

Post-translational modification prediction via prompt-based fine-tuning of a GPT-2 model

Received: 11 March 2024

Accepted: 29 July 2024

Published online: 07 August 2024

 Check for updates


Palistha Shrestha ^{1,5}, Jeevan Kandel ^{2,5}, Hilal Tayara ³  & Kil To Chong ^{1,4} 

Post-translational modifications (PTMs) are pivotal in modulating protein functions and influencing cellular processes like signaling, localization, and degradation. The complexity of these biological interactions necessitates efficient predictive methodologies. In this work, we introduce PTMGPT2, an interpretable protein language model that utilizes prompt-based fine-tuning to improve its accuracy in precisely predicting PTMs. Drawing inspiration from recent advancements in GPT-based architectures, PTMGPT2 adopts unsupervised learning to identify PTMs. It utilizes a custom prompt to guide the model through the subtle linguistic patterns encoded in amino acid sequences, generating tokens indicative of PTM sites. To provide interpretability, we visualize attention profiles from the model's final decoder layer to elucidate sequence motifs essential for molecular recognition and analyze the effects of mutations at or near PTM sites to offer deeper insights into protein functionality. Comparative assessments reveal that PTMGPT2 outperforms existing methods across 19 PTM types, underscoring its potential in identifying disease associations and drug targets.

Proteins, the essential workhorses of the cell, are modulated by post-translational modifications (PTM), a process vital for their optimal functioning. With over 400 known types of PTMs¹, they enhance the functional spectrum of the proteome. Despite originating from a foundational set of approximately 20,000 protein-coding genes², the dynamic realm of PTMs is estimated to expand the human proteome to over a million unique protein species. This diversity echoes the complexity inherent in human languages: individual amino acids assemble into 'words' to form functional 'sentences' or domains. This linguistic parallel extends to the dense, information-rich structure of both protein sequences and human languages. The advancing field of Natural Language Processing (NLP) not only unravels the intricacies of human communication but is also increasingly applied to decode the complex language of proteins. Our study leverages the transformative

capabilities of generative pretrained transformers (GPT) based models, a cornerstone in NLP, to interpret and predict the complex landscape of PTMs, highlighting an intersection where computational linguistics meets molecular biology.

In the quest to predict PTM sites, the scientific community has predominantly relied on supervised methods, which have evolved significantly over the years³⁻⁵. These methods typically involve training algorithms on datasets where the modification status of each site is known, allowing the model to learn and predict modifications on new sequences. B.Trost and A.Kusalik⁶ initially focus on methods like Support Vector Machines and decision trees, which classify amino acid sequences based on geometric margins or hierarchical decision rules. Progressing towards more sophisticated approaches, Zhou, F. et al.⁷ discuss the utilization of Convolutional Neural Networks and

¹Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju, Jeollabuk-do, Republic of Korea. ²Graduate School of Integrated Energy-AI, Jeonbuk National University, Jeonju, Jeollabuk-do, Republic of Korea. ³School of International Engineering and Science, Jeonbuk National University, Jeonju, Jeollabuk-do, Republic of Korea. ⁴Advances Electronics and Information Research Center, Jeonbuk National University, Jeonju, Jeollabuk-do, Republic of Korea. ⁵These authors contributed equally: Palistha Shrestha, Jeevan Kandel.  e-mail: hilaltayara@jbnu.ac.kr; kitchong@jbnu.ac.kr

Recurrent Neural Networks, adept at recognizing complex patterns and capturing temporal sequence dynamics. DeepSucc⁸ proposed a specific deep learning architecture for identifying succinylation sites, indicative of the tailored application of deep learning in PTM prediction. Smith, L. M. and Kelleher, N. L.⁹ highlight the challenges in data representation and quality in proteomics. Building upon these developments, Smith, D. et al.¹⁰ further advance the field, presenting a deep learning-based approach that achieves high accuracy in PTM site prediction. Their method involves intricate neural network architectures optimized for analyzing protein sequences, signifying a refined integration of deep learning in protein sequence analysis.

On the unsupervised learning front, methods like those developed by Chung, C. et al.¹¹ have significantly advanced the field. Central to their algorithm is the clustering of similar features and the recognition of patterns indicative of PTM sites. Lee, Tzong-Yi, et al.¹² utilized distant sequence features in combination with Radial Basis Function Networks, an approach that effectively identifies ubiquitin conjugation sites by integrating non-local sequence information. However, the complexity of PTM processes, often characterized by subtle and context-dependent patterns, pose challenges to these methods. Despite their advancements, they often grapple with issues like data imbalance, where certain PTMs are underrepresented, and the dependency on high-quality annotated datasets. These approaches can be challenged by the intricate and subtle nature of PTM sites, potentially overlooking crucial biological details. This landscape of PTM site prediction is ripe for innovation through generative transformer models, particularly in the domain of unsupervised learning. Intrigued by this possibility, we explored the potential of generative transformers, exemplified by the GPT architecture, for predicting PTM sites.

Here, we introduce PTMGPT2, a suite of models capable of generating tokens that signify modified protein sequences, crucial for identifying PTM sites. At the core of this platform is PROTGPT2¹³, an autoregressive transformer model. We have adapted PROTGPT2, utilizing it as a pre-trained model, and further fine-tuned it for the specific task of generating classification labels for a given PTM type. PTMGPT2 utilizes a decoder-only architecture, which eliminates the need for a task-specific classification head during training. Instead, the final layer of the decoder functions as a projection back to the vocabulary space, effectively generating the next possible token based on the learned patterns among tokens in the input prompt. When provided with a prompt, the model is faced with a protein sequence structured in a fill-in-the-blank format. Impressively, even without any hyperparameter optimization procedures, our model has demonstrated an average 5.45% improvement in Matthews Correlation Coefficient (MCC) over all other competing methods. The webserver and models that underpin PTMGPT2 are available at <https://nscbio.jbnu.ac.kr/tools/ptmgpt2>. Given the critical role of PTM in elucidating the mechanisms of various biological processes, we believe PTMGPT2 represents a significant stride forward in the efficient prediction and analysis of protein sequences.

Results

PTMGPT2 implements a prompt-based approach for PTM prediction

We introduce an end-to-end deep learning framework depicted in Fig. 1, utilizing a GPT as the foundational model. Central to our approach is the prompt-based finetuning of the PROTGPT2 model in an unsupervised manner. This is achieved by utilizing informative prompts during training, enabling the model to generate accurate sequence labels. The design of these prompts is a critical aspect of our architecture, as they provide essential instructional input to the pre-trained model, guiding its learning process. To enhance the explanatory power of these prompts, we have introduced four custom tokens to the pre-trained tokenizer, expanding its vocabulary size from 50,257

to 50,264. This modification is particularly significant due to the tokenizer's reliance on the Byte Pair Encoding (BPE) algorithm¹⁴. A notable consequence of this approach is that our model goes beyond annotating individual amino acid residues. Instead, it focuses on annotating variable-length protein sequence motifs. This strategy is pivotal as it ensures the preservation of evolutionary biological functionalities, allowing for a more nuanced and biologically relevant interpretation of protein sequences.

In the PTMGPT2 framework, we employ a prompt structure that incorporates four principal tokens. The first, designated as the 'SEQUENCE:' token, represents the specific protein subsequence of interest. The second, known as the 'LABEL:' token, indicates whether the subsequence is modified ('POSITIVE') or unmodified ('NEGATIVE'). This token-driven prompt design forms the foundation for the fine-tuning process of the PTMGPT2 model, enabling it to accurately generate labels during inference. A key aspect of this model lies in its architectural foundation, which is based on GPT-2¹⁵. This architecture is characterized by its exclusive use of decoder layers, with PTMGPT2 utilizing a total of 36 such layers, consistent with the pretrained model. This maintains architectural consistency while fine-tuning for our downstream task of PTM site prediction. Each of these layers is composed of masked self-attention mechanisms¹⁶, which ensure that during the training phase, the protein sequence and custom tokens can be influenced only by their preceding tokens in the prompt. This is essential for maintaining the autoregressive property of the model. Such a method is fundamental for our model's ability to accurately generate labels, as it helps preserve the chronological integrity of biological sequence data and its dependencies with custom tokens, ensuring that the predictions are biologically relevant.

A key distinction in our approach lies in the methodology we employed for prompt based fine-tuning during the training and inference phases of PTMGPT2. During the training phase, PTMGPT2 is engaged in an unsupervised learning process. This approach involves feeding the model with input prompts and training it to output the same prompt, thereby facilitating the learning of token relationships and context within the prompts themselves. This process enables the model to generate the next token based on the patterns learned during training between protein subsequences and their corresponding labels. The approach shifts during the inference phase, where the prompts are modified by removing the 'POSITIVE' and 'NEGATIVE' tokens, effectively turning these prompts into a fill-in-the-blank exercise for the model. This strategic masking triggers PTMGPT2 to generate the labels independently, based on the patterns and associations it learned during the training phase. An essential aspect of our prompt structure is the consistent inclusion of the '<startoftext>' and '<endoftext>' tokens. These tokens are integral to our prompts, signifying the beginning and end of the prompt helping the model to contextualize the input more effectively. This interplay of training techniques and strategic prompt structuring enables PTMGPT2 to achieve high prediction accuracy and efficiency. Such an approach sets PTMGPT2 apart as an advanced tool for protein sequence analysis, particularly in predicting PTMs.

