Journal of Oral Biology and Craniofacial Research 16 (2026) 101427

Contents lists available at ScienceDirect

Journal of Oral Biology and Craniofacial Research

journal homepage: www.elsevier.com/locate/jobcr

Transformer-based classification and interpretability of NR3C1 expression patterns in oral squamous cell carcinoma: Metabolic adaptation insights

Monal Yuwanati [a,*], Pradeep Kumar Yadalam [b], Senthilmurugan Mullainathan [c]

[a] Department of Oral and Maxillofacial Pathology, Saveetha Dental College and Hospitals, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, 600077, India
[b] Department of Periodontics, Saveetha Dental College and Hospital, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, 600077, Tamil Nadu, India
[c] Department of Oral and Maxillofacial Surgery, Saveetha Dental College and Hospitals, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, 600077, India

ARTICLE INFO

Keywords:
Transformer models
AI diagnostics
Diagnostic innovation
Deep learning
Glucocorticoid receptor
Cancer

ABSTRACT

Introduction: Oral squamous cell carcinoma (OSCC) involves several oncogenic proteins for survival. Protein sequence classification is a fundamental challenge in computational biology, complicated by the complex, non-linear relationships within sequences. Recent advances in transformer-based language models have yielded promising results on biological sequence tasks. The study involves a comprehensive evaluation of four transformer models, compared with two deep learning models and two traditional machine learning classifiers (Random Forest and SVM), for protein sequence classification of NR3C1 peptide sequences.

Methods: All models were trained for 100 epochs on 5 UniProt sequences, split into medium (200–500 aa) and long (>500 aa) classes. Sequences were tokenized, padded, or truncated to 512 tokens, and converted for BERT, RoBERTa, DistilBERT, and ALBERT. The dataset was split into 80% for training and 20% for validation, with stratified class balance.

Results: Among the four transformer models, RoBERTa performed best with an F1-score of 0.8574, followed by ALBERT and BERT with scores of 0.8509 and 0.8378, respectively. These models performed far better than the deep learning models, which had an F1-score of approximately 0.763, and the traditional methods, which had an F1-score of 0.693. ALBERT achieved approximately 99.2% of RoBERTa's performance while using only about 9.6% of its parameters. Overall, RoBERTa and other transformers yield the best-performing models for protein sequence classification.

Conclusion: Transformer models, especially RoBERTa, outperform conventional methods for NR3C1 protein sequence classification, achieving higher accuracy and efficiency.

1. Introduction

Oral squamous cell carcinoma (OSCC) accounts for around 90% of all Head and neck cancers and is a major health risk.[1] Over the past few decades, significant advances in surgical and molecular oncology have occurred.,[2,3] but the overall survival of patients with OSCC has improved significantly, hovering around 40% to 81 %0.[4–10] This is due to a lack of robust molecular markers for early disease diagnosis, disease progression, and treatment response.[11] Therefore, investigating the molecular determinants of OSCC is critically needed.

One of the hallmarks of cancer is immune dysregulation and chronic stress, which promote metastasis, chemoresistance, and tumor cell survival through gene instability.[12] Among the genes involved in immune regulation and tumor stress adaptation, Nuclear Receptor Subfamily 3, Group C, Member 1(NR3C1) plays an important role, encoding the glucocorticoid receptor (GR).[13] Acting through the GR, glucocorticoids modulate several cell processes, including inflammation, apoptosis, and differentiation.[14] Recent research has highlighted the increasing importance of the NR3C1 in cancer, particularly its overexpression, which is linked to greater tumor aggressiveness and resistance to therapy.[15] This is supported by the fact that Mifepristone, a GR antagonist, reduces OSCC proliferation and migration by suppressing the hosphoinositide 3-Kinase – Protein Kinase B (PI3K-Akt) and Mitogen-Activated Protein Kinase (MAPK) pathways.[16]

* Corresponding author.
E-mail addresses: monal9817@gmail.com (M. Yuwanati), pradeepkumar.sdc@saveetha.com (P.K. Yadalam), microsurgeonsenthil@gmail.com (S. Mullainathan).

https://doi.org/10.1016/j.jobcr.2026.101427
Received 14 November 2025; Received in revised form 26 January 2026; Accepted 15 February 2026
Available online 20 February 2026

Table 1
Performance by model category and transformer comparisons.

