group: $F > 58$ (positive: 365, negative: 111)
  - Low-norm group: $F < 40$ (positive: 109, negative: 500)
- **Tyrosine (Y) sites:**
  - High-norm group: $F > 64$ (positive: 183, negative: 44)
  - Low-norm group: $F < 41$ (positive: 50, negative: 139)

Subsequently, we balanced the dataset by reducing the larger group, facilitating the subsequent experiments.

The classification results indicated a clear distinction between the two groups. To further visualize the effectiveness of this distinction, we employed 3D UMAP, which provided an intuitive representation of the data (see Fig. 5).

Subsequently, we performed predictions on both norm groups. The results showed significant differences in performance, highlighting that the high-norm group had markedly better predictive capabilities compared to the low-norm group (see in Fig. 5).

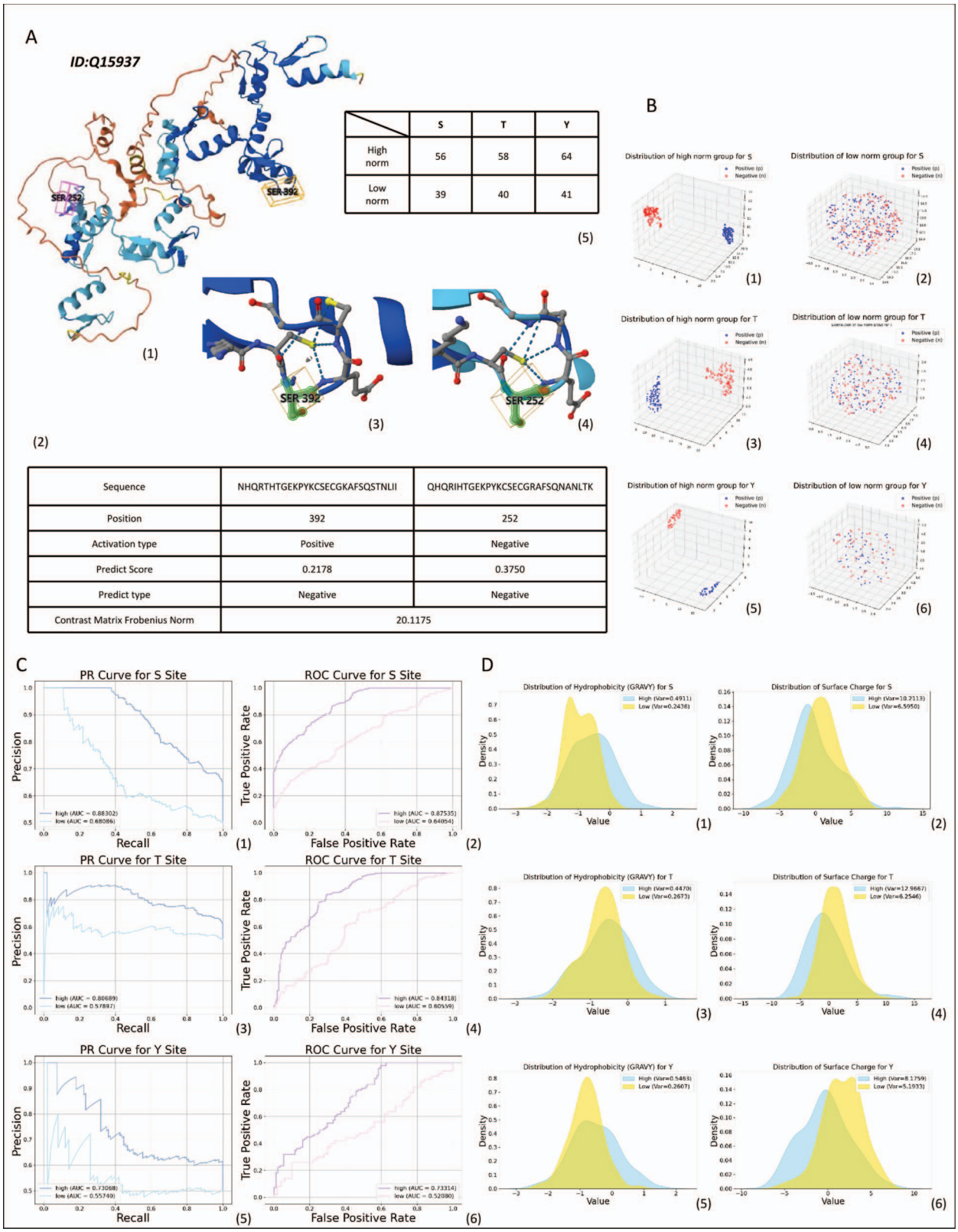
Further analysis of the distribution of chemical properties revealed notable variance in the density plots (see in Fig. 5), validating our hypothesis that the similarity among certain phosphorylation sites could hinder the model's ability to make accurate predictions. This suggests that the quality of the dataset plays a crucial role in model performance.

