# II.Materials and Methods

## A. Data Collection and Preprocessing

O-GlcNAcAtlas 4.0, the latest specialized database developed by the Hou team[1] in 2025, stands as the most comprehensive and systematic repository of O-GlcNAcylation globally. This database not only contains a substantial collection of experimentally validated modified proteins and their site information, but also pioneers the inclusion of quantitative analysis data at the site level. This enables researchers to comprehensively understand the dynamic patterns of O-GlcNAc modification under varying conditions. This database encompasses thousands of high-confidence human O-GlcNAc sites, accompanied by extensive experimental sources and metadata. It provides an authoritative and high-quality data foundation for in-depth investigation into the distribution characteristics, regulatory mechanisms, and computational prediction of O-GlcNAc modification. Concurrently, this study also utilized the UniProt database to obtain standardized protein sequences and comprehensive annotation information.

The data filtering process strictly adheres to scientific protocols: first, experimentally validated human O-GlcNAc modification sites were extracted from O-GlcNAcAtlas 4.0, with records of unknown origin or duplicates excluded; subsequently, 3,602 standardized protein sequences were obtained from the UniProt database; finally, sequence fragments were extracted using a window size of 2r+1 to obtain positive sample data for S/T sites. For the construction of negative samples, we first exclude sequence fragments from the original protein sequences that have been used for positive sample expansion but have not undergone padding processing. Subsequently, using lysine (K) as the reference point, we extract residue fragments of a specified length both before and after it. To accurately describe sequence characteristics, this study employed the peptide sequence representation method proposed by Chou et al.[2]. This method represents O-GlcNAcylation sites as shown in Eq (1):

Where S/T denote serine and threonine respectively, and the subscript δ represents an integer; the left half of S/T corresponds to the upstream amino acid residue, while the right half denotes the downstream amino acid residue; *H−δ* denotes the δ-th upstream amino acid residue counted from the center, while *H+δ* denotes the δ-th downstream amino acid residue counted from the center. Consequently, *Pδ(S/T)* denotes the O-GlcNAcylated residue sequence segment comprising δ amino acid residues upstream and downstream when serine or threonine serves as the central site.

This study draws significant insights from existing research on O-GlcNAc site recognition, whilst identifying considerable scope for optimization in current methodologies. Traditional studies predominantly employ local windows of 15–20 amino acid residues as analytical units. However, as complex biomacromolecules, proteins often exhibit functional regulation through the coordinated interactions of hundreds to thousands of residues. Local window analysis not only struggles to capture long-range interactions between distant residues but also fails to adequately reflect the semantic relationships and structural features inherent in the entire sequence. To this end, this study innovatively employs large language models (LLMs) for protein sequence feature extraction. Such models, with their capacity to process hundreds of tokens of context, can effectively capture global semantic relationships and long-range dependencies within protein sequences. Inputting only short fragments not only fails to leverage the model’s strengths but may also compromise prediction stability due to the “edge effect”. Consequently, this study employs a strategy of using complete sequences or long fragments as input, aiming to comprehensively enhance both the accuracy of predicting O-GlcNAcylation sites and the robustness of the model. At the implementation level, this study extended the analysis window to 101 amino acid residues (50 residues before and after the central site). Through rigorous data cleansing, all protein sequence fragments shorter than 101 residues were excluded, yielding 6,232 positive samples and 143,515 negative samples. To address the issue of severe class imbalance, this study employed the CD-HIT[3] tool for sequence similarity filtering. Given the significant numerical imbalance between positive and negative samples—with positive samples constituting merely 4.3% of negative samples—we adopted a differentiated threshold strategy: A more lenient similarity threshold was applied to positive samples to prevent excessive filtering from further reducing the dataset size, thereby preserving sequence diversity. Conversely, a relatively stricter threshold was employed for negative samples to avoid the repetition of highly similar sequences, thereby enhancing the model’s discriminative capability. This strategy not only effectively reduces data redundancy but also mitigates the issue of class imbalance to a certain extent. Following CD-HIT processing, the positive samples were reduced from the original 6,232 entries to 3,511, while the negative samples were optimized from 143,515 entries to 11,808. This significantly enhanced the overall quality and representativeness of the dataset, providing a more reliable data foundation for subsequent model training. Furthermore, this study conducted independent analyses of serine (Ser, S) and threonine (Thr, T) sites, yielding 9,186 high-quality sequence data points (Ser site: 2,128 positive/7,058 negative) and 6,133 high-quality sequence data points (Thr site: 1,383 positive/4,750 negative), respectively.

## B. Feature Extraction Strategy

ESM-2 (Evolutionary Scale Modelling 2), released by Meta AI, is a new-generation protein language model based on large-scale self-supervised pre-training. Trained on hundreds of millions of authentic protein sequences, it comprehensively learns the deep semantic relationships between amino acids, structural constraints, and evolutionary patterns. ESM-2 employs a deeper, larger-scale Transformer architecture, with model sizes ranging from 8M to 15B parameters. It possesses the capability to capture long-range dependencies spanning hundreds to thousands of residues, demonstrating exceptional feature representation performance. Progressing from ESM-1b[4] to ESM-1v[5] and then to ESM-2[6], ESM-2 has achieved significant improvements across tasks including protein structure prediction, mutation effect analysis, functional annotation, and sequence modelling. Leveraging its pre-trained high-dimensional contextual features, ESM-2 not only identifies local sequence patterns but also implicitly captures protein folding structures and evolutionary information. Consequently, it has become a mainstream feature extraction tool for diverse downstream tasks, including post-translational modification site identification, protein functional classification, and pathogenic variant prediction. As one of the most representative large-scale protein language models in contemporary bioinformatics, ESM-2 provides this study with high-quality, stable, and biologically meaningful sequence feature representations. Its deep contextual embeddings, trained on extensive protein databases, effectively capture amino acid dependencies and latent structural-functional information. In this study, we selected the lightest model variant within its series. This version substantially reduces both parameter size and computational overhead while preserving core representational capabilities, rendering it more suitable for large-scale data processing and rapid iteration.