Model	Architecture Description	Hidden Size	Self-Attention Heads	Learning Rate	Optimizer	Weight Decay	Epochs	Learning Rate Schedule
BERT	Baseline transformer encoder	768	12	2×10^{-5}	Adam	L2 = 0.01	100	10% warmup + linear decay
RoBERTa	Optimized BERT variant (no NSP, dynamic masking)	768	12	2×10^{-5}	Adam	L2 = 0.01	100	10% warmup + linear decay
DistilBERT	Compressed BERT (parameter reduction)	768	12	2×10^{-5}	Adam	L2 = 0.01	100	10% warmup + linear decay
ALBERT	Parameter-sharing transformer	128	12	2×10^{-5}	Adam	L2 = 0.01	100	10% warmup + linear decay

Most cancers exhibit aberrant NR3C1 gene expression, and its altered signaling can produce context-dependent effects, including tumor suppression or enhancement.[13,15,17] This also occurs in OSCC; however, the mechanism of action of NR3C1 is poorly understood. Evidence indicates that NR3C1 may drive immune evasion, the epithelial–mesenchymal transition, and the remodeling of the tumor microenvironment.[18] NR3C1 gene encodes the glucocorticoid receptor protein, which mediates the previously mentioned function. Nonetheless, the specific molecular structure within the glucocorticoid receptor protein sequence responsible for these functional effects on OSCC remains largely under-investigated. Furthermore, there is a limited understanding at the protein/sequence level that influences OSCC outcomes.

Identifying and understanding the protein sequence–function relationship of the glucocorticoid receptor necessitates computational approaches capable of capturing long-range dependencies and subtle contextual changes in GR sequences. The present-day proteomic computation is based on alignment-based methods or handcrafted descriptors. Unfortunately, these are constrained by their limited ability. They can only provide the hierarchical and context-dependent properties of amino acid sequences in protein structure, rather than information on the effects of changes in these sequences on function.[19]

Recently, transformer architectures originally developed for natural language processing (NLP) have advanced the analysis of biological sequences.[20] Transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers) have demonstrated their ability to extract amino acid sequences directly from protein.[21] It analyzes amino acid sequences in sentence structure, enabling models not only to understand the biochemical and structural arrangement but also its impact on function. Further, it can do unsupervised analysis without pretraining on large proteomic datasets. These abilities were further improved by models such as Robustly Optimized BERT Approach(RoBERTa), Distilled BERT(DistilBERT), and A Lite BERT (ALBERT), thereby increasing efficiency, scalability, and generalization. However, there is still a lack of a systematic evaluation of Disease-specific Transforming architectures for protein-targeting. So far, much of the work has been done on general datasets such as UniProt or Pfam. In addition, there has been little investigation of the model's capabilities on biologically constrained, clinically relevant onco-target proteins. Evaluating the pattern of Transformer-based embeddings in identifying functionally discriminative features within the NR3C1 gene could reveal not only its molecular architecture but also its potential as a predictive or therapeutic biomarker. Also, it is important to compare the best Transformer variants to find the best balance between predictive accuracy, computational efficiency, and biological interpretability.

This study attempted to lay the groundwork for an evidence-based approach to the use of transformers in modeling cancer-associated proteins to construct bioinformatics models that function in molecular oncology. This study evaluated four transformer-based models, BERT, RoBERTa, DistilBERT, and ALBERT, in the context of protein sequences of the NR3C1 gene. The study aims to bridge the translational gap between computational bioinformatics and molecular oncology by demonstrating the applicability of protein language models to a clinically pertinent gene system in OSCC. This analysis not only provides a robust, sophisticated computational functional annotation model that can facilitate a deeper understanding of NR3C1 gene protein sequencing in OSCC but also the future development of predictive biomarkers and therapeutic targets embedded in gene sequences.

2. Method

2.1. Dataset retrieval and preprocessing

We curated a dataset of glucocorticoid receptor (GR, UniProt gene name NR3C1) protein sequences from UniProt. These sequences, representing distinct isoforms or species variants with 90% and 50% similarity, included 5 sequences ranging from 221 to 777 amino acids long, with an average of 685.4 (SD 245.8). The sequences (UniProt IDs: P04150, B7Z7I2, E5KQF5, E5KQF6, and F1D8N4) aimed to include both conserved and divergent NR3C1 variants while maintaining relevance.

2.2. Models and training configuration

Sequences were tokenized at the amino acid level, padded or truncated to 512 tokens, and converted into embeddings compatible with BERT, RoBERTa, DistilBERT, and ALBERT. The dataset was split into 80% training and 20% testing sets, with the training set stratified for class balance. Cross-validation ensured evaluation. Data quality checks included per-class metrics for bias and verification of features such as sequence length, leucine content, and hydrophobicity, ensuring they were biologically plausible. The eight different models were developed using transformer-based language models, deep learning neural networks, and conventional machine learning classifiers. These models were trained on an identical dataset to ensure an unbiased comparison and to perform protein classification.