In conclusion, our findings indicate that the similarities among certain phosphorylation sites like chemical properties, structures significantly impact predictive capabilities.

Evaluation of PhosF3C across various protein tasks

To evaluate the generalization ability of the PhosF3C model across various protein-related tasks, we conducted experiments on three significant bioactivities: histone lysine crotonylation (Kcr) [46], methylation, and the sequential and spatial methylation fusion network (SSMFN) [47]. Each task represents distinct biological challenges. Kcr prediction focuses on identifying histone lysine crotonylation sites, which play key roles in cellular regulation and are implicated in various human diseases. Methylation prediction involves detecting glutamine methylation sites, a process crucial for gene regulation and cancer progression. The SSFMN task, designed to predict protein methylation sites, leverages neural networks to improve the efficiency and accuracy of PTM site identification.

The model's performance was evaluated using several important metrics. The results highlight that PhosF3C demonstrates significant improvements over other models in specific tasks. For example, in the SSMFN task, PhosF3C outperforms with an AUC



	S	T	Y
High norm	56	58	64
Low norm	39	40	41

(5)

Sequence	NHQRTHTGEKPYKCSECGKAFSQSTNLII	QHQRIHTGEKPYKCSECGRAFSQNANLTK
Position	392	252
Activation type	Positive	Negative
Predict Score	0.2178	0.3750
Predict type	Negative	Negative
Contrast Matrix Frobenius Norm	20.1175	

Figure 5. A presents a set of cases with lower and upper bounds of norm values for the high and low norm groups respectively, accompanied by a structural diagram. The norm information indicates that the protein differences within the high norm group will not be lower than the lower bound, and vice versa for the low norm group. B shows the distribution of the high and low norm groups after dimensionality reduction to 3D using UMAP. C presents the PR and ROC curves for the model's predictions on the high and low norm groups. D shows the density distribution and variance of the properties hydrophobicity (GRAVY) and isoelectric point predicted by biopython for the high and low norm groups.