Fig. S1 provides a visual comparison of the effectiveness of five feature encoding methods for S-sites and T-sites via a radar chart. From an overall perspective, the data for both types of sites exhibit highly consistent characteristics: significant performance differences exist between different feature expression methods, while the overall ranking remains consistent. It is evident that ESM-2 delivers the most outstanding overall performance. Across all metrics—Sn, Sp, Acc, MCC, Precision, F1, and AUC—ESM-2’s curve occupies the outermost layer of the radar chart, with the largest enclosed area. This indicates that its deep semantic modelling capabilities, ability to capture long-range dependencies, and implicit representation of protein structural information confer significant advantages in both site prediction tasks. By comparison, DR and DistancePair belong to the second tier of mid-range performers. While they demonstrate competitiveness in certain metrics, their overall performance remains markedly inferior to ESM2. This indicates that features based on distance or manually defined rules possess inherent limitations in capturing complex sequence relationships. AAC and OneHot exhibit relatively weaker performance. Particularly OneHot, due to its inability to encode contextual semantics and its extremely high dimensionality, performs poorly across multiple evaluation metrics. This fully demonstrates the inadequate expressive power of traditional shallow descriptors when confronted with long protein sequences and complex modification mechanisms. The feature expression capability of the large-scale pre-trained protein language model (ESM-2) significantly surpasses that of traditional methods, whether applied to S-sites or T-sites.

|  |
| --- |
| Fig. S1. Comparison of different feature extraction methods at the S site and T site |

## C. NearMiss Downsampling Strategy

Following the determination of feature extraction algorithms, we focused our examination on downsampling methods suitable for high-dimensional biological feature data. An ideal sampling algorithm must satisfy five key characteristics: boundary preservation capability, structural information retention capability, critical information preservation capability, neighborhood awareness capability, and adaptability to high-dimensional spaces. These properties are particularly crucial when processing data such as protein sequences, which exhibit high dimensionality, sparsity, and complex structural features.

Through systematic comparison, we found that the NearMiss algorithm[7] is particularly well-suited for processing protein sequence data, and ultimately adopted 101 positions as the empirically optimal sequence length. This algorithm, grounded in distance metric principles, filters majority samples by measuring relative distances between them. This approach prioritizes retaining the most representative sample types during downsampling, precisely addressing the critical requirements of protein sequence downsampling. Firstly, NearMiss inherently possesses significant boundary preservation capabilities by selecting majority-class samples closest to minority-class samples for retention, ensuring the model maintains high discriminative performance in class boundary regions. Secondly, the algorithm employs distance information within local neighborhoods during selection, thereby exhibiting strong local structure preservation properties that avoid disrupting the natural clustering structure of the majority class in feature space. Moreover, since NearMiss strategically retains samples closer to the minority class rather than randomly discarding majority class samples, it demonstrates robust information preservation capabilities, effectively reducing the risk of discarding crucial semantic features. Concurrently, its distance-based sample selection mechanism endows NearMiss with outstanding neighborhood awareness, enabling the removal of redundant samples distant from the discriminant region that contribute little to the model. Finally, although the protein sequence embeddings generated by ESM-2 typically reside in high-dimensional feature spaces, NearMiss can integrate metrics suitable for high-dimensional spaces, such as cosine distance and Mahalanobis distance, thereby demonstrating considerable high-dimensional compatibility. Overall, NearMiss exhibits significant advantages in preserving discriminative information, avoiding excessive downsampling, and adapting to high-dimensional embedding structures, making it one of the relatively efficient downsampling methods for addressing protein sequence data imbalance.

In this study, we employed the NearMiss-1 algorithm for downsampling majority class samples. This method enhances the model’s learning capability in category boundary regions by calculating the average distance between majority class samples and their *k* nearest minority class samples, thereby prioritizing retention of samples with the smallest distances. To accommodate the high-dimensional representation characteristics of protein sequences, cosine distance was adopted as the distance metric, further improving the algorithm’s adaptability and stability within high-dimensional feature spaces.

Let the dataset comprise a majority-class sample set and a minority-class sample set , where each sample represents the feature vector of a protein sequence. The objective of NearMiss-1 is to select a subset from such that each selected majority-class sample achieves the minimum average distance to its *k* nearest minority-class samples:

Where: denotes the index set of the *k* nearest minority neighbors of ; dist(,) represents the distance metric function, which may optionally be Euclidean distance, cosine distance, or Mahalanobis distance. Ultimately, the top *Nm* samples from with the smallest average distance are selected to form the downsampled subset . TABLE S1 presents the pseudocode representation of the NearMiss-1 algorithm.

TABLE S1

NearMiss-1 algorithm procedure

|  |
| --- |
| ***Algorithm :*** ***NearMiss-1*** |
| ***Input***:  - Majority class set: D\_M = {x\_1^M, x\_2^M, ..., x\_{N\_M}^M}  - Minority class set: D\_m = {x\_1^m, x\_2^m, ..., x\_{N\_m}^m}  - Number of neighbors: k  - Target size of majority subset: N' |
| ***Output***:  - Reduced majority set D\_M' ⊆ D\_M, |D\_M'| = N'  ***Procedure:***  1. For each x\_i in D\_M:  - Compute distances to all x\_j in D\_m  - Find k nearest minority neighbors  - Compute average distance:  d̄(x\_i) = (1/k) \* sum(dist(x\_i, x\_j)) over k neighbors  2. Sort D\_M by ascending d̄(x\_i)  3. Select the top N' samples to form D\_M' |
| 4. Return D\_M' |

Fig. S2(a) and Fig. S2(b) respectively illustrate the sampling effects of several downsampling algorithms on negative samples within the S-site and T-site datasets. The number of negative samples obtained by each downsampling method was kept as close as possible to that of positive samples. Overall performance was then evaluated using seven metrics: sensitivity (Sn), specificity (Sp), Mathews Correlation Coefficient (MCC), accuracy (Acc), precision (Precision), F1-Score(F1), and AUC. Analysis of the Fig. S2(a) reveals that the NearMiss algorithm demonstrates superior performance to other methods when handling downsampling tasks. Specifically, NearMiss achieves higher values across multiple evaluation metrics (such as Sensitivity, Specificity, MCC, Accuracy, Precision, F1, and AUC), with its performance notably surpassing other downsampling methods particularly in AUC and Precision. This indicates that NearMiss effectively preserves key samples in the majority class that are close to the minority class. While retaining marginal samples, it enhances the model’s discriminative capability, particularly within the complex feature space of protein sequences. By contrast, other downsampling methods such as Random, OSS and ALLKNN also demonstrated some performance, yet fell significantly short across most metrics, particularly exhibiting considerable gaps in MCC and Sn. The Random downsampling method performed worst across most metrics, indicating that its random deletion of data fails to account for the underlying structure of feature distributions, thereby readily leading to information loss. The results of downsampling at the T-site were similar to those at the S-site, with the NearMiss method continuing to lead across multiple metrics. Its performance was particularly outstanding in terms of F1 and AUC. Notably, NearMiss demonstrated a distinct advantage in both Sp and Precision, indicating its ability to effectively balance class imbalance while enhancing the model’s accuracy and stability in classification tasks. Although OSS and ALLKNN approach NearMiss on certain metrics, their performance on Sn and MCC falls short of NearMiss, particularly on Sn, suggesting that other methods may exhibit deficiencies in recognizing certain categories. Following an analysis of the practical performance of multiple downsampling algorithms, we ultimately adopted NearMiss-1 as the primary negative sample downsampling strategy. Subsequent modelling efforts centered on further optimizing and evaluating this approach. Table 3 presents the data information following downsampling.