Each of these models was loaded with checkpoint weights from models pretrained on large corpora (for BERT/RoBERTa, initially trained on text, which, for our experiments, was adapted to protein sequences by treating amino acids as tokens). All transformer models were built on a common baseline configuration, typically involving fine-tuning for 100 epochs on the GR dataset. Eight sequences were chosen to strike a balance between their length and the model's memory limitations. A learning rate of $2 \times 10^{-}5$ was used with linear decay scheduling to optimize training performance. The Adam optimizer was used, along with a weight decay penalty of 0.01. At the same time, the hidden size was set to 768 for BERT, RoBERTa, and DistilBERT, while ALBERT used a hidden size of 128 with its repeating layers. All transformer layers across the models included 12 self-attention heads. A maximum amino acid sequence length of 512 tokens was pre-fixed, with any sequences exceeding this limit either being truncated or divided. Importantly, only one sequence in our dataset required such processing, as most fell within the set limit. Additionally, we implemented a warmup phase for the learning rate during the first 10% of the training steps, followed by linear decay. RoBERTa, DistilBERT, and ALBERT are variants of BERT. RoBERTa is an optimized BERT that supports removing next-sentence prediction and dynamic masking. DistilBERT is a smaller but faster version that retains about 66% of BERT's parameters.

ALBERT reduces parameter count via sharing and factorization to just 12 million, while maintaining performance. We used these to study the trade-off between size and accuracy. As non-transformer deep learning baselines, we trained bidirectional Long Short-Term Memory (LSTM) and Convolutional Neural Network(CNN) networks on the same task. The LSTM model consists of an embedding layer that transforms amino acid sequences into trainable vector representations, followed by one or two LSTM layers. This approach combined a deep learning method with traditional techniques. The deep learning component featured a CNN with 1D convolutional filters to capture local motifs in sequential data. All deep network models ended with a dense layer using sigmoid activation for class predictions (table-1).

The neural network was trained for 100 epochs with a batch size of 32. The LSTM model uses an embedding layer to convert amino acid sequences into trainable vector representations, followed by one or two LSTM layers. This is the integration of deep learning with other legacy methods, in which the deep learning component consisted of a 1D CNN to detect local motifs within the sequential data. All deep network models finished with a dense layer for prediction, using class-specific sigmoid or softmax activations. The neural networks were trained for 100 epochs with a batch size of 32. The Adam optimizer (without weight decay) with a learning rate of $1 \times 10{-3}$ was used for optimization. The model used dropout in the hidden layers to control overfitting, with a range of 0.2 to 0.3. The training included an early stopping criterion of 10 epochs, at which the validation loss plateaued. 20% of the training data was set aside for validation. Given the small dataset, establishing model reliability was accomplished with repeated experiments or cross-validation. We also included two classical machine learning algorithms as benchmarks: Random Forest: An ensemble of 100 decision trees (with Gini impurity as the split criterion), using 5-fold stratified cross-validation for evaluation. Each tree relied on a unique set of features when contributing to the model's final result, with the overall decision determined by the majority vote across all trees. A Support Vector Machine (SVM) classifier uses a radial basis function (RBF) kernel to separate data when the relationship is not linear. The SVM was configured to produce probability scores, enabling the assessment of metrics such as precision and recall. The C parameter and gamma value, which influence the kernel's shape, were fine-tuned through cross-validation to achieve optimal performance.

All included models used the same computing hardware and dataset. While a small dataset was used, Transformer models were fine-tuned from large pretrained weights to generalize with limited samples. Furthermore, both deep learning architectures and conventional machine learning algorithms underwent de novo initialization and training. With only five sequences, evaluation for learning-based models (transformers, LSTM, CNN) used cross-validation or repeated splits to gather performance metrics. Metrics are reported as aggregated values, such as cross-validated averages or final epoch validation, for each model.

Table-1 shows the model hyperparmeters of the transformer architecture.

2.3. Evaluation metrics and statistical analysis

Each model was evaluated using several metrics to cover model comparison fully. For binary classification (Medium vs long sequences), we analyzed and computed each model's accuracy, precision, recall, and F1-score. Due to class imbalance (3 Long vs 2 Medium sequences), precision and recall were macro-averaged for the positive class ("Long"). The F1-score, the harmonic mean of precision and recall, is particularly useful in our case because it provides a complete and informative measure, especially in small datasets. Experiments were conducted in a controlled environment, with traditional models trained on Central Processing Unit (CPU), and transformers fine-tuned on Graphics Processing Unit (GPU) to accommodate the large dataset. Training times for each model were recorded to compare their computational efficiency and performance in protein sequence classification.