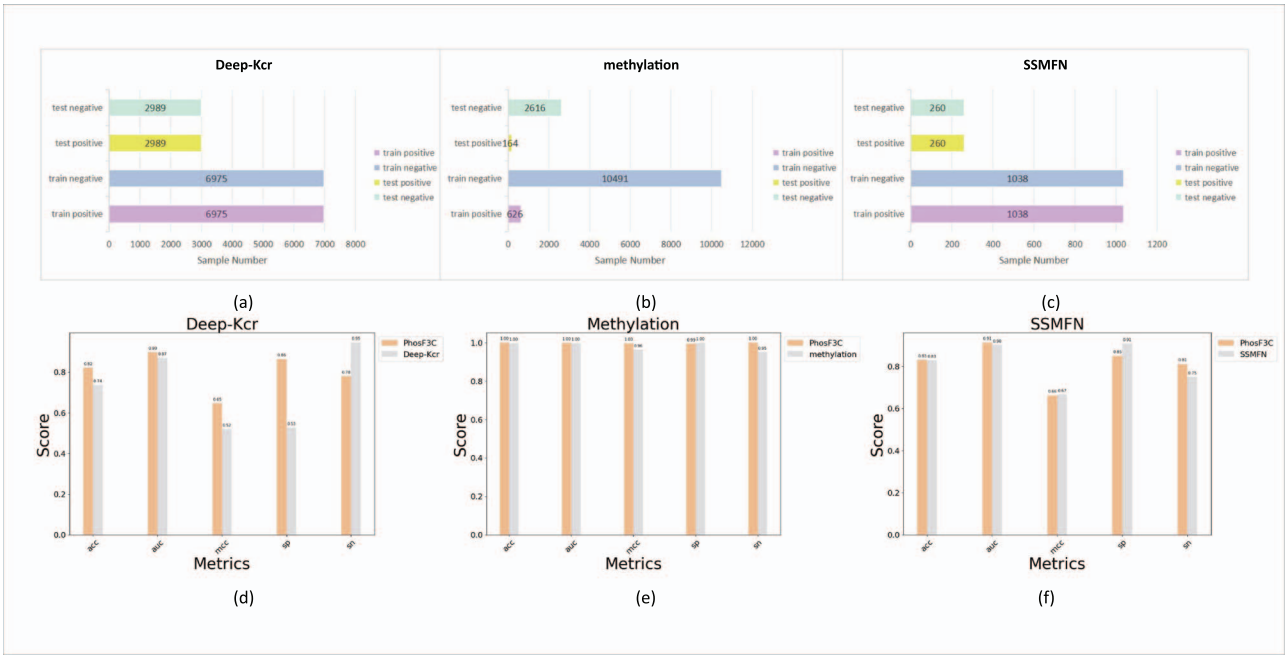


Figure 6. Figures (a–c) display the data information for three protein-related tasks: histone lysine crotonylation (Kcr), methylation, and the SSMFN, along with the distribution of positive and negative samples for each dataset. Figures (d–f) show the comparison of the PhosF3C model's performance with baseline models in these tasks.

of 0.9128 and accuracy of 0.8308, compared to SSMFN's AUC of 0.9011 and accuracy of 0.8288. In the Deep-Kcr task, PhosF3C also excels, achieving an AUC of 0.8989 and accuracy of 0.8225, surpassing Deep-Kcr's AUC of 0.8698 and accuracy of 0.7364. These results confirm that PhosF3C not only excels in individual tasks but also showcases robust generalization across a wide range of protein-related predictions. More details about the model performance were shown in Fig. 6.

Conclusion

In this study, we proposed a novel framework for phosphorylation site prediction that integrates CNN and transformer architectures to capture both local and global sequence features. By using ESM2 [24] embeddings, our model effectively represents protein sequences, enabling robust phosphorylation site prediction with state-of-the-art performance.

We also introduced FCU to enhance feature interaction and further improve model performance. Additionally, we applied LRT to evaluate the model's feature extraction capabilities. The results showed that the model captures relevant biochemical signals, with larger window sizes providing a better representation of physicochemical distributions.

Contributions efficient feature extraction and prediction: we developed a model that uses LoRA fine-tuning [22]} and the Conformer architecture. This combination enables efficient feature extraction, accurate prediction of short protein segments, and significant reductions in both storage and computational resource requirements.

High generalizability: our framework has been successfully applied to various protein-related tasks, demonstrating its versatility and adaptability across different protein prediction challenges. This highlights its reliability for broader applications in bioinformatics.

Interpretability enhancement: We introduced a top-down analysis method called LRT to examine protein representations derived from ESM2 [24] embeddings. This approach employs traditional machine learning techniques to identify the principal directions of properties within the representations, providing valuable insights that enhance model interpretability.

However, challenges remain. One key issue is the imbalance in datasets, which can introduce biases in large-scale protein analysis. Current sampling strategies may not fully address this concern. Additionally, the reliance on high-quality sequence embeddings limits the applicability of the model to incomplete protein sequences. Future work should focus on incorporating more experimental data, improving interpretability, and extending the framework to support multi-modal inputs, such as structural or interaction data.

In conclusion, our framework advances phosphorylation site prediction and holds potential for broader biological insights. Its applications in drug development and personalized medicine make it a significant step forward in computational biology and AI-driven research.

Key Points

- Model architecture: This study presents a novel phosphorylation site prediction model that combines LoRA fine-tuning, and the Conformer architecture. This integration enables efficient feature extraction and highly accurate predictions of short protein segments.
- Feature extraction with LRT: We utilized linear regression tomography (LRT) to assess the model's feature extraction capabilities. LRT effectively demonstrates how the model captures critical biochemical signals, while also improving overall interpretability.
- Analysis of structural and biochemical properties: our in-depth analysis investigates the role of structural, physicochemical, and biochemical properties in influencing phosphorylation site prediction. This approach

provides valuable insights into the underlying mechanisms of the task.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

None declared.

Data availability

All processed datasets are accessible for download at https://github.com/SkywalkerLuke/PhosF3C/tree/main/data. The code is accessible through GitHub via https://github.com/SkywalkerLuke/PhosF3C/.