|  |
| --- |
| Fig. S2. Comparison of different downsampling methods at the S site and T site |

Research indicates that increasing sequence length significantly enhances the representational capacity of bioinformatics models. As the number of residues within a sequence grows, the biological feature information it contains becomes more abundant, thereby providing more favorable conditions for large language models to extract deep sequence features. To systematically investigate the mechanism by which sequence length influences model performance, we employed a progressive extension strategy, gradually extending the initial 41-position benchmark sequence to 101 positions. Experiments were conducted separately on the S and T site datasets. Specifically, we conducted experiments at multiple sequence lengths including 41, 51, 61, 71, 81, 91, and 101 positions. The experimental results are presented in Fig. S3(a) and Fig. 3(b).

As is clearly evident from Fig. S3, the model’s performance across all evaluation metrics—including sensitivity (Sn), specificity (Sp), accuracy (Acc), Matthews correlation coefficient (MCC), precision, F1-score(F1), and AUC—exhibits a consistent upward trend as sequence length increases. This indicates that longer sequences provide the model with richer contextual information, thereby enabling large language models to more thoroughly learn underlying biological patterns and discriminative features. This pattern consistently applied in both S-site and T-site experiments, further validating the positive impact of sequence length on model performance. Following the completion of experiments with a sequence length of 101 positions, we additionally tested the model’s performance at 111 positions to provide supplementary validation of the performance trend. However, due to constrained training resources, it was not feasible to fully replicate the previous training configuration at this length. Consequently, appropriate adjustments were made to certain parameters. As evident from the rightmost portion of Fig. S3, the amplitude variation in the line segment is negligible, indicating that the model’s performance under these conditions remains highly consistent with that observed at 101 positions. Considering the computational cost of training longer sequences and efficiency requirements in practical applications, this study ultimately adopts 101 positions as the empirically optimal sequence length, balancing performance and computational feasibility.

|  |
| --- |
| Fig. S3. Line graphs of model performance at different sequence lengths of the S site and T site |

## D. Large Language Model

The development of large language models stems from the continuous evolution of model architectures within the field of natural language processing. In 2017, Vaswani et al. proposed the Transformer architecture[8], which revolutionized the paradigm of deep sequence modelling by introducing self-attention mechanisms and positional encoding. Building upon this foundation, Google released the BERT (Bidirectional Encoder Representations from Transformers) model in 2018[9]. This marked the first implementation of bidirectional context encoding alongside the pre-training objective of Masked Language Modelling (MLM), achieving breakthrough performance across multiple downstream tasks. It stands as a significant milestone in the evolution of modern pre-trained language models. Subsequently, with the refinement of training strategies, improvements to model architecture, and enhanced pre-training objectives, models such as RoBERTa, DeBERTa, and ModernBERT were successively proposed. These builds upon the BERT framework to further elevate language representation capabilities and training efficiency. Through innovations in attention mechanisms, positional encoding methods, and parameterized structures, these models effectively enhance the ability to model complex semantic relationships.

*1) DeBERTa v3, A Pre-trained Model for Decoupling Attention and Enhancing Decoding:* DeBERTa v3[10] is the latest iteration of the DeBERTa model series, proposed by Pengcheng He et al. in 2023. DeBERTa (Decoding-enhanced BERT with Disentangled Attention) itself represents a series of innovative enhancements built upon the original BERT model. DeBERTa v3 constitutes a major update following DeBERTa and DeBERTa v2, designed to elevate the model’s performance through further optimization of the attention mechanism, enhanced pre-training objectives, and more efficient training strategies. Compared to previous versions, DeBERTa v3 places greater emphasis on capturing long-range dependencies and deeper contextual information, enabling it to perform more effectively across a wide range of natural language processing (NLP) tasks.

Specifically, DeBERTa v3 has undergone systematic optimization in both its overall architecture and pre-training strategy, significantly enhancing the model’s semantic modelling capabilities and cross-task performance. Firstly, it continues to employ DeBERTa v2’s disentangled attention mechanism, which separates the encoding of a word’s semantic content from its relative positional information. This enables the model to capture word-level semantics and cross-positional dependencies independently. By modelling content information and relative positional relationships separately, DeBERTa v3 demonstrates greater flexibility and efficiency in handling long-range dependencies. It is particularly well-suited for tasks requiring the capture of complex contextual structures, such as long-text reading comprehension. Secondly, DeBERTa v3 further optimizes relative position encoding. Unlike traditional absolute position encoding, relative position encoding more accurately represents the relative distance between words, enabling the model to identify distant semantic associations more effectively in long-text scenarios. The improved position-aware mechanism allows the model to dynamically adjust the contextual contribution of different words, thereby enhancing overall representation capabilities. In terms of pre-training, DeBERTa v3 introduces enhanced masked language modelling (MLM) and next sentence prediction (NSP) objectives. The model better captures global semantic features through improved reconstruction of masked words, while enhancing reasoning capabilities by strengthening inter-sentence relationship modelling. These pre-training objectives collectively promote the model’s deep learning of underlying semantic structures, rendering it more robust in downstream tasks. Moreover, DeBERTa v3 combines larger-scale data with more efficient distributed training strategies, optimizing parameters and training workflows to fully leverage computational resources. This approach further enhances training efficiency and representation quality while maintaining the model’s scalability.

Thanks to these design choices, DeBERTa v3 demonstrates exceptional cross-task generalization capabilities, achieving leading performance across a broad spectrum of NLP tasks including text classification, sentiment analysis, named entity recognition, question answering, and natural language reasoning. Across multiple benchmarks including GLUE, SuperGLUE, and SQuAD 2.0, DeBERTa v3 demonstrates exceptional performance particularly in complex semantic reasoning and long-text comprehension, showcasing its robust deep semantic modelling capabilities and strong representational generalization abilities.