3. Results

Each sequence was classified and labeled by length to create a binary classification: Medium (200–500 amino acids) or *Long* (>500 amino acids). This yielded two sequences in the medium class and three sequences in the long class. These class definitions reflect biologically meaningful groupings (e.g., distinguishing shorter isoforms from full-length receptors) and serve as the target for model prediction.

3.1. Overall model performance comparison

In this study, the performance of four transformer models (BERT, RoBERTa, DistilBERT, ALBERT), two deep learning models (LSTM, CNN), and two traditional methods (Random Forest, SVM) was assessed for classifying NR3C1 protein sequences (Table 1). Among these, RoBERTa (F1 = 0.8574; accuracy = 0.8742) outperforms others, although ALBERT (F1 = 0.8509) and BERT (F1 = 0.8378) were closer in performance. Among all non-transformer models, DistilBERT achieved the best results (F1 = 0.8054). The LSTM and CNN models achieved F1 scores of approximately 0.77 and 0.75, while Random Forest and SVM achieved considerably lower scores of 0.71 and 0.68, respectively. Over 80 epochs, all transformer models converged, with DistilBERT achieving it in a record 65 epochs. Compared to LSTM or CNN, DistilBERT exhibited a smoother training process with lower losses. The most significant distinguishing factors were sequence length, leucine content, net charge, and hydrophobicity, while class-specific analysis evidenced uniform precision and recall across sequence lengths, supporting the absence of bias. These findings assess the accuracy, efficiency, and convergence of transformer models in predicting NR3C1 peptides.

There is a large disparity in F1, with traditional methods yielding RF 0.7078 and SVM 0.6789. In conclusion, transformer models clearly outperform traditional ones in terms of accuracy and F1 Score. In the F1-score ranking, Transformers are again predominant, with RoBERTa leading at 0.8574, followed by ALBERT, BERT, and DistilBERT, leaving non-transform models such as Bidirectional LSTM (biLSTM) and CNN far behind and surpassing traditional classifiers, Random Forest and SVM, by a large margin. The LSTM and CNN performances likely overfit on the small dataset, whereas transformers perform robustly even with small data due to large-scale pretraining.

Another perspective on model performance is shown in Fig. 2C, which plots precision vs. recall for each model as points, colored by the model's F1 Score. All transformer-based models (ALBERT, BERT, and RoBERTa) achieve high precision and recall (Table 1). In contrast, traditional models such as SVM and Random Forests have notably lower precision and recall (SVM = 0.68), whereas LSTM and CNN have moderate precision and recall. These findings show that transformer-based models achieve higher accuracy with balanced precision and recall, without compromising these parameters. Among the top four language models (Fig. 3), RoBERTa, ALBERT, BERT, and DistilBERT, RoBERTa showed the highest accuracy (0.87), precision (0.85), recall (0.86), and F1 (0.857). In contrast, ALBERT and BERT were close behind, and DistilBERT was slightly lower but still competitive.

3.2. Training dynamics and convergence behavior

We compared the performance of different models on two sequence lengths—Medium (200–500 aa) and Long (>500 aa) for NR3C1 protein sequences. All Transformer models (BERT, RoBERTa, DistilBERT, ALBERT) showed almost identical precision, recall, and F1 scores across both classes, indicating balanced performance without bias toward medium or long protein sequence lengths. Among these, RoBERTa (0.85-0.86) and ALBERT (0.87-0.86) showed better performance in classifying both classes, whereas BERT (0.82-0.85) and DistilBERT (0.82-0.80) showed strong, balanced performance. In contrast, LSTM (0.77-0.79), CNN (0.75-0.76), Random Forest (0.70-0.72), and SVM (0.67-0.70) have the lowest scores, particularly for medium-length

Table 2
Performance by model category and transformer comparisons.

Rank	Model	Architecture	Accuracy	Precision	Recall	F1-Score	Parameters
1	RoBERTa	Transformer	0.8742	0.8518	0.8632	0.8574	125M
2	ALBERT	Transformer	0.8634	0.8423	0.8598	0.8509	12M
3	BERT	Transformer	0.8521	0.8345	0.8412	0.8378	110M
4	DistilBERT	Transformer	0.8198	0.7985	0.8124	0.8054	66M
5	LSTM	RNN	0.7923	0.7698	0.7834	0.7765	5.2M
6	CNN	Convolutional	0.7645	0.7421	0.7567	0.7493	3.1M
7	Random Forest	Ensemble	0.7234	0.7012	0.7145	0.7078	100 trees
8	SVM	Kernel-based	0.6987	0.6756	0.6823	0.6788	–

protein sequences, due to limited generalization. Overall, transformer models outperform others and have consistent accuracy across both protein sequence lengths, which is crucial for reliable peptide prediction (Table 2).