Effect of prompt design and fine-tuning on PTMGPT2 performance

We designed five prompts with custom tokens ('SEQUENCE:', 'LABEL:', 'POSITIVE', and 'NEGATIVE') to identify the most efficient one for capturing complexity, allowing PTMGPT2 to learn and process specific sequence segments for more meaningful representations. Initially, we crafted a prompt that integrates all custom tokens with a 21-length protein subsequence. Subsequent explorations were conducted with 51-length subsequence and 21-length subsequence split into groups of k-mers, with and without the custom tokens. Considering that the pre-trained model was originally trained solely on protein sequences, we

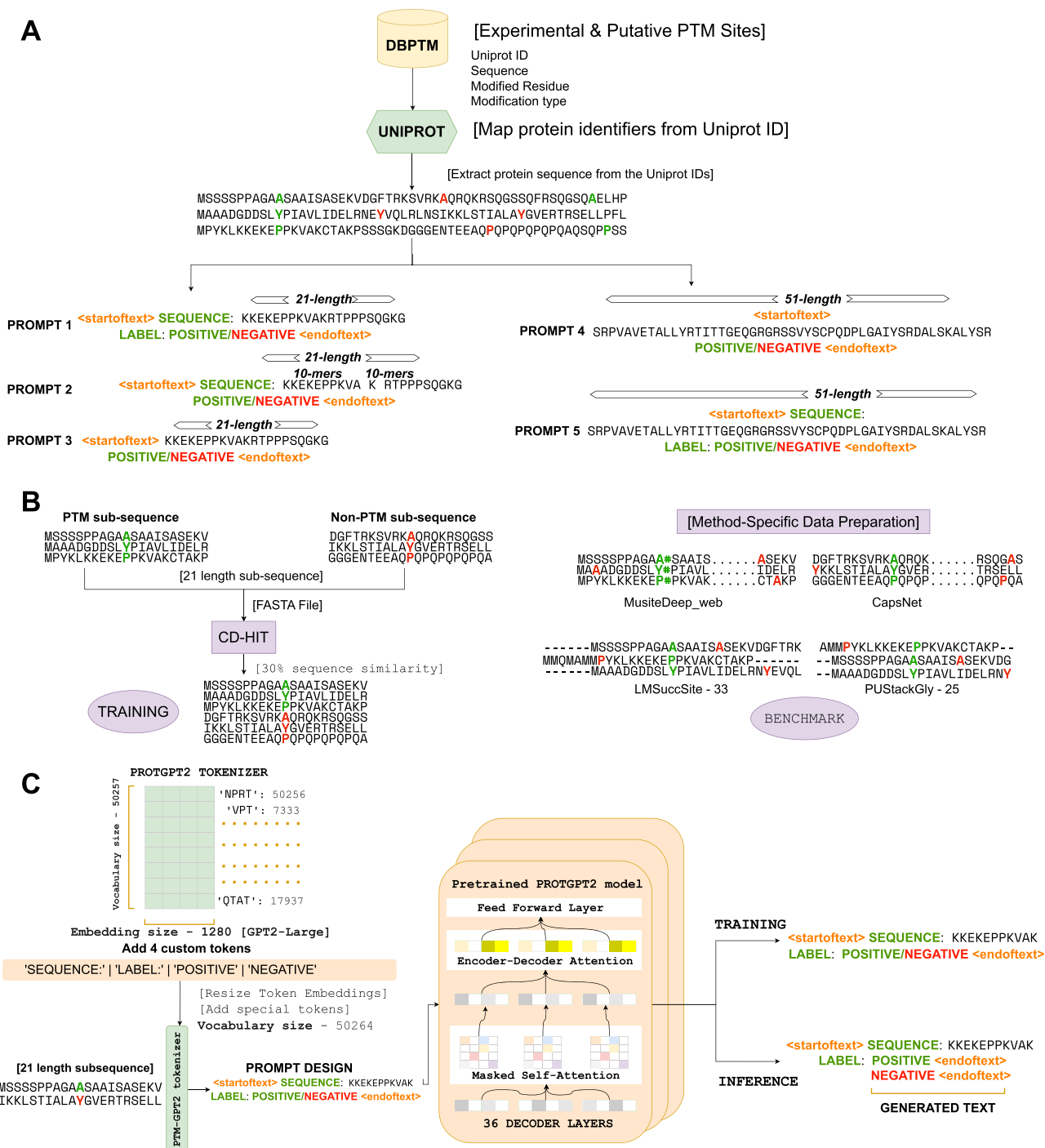


Fig. 1 | Schematic representation of the PTMGPT2 framework. A Preparation of inputs for PTMGPT2, detailing the extraction of protein sequences from Uniprot and the generation of five distinct prompt designs. **B** Method-specific data preparation process for benchmark, depicting both modified and unmodified sub-sequence extraction, followed by the creation of a training dataset using CD-HIT with 30% sequence similarity. **C** Architecture of the PTMGPT2 model and the training and inference processes. It highlights the integration of custom tokens into the tokenizer, the resizing of token embeddings, and the subsequent prompt design utilized during training and inference to generate predictions.

fine-tuned it with prompts both with and without the tokens to ascertain their actual contribution to improving PTM predictions.

Upon fine-tuning PTMGPT2 with training datasets for arginine(R) methylation and tyrosine(Y) phosphorylation, it became evident that the prompt containing the 21-length subsequence and the four custom tokens yielded the best results in generating accurate labels, as shown in Table 1. For methylation (R), the MCC, F1 Score, precision, and recall were reported as 80.51, 81.32, 95.14, and 71.01, respectively. Similarly, for phosphorylation (Y), the MCC, F1 Score, precision, and recall were

48.83, 46.98, 30.95, and 97.51, respectively. So, for all the experiments, we used the 21-length sequence with custom tokens. The inclusion of 'SEQUENCE:' and 'LABEL:' tokens provided clear contextual cues to the model, allowing it to understand the structure of the input and the expected output format. This helped the model differentiate between the sequence data and the classification labels, leading to better learning and prediction accuracy. The 21-length subsequence was an ideal size for the model to capture the necessary information without being too short to miss important context or too long to introduce

Table 1 | Benchmark results of PTMGPT2 after fine-tuning for optimal prompt selection

PTM	Prompt	MCC	F1Score	Precision	Recall
Methylation (R)	21-length w/ tokens [Proposed]	80.51	81.32	95.14	71.01
	21-length k-mer w/o tokens	77.16	78.53	94.98	65.03
	51-length w/o tokens	58.25	63.37	85.02	33.56
	51-length w/ tokens	60.85	66.98	89.34	45.77
Phosphorylation (Y)	21-length w/ tokens [Proposed]	48.83	46.98	30.95	97.51
	21-length w/o tokens	45.27	44.07	30.04	90.47
	51-length w/o tokens	27.48	31.25	20.32	67.56
	51-length w/ tokens	31.01	32.76	20.71	78.37

Top values are represented in bold.

noise. By framing the task clearly with the ‘SEQUENCE’ and ‘LABEL’ tokens, the model faced less ambiguity in generating predictions, which can be particularly beneficial for complex tasks such as PTM site prediction.

Comparative benchmark analysis reveals PTMGPT2's dominance

To validate PTMGPT2's performance, benchmarking against a database that encompasses a broad spectrum of experimentally verified PTMs and annotates potential PTMs for all UniProt¹⁷ entries was imperative. Accordingly, we chose the DBPTM database¹⁸ for its extensive collection of benchmark datasets, tailored for distinct types of PTMs. The inclusion of highly imbalanced datasets from DBPTM proved to be particularly advantageous, as it enabled a precise evaluation of PTMGPT2's ability to identify unmodified amino acid residues. This capability is crucial, considering that the majority of residues in a protein sequence typically remain unmodified. For a thorough assessment, we sourced 19 distinct benchmarking datasets from DBPTM, each containing a minimum of 500 data points corresponding to a specific PTM type.

Our comparative analysis underscores PTMGPT2's capability in predicting a variety of PTMs, marking substantial improvements when benchmarked against established methodologies using the MCC as the metric as shown in Table 2. For instance, in the case of lysine(K) succinylation, Succ-PTMGPT2 achieved a notable 7.94% improvement over LM-SuccSite. In the case of lysine(K) sumoylation, Sumoy-PTMGPT2 surpassed GPS Sumo by 5.91%. The trend continued with N-linked glycosylation on asparagine(N), where N-linked-PTMGPT2 outperformed Musite-Web by 5.62%. RMethyl-PTMGPT2, targeting arginine(R) methylation, surpassed Musite-Web by 12.74%. Even in scenarios with marginal gains, such as lysine(K) acetylation where KAcetyl-PTMGPT2 edged out Musite-web by 0.46%, PTMGPT2 maintained its lead. PTMGPT2 exhibited robust performance for lysine(K) ubiquitination, surpassing Musite-Web by 5.01%. It achieved a 9.08% higher accuracy in predicting O-linked glycosylation on serine(S) and threonine(T) residues. For cysteine(C) S-nitrosylation, the model outperformed PresSNO by 4.09%. In lysine(K) malonylation, PTMGPT2's accuracy exceeded that of DL-Malosite by 3.25%, and for lysine(K) methylation, it achieved 2.47% higher accuracy than MethylSite. Although PhosphoST-PTMGPT2's performance in serine-threonine (S, T) phosphorylation prediction was 16.37%, lower than Musite-Web, it excelled in tyrosine(Y) phosphorylation with an accuracy of 48.83%, which was notably higher than Musite-Web's 40.83% and Capsnet's 43.85%. In the case of cysteine (C) glutathionylation and lysine (K) glutarylation, GlutathioPTMGPT2 and Glutary-PTMGPT2 exhibited improvements of 7.51% and 6.48% over DeepGSH and ProtTrans-Glut, respectively. In the case of valine (V) amidation and cysteine (C) s-palmitoylation, Ami-PTMGPT2 and Palm-PTMGPT2 surpassed prAS and CapsNet by 4.78% and 1.56%, respectively. Similarly, in the cases of proline (P) hydroxylation, lysine (K) hydroxylation, and lysine (K)

formylation, PTMGPT2 achieved superior performance over CapsNet by 11.02%, 7.58%, and 4.39%, respectively. Collectively, these results demonstrate the significant progress made by PTMGPT2 in advancing the precision of PTM site prediction, thereby solidifying its place as a leading tool in proteomics research.