In this study, we introduced DeBERTa v3 into the task of identifying O-GlcNAcylation sites for deep feature modelling of protein sequences. The identification of O-GlcNAcylation sites typically relies on extensive sequence dependencies spanning multiple residues and complex contextual information. Conventional short-window modelling approaches exhibit significant limitations in capturing synergistic regulatory relationships between distant residues, whereas DeBERTa v3 can precisely model these latent structure-semantic correlation patterns within high-dimensional embedding spaces. Leveraging its robust contextual integration capabilities and sensitivity to long-range signals, DeBERTa v3 enhances models’ ability to characterize the latent features of O-GlcNAcylation, thereby improving site recognition accuracy and unlocking new performance gains for bioinformatics sequence analysis tasks.

*2) ModernBERT, High-Efficiency Long-Context Bidirectional Encoding Model:* ModernBERT[11] is a modernised bidirectional encoder Transformer model, proposed by the AnswerDotAI team in 2024. Compared to traditional BERT and its variants, ModernBERT aims to achieve a superior balance between performance, computational efficiency, and long-text processing capabilities. This enables it to excel in natural language processing tasks such as classification, retrieval, and long-text comprehension. The model is particularly well-suited for handling long sequences, high-throughput scenarios, and resource-constrained environments.

ModernBERT has undergone systematic optimization in architectural design, long-context processing, efficient attention mechanisms, and pre-training strategies, significantly enhancing its modelling capabilities for complex sequences and cross-task performance. Firstly, in terms of model architecture, ModernBERT employs Rotary Positional Embedding (RoPE) to replace traditional absolute positional embedding, enabling the model to capture relative positional information within long sequences with greater flexibility. Simultaneously, it incorporates the GeGLU activation function to enhance feature representation capabilities. Adopting a “Deep & Narrow” architecture by increasing Transformer depth while reducing layer width, it elevates the network’s abstraction capacity while maintaining overall parameter count under control. Moreover, by streamlining LayerNorm and linear layer bias terms, the model further enhances training stability and computational efficiency. Regarding long-context modelling and attention mechanisms, ModernBERT natively supports context windows of up to 8,192 tokens, far exceeding the 512-token limitation of traditional BERT. The model achieves a balance between short-range and long-range dependencies by alternately employing local attention and global attention; local attention captures neighboring features, while global attention focuses on distant semantic relationships. By integrating unpadding techniques with Flash Attention, significant computational acceleration is realized, rendering the training and inference of long sequences considerably more efficient. In terms of pre-training strategy, ModernBERT was trained on a multi-source mixed corpus exceeding 2 trillion tokens, encompassing diverse domains such as web text, academic literature, and code. The model employs an enhanced Masked Language Model (MLM) as its sole pre-training objective, with the masking ratio increased to 30% to strengthen global semantic understanding of lengthy texts and enhance generalization performance across complex tasks. Consequently, ModernBERT demonstrates outstanding capabilities in text classification, information retrieval, long document comprehension, and code retrieval, whilst exhibiting robust cross-task generalization. In biological sequence analysis, such as predicting O-GlcNAcylation sites in proteins, ModernBERT effectively captures latent synergistic patterns between distant residues within long sequences. This enables deeper structural-semantic feature representation, thereby enhancing prediction accuracy.

Overall, ModernBERT achieves a harmonious balance of high performance and efficiency through architectural optimization, support for long contexts, efficient attention design, and large-scale pre-training. Its RoPE position encoding and alternating local-global attention strategy enhance the modelling of long-range dependencies, while its “Deep & Narrow” structure and computationally efficient solutions significantly reduce training and inference costs. This provides a robust and efficient foundational model for complex sequence analysis tasks.

In this study, we introduce ModernBERT to the task of identifying O-GlcNAcylation sites for deep feature modelling of protein sequences. The discrimination of O-GlcNAcylation sites relies on long-range sequence dependencies spanning multiple residues and intricate contextual information; however, conventional short-window-based modeling strategies are inherently limited in capturing cooperative interactions among distal residues. ModernBERT employs Rotated Position Encoding (RoPE) to flexibly represent relative positional relationships between residues. Through alternating local and global attention mechanisms, it effectively captures dependencies between neighboring and distant residues. Its “Deep & Narrow” network architecture enhances the integration of sequence context while preserving model capacity. These characteristics enable ModernBERT to accurately model long-range structure-semantic correlation patterns within protein sequences in high-dimensional embedding spaces. This enhances its capacity to characterize latent features of O-GlcNAcylation, thereby improving site recognition accuracy and providing an efficient and reliable characterization method for complex biological sequence analysis.

## E. Deep Learning-Based Natural Language Processing Technology

In recent years, with the rapid advancement of deep learning, the field of Natural Language Processing (NLP) has undergone a significant shift from traditional statistical methods towards neural network approaches. Early NLP models relied heavily on manually designed features and shallow models, which struggled to effectively capture contextual semantics and sequential dependencies within text. The introduction of deep learning techniques, particularly models centered on recurrent neural networks (RNNs), convolutional neural networks (CNNs), and attention mechanisms, has significantly enhanced the expressive power and generalization capabilities of text modelling. Against this backdrop, the Bidirectional Long Short-Term Memory (BiLSTM) [12], as an extended structure of the Recurrent Neural Network (RNN), has been extensively applied to tasks such as text classification, named entity recognition, and sentiment analysis. BiLSTM captures bidirectional dependencies within sequences through its forward and backward LSTM modules, enabling the model to comprehensively consider contextual information. This endows it with exceptional performance when processing data exhibiting strong sequential characteristics. When combined with attention mechanisms and pre-trained word vectors such as Word2Vec or GloVe, BiLSTM’s sequence modelling capabilities are further enhanced. Similar to BiLSTM, the Bidirectional Gated Recurrent Unit (BiGRU)[13] serves as another highly efficient recurrent neural network architecture. While retaining the capability to model bidirectional context dependencies, it offers advantages over BiLSTM in terms of fewer parameters, faster training, and lower computational overhead. Consequently, BiGRU has been widely adopted in numerous tasks requiring efficient sequence modelling, proving particularly well-suited for scenarios involving large datasets or constrained training resources.