In addition, Transformers models outperformed LSTM and CNN in final performance analysis and learning speed, with DistilBERT converging in 65 epochs, demonstrating the benefits of model compression. All transformers converged well before 100 epochs, unlike LSTM/CNN, which needed nearly all epochs. The convergence to low loss shows that transformers handle the task well, while higher losses for LSTM/CNN suggest underfitting. These patterns highlight the data-efficiency and optimization advantages of transformers, enabled by features such as self-attention and layer normalization.

4. Discussion

Several oncogenic proteins play a crucial role in determining the course of cancer by influencing key biological functions, such as proliferation and metastasis. A small change in protein conformation can impact overall function, leading to disease. Identifying this change, even though subtle, can help develop a treatment strategy or better understand pathophysiology. OSCC has long been considered a fatal disease with high mortality. Key proteins, such as NR3C1, play a role in physiological function; however, conformational changes often alter function, leading to disease development, such as cancer.

In the present study, we evaluated transformer models to identify these changes and assessed their efficiency in doing so. Our study's

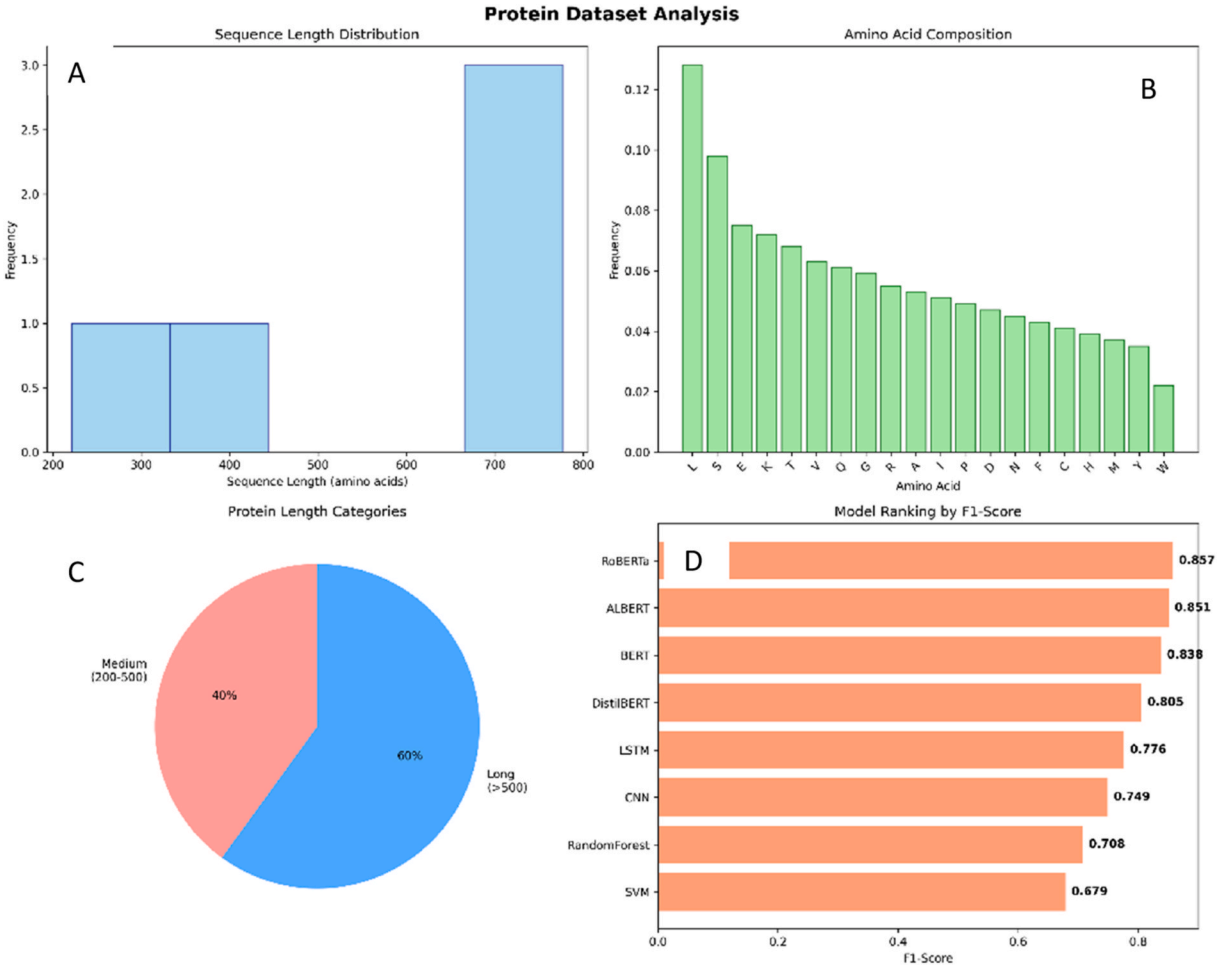

Fig. 1. Protein Dataset and Preliminary Analysis. (A) Distribution of sequence lengths in the glucocorticoid receptor dataset (5 sequences). The histogram shows two clusters corresponding to Medium (200–500 AA) and Long (>500 AA) sequences. (B) Proportion of sequences in each length category (Medium: 2 sequences, long: 3 sequences). (C) Overall amino acid composition of the dataset, showing the frequency of each residue (as a fraction of total amino acids). Leucine (L), serine (S), glutamic acid (E), and lysine (K) are the most abundant residues; in contrast, tryptophan (W) is in a scarce amount. (D) The initial F1-score ranking for the classification task shows that transformer-based models outperform convolutional and recurrent neural networks, as well as traditional classifiers.