PTMGPT2 captures sequence-label dependencies through an attention-driven interpretable framework

To enable PTMGPT2 to identify critical sequence determinants essential for protein modifications, we designed a framework depicted in Fig. 2A that processes protein sequences to extract attention scores from the model's last decoder layer. The attention mechanism is pivotal as it selectively weighs the importance of different segments of the input sequence during prediction. Particularly, the extracted attention scores from the final layer provided a granular view of the model's focus across the input sequence. By aggregating the attention across 20 attention heads (AH) for each position in the sequence, PTMGPT2 revealed which amino acids or motifs the model deemed crucial in relation to the ‘POSITIVE’ token. The Position Specific Probability Matrix (PSPM)¹⁹, characterized by rows representing sequence positions and columns indicating amino acids, was a key output of this analysis. It sheds light on the proportional representation of each amino acid in the sequences, as weighted by the attention scores. PTMGPT2 thus offers a refined view of the probabilistic distribution of amino acid occurrences, revealing key patterns and preferences in amino acid positioning.

Motifs K**A*A and C**K were identified in AH 10, while motifs K***K****A, *KH*, and K***K were detected in AH 7. In AH 19, motifs K*C and C*K motifs were observed, and the *GK* motif was found in AH 15. Furthermore, motifs *EK* and KL**ER were identified in AH 5, motifs H***K, D**K, and *FK* were detected in AH 4. The A**K motif was observed in AH13, A*K motif in AH2, and *KM* motif in AH 11. To validate the predictions made by PTMGPT2 for lysine (K) acetylation, as shown in Fig. 2B, we compared these with motifs identified in prior research that has undergone experimental validation. Expanding our analysis to protein kinase domains, we visualized motifs for the CMGC and AGC kinase families, as shown in Fig. 3A, B. Additionally, the motifs for the CAMK kinase family and general protein kinases are shown in Fig. 4A, B, respectively. The CMGC kinase family named after its main members, CDKs (cyclin-dependent kinases), MAPKs (mitogen-activated protein kinases), GSKs (glycogen synthase kinases), and CDK-like kinases is involved in cell cycle regulation, signal transduction, and cellular differentiation²⁰. PTMGPT2 identified the common motif *P*SP* (Proline at positions -2 and +1 from the phosphorylated serine residue) in this family. The AGC kinase family, comprising key serine/threonine protein kinases such as PKA (protein kinase A), PKG (protein kinase G), and PKC (protein kinase C), plays a critical role in regulating metabolism, growth, proliferation, and survival²¹. The predicted common motif in this family was R**SL (Arginine at position -2 and leucine at position +1 from either a phosphorylated serine or threonine). The

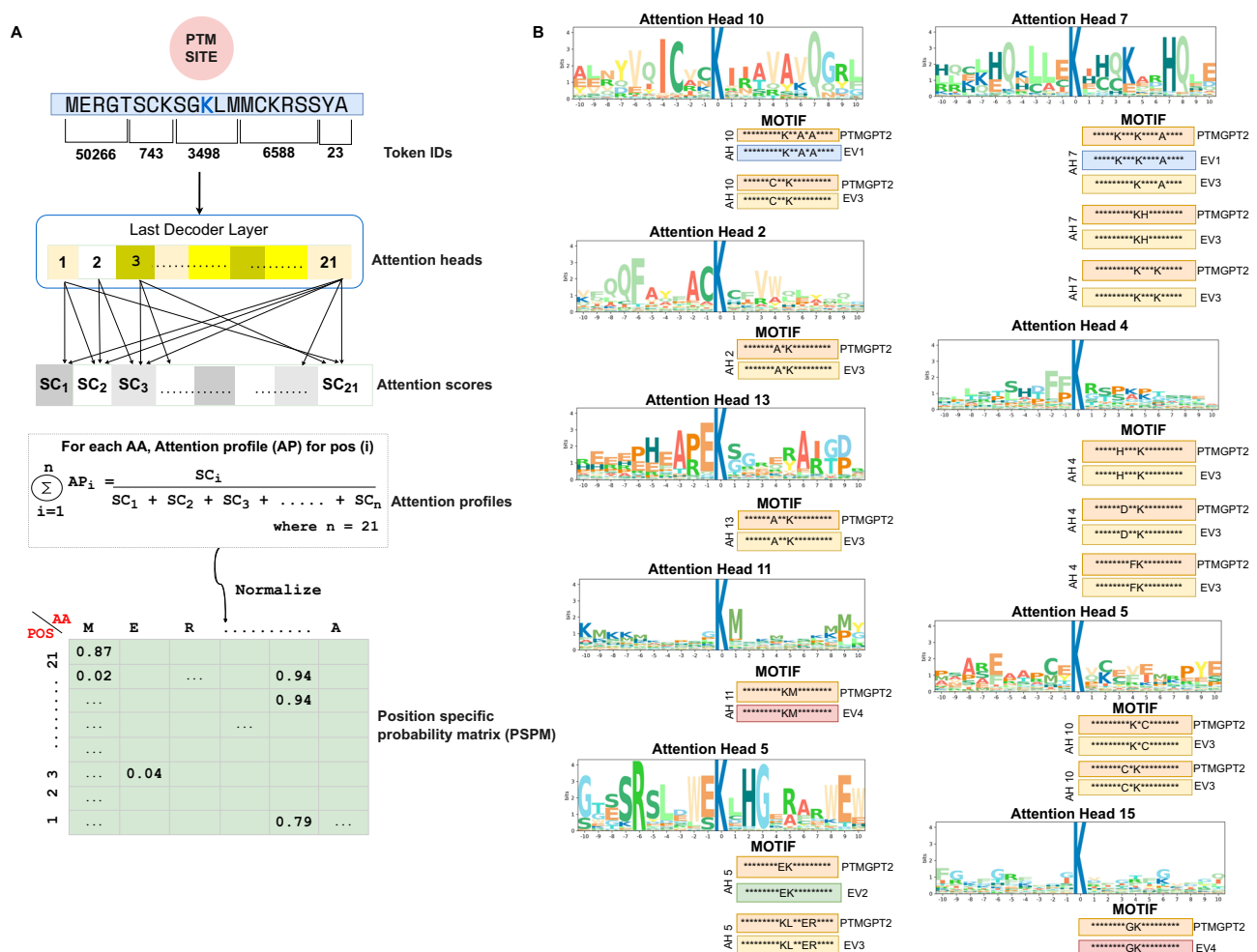
Table 2 | Benchmark dataset results

PTM	Model	MCC	F1Score	Precision	Recall
Succinylation (K)	LM-SuccSite ³⁴	43.94	62.26	53.05	45.34
	Psuc-EDBAM ³⁵	43.15	58.23	51.03	47.8
	SuccinSite ³⁶	15.87	34.47	64.95	23.46
	Deep SuccinylSite ³⁷	39.77	59.4	70.51	48.32
	Deep-KSuccSite ³⁸	36.88	55.89	47.41	48.06
	Succ-PTMGPT2	51.88	60.74	79.15	49.27
Sumoylation (K)	Musite-Web ³⁹	42.84	41.73	63.96	27.21
	CapsNet ⁴⁰	37.05	33.35	68.75	20.45
	ResSumo ⁴¹	21.07	25.59	15.56	61.97
	GPS Sumo ⁴²	61.02	65.76	70.38	25.67
	Sumoy-PTMGPT2	66.93	69.21	75.87	63.63
N-linked Glycosylation (N)	Musite-Web ³⁹	65.30	68.56	61.42	87.08
	CapsNet ⁴⁰	50.38	56.18	66.04	45.65
	LMNglyPred ⁴³	39.34	35.17	57.49	14.58
	N-linked-PTMGPT2	70.92	74.21	64.35	87.65
Methyl Arginine (R)	Musite-Web ³⁹	67.77	68.99	85.41	57.86
	CapsNet ⁴⁰	65.39	65.12	90.03	51.08
	DeepRMethylSite ⁴⁴	34.02	52.49	79.89	39.08
	PRmePred ⁴⁵	53.76	54.33	76.45	36.22
	CNNArginineMe ⁴⁶	34.55	36.98	97.42	22.82
	RMethyl-PTMGPT2	80.51	81.32	95.14	71.01
Lysine Acetylation (K)	Musite-Web ³⁹	21.63	36.14	94.86	22.32
	CapsNet ⁴⁰	21.62	36.54	93.46	22.71
	GPS-PAIL ⁴⁷	12.35	16.14	72.63	9.07
	KAcetyl-PTMGPT2	22.09	40.51	96.06	25.67
Ubiquitination (K)	Musite-Web ³⁹	26.24	27.67	75.67	16.93
	DL-Ubiq ⁴⁸	10.91	33.37	36.67	23.76
	Ubiq-PTMGPT2	31.25	35.74	80.46	22.97
O-linked-Glycosylation (S,T)	Musite-Web ³⁹	49.89	50.29	61.64	40.38
	OGlyThr ⁴⁹	41.6	50.71	52.82	60.78
	GlyCopp ⁵⁰	41.05	14.38	50.85	46.67
	O-linked-PTMGPT2	58.97	61.80	58.79	65.14
S-Nitrosylation (C)	PCysMod ⁵¹	48.53	66.67	62.50	71.42
	DeepNitro ⁵²	59.45	62.50	79.55	45.45
	PresSNO ⁵³	69.52	74.21	84.19	72.72
	pIMSNOSite ⁵⁴	55.91	17.60	14.74	21.82
	SNitro-PTMGPT2	73.61	80.49	84.30	77.01
Malonylation (K)	DL-Malosite ⁵⁵	69.78	77.12	72.83	81.94
	Maloy-PTMGPT2	73.03	78.14	82.95	73.85
Methyl Lysine (K)	Musite-Web ³⁹	14.97	10.13	72.44	5.44
	MethylSite ⁵⁶	38.60	35.54	39.41	66.66
	KMethyl-PTMGPT2	41.07	36.35	88.68	22.86
Phosphorylation (S, T)	Musite-Web ³⁹	17.73	12.22	6.67	72.74
	CapsNet ⁴⁰	16.23	12.28	6.71	71.94
	PhosphoST-PTMGPT2	16.37	15.35	9.14	46.97
Phosphorylation (Y)	Musite-Web ³⁹	40.83	43.03	29.82	77.27
	CapsNet ⁴⁰	43.85	46.61	36.58	68.18
	PhosphoY-PTMGPT2	48.83	46.98	30.95	97.51
Glutathionylation (C)	DeepGSH ⁵⁷	70.77	75.85	74.23	75.28
	Glutathio-PTMGPT2	78.28	82.37	90.07	75.89
Glutarylation (K)	ProtTrans-Glutar ⁵⁸	62.99	59.45	68.09	79.16
	Glutary-PTMGPT2	69.47	73.78	71.02	76.76
Amidation (V)	PrAS ⁵⁹	76.00	NR	NR	81.2
	Ami-PTMGPT2	80.78	76.59	66.67	86.76
S-Palmitoylation (C)	Musite-Web ³⁹	35.69	47.82	79.78	49.45