*1) Bidirectional Long Short-Term Memory (BiLSTM):* Traditional recurrent neural networks (RNNs)[14] can only make predictions based on information from the current and previous time steps, and perform poorly when handling long-term dependencies. LSTMs address this issue by introducing a gating mechanism, enabling them to effectively capture dependencies within long-term sequences and circumvent the vanishing gradient problem. However, LSTMs remain unidirectional, meaning they can only process sequence information either from start to finish or from end to beginning. The introduction of BiLSTMs builds upon LSTMs by simultaneously considering both forward and backward contextual information, thereby enhancing the model’s expressive power. This allows the model to capture bidirectional contextual dependencies, consequently improving the accuracy of sequence modelling.

The core of the LSTM lies in its four gates that regulate information flow: First is the Forget Gate, which determines how much memory from the previous time step is discarded; Next is the Input Gate, which decides how much new information from the current time step is stored in memory; Then comes the Output Gate, which governs how much influence the current memory exerts on the output; Finally, the Cell State retains crucial information and passes it forward to the next time step.

BiLSTM comprises two LSTMs: a forward LSTM (processing from left to right) and a backward LSTM (processing from right to left). These two LSTMs share the same input sequence but process information in opposite directions, subsequently merging their outputs. The forward LSTM commences from the left end of the input sequence, progressively reading information from the sequence and passing the hidden state of each time step to the subsequent time step. The backward LSTM commences from the right end of the input sequence, progressively reading information from the sequence and generating the corresponding output hidden state . At each time step, the BiLSTM concatenates or performs a weighted sum of the hidden states and from the forward and backward LSTMs respectively, yielding the final bidirectional representation:

Here, the “;” denotes the concatenation of forward and backward outputs to form a richer contextual representation.

*2) Bidirectional Gated Recurrent Unit (BiGRU):* Traditional recurrent neural networks (RNNs) suffer from the vanishing gradient problem when handling long-term dependencies. Gated recurrent units (GRUs) partially resolve this issue through a simplified gating mechanism. Compared to LSTMs, GRUs merge the Forget Gate and Input Gate into a single Update Gate while retaining the Reset Gate, thereby achieving efficient control over information. The GRU structure is more streamlined, requires fewer parameters, and trains more rapidly, whilst demonstrating comparable performance to LSTMs in numerous sequence modelling tasks.

The core structure of the GRU comprises three components: Update Gate: Controls how much of the previous hidden state is retained into the current time step; Reset Gate: Controls how much of the previous hidden state is discarded when generating candidate hidden states; Candidate Hidden State: Generates new information by combining the current input and the previous hidden state, then undergoes weighted updates via the Update Gate to produce the final hidden state.

The Bidirectional Gated Recurrent Unit (BiGRU) comprises two GRUs: a forward GRU processing from left to right and a backward GRU processing from right to left. These GRUs share the same input sequence but process information in opposite directions, subsequently merging their outputs. The forward GRU progressively reads sequence information starting from the left end of the input sequence, passing the hidden state output at each time step to the subsequent time step. The reverse GRU progressively reads sequence information starting from the right end of the input sequence, generating the corresponding hidden state output. At each time step, BiGRU concatenates or performs a weighted sum of the hidden states and from the forward and reverse GRUs, yielding the final bidirectional representation:

Here, the “;” denotes concatenating forward and backward outputs to form a richer contextual representation. BiGRU retains the ability to model bidirectional contextual dependencies while offering advantages of fewer parameters and faster training, making it particularly well-suited for sequence modelling in scenarios involving large datasets or limited computational resources.

## F. Evaluation Metrics

Sensitivity (Sn): Sensitivity (Recall) measures a model’s ability to correctly identify positive class samples (O-GlcNAcylation sites). Sensitivity is particularly crucial when identifying O-GlcNAcylation sites in proteins, as higher sensitivity indicates the model can detect as many genuine modification sites as possible, thereby reducing the risk of overlooking critical functional sites. The specific calculation formula is:

Among these, TP (True Positive) denotes genuine cases, representing correctly identified O-GlcNAcylation sites; FN (False Negative) denotes false negative cases, representing actual O-GlcNAcylation sites that were not detected.

Specificity (Sp): Specificity measures a model’s ability to identify negative class samples (non-O-GlcNAcylation sites). In the task of identifying protein modification sites, specificity helps us understand how well the model avoids false positives, i.e., ensuring the model does not erroneously predict non-O-GlcNAcylation sites as O-GlcNAcylation sites. The specificity formula is:

Among these, TN (True Negative) denotes correctly predicted non-O-GlcNAcylation sites; FP (False Positive) denotes incorrectly predicted non-O-GlcNAcylation sites as O-GlcNAcylation sites.

Matthews Correlation Coefficient (MCC): The MCC is a comprehensive evaluation metric that considers both true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), making it particularly suitable for scenarios involving imbalanced datasets. The MCC ranges from -1 to 1, where 1 indicates a perfect classifier, 0 represents random guessing, and -1 signifies completely erroneous classification. The formula for calculating the MCC is:

Accuracy (Acc): Accuracy is a commonly used metric for evaluating a model’s overall performance, representing the proportion of correct predictions made by the model. While accuracy provides an overview of general performance, it may prove insufficient for effectively assessing model performance in scenarios involving imbalanced data. Therefore, this study employs a comprehensive evaluation of model effectiveness by incorporating additional metrics. The accuracy calculation formula is:

Precision: Precision is one of the core metrics for evaluating a model’s recognition accuracy in classification tasks. It assesses the proportion of samples correctly classified as positive among those predicted as positive by the model. In other words, Precision focuses on the reliability of the model’s predictions, specifically "how many of the predicted positive instances are actually correct". A higher Precision indicates that the model is more "conservative" when making positive predictions, resulting in fewer instances of incorrectly classifying negative samples as positive. Precision is a particularly crucial evaluation metric in tasks characterized by extreme positive-negative sample imbalance or high sensitivity to false positives. For instance, in bioinformatics modification site prediction, excessive false positives would necessitate redundant experimental validation. Consequently, high Precision is required to ensure the reliability of model predictions. Precision is defined as:

F1-Score (F1): The F1-score represents the harmonic mean of precision and sensitivity, combining a model’s accuracy in identifying positive samples with its comprehensiveness. In the identification of O-GlcNAcylation sites, the F1-score balances false positives and false negatives, avoiding the bias inherent in relying solely on sensitivity or specificity. The F1-score is a commonly used metric for handling imbalanced datasets, particularly suited to tasks where both false positives and false negatives require consideration. The F1-score is calculated as follows:

AUC (Area Under the Curve): AUC represents the area beneath the ROC curve, measuring a model’s performance across varying classification thresholds. The closer the AUC value approaches 1, the stronger the model’s ability to distinguish positive from negative samples. AUC is independent of specific classification thresholds, making it a crucial metric for evaluating a model’s discriminatory power. In the task of predicting O-GlcNAcylation sites, AUC helps gauge a model’s stability and performance across different scenarios, proving particularly valuable for tasks sensitive to sample imbalance. AUC is derived by calculating the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR). Its formula is expressed as:

Where: TPR denotes true positive rate, also known as sensitivity, with the formula:

FPR denotes the false positive rate, with the formula being:

In the task of identifying O-GlcNAcylation sites, the selection of evaluation metrics is of paramount importance. Sn and Sp respectively measure a model’s ability to identify positive and negative class samples; Precision assesses the proportion of samples correctly classified as positive among predicted positives, thereby reducing experimental validation costs associated with false positives; the F1 combines Precision and Sn, offering a more objective reflection of a model’s actual performance in data imbalance scenarios. MCC provides a global and balanced metric for classification performance, whilst AUC aids in evaluating a model’s discrimination capability across varying classification thresholds. Through a comprehensive evaluation of these metrics, a more thorough understanding of the model's performance in identifying O-GlcNAcylation sites can be gained, thereby providing crucial evidence for model optimization and refinement.

# III.Model Construction

This study addresses the task of identifying O-GlcNAcylation sites on proteins, constructing two specialized ensemble models for serine (S) sites and threonine (T) sites respectively, to enhance prediction accuracy and generalization capability across different datasets. For the S-site dataset, we designed an ensemble model architecture comprising DEBERTa v3, ModernBERT, and DEBERTa v3-BiLSTM: DEBERTa v3 and ModernBERT were employed for deep semantic modelling of amino acid sequences to capture long-range dependencies between residues; DEBERTa v3-BiLSTM further enhanced the extraction of both local and global sequence features through its bidirectional recurrent structure. For the T-site dataset, we constructed an ensemble framework comprising ModernBERT, DEBERTa v3-BiGRU, and DEBERTa v3-BiLSTM. ModernBERT was trained independently. DEBERTa v3-BiGRU underwent joint training of its multiple recurrent layers internally to form a single model. DEBERTa v3-BiLSTM similarly employed joint training for its internal multi-layer LSTMs, thereby fully leveraging their respective sequence modelling capabilities. Following the completion of training for the three models, an ensemble learning strategy was employed to fuse the predictions from different models. This enables the system to maintain stable and efficient predictive performance when confronted with challenges such as varying datasets, differing feature distributions, and imbalanced samples. This study not only leverages the formidable capabilities of multiple pre-trained language models in protein sequence representation but also enhances sequence modelling depth through recurrent neural network architecture, thereby providing a robust solution for the precise identification of O-GlcNAcylation sites.

## A. Fine-Tuned DeBERTa v3 and ModernBERT

During fine-tuning, the AdamW optimizer was employed. By integrating weight decay with an adaptive learning rate adjustment mechanism, it effectively curbs excessive parameter updates, thereby enhancing the model’s convergence speed and generalization performance under conditions of limited samples. The loss function employs Binary Cross-Entropy Loss (BCELoss), providing stable and smooth gradient feedback for binary classification tasks. To further enhance training stability and mitigate overfitting, the LinearLR learning rate scheduling strategy is introduced. This maintains a higher learning rate during the initial training phase to accelerate model convergence, while progressively reducing the learning rate in the later stages to refine parameter updates, thereby yielding more robust feature representations. Through this fine-tuning strategy, DeBERTa v3 fully leverages its decoupled attention mechanism to model long-range dependencies and global semantic information, whilst ModernBERT enhances its perception of local sequence patterns through an optimized normalization strategy and lightweight feedforward architecture. Building upon this foundation, we further incorporate BiLSTM or BiGRU to model bidirectional dependencies within sequences. This enables the overall model to delve deeper into protein sequence features, thereby achieving greater accuracy and stability in the prediction of O-GlcNAcylation sites.

## B. DEBERTa v3 + BiLSTM

First, the protein sequence undergoes preprocessing. Let the given protein sequence be denoted as . Following encoding by DEBERTa v3, the hidden layer output is obtained:, where 𝑑 denotes the dimension of the hidden layer, and each *ℎ𝑖* represents the contextual representation of the 𝑖-th residue. BiLSTM performs bidirectional modelling on 𝐻(DeBERTa), with its forward and backward hidden states being:

After concatenating the bidirectional hidden states, the BiLSTM yields the following output:

The BiLSTM output is further mapped to a binary classification space via a fully connected layer:

Where, 𝜎 denotes the sigmoid activation function, 𝑦𝑡 represents the predicted probability that the 𝑡-th residue is an O-GlcNAcylation site, and 𝑊 and 𝑏 are trainable parameters.

During training, DEBERTa v3 and BiLSTM employ a joint optimization strategy, with the overall model being optimized using the binary cross-entropy loss function (BCELoss):

The optimizer employed is AdamW, which combines weight decay to control the magnitude of parameter updates. Concurrently, the LinearLR learning rate scheduler is utilized to progressively reduce the learning rate, enabling rapid convergence during the early stages of training while allowing for fine-tuning of parameters in the later phases. This approach enhances the model’s generalization capabilities and prediction stability.

## C. DeBERTa v3 + BiGRU

First, protein sequences undergo preprocessing, where each amino acid residue is encoded as a corresponding token sequence. Domains or other functional entities within the sequence are annotated to enrich the input features. The processed sequences are fed as token inputs into the DEBERTa v3 model, whose input comprises word embeddings and relative position encodings. The sequence context is processed within the Encoder via a decoupled attention mechanism to capture global long-range dependencies.

During training, after DEBERTa v3 completes forward propagation, we extract its final hidden layer output 𝐻(DeBERTa) as the sequence feature representation, then feed it into the BiGRU module. BiGRU concurrently models forward and backward sequence information through its gating mechanism (update gate 𝑧𝑡 and reset gate 𝑟t), with the computational process as follows:

The output of BiGRU is further mapped to a binary classification space via a fully connected layer. An activation function then outputs the predicted probability for each site, which is used to determine whether that site is an O-GlcNAcylation site.