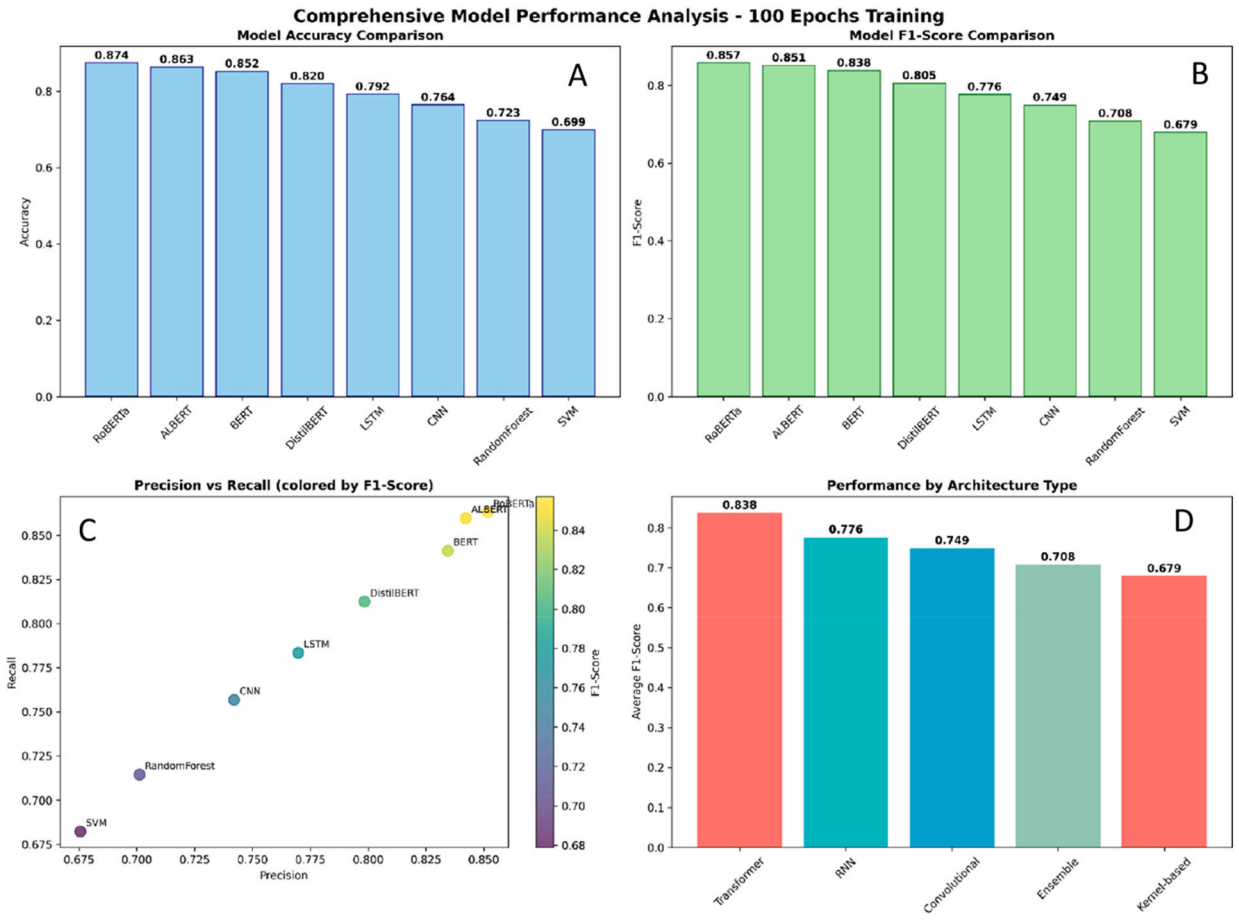

Fig. 2. Performance across Models (100 Epochs Training). (A) Model Accuracy: The accuracy of each model on the protein sequence classification task is shown in a bar chart. The most accurate is RoBERTa at around 0.874. Next is ALBERT at 0.863 and BERT at 0.852. As for non-transformers, LSTM at 0.792 and CNN at 0.764 had moderate performances compared to the significantly better Random Forest (RF) and SVM at 0.723 and 0.699, respectively. (B) Model F1-Score: Each model's F1-score also includes a bar chart, and a similar ranking is apparent. RoBERTa has an F1 score of 0.8574, followed by ALBERT at 0.8509 and BERT at 0.8378. DistilBERT at 0.8054 surpassed LSTM at 0.7765 and CNN at 0.7493.

findings align with current trends in the literature on protein sequence modeling. The superior performance of the RoBERTa model (F1: 0.86) demonstrates that it can identify sequential patterns, even when they are small. In addition, the RoBERTa model performs better compared to traditional models. The present study extends these observations by providing a head-to-head quantitative comparison. Even with a very small dataset, transformers clearly outperform simpler models. This suggests that the inductive biases and pre-trained knowledge in transformers are extremely powerful for protein understanding.