Table 2 (continued) | Benchmark dataset results

PTM	Model	MCC	F1Score	Precision	Recall
	CapsNet ⁴⁰	39.81	47.67	73.89	43.94
	GPS-Palm ⁶⁰	24.05	28.49	16.69	97.19
	Palm-PTMGPT2	41.37	48.34	42.73	55.64
Proline Hydroxylation (P)	Musite-Web ³⁹	78.08	80.58	98.32	67.47
	CapsNet ⁴⁰	78.87	85.92	94.66	74.79
	ProHydroxy-PTMGPT2	89.89	92.30	96.18	88.73
Lysine Hydroxylation (K)	Musite-Web ³⁹	57.67	63.76	76.33	72.12
	CapsNet ⁴⁰	58.87	65.92	74.66	74.79
	LysHydroxy-PTMGPT2	66.45	68.18	88.23	55.56
Formylation (K)	CapsNet ⁴⁰	40.19	33.33	24.09	68.34
	Musite-Web ³⁹	39.55	28.47	23.24	88.88
	Formy-PTMGPT2	44.58	39.97	26.92	77.78

Top values for each PTM are represented in bold

**Fig. 2 | Attention head analysis of lysine (K) acetylation by PTMGPT2.**

A Computation of attention scores from the model's last decoder layer, detailing the process of generating a Position-Specific Probability Matrix (PSPM) for a targeted protein sequence. 'SC' denotes attention scores, 'AP' denotes attention

profiles, 'AA' represents an amino acid, and 'n' is the number of amino acids in a subsequence. **B** Sequence motifs validated by experimentally verified studies—EV1⁶¹, EV2⁶², EV3⁶³, EV4⁴⁷. 'AH' denotes attention head.

CAMK kinase family, which includes key members like CaMK2 and CAMKL, is crucial in signaling pathways related to neurological disorders, cardiac diseases, and other conditions associated with calcium signaling dysregulation²². The common motif identified by PTMGPT2 in CAMK was R**S (Arginine at position -2 from either a

phosphorylated serine or threonine). Further analysis of general protein kinases revealed distinct patterns: DMPK kinase exhibited the motif RR*T (Arginine at positions -2 and -3), MAPKAPK kinase followed the R*LS motif (Arginine at position -3 and leucine at position -1), AKT kinase was characterized by the R*RS motif (Arginine at

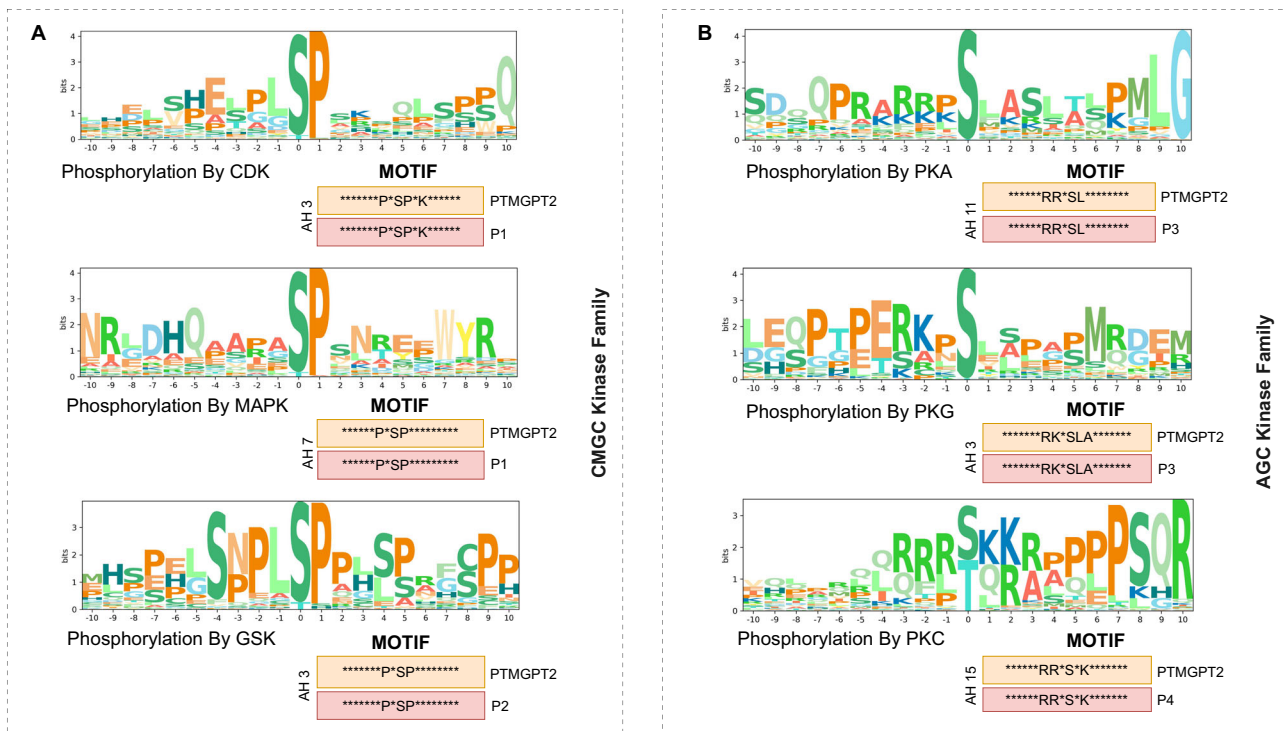


Fig. 3 | Attention head analysis of the CMGC kinase family and the AGC kinase family by PTMGPT2. A Motifs from CMGC kinase family validated against P1⁶⁴ and P2⁶⁵. **B** AGC kinase family motifs validated against P3⁶⁶ and P4⁶⁷. The 'P*SP' motif is

common in the CMGC kinase family, whereas the 'R**S' motif is common in the AGC kinase family. 'AH' denotes attention head.

positions -1 and -3), CK1 kinase showed K*K**S/T (Lysine at positions -3 and -5), and CK2 kinase was defined by the SD*E motif (Aspartate at position +1 and glutamate at position +3). These comparisons underscored PTMGPT2's ability to accurately identify motifs associated with diverse kinase groups and PTM types. PSPM matrices, corresponding to 20 attention heads across all 19 PTM types, are detailed in Supplementary Data 1. These insights are crucial for deciphering the intricate mechanisms underlying protein modifications. Consequently, this analysis, driven by the PTMGPT2 model, forms a core component of our exploration into the contextual relationships between protein sequences and their predictive labels.

Recent uniprot entries validate PTMGPT2's robust generalization abilities

To demonstrate PTMGPT2's robust predictive capabilities on unseen datasets, we extracted proteins recently released on UniProt, strictly selecting those added after June 1, 2023, to validate the model's performance. We ensured these proteins were not present in the training or benchmark datasets from DBPTM (version May 2023), which was a crucial step in the validation process. A total of 31 proteins that met our criteria were identified, associated with PTMs such as phosphorylation (S, T, Y), methylation (K), and acetylation (K). The accurate prediction of PTMs in recently identified proteins not only validates the effectiveness of our model but also underscores its potential to advance research in protein biology and PTM site identification. These predictions are pivotal for pinpointing the precise locations and characteristics of modifications within the protein sequences, which are crucial for verifying PTMGPT2's performance. The predictions for all 31 proteins, along with the ground truth, are detailed in Supplementary Table S1-S5.