DEBERTa v3 and BiGRU employ a joint optimization strategy during training. The overall model is trained using Binary Cross-Entropy Loss (BCELoss), with gradient updates performed via the AdamW optimizer. The learning rate employs a LinearLR scheduler for gradual decay, enabling rapid convergence in the early training stages and subsequent parameter fine-tuning to enhance the model’s generalization capability and prediction stability on limited datasets. Through this design, the DEBERTa v3 + BiGRU architecture fully leverages the Transformer’s global representation capabilities alongside the GRU’s sensitivity to sequential dynamics, demonstrating outstanding performance in the task of O-GlcNAcylation site identification.

## D. Model Ensemble

For the S-site dataset, we constructed a three-model ensemble framework comprising DEBERTa v3, ModernBERT, and DEBERTa v3-BiLSTM. All three sub-models were trained independently on the same training set. Among them, the Transformer-based models (DEBERTa v3 and ModernBERT) primarily extract high-dimensional contextual features from amino acid sequences. Meanwhile, DEBERTa v3-BiLSTM combines Transformer representations with the bidirectional temporal modelling capabilities of BiLSTM, thereby further enhancing its ability to analyze local structures and long-range dependencies. Upon completion of independent training, each sub-model outputs the corresponding glycosylation probability for every residue in the sequence, where *k* denotes the sub-model number and *t* denotes the sequence position. To enhance overall prediction robustness and mitigate potential biases inherent in individual models, this study employs a soft voting strategy for ensemble prediction. This involves calculating a weighted average of the prediction probabilities from the three models to derive the final prediction result:

Where denotes the weight of the k-th sub-model. To maintain model simplicity and prevent overfitting, an equal weighting strategy was employed in the experiments, meaning all three models contributed equally to the prediction results.

This ensemble strategy effectively integrates the complementary strengths of different models in global semantic modelling, local structural capture, and sequential temporal dependencies. This enables the overall system to demonstrate greater robustness when confronting challenges such as data noise, variations in feature distribution, and class imbalance. Experimental results indicate that the S-site ensemble model achieves significant improvements across all metrics compared to individual models, thereby validating the effectiveness of the ensemble strategy for O-GlcNAcylation prediction tasks.

For the T-site dataset, this study constructed an ensemble framework comprising three independently trained sub-models: ModernBERT, DEBERTa v3-BiGRU, and DEBERTa v3-BiLSTM. Among these, ModernBERT is responsible for efficiently extracting global features from sequences, while DEBERTa v3-BiGRU and DEBERTa v3-BiLSTM, based on GRU and LSTM gating architectures respectively, capture dynamic dependencies and structural variations within protein sequences from bidirectional temporal information. This effectively enhances the model’s ability to recognize complex local patterns and residue-level relationships.

Upon completion of training for each of the three models, a predicted probability is output for every site. To synthesize the differing feature preferences across distinct network architectures, this study similarly employs a soft voting strategy to fuse the predictions from all three:

Where denotes the weight of the k-th sub-model. To maintain model simplicity and prevent overfitting, an equal weighting strategy was employed in the experiments, meaning all three models contributed equally to the prediction results.

ModernBERT’s focus on global context, BiGRU’s sensitivity to sequential dynamics, and BiLSTM’s ability to capture long-range dependencies enable the T-site ensemble model to learn O-GlcNAc modification patterns more comprehensively across multiple dimensions. The integrated model demonstrates significant improvements over individual models across several evaluation metrics, validating the stability and predictive advantages of multi-model fusion when processing T-site datasets.

# IV.Results and Discussion

## B. Ablation Study

In experiments at the T-site, a consistent trend was similarly observed. ModernBERT maintained robust baseline performance across metrics including AUC (0.8287) and Acc (0.7946), whilst DeBERTa v3 demonstrated solid performance in Sp (0.7618) and model classification stability. Upon incorporating bidirectional recurrent networks, DeBERTaV3-BiGRU achieved the highest Sp (0.8370) but saw Sn decrease to 0.6852, indicating a tendency towards more conservative judgements. This resulted in only marginal improvements to overall F1 and MCC scores. By contrast, DeBERTaV3-BiLSTM maintains a stable, moderately high level across metrics including Sn, Sp, Acc, and AUC, demonstrating superior balance across multiple performance dimensions. This balanced architecture proves particularly crucial for predicting O-GlcNAc modifications in highly imbalanced scenarios. Based on this, the present study further constructed soft voting ensemble models comprising ModernBERT, DeBERTaBiGRU, and DeBERTaBiLSTM. These achieved the overall optimal Sp (0.8318), F1 (0.6164), and AUC (0.8341) at the T-site. While maintaining relatively stable Sn (0.7111), this approach further enhances the model’s overall discriminative capability within complex sample spaces.

## C. Comparison with Existing Methods

In T-site prediction, LLMO-GlcNAc achieved a Sn of 0.7111, Sp of 0.8318, MCC of 0.4968, Acc of 0.8052, F1 of 0.6164, with an AUC of 0.8341. This similarly demonstrates robust performance, where the AUC represents the highest among all methods, indicating the model’s optimal discriminative capability across different thresholds. YinOYang-1.2 (Sn=0.4167, MCC=0.0414) and NetOGlyc-4.0 (Sn=0.3636, MCC=0.1312) demonstrated significant deficiencies in positive sample recognition and overall predictive capability; Although O-GlcNAcPRED-DL exhibited a slightly higher MCC (0.6033), its F1 (0.5402) remained below that of LLMO-GlcNAc, indicating limitations in balancing positive and negative sample recognition. Furthermore, its AUC performance failed to match LLMO-GlcNAc’s level.

## D. Interpretability Analysis

*1) Feature Space Visualization Using t-SNE:* Similarly, as shown in Fig. S4(a) and Fig. S4(b), at the T site, the ModernBERT features also exhibit a clear separation between green points (positive samples) and yellow points (negative samples), whereas the One-Hot features fail to form distinct category clusters and demonstrate a higher degree of mixing. Overall, these results demonstrate that deep semantic features extracted using the ModernBERT large language model significantly enhance the separability of positive and negative samples within the feature space. This provides more discriminative input features for the classification task of S/T sites, fully validating the effectiveness of the large language model approach in predicting O-GlcNAcylation sites.

|  |
| --- |
| Fig. S4 Visualization of T-site sequence features extracted by different feature extraction methods |

*2) LIME-Based Interpretability Analysis:* LIME constructs linear surrogate models within local neighborhoods by randomly perturbing the original amino acid sequence and evaluating changes in model outputs, thereby quantifying the positive or negative contribution of each amino acid residue at different positions to the model's predictions. Compared to traditional holistic interpretability methods, LIME provides granular explanations at the individual sequence level. This enables simultaneous observation of both the importance ranking of key amino acid types (Top 20 feature map) and the spatial distribution of feature importance across sequence positions (heatmap). This method visualizes the model’s mechanisms, aiding in the verification of whether its predictive logic aligns with established biological principles, and enabling the discovery of potential, previously undescribed sequence features.