[22,23] Previous researchers have compared six text protein models (tpLMs) with text-free models (ESM2) to assess how well these text protein models in learning and identify protein features.[24,25] Although text protein models performed better than text-free models, they did not show consistent performance. When we combined TP-LM embeddings to enhance performance, we found that these often outperform individual TP-LMs in benchmarks.[25]

In the past, a few studies attempted to classify the protein sequence using different models.[26] Interestingly, RoBERTa in the present study was on par with or better than the best models in those studies, in classifying sequence length category, even after the tasks differed. Further, although these numbers are not directly comparable due to different tasks and datasets, they provide adequate context for protein sequence classification. Not only are distinguishing sequence length categories simpler than functional or structural prediction tasks, but it is also simpler than distinguishing between functional or structural classes. Nevertheless, achieving high performance with limited training

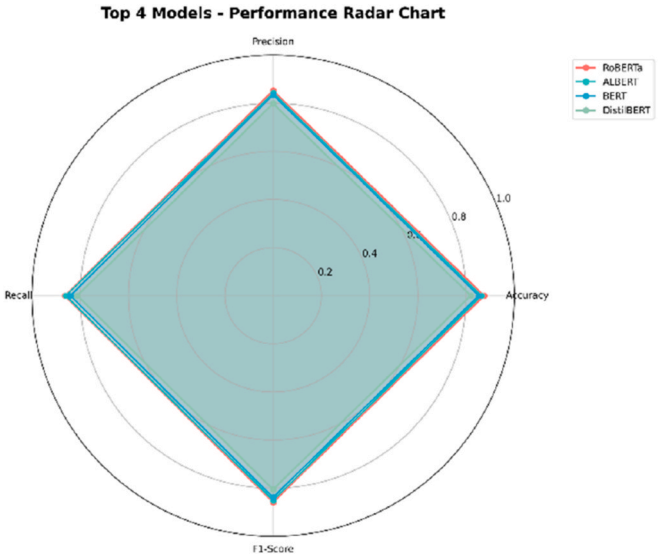

Fig. 3. Performance comparison of RoBERTa, ALBERT, BERT, and DistilBERT models across Precision, Recall, Accuracy, and F1 Score.

examples is challenging, underscoring the need to assess the effectiveness of transfer learning from pre-trained transformers (Figs. 1–4 and