References

1. Ardito F, Giuliani M, Perrone D. *et al.* The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy. *Int J Mol Med* 2017;**40**:271–80. https://doi.org/10.3892/ijmm.2017.3036
2. Rives A, Meier J, Sercu T. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci* 2021;**118**:e2016239118. https://doi.org/10.1073/pnas.2016239118
3. Arora A, Scholar EM. Role of tyrosine kinase inhibitors in cancer therapy. *J Pharmacol Exp Ther* 2005;**315**:971–9. https://doi.org/10.1124/jpet.105.084145
4. Madhusudan S, Ganesan TS. Tyrosine kinase inhibitors in cancer therapy. *Clin Biochem* 2004;**37**:618–35. https://doi.org/10.1016/j.clinbiochem.2004.05.006
5. Noble W, Hanger DP, Miller CCJ. *et al.* The importance of tau phosphorylation for neurodegenerative diseases. *Front Neurol* 2013;**4**:83.
6. Hanger DP, Anderton BH, Noble W. Tau phosphorylation: the therapeutic challenge for neurodegenerative disease. *Trends Mol Med* 2009;**15**:112–9. https://doi.org/10.1016/j.molmed.2009.01.003
7. Hambrecht R, Adams V, Erbs S. *et al.* Regular physical activity improves endothelial function in patients with coronary artery disease by increasing phosphorylation of endothelial nitric oxide synthase. *Circulation* 2003;**107**:3152–8. https://doi.org/10.1161/01.CIR.0000074229.93804.5C
8. O'Rourke B, Van Eyk JE, Brian D. *et al.* Mitochondrial protein phosphorylation as a regulatory modality: implications for mitochondrial dysfunction in heart failure. *Congest Heart Fail* 2011;**17**:269–82. https://doi.org/10.1111/j.1751-7133.2011.00266.x
9. Cohen P. The role of protein phosphorylation in human health and disease. *Eur J Biochem* 2001;**268**:5001–10. https://doi.org/10.1046/j.0014-2956.2001.02473.x
10. Zittlau K, Nashier P, Cavarischia-Rega C. *et al.* Recent progress in quantitative phosphoproteomics. *Expert Rev Proteomics* 2023;**20**:469–82. https://doi.org/10.1080/14789450.2023.2295872
11. Xiaofeng W, Liu Y-K, Iliuk AB. *et al.* Mass spectrometry-based phosphoproteomics in clinical applications. *TrAC Trends Anal Chem* 2023;**163**:117066. https://doi.org/10.1016/j.trac.2023.117066
12. Lancaster NM, Sinitcyn P, Forny P. *et al.* Fast and deep phosphoproteome analysis with the orbitrap astral mass spectrometer. *Nat Commun* 2024;**15**:7016.
13. Zhang G, Zhang C, Cai M. *et al.* Funcphos-str: an integrated deep neural network for functional phosphosite prediction based on alphafold protein structure and dynamics. *Int J Biol Macromol* 2024;**266**:131180. https://doi.org/10.1016/j.ijbiomac.2024.131180
14. Esmaili F, Pourmirzaei M, Ramazi S. *et al.* A review of machine learning and algorithmic methods for protein phosphorylation site prediction. *Genom Proteom Bioinform* 2023;**21**:1266–85. https://doi.org/10.1016/j.gpb.2023.03.007
15. Varshney N, Mishra AK. Deep learning in phosphoproteomics: methods and application in cancer drug discovery. *Proteomes* 2023;**11**:16.
16. Guo L, Wang Y, Xiangnan X. *et al.* Deeppsp: a global–local information-based deep neural network for the prediction of protein phosphorylation sites. *J Proteome Res* 2020;**20**:346–56. https://doi.org/10.1021/acs.jproteome.0c00431
17. Wang D, Zeng S, Chunhui X. *et al.* Musitedeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics* 2017;**33**:3909–16. https://doi.org/10.1093/bioinformatics/btx496
18. Elnaggar A, Heinzinger M, Dallago C. *et al.* Prottrans: toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* 2021;**44**:7112–27.
19. Weissenow K, Heinzinger M, Rost B. Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. *Structure* 2022;**30**:1169–1177.e4. https://doi.org/10.1016/j.str.2022.05.001
20. Ziyang X, Zhong H, He B. *et al.* Ptransips: identification of phosphorylation sites enhanced by protein plm embeddings. *IEEE J Biomed Health Inform* 2024;1–10.
21. Brandes N, Ofer D, Peleg Y. *et al.* Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics* 2022;**38**:2102–10. https://doi.org/10.1093/bioinformatics/btac020
22. Edward JH, Shen Y, Wallis P. *et al.* LoRA: low-rank adaptation of large language models. In: *International Conference on Learning Representations*, 2022.
23. Ding N, Qin Y, Yang G. *et al.* Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat Mach Intell* 2023;**5**:220–35. https://doi.org/10.1038/s42256-023-00626-4
24. Verkuil R, Kabeli O, Yilun D. *et al.* Language models generalize beyond natural proteins. *BioRxiv* 2022;2022–12.
25. Peng Z, Huang W, Shanzhi G. *et al.* Conformer: Local features coupling global representations for visual recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 367–76, 2021.
26. Leo Breiman. *Random Forests Machine Learning*, **45**:5–32, 2001, https://doi.org/10.1023/A:1010933404324.
27. Tranmer M, Elliot M. Multiple linear regression. *The Cathie Marsh Centre for Census and Survey Research (CCSR)* 2008;**5**:1–5.
28. Zou A, Phan L, Chen S. *et al.* Representation engineering: a top-down approach to AI transparency . arXiv preprint, arXiv:2310.01405. 2023.
29. Park K, Choe YJ, Veitch V. The linear representation hypothesis and the geometry of large language models. arXiv preprint, arXiv:2311.03658. 2023.