PTMGPT2 identifies mutation hotspots in phosphosites of *TP53*, *BRAF*, and *RAF1* genes

Protein PTMs play a vital role in regulating protein function. A key aspect of PTMs is their interplay with mutations, particularly near

modification sites, where mutations can significantly impact protein function and potentially lead to disease. Previous studies²³⁻²⁵ indicate a strong correlation between pathogenic mutations and proximity to phosphoserine sites, with over 70% of PTM-related mutations occurring in phosphorylation regions. Therefore, our study primarily targets phosphoserine sites to provide a more in-depth understanding of PTM-related mutations. This study aims to evaluate PTMGPT2's ability to identify mutations within 1-8 residues flanking a phosphoserine site, without explicit mutation site annotations during training. For this, we utilized the dbSNP database²⁶, which includes information on human single nucleotide variations linked to both common and clinical mutations. *TP53*²⁷ is a critical tumor suppressor gene, with mutations in *TP53* being among the most prevalent in human cancers. When mutated, *TP53* may lose its tumor-suppressing function, leading to uncontrolled cell proliferation. *BRAF*²⁸ is involved in intracellular signaling critical for cell growth and division. *BRAF* mutations, especially the V600E mutation, are associated with various cancers such as melanoma, thyroid cancer, and colorectal cancer. *RAF1*²⁹ plays a role in the RAS/MAPK signaling pathway. While *RAF1* mutations are less common in cancers compared to *BRAF*, abnormalities in *RAF1* can contribute to oncogenesis and genetic disorders like Noonan syndrome, characterized by developmental abnormalities.

PTMGPT2's analysis of the *TP53* gene revealed a complex pattern of phosphosite mutations depicted in Fig. 5A, including G374, K370, and H368, across multiple cancer types²⁵. This is validated by dbSNP data, indicating that 21 of the top 28 mutations with the highest number of adjacent modifications occur in the tumor suppressor protein TP53. The *RAF1* gene, a serine/threonine kinase, exhibits numerous mutations, many of which are associated with disrupted MAPK activity due to altered recognition and regulation of PTMs. In our analysis of *RAF1* S259 phosphorylation, PTMGPT2 precisely identified mutations directly on S259 and in adjacent hotspots at residues S257, T258, and P261 depicted in Fig. 5B. These findings are consistent with genetic studies^{29,30} linking *RAF1* mutations near S259 to Noonan

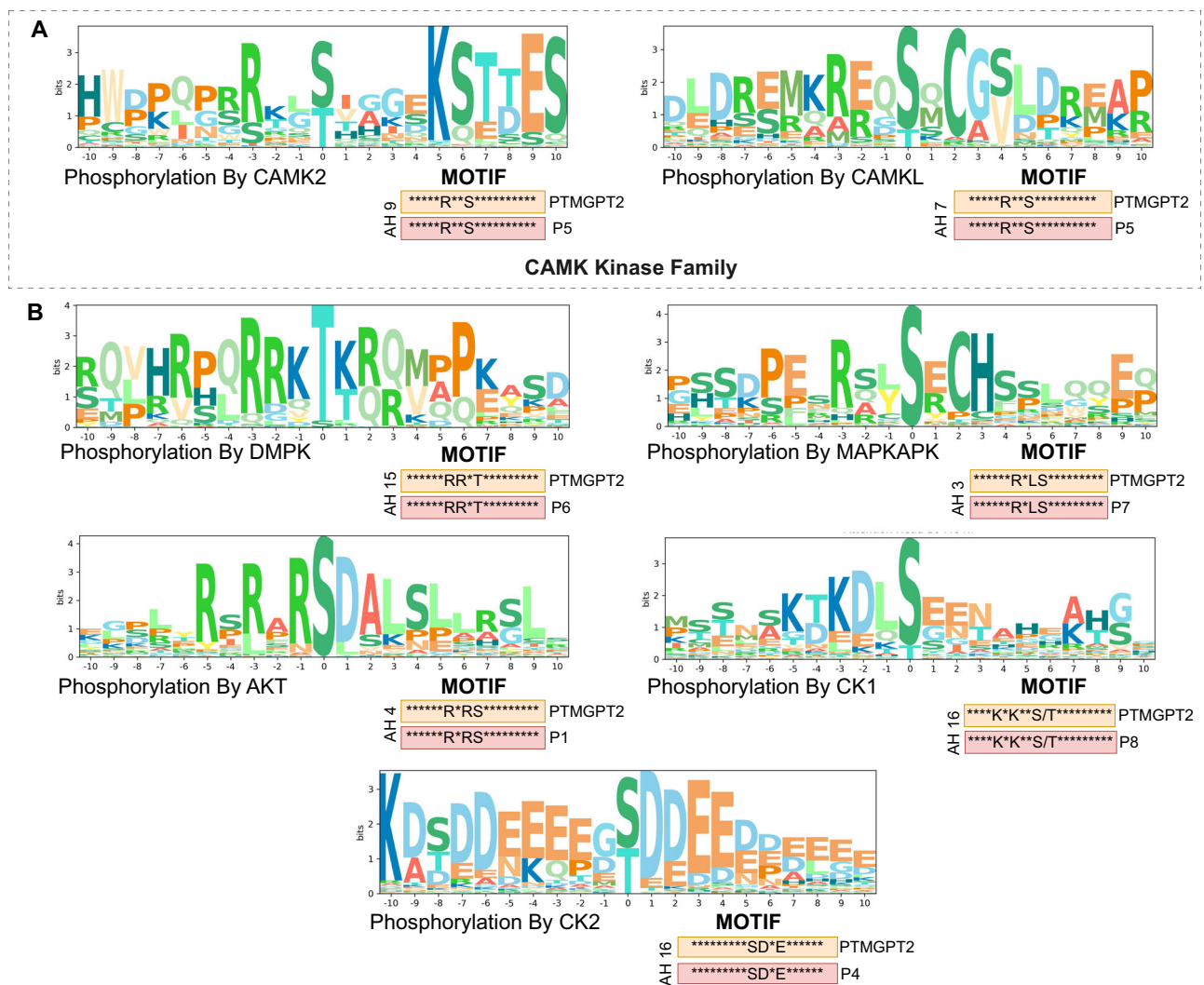


Fig. 4 | Attention head analysis of the CAMK kinase family and general protein kinases by PTMGPT2. A CAMK kinase family motifs validated against P5⁶⁸. **B** General protein kinase motifs validated against P6⁶⁹, P7⁷⁰, and P8⁷¹. The 'R**S' motif is common in the CAMK kinase family. 'AH' denotes attention head.

and LEOPARD Syndrome. Furthermore, in *BRAF*, another serine/threonine kinase, PTMGPT2's analysis of the S602 phosphorylation site revealed mutations in flanking residues (1–7 positions) such as D594N, L597Q, V600E, V600G, and K601E²³ shown in Fig. 5C. These mutations, particularly those activating *BRAF* functions, are found in over 60% of melanomas²⁸. Heatmap plots and line plots for remaining genes in dbSNP, and a bar chart depicting the selected genes for analysis, are provided in Supplementary Figs. S1–S19. These results demonstrate PTMGPT2's proficiency not only in predicting PTM sites but also in identifying potential mutation hotspots around these sites.

Discussion

GPT models have significantly advanced the state of NLP by demonstrating the power of the transformer architecture, the efficacy of pre-training on a large corpus of data, and the versatility of language models across a range of tasks through transfer learning. In this paper, we proposed PTMGPT2 for protein PTM site prediction in a way that reformulates protein classification tasks as protein label generation. We discovered that PTMGPT2, when subjected to prompt-based fine-tuning on large-scale datasets and tested on external benchmarks, demonstrates notable improvement in prediction accuracy, with an average improvement of 5.45% in MCC, underscoring the efficacy of protein language as a robust yet powerful descriptor for PTM site prediction. The crucial role of structuring an informative prompt,

which accurately captured the dependencies between protein sequence and its corresponding label, played a significant role in building an accurate generative model. Along with generating correct sequence labels for each PTM type, PTMGPT2 was able to precisely interpret attention scores as motifs and analyze the significance of having a pathogenic mutation directly on the site of modification or within a possible recognition area of the site of modification. This ability aids in exploring internal data distribution related to biochemical significance.