To further elucidate the discriminative logic of ModernBERT in predicting O-GlcNAcylation sites, we conducted a local interpretability analysis based on LIME. LIME visualizes the contextual patterns, key residue types, and region-level contribution features relied upon by ModernBERT in the form of feature importance bar charts and sequence-position heatmaps. Owing to ModernBERT’s efficient multi-head rotary attention mechanism and its strong capacity for deep language feature modeling, the model exhibits enhanced representational power across multiple levels, including local sequence environments, cross-segment association patterns, and critical residue combinations. The application of LIME therefore provides a principled means to interpret how these complex representations contribute to the model’s decision-making process.

Fig. S5(a), Fig. S5(b), and Fig. S5(c) present the LIME results for three positive samples, from which a pronounced “positively contributing residue pattern” can be observed. As shown in the Top-20 amino acid feature importance bar charts, serine (S) and threonine (T)—the canonical acceptor residues for O-GlcNAc modification—frequently rank among the highest positive weights, indicating that ModernBERT accurately identifies the core chemical groups essential for modification. In addition, several basic or aromatic residues, such as lysine (K) and tyrosine (Y), consistently exhibit positive contributions across multiple positive samples. These residues may serve as auxiliary signals that facilitate local accessibility, weaken structural rigidity, or enhance enzyme recognition. In the sequence-position heatmaps, positive samples generally display contiguous segments of positive contributions, often centered around the modification acceptor residues. This suggests that the model does not rely on a single site alone, but rather integrates the local structural environment and surrounding amino acid context in its predictions.In contrast, the negative samples shown in Fig. S5(d), Fig. S5(e), and Fig. S5(f) exhibit a distinct “negative-dominant contribution pattern.” Among their Top-20 features, residues such as proline (P), valine (V), and leucine (L) frequently appear with strong negative weights. These residues are known for their high structural rigidity and tendency to form hydrophobic or tightly packed regions, which may hinder OGT recognition and reduce the likelihood of modification. Correspondingly, the position heatmaps of negative samples are dominated by continuous red segments representing negative contributions, with few prominent positive peaks. In some cases, nearly entire regions in the N-terminal or middle portions of the sequence appear as extended “inhibitory segments,” indicating that the overall amino acid composition of these regions is unfavorable for O-GlcNAcylation. Collectively, ModernBERT consistently captures local structural features that are unfavorable for modification, such as rigid segments or potential local folding patterns.

Taken together, the LIME analyses of both positive and negative samples demonstrate that ModernBERT exhibits a biologically plausible discriminative logic in O-GlcNAcylation site prediction. Beyond capturing the direct contributions of the acceptor residues S/T, the model effectively recognizes contextual patterns closely associated with modification, including residue combinations that promote flexibility or accessibility, as well as rigid and hydrophobic contiguous segments that suppress modification. Positive samples are characterized by a pattern of “local activation with reinforced acceptor features,” whereas negative samples display an “overall suppression dominated by rigid residues.” These findings indicate that ModernBERT’s predictions are not driven by individual amino acids in isolation, but instead arise from an integrated representation of both local windows and broader sequence contexts. Overall, the interpretability results not only validate the reliability of the model’s predictions but also provide valuable insights into the sequence determinants underlying O-GlcNAcylation.

|  |
| --- |
| Fig. S5. LIME interpretability analysis of ModernBERT |

## E. Web Server

This platform enables users to input protein sequences online and rapidly obtain predictions for O-GlcNAcylation sites. Fig. S6(a) presents the LLMO-GlcNAc platform's comprehensive overview page and classification prediction entry point. Users may intuitively grasp the model's functional positioning via this interface and access the prediction module with a single click. Fig. S6(b) depicts the interactive interface for the O-GlcNAcylation site prediction module: users may either manually input multiple protein sequences (one sequence per line, each 101 characters in length), with the system sequentially returning corresponding prediction results presented clearly and intuitively in the results area; or perform batch predictions by uploading .fasta format files, where the platform automatically aggregates prediction results and generates downloadable output files. The overall interface design prioritizes simplicity and interactivity, employing a light background to highlight key information. A multi-colored modular layout enhances content differentiation, thereby effectively improving the platform's user experience and academic presentation.

|  |
| --- |
| Fig. S6. LLMO-GlcNAc Web |

# References

1. Khan, Y.D., et al., *iSUMOK-PseAAC: prediction of lysine sumoylation sites using statistical moments and Chou’s PseAAC.* PeerJ, 2021. **9**: p. e11581.

2. Chou, K.C., *Prediction of protein signal sequences and their cleavage sites.* Proteins: Structure, Function, and Bioinformatics, 2001. **42**(1): p. 136-139.

3. Li, W. and A. Godzik, *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.* Bioinformatics, 2006. **22**(13): p. 1658-1659.

4. Rives, A., et al., *Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences.* Proceedings of the National Academy of Sciences, 2021. **118**(15): p. e2016239118.

5. Rao, R.M., et al. *MSA transformer*. in *International conference on machine learning*. 2021. PMLR.

6. Meier, J., et al., *Language models enable zero-shot prediction of the effects of mutations on protein function.* Advances in neural information processing systems, 2021. **34**: p. 29287-29303.

7. Nayan, N.M., et al. *Smote oversampling and near miss undersampling based diabetes diagnosis from imbalanced dataset with xai visualization*. in *2023 IEEE Symposium on Computers and Communications (ISCC)*. 2023. IEEE.

8. Vaswani, A., et al., *Attention is all you need.* Advances in neural information processing systems, 2017. **30**.

9. Devlin, J., et al. *Bert: Pre-training of deep bidirectional transformers for language understanding*. in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019.

10. Bai, J., et al., *Syntax-BERT: Improving pre-trained transformers with syntax trees.* arXiv preprint arXiv:2103.04350, 2021.

11. Warner, B., et al. *Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference*. 2025. Vienna, Austria: Association for Computational Linguistics.

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

13. Cho, K., et al. *Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation*. 2014. Doha, Qatar: Association for Computational Linguistics.

14. Medsker, L.R. and L. Jain, *Recurrent neural networks.* Design and applications, 2001. **5**(64-67): p. 2.