30. Bairoch A, Apweiler R, Wu CH. *et al.* The universal protein resource (uniprot). *Nucleic Acids Res* 2005;**33**:D154–9. https://doi.org/10.1093/nar/gki070
31. Hornbeck PV, Zhang B, Murray B. *et al.* Phosphositeplus, 2014: mutations, PTMS and recalibrations. *Nucleic Acids Res* 2015;**43**:D512–20.
32. Lee T-Y, Huang H-D, Hung J-H. *et al.* Dbptm: an information repository of protein post-translational modification. *Nucleic Acids Res* 2006;**34**:D622–7.
33. Dinkel H, Chica C, Via A. *et al.* Phospho. Elm: a database of phosphorylation sites—update 2011. *Nucleic Acids Res* 2010;**39**:D261–7.
34. Lv H, Dao F-Y, Zulfiqar H. *et al.* Deepips: comprehensive assessment and computational identification of phosphorylation sites of sars-cov-2 infection using a deep learning-based approach. *Brief Bioinform* 2021;**22**:bbab244.
35. Ziyuan Y, Jialin Y, Wang H. *et al.* Phosaf: an integrated deep learning architecture for predicting protein phosphorylation sites with alphafold2 predicted structures. *Anal Biochem* 2024;**690**:115510.
36. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9. https://doi.org/10.1093/bioinformatics/btl158
37. Kim Y. Convolutional neural networks for sentence classification. In: *Conference on Empirical Methods in Natural Language Processing*, 2014.
38. Vaswani A, Shazeer N, Parmar N. *et al.* Attention is All You Need. *Adv Neural Inf Process Syst* 2017;**30**. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
39. Dosovitskiy A, Beyer L, Kolesnikov A. *et al.* An image is worth 16x16 words: transformers for image recognition at scale. *ICLR* 2021.
40. Wang X, Zhang Z, Zhang C. *et al.* Transphos: a deep-learning model for general phosphorylation site prediction based on transformer-encoder architecture. *Int J Mol Sci* 2022;**23**:4263. https://doi.org/10.3390/ijms23084263
41. McInnes L, Healy J, Melville J. Umap: uniform manifold approximation and projection for dimension reduction. arXiv preprint, arXiv:1802.03426. 2018.
42. Johnson JL, Yaron TM, Huntsman EM. *et al.* An atlas of substrate specificities for the human serine/threonine kinome. *Nature* 2023;**613**:759–66. https://doi.org/10.1038/s41586-022-05575-3
43. Cock PJA, Antao T, Chang JT. *et al.* Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009;**25**:1422.
44. Strait BJ, Gregory T, Dewey. The Shannon information entropy of protein sequences. *Biophys J* 1996;**71**:148–55. https://doi.org/10.1016/S0006-3495(96)79210-X
45. Haldane A, Flynn WF, He P. *et al.* Structural propensities of kinase family proteins from a Potts model of residue co-variation. *Protein Sci* 2016;**25**:1378–84. https://doi.org/10.1002/pro.2954
46. Lv H, Dao F-Y, Guan Z-X. *et al.* Deep-kcr: accurate detection of lysine crotonylation sites using deep learning method. *Brief Bioinform* 2021;**22**:bbaa255.
47. Lumbanraja FR, Mahesworo B, Cenggoro TW. *et al.* SSMFN: a fused spatial and sequential deep learning model for methylation site prediction. *Peer J Comput Sci* 2021;**7**:e683.

Briefings in Bioinformatics, 2025, **26(3)**, bbaf242
https://doi.org/10.1093/bib/bbaf242
Problem Solving Protocol