We employed PTMGPT2 for 19 different PTM types, including phosphorylation, N-linked glycosylation, N6-acetyllysine, methyl-arginine, succinylation, sumoylation, lysine acetylation, ubiquitination, O-linked glycosylation, S-nitrosylation, malonylation, methyl-lysine, glutathionylation, glutarylation, amidation, S-palmitoylation, hydroxylation, and formylation. The comparative results demonstrate that PTMGPT2 outperforms existing deep-learning methods and tools in most cases, offering promising prospects for practical applications. PTMGPT2's performance, categorized by species and evaluated using MCC, F1 Score, Precision, and Recall for all 19 PTMs, is listed in Supplementary Tables S6–S23. One limitation of our approach is the constrained exploration of prompt designs for certain PTM types, particularly in instances where PTMGPT2 did not surpass the performance of competing methods. We plan to investigate additional prompts and model tuning specifically designed

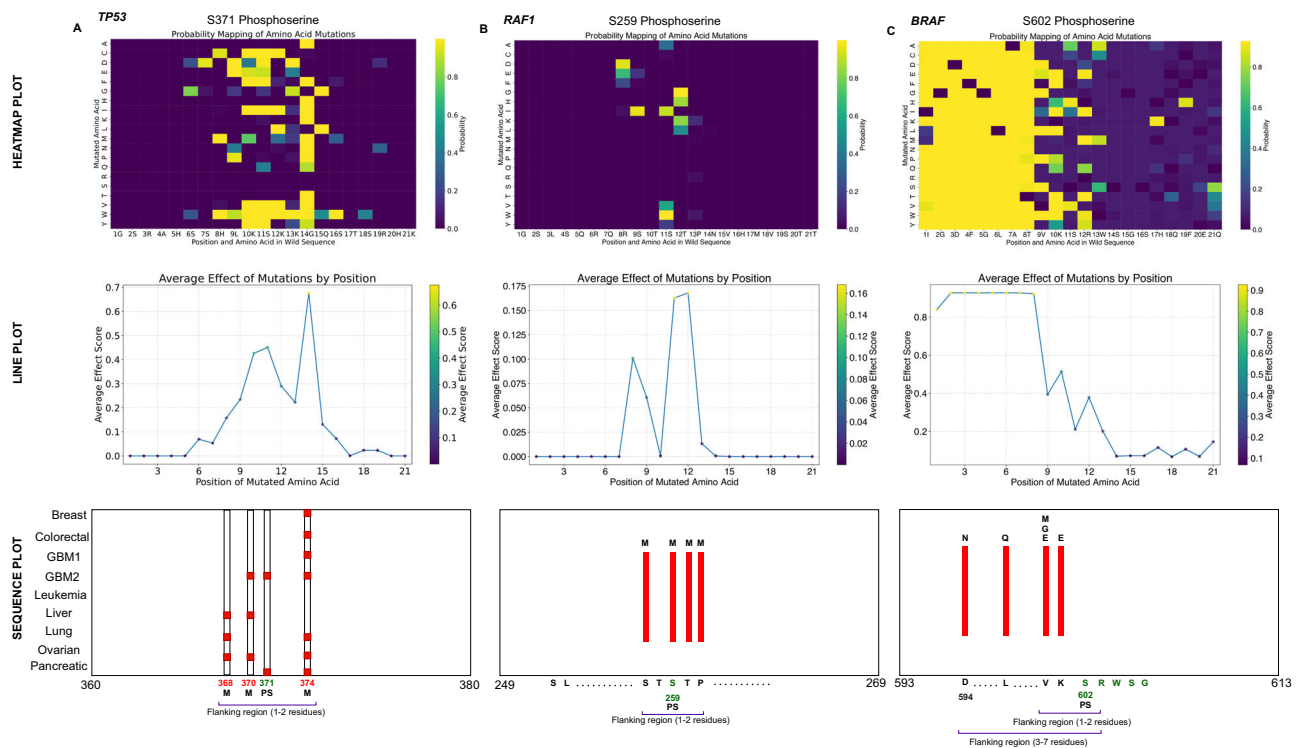


Fig. 5 | PTMGPT2 analysis of mutation distribution around PTM sites. Heatmaps and corresponding line plots illustrate the probability and impact of mutations within the recognition sites of phosphoserines across *TP53*, *RAF1*, and *BRAF* genes. The heatmaps' X-axes display the wild-type sequence while the Y-axes represent the 20 standard amino acids; yellow indicates the presence of mutations and darker shades indicate their absence. **A** For *TP53* S371 phosphoserine, PTMGPT2 predicts mutations predominantly in the 1–2 flanking residues. The line plot shows the average effect of these mutations by position, and the sequence plot reveals

predicted mutation hotspots directly on S371 and 1–2 residues from S371 across multiple human diseases. **B** Analysis of *RAF1* S259 phosphoserine, showing a concentrated mutation effect at the S259 and its immediate vicinity. **C** For *BRAF* S602 phosphoserine, PTMGPT2 identifies a broader distribution of mutations within the 1–7 flanking residues, with the line plot indicating significant mutation impacts at positions close to the S602. In the sequence plots, 'M' represents a mutation, and 'PS' indicates a phosphorylated serine residue. Source data are provided as a Source Data file.

for PTMs such as phosphorylation (S,T) in future versions of our work. Another limitation associated with PTM sites is the presence of false negatives in the training data. To mitigate the impact of false negatives, we plan in our future work to incorporate methods similar to outlier detection and anomaly detection, which focus on modeling the distribution of positive data, instances where PTMs are known to occur. By not relying on negative data, which can be sparsely labeled or misclassified in biological datasets, it is possible to avoid the common pitfalls of training models with false negatives. The reason behind implementing task-specific tuning for each PTM type, rather than utilizing shared weights among these PTMs, was that a single model using shared weights would require specially designed prompts to accurately handle multiple concurrent PTM types on the same residue. Moreover, in our motif identification analysis, utilizing a shared-weight model could complicate the interpretation of which residues influenced specific types of modifications. PTMGPT2 marks a major leap in protein PTM site prediction, laying the groundwork for future studies on protein sequence and function through the analysis of complex GPT-based attention mechanisms and their real-world applications. The autoregressive nature of PTMGPT2, which predicts the next token based on all previously observed tokens in a prompt, is particularly suited for tasks involving sequential data like protein sequences. Regarding the potential application of other large language models, it is plausible that similar enhancements could be achieved if these models are properly trained and fine-tuned for specific tasks in bioinformatics. Future efforts will focus on exploring and restructuring prompts, a strategy aimed at improving the accuracy of PTM site predictions within the protein domain.

Methods

Prompt design

In designing the prompt, our aim was to enable PTMGPT2 to distinguish between modified and unmodified protein sequences and design a flexible approach that can generalize to new classification tasks without the need for task-specific classification heads. The key was selecting tokens that accurately represented the protein subsequence and capture its contextual dependency with generated tokens. We incorporated two special tokens: <startoftext> and <endoftext>, marking the beginning and end of the prompt, respectively. Additionally, we integrated four custom tokens: 'SEQUENCE:' for the protein subsequence, 'LABEL:' for the PTM ground truth, 'POSITIVE' for modified sequences, and 'NEGATIVE' for unmodified sequences. Prompt 1 comprised four custom tokens flanking a 21-length protein subsequence. Prompt 2 arranged the 21-length subsequence into two 10-mers, delineating either a modified or unmodified central residue. In contrast, Prompt 3 employed only special tokens, omitting the custom tokens, yet it maintained the same 21-amino acid subsequence. Prompt 4 incorporated only special tokens but extended the sequence to 51 amino acids. Lastly, Prompt 5 combined both special and custom tokens with a 51-length protein sequence. All the prompts used for finetuning PTMGPT2 are displayed in Fig. 1A. Inference prompts are displayed in Supplementary Fig. S20.

Dataset preparation

Our dataset was compiled from DBPTM, a resource offering both experimentally verified training and non-homologous benchmark datasets. We extracted information regarding the UniProt ID, modified residue, and modification type for 19 PTM types from DBPTM, which

were then used to retrieve full-length protein sequences from Uniprot, available at <https://www.uniprot.org/id-mapping>. This step was necessary, as DBPTM does not provide full-length sequences. The training and benchmark datasets were compiled based on the categorizations pre-established by DBPTM. These specific PTMs were selected due to their high data volume, aligning with the requirement of GPT-based models for substantial data to develop an effective predictive model. Positive samples were prepared by extracting 21-length subsequences centered on modified residues. Negative samples were prepared by extracting 21-length subsequences from unmodified residues (amino acid positions not annotated as modified) for each distinct PTM type, using an approach identical to that used for positive samples. After combining both positive and negative subsequences for each PTM, we compiled them into a fasta file. To reduce sequence redundancy, sequences with over 30% similarity were removed using CD-hit³¹ for each PTM as shown in Fig. 1B. This process resulted in highly imbalanced yet refined training datasets for 19 PTMGPT2 models. Performance results for PTMGPT2 with CD-hit similarity cut-offs of 40% and 50% have been included in Supplementary Table S24 to provide a detailed analysis of how different thresholds affect the model's performance. To benchmark against models that require specific input formats, our data preparation had to align with their unique requirements, whether they needed variable-length sequences or symbol-based inputs. UniProt identifiers were mapped to acquire full-length sequences, from which method-specific subsequences were generated. Subsequently, we excluded any protein subsequences in the benchmark dataset that overlapped with our training dataset. Detailed data statistics for each PTM type are provided in Supplementary Tables S25 and S26. Data preparation pipelines for both training and benchmark datasets are shown in Supplementary Figs. S21 and S22.

Vocabulary encoding

We utilized the pre-trained tokenizer, fine-tuned on the BPE sub-word tokenization algorithm, to encode protein sequences. Unlike traditional methods that assign a unique identifier to each residue, this approach recognizes entire motifs as unique identifiers during tokenization. This not only retains evolutionary information but also highlights conserved motifs within sequences. As a result, this technique offers advantages over one-hot tokenization and alleviates out-of-vocabulary issues. On average, a token is represented by a motif of four amino acids. We expanded the pre-trained tokenizer's vocabulary to incorporate four custom tokens for prompt fine-tuning: 'POSITIVE', 'NEGATIVE', 'SEQUENCE:', and 'LABEL:', in addition to two special tokens: '<startoftext >' and '<endoftext >'. These tokens play a crucial role in constructing the prompts for output generation, where 'POSITIVE' and 'NEGATIVE' are the tokens generated by PTMGPT2, represented by token IDs 50262 and 50263, respectively. Furthermore, 'SEQUENCE:' and 'LABEL:' are instrumental in the finetuning process, represented by token IDs 50260 and 50261, respectively. This process resulted in a comprehensive vocabulary for PTMGPT2, with a total of 50,264 tokens.

Model training and inference

We began by instantiating the PROTGPT2 model and tokenizer, sourced from HuggingFace³². We initialized the model with pre-trained weights and fine-tuned it for our label generation task using an unsupervised approach, retaining tokens during training to learn sequence-label dependency. During inference, the 'POSITIVE', 'NEGATIVE', and '<endoftext >' tokens were excluded, allowing the model to generate labels that identify whether a protein sequence is modified or unmodified as shown in Fig. 1C. We utilized HuggingFace's trainer object to establish the training loop. Training each PTM model lasted 200 epochs, using a batch size of 128 per device, a weight decay of 0.01 for the Adam optimizer, and a learning rate of 1e-03, consistent with the

original pre-trained model. Negative log-likelihood was utilized as the loss function for fine-tuning and checkpoints were saved every 500 steps. The primary objective during inference was to remove the 'POSITIVE' and 'NEGATIVE' tokens and utilize greedy sampling to generate the most probable token associated to a modification, which was then compared with the ground truth. To evaluate the performance of PTMGPT2, we selected the best performing checkpoint on the basis of MCC, F1Score, precision, and recall on external benchmark dataset for each PTM type. All PTMGPT2 models were trained and bench-marked using NVIDIA A100 80GB and NVIDIA RTX A6000 48GB GPUs.

Benchmark comparison

For CapsNet, developed in 2018, we retrained models for sumoylation, N-linked glycosylation, methylation, acetylation, phosphorylation, and S-palmitoylation. Similarly, for MusiteDeep, introduced in 2020, the models were retrained for sumoylation, N-linked glycosylation, methylation, acetylation, phosphorylation, S-palmitoylation, ubiquitination, and O-linked glycosylation. The rationale for retraining these models included the availability of their training code and their documentation of multiple PTMs, which facilitated the comparison of varied PTMs. Additionally, the presence of overlapping protein sequences between our benchmark data and their training data necessitated the removal of overlapping data points and the retraining of their models from scratch using unique training sequences. This approach ensured a clearer demonstration of each method's intrinsic advantages without the influence of data overlap. To address the potential implications of outdated data, although the original publications for CapsNet and MusiteDeep did not include models for hydroxylation and formylation, we trained models from scratch using updated training data from DBPTM for these PTMs. This allowed for comparison with PTMGPT2 and was also required by the absence of other functional existing methods for these modifications. Methods introduced from 2021 onwards were not retrained, as they utilized more recent protein sequence data; therefore, they were compared using our common benchmark data. For other methods published before 2021, whose training code was not publicly available, we resorted to using their online web servers for comparison. Bar charts for benchmark comparison are provided in Supplementary Figs. S23–S41.

PSPM and attention head visualization

We proceeded by preparing the input sequence, prefixed and suffixed with custom tokens that delineated the sequence and label within the prompt. This formatted text was tokenized into a tensor of token IDs, which the model processed to produce attention scores. Since tokens represent motifs and the self-attention scores in GPT2 decoders operate between tokens, we calculate the attention scores for individual amino acids. To achieve this, we disaggregate the attention assigned to each token back down to the constituent amino acids. This disaggregation is performed such that each amino acid within a token is assigned the same attention score as the token's attention score. The process of transforming attention scores into attention profiles began with the initialization of a matrix to collate attention scores across the 20 standard amino acids for 21 sequence positions. The attention scores, derived from the previous step, were iteratively accumulated for each amino acid at each position in the sequence, skipping any instances of the amino acid 'X', typically indicative of an unknown amino acid. This resulted in a matrix reflecting the weighted significance of each amino acid at every position. Normalization of this matrix against the sum of attention scores per position yielded a relative attention profile. Next step was to transform sequences and their corresponding attention profiles into a PSPM Matrix. Initially, this involved accumulating attention profiles for each amino acid at their specific positions in the sequences. This was achieved by iterating over the sequences, ensuring that the length of each sequence matches its

associated attention profiles to maintain data integrity. The Logomaker Python package³³ was used to generate protein sequence logos. PSPM was provided as input to the Logomaker object. The algorithm for generating a PSPM is detailed in Supplementary Fig. S42.

Mutation analysis

Data for mutation analysis was compiled from dbSNP, which annotates pathogenic mutations occurring directly at or near modification sites. Uniprot identifiers for each gene were mapped using <https://www.uniprot.org/id-mapping> to extract the full-length protein sequences used in our analysis. Subsequently, wild-type and mutant protein subsequences were prepared. For each wild-type subsequence (with the modified amino acid residue at the center), 399 mutant sequences were generated. This was achieved by substituting each amino acid residue in the wild-type sequence with one of the 20 standard amino acids, excluding the original residue. Given the 21-length of the wild-type subsequences, the total possible combinations to generate point-mutated subsequences amounted to 21×19 . Our method produced all possible single amino acid substitutions for each wild-type sequence at every position. Inference prompts for both wild and mutant subsequences were then processed through the model to generate probability scores. These scores indicate whether any point mutations in the wildtype subsequence affected the model's output by altering PTM regulation. Initially, the probability score for the wild-type subsequence was obtained (the probability of the model generating a 'POSITIVE' token signifying modification). Subsequently, we analyzed the probability scores for all mutant subsequences to determine whether specific point mutations caused a significant change in PTM regulation (probability of the model generating a 'NEGATIVE' token, signifying PTM downregulation). Heatmaps and line plots were utilized to analyze the average effect of mutations by position, based on the output probabilities generated by PTMGPT2.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Training and benchmark datasets for all 19 PTMs are publicly available at Zenodo: <https://doi.org/10.5281/zenodo.11377398>. Trained PTMGPT2 models are available at <https://zenodo.org/records/11362322> and <https://doi.org/10.5281/zenodo.11371883>. Source data are provided with this paper.

Code availability

The source code for PTMGPT2 is publicly accessible at <https://github.com/pallucs/PTMGPT2>, and the repository includes files essential for conducting training and inference procedures. **Model:** This folder hosts a sample model designed to predict PTM sites from given protein sequences, illustrating PTMGPT2's application. **Tokenizer:** This folder contains a sample tokenizer responsible for tokenizing protein sequences, including handcrafted tokens for specific amino acids or motifs. **Inference.ipynb:** This file provides executable code for applying PTMGPT2 model and tokenizer to predict PTM sites, serving as a practical guide for users to apply the model to their datasets.

References

- Hong, X. et al. PTMint database of experimentally verified PTM regulation on protein–protein interaction. *Bioinformatics* **39**, btac823 (2023).
- Pray, L. Eukaryotic genome complexity. *Nat. Educ.* **1**, 96 (2008).
- Ramazi, S. & Zahiri, J. Post-translational modifications in proteins: resources, tools and prediction methods. *Database* **2021**, baab012 (2021).
- Virág, D. et al. Current trends in the analysis of post-translational modifications. *Chromatographia* **83**, 1–10 (2020).
- Meng, L. et al. Mini-review: recent advances in post-translational modification site prediction based on deep learning. *Comput. Struct. Biotechnol. J.* **20**, 3522–3532 (2022).
- Trost, B. & Kusalik, A. Predicting protein post-translational modification sites: an overview. *Comput. Biol. Chem.* **35**, 1–13 (2011).
- Zhou, F., Xue, Y., Chen, G. & Yao, X. Deep learning approaches for predicting post-translational modification sites in proteins. *Brief. Bioinform.* **21**, 615–630 (2019).
- Nguyen, Q. N., Huang, K. Y. & Ho, S. Y. DeepSucc: a deep learning architecture for succinylation site prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **18**, 685–693 (2019).
- Smith, L. M. & Kelleher, N. L. Proteoform: a single term describing protein complexity. *Nat. Methods* **15**, 186–187 (2018).
- Smith, D. Protein sequence analysis using deep learning: achieving accurate prediction of post-translational modification sites. *Nat. Methods* **17**, 779–787 (2020).
- Chung, C., Liu, J., Emili, A. & Frey, B. J. Computational refinement of post-translational modifications predicted from tandem mass spectrometry. *Bioinformatics* **27**, 797–806 (2011).
- Lee, T.-Y. et al. Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. *PLoS ONE* **4**, e4160 (2009).
- Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **13**, 4348 (2022).
- Gage, P. A new algorithm for data compression. *C Users J.* **12**, 23–38 (1994).
- Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
- Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 6000–6010 (2017).
- Boutet, E. et al. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Plant Bioinforma. Methods Protoc.* **1374**, 23–54 (2016).
- Li, Z. et al. dbPTM in 2022: an updated database for exploring regulatory networks and functional associations of protein post-translational modifications. *Nucleic Acids Res.* **50**, D471–D479 (2022).
- Henikoff, J. G. & Henikoff, S. Using substitution probabilities to improve position-specific scoring matrices. *Bioinformatics* **12**, 135–143 (1996).
- Varjosalo, M. et al. The protein interaction landscape of the human CMGC kinase group. *Cell Rep.* **3**, 1306–1320 (2013).
- Pearce, L. R., Komander, D. & Alessi, D. R. The nuts and bolts of AGC protein kinases. *Nat. Rev. Mol. Cell Biol.* **11**, 9–22 (2010).
- Swulius, M. T. & Waxham, M. N. Ca²⁺/calmodulin-dependent protein kinases. *Cell. Mol. Life Sci.* **65**, 2637–2657 (2008).
- Peng, D. et al. PTMsnp: a web server for the identification of driver mutations that affect protein post-translational modification. *Front. Cell Dev. Biol.* **8**, 593661 (2020).
- Holehouse, A. S. & Naegle, K. M. Reproducible analysis of post-translational modifications in proteomes—Application to human mutations. *PLoS ONE* **10**, e0144692 (2015).
- Reimand, J. & Bader, G. D. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.* **9**, 637 (2014).
- Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
- Olivier, M., Hollstein, M. & Hainaut, P. TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb. Perspect. Biol.* **2**, a001008 (2010).
- Davies, H. et al. Mutations of the BRAF gene in human cancer. *Nature* **417**, 949–954 (2002).

29. Pandit, B. et al. Gain-of-function RAF1 mutations cause Noonan and LEOPARD syndromes with hypertrophic cardiomyopathy. *Nat. Genet* **39**, 1007–1012 (2007).
30. Kobayashi, T. et al. Molecular and clinical analysis of RAF1 in Noonan syndrome and related disorders: dephosphorylation of serine 259 as the essential mechanism for mutant activation. *Hum. Mutat.* **31**, 284–294 (2010).
31. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
32. Wolf, T. et al. Transformers: State-of-the-Art Natural Language Processing. In *Proc. 2020 Conf. Empirical Methods Nat. Lang. Process.: Syst. Demonstrations* 38–45 (2020).
33. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2020).
34. Pokharel, S., Pratyush, P., Heinzinger, M., Newman, R. H. & KC, D. B. Improving protein succinylation sites prediction using embeddings from protein language model. *Sci. Rep.* **12**, 16933 (2022).
35. Jia, J., Wu, G., Li, M. & Qiu, W. pSuc-EDBAM: Predicting lysine succinylation sites in proteins based on ensemble dense blocks and an attention module. *BMC Bioinform.* **23**, 1–16 (2022).
36. Kao, H.-J., Nguyen, V.-N., Huang, K.-Y., Chang, W.-C. & Lee, T.-Y. SuccSite: incorporating amino acid composition and informative k-spaced amino acid pairs to identify protein succinylation sites. *Genom. Proteom. Bioinform.* **18**, 208–219 (2020).
37. Thapa, N. et al. DeepSuccinylSite: a deep learning based approach for protein succinylation site prediction. *BMC Bioinform.* **21**, 1–10 (2020).
38. Liu, X. et al. Deep_KsuccSite: a novel deep learning method for the identification of lysine succinylation sites. *Front. Genet.* **13**, 1007618 (2022).
39. Wang, D. et al. MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic Acids Res.* **48**, W140–W146 (2020).
40. Wang, D., Liang, Y. & Xu, D. Capsule network for protein post-translational modification site prediction. *Bioinformatics* **35**, 2386–2394 (2019).
41. Zhu, Y., Liu, Y., Chen, Y. & Li, L. ResSUMO: a deep learning architecture based on residual structure for prediction of lysine SUMOylation sites. *Cells* **11**, 2646 (2022).
42. Zhao, Q. et al. GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs. *Nucleic Acids Res.* **42**, W325–W330 (2014).
43. Pakhrin, S. C. et al. LMNglyPred: prediction of human N-linked glycosylation sites using embeddings from a pre-trained protein language model. *Glycobiology* **33**, 411–422 (2023).
44. Chaudhari, M. et al. DeepRMethylSite: a deep learning based approach for prediction of arginine methylation sites in proteins. *Mol. Omics* **16**, 448–454 (2020).
45. Kumar, P., Joy, J., Pandey, A. & Gupta, D. PRmePRed: a protein arginine methylation prediction tool. *PLoS ONE* **12**, e0183318 (2017).
46. Zhao, J. et al. CNNArginineMe: a CNN structure for training models for predicting arginine methylation sites based on the One-Hot encoding of peptide sequence. *Front. Genet.* **13**, 1036862 (2022).
47. Deng, W. et al. GPS-PAIL: prediction of lysine acetyltransferase-specific modification sites from protein sequences. *Sci. Rep.* **6**, 39787 (2016).
48. Wang, H., Wang, Z., Li, Z. & Lee, T.-Y. Incorporating deep learning with word embedding to identify plant ubiquitylation sites. *Front. Cell Dev. Biol.* **8**, 572195 (2020).
49. Tang, H., Tang, Q., Zhang, Q. & Feng, P. O-GlyThr: prediction of human O-linked threonine glycosites using multi-feature fusion. *Int. J. Biol. Macromol.* **242**, 124761 (2023).
50. Chauhan, J. S., Bhat, A. H., Raghava, G. P. S. & Rao, A. GlycoPP: a webserver for prediction of N- and O-glycosites in prokaryotic protein sequences. *PLoS ONE* **7**, e40155 (2012).
51. Li, S. et al. pCysMod: prediction of multiple cysteine modifications based on deep learning framework. *Front. Cell Dev. Biol.* **9**, 617366 (2021).
52. Xie, Y. et al. DeepNitro: prediction of protein nitration and nitrosylation sites by deep learning. *Genom. Proteom. Bioinform.* **16**, 294–306 (2018).
53. Hasan, M. M., Manavalan, B., Khatun, M. S. & Kurata, H. Prediction of S-nitrosylation sites by integrating support vector machines and random forest. *Mol. Omics* **15**, 451–458 (2019).
54. Pratyush, P., Pokharel, S., Saigo, H. & KC, D. B. pLMSNOSite: an ensemble-based approach for predicting protein S-nitrosylation sites by integrating supervised word embedding and embedding from pre-trained protein language model. *BMC Bioinform.* **24**, 41 (2023).
55. Thapa, N. et al. RF-MaloSite and DL-Malosite: methods based on random forest and deep learning to identify malonylation sites. *Comput. Struct. Biotechnol. J.* **18**, 852–860 (2020).
56. Biggar, K. K. et al. MethylSight: taking a wider view of lysine methylation through computer-aided discovery to provide insight into the human methyl-lysine proteome. *bioRxiv* 274688 (2018).
57. Li, S. et al. Deep learning based prediction of species-specific protein S-glutathionylation sites. *Biochim. et. Biophys. Acta (BBA)-Proteins Proteom.* **1868**, 140422 (2020).
58. Indriani, F., Mahmudah, K. R., Purnama, B. & Satou, K. Prottrans-glutar: Incorporating features from pre-trained transformer-based models for predicting glutarylation sites. *Front. Genet.* **13**, 885929 (2022).
59. Wang, T. et al. PrAS: prediction of amidation sites using multiple feature extraction. *Comput. Biol. Chem.* **66**, 57–62 (2017).
60. Ning, W. et al. GPS-Palm: a deep learning-based graphic presentation system for the prediction of S-palmitoylation sites in proteins. *Brief. Bioinform.* **22**, 1836–1847 (2021).
61. Zhang, H. et al. Quantitative proteomic analysis of the lysine acetylome reveals diverse SIRT2 substrates. *Sci. Rep.* **12**, 3822 (2022).
62. Zhang, X. et al. Widespread protein lysine acetylation in gut microbiome and its alterations in patients with Crohn’s disease. *Nat. Commun.* **11**, 4120 (2020).
63. Yuan, B. et al. Comprehensive proteomic analysis of lysine acetylation in *Nicotiana benthamiana* after sensing CWMV infection. *Front. Microbiol.* **12**, 672559 (2021).
64. Schwartz, D. & Gygi, S. P. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat. Biotechnol.* **23**, 1391–1398 (2005).
65. Ryu, G.-M. et al. Genome-wide analysis to predict protein sequence variations that change phosphorylation sites or their corresponding kinases. *Nucleic Acids Res.* **37**, 1297–1307 (2009).
66. Tegge, W., Frank, R., Hofmann, F. & Dostmann, W. R. G. Determination of cyclic nucleotide-dependent protein kinase substrate specificity by the use of peptide libraries on cellulose paper. *Biochemistry* **34**, 10569–10577 (1995).
67. Kreegipuu, A., Blom, N. & Brunak, S. PhosphoBase, a database of phosphorylation sites: release 2.0. *Nucleic Acids Res.* **27**, 237–239 (1999).
68. Viengkhou, B., White, M. Y., Cordwell, S. J., Campbell, I. L. & Hofer, M. J. A novel phosphoproteomic landscape evoked in response to type I interferon in the brain and in glial cells. *J. Neuroinflamm.* **18**, 1–20 (2021).
69. Wansink, D. G. et al. Alternative splicing controls myotonic dystrophy protein kinase structure, enzymatic activity, and subcellular localization. *Mol. Cell Biol.* **23**, 5489–5501 (2003).
70. Manke, I. A. et al. MAPKAP kinase-2 is a cell cycle checkpoint kinase that regulates the G2/M transition and S phase progression in response to UV irradiation. *Mol. Cell* **17**, 37–48 (2005).
71. Fulcher, L. J. & Sapkota, G. P. Functions and regulation of the serine/threonine protein kinase CK1 family: moving beyond promiscuity. *Biochem. J.* **477**, 4603–4621 (2020).

Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C2005612 to K.T.C.) and (No. 2022R1G1A1004613 to H.T.), in part by the Human Resources Program in Energy Technology of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) with financial resources from the Ministry of Trade, Industry & Energy, Republic of Korea (No. 20204010600470 to K.T.C.), and in part by the Korea Big Data Station (K-BDS) with computing resources and technical support.

Author contributions

P.S., J.K., H.T., and K.T.C.: conceptualization, methodology, investigation, writing review & editing and validation. P.S. and J.K.: writing original draft, Data Curation, implementation, software and website development, and visualization. H.T., and K.T.C.: Supervision and project administration.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-51071-9>.

Correspondence and requests for materials should be addressed to Hilal Tayara or Kil To Chong.

Peer review information *Nature Communications* thanks Dong Xu and the other, anonymous, reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024