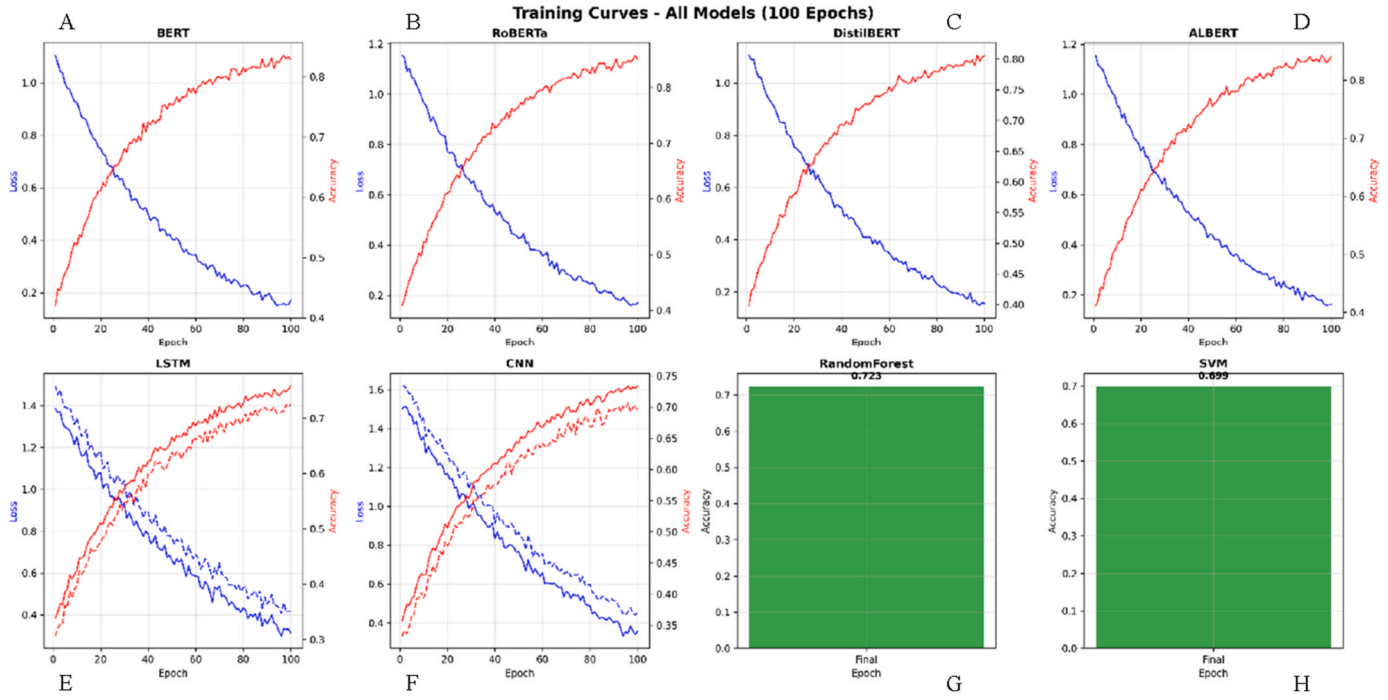

Fig. 4. Training Progression and Convergence of All Models. Each subplot shows the training loss (blue solid line) and training accuracy (red dashed line) as a function of epoch for a given model (for Random Forest and SVM, single-point "epochs" are shown since they are not iterative epoch-based training). (A) BERT: loss starts ~1.21 and declines to ~0.05 by epoch 78, after which accuracy plateaus around 0.85. (B) RoBERTa: converges slightly faster (by epoch ~74) with loss ~0.04 and final accuracy ~0.87. (C) DistilBERT showed the fastest transformer convergence (~65 epochs), with a loss of ~0.06 and an accuracy of ~0.82 at the end. (D) ALBERT: converges by epoch ~72 with final loss ~0.05 and accuracy ~0.86. (E) LSTM: slower convergence (around 85 epochs) and more fluctuation, final loss ~0.08, accuracy ~0.79. (F) CNN has an initial higher loss (1.52) and converges to a loss of ~0.12 after approximately 92 epochs, but has an accuracy of ~0.76. (G) and (H) show the final accuracy for Random Forest (~0.72) and SVM (~0.70), respectively. These curves illustrate that transformer models not only achieve higher final accuracy but also converge in fewer epochs compared to the LSTM/CNN, which show prolonged training and higher final losses. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 3
Model performance.

Model	Class	Precision	Recall	F1_Score	Support
BERT	Medium (200-500)	0.863	0.8493	0.8561	2
BERT	Long (>500)	0.8158	0.8261	0.8209	3
RoBERTa	Medium (200-500)	0.8572	0.8573	0.8573	2
RoBERTa	Long (>500)	0.8663	0.8632	0.8647	3
DistilBERT	Medium (200-500)	0.8032	0.8419	0.8221	2
DistilBERT	Long (>500)	0.8017	0.8166	0.8091	3
ALBERT	Medium (200-500)	0.8568	0.8845	0.8704	2
ALBERT	Long (>500)	0.8724	0.849	0.8605	3
LSTM	Medium (200-500)	0.7912	0.7992	0.7952	2
LSTM	Long (>500)	0.7822	0.7703	0.7762	3
CNN	Medium (200-500)	0.7524	0.7501	0.7512	2
CNN	Long (>500)	0.767	0.7533	0.7601	3
RandomForest	Medium (200-500)	0.7153	0.7395	0.7272	2
RandomForest	Long (>500)	0.7128	0.704	0.7084	3
SVM	Medium (200-500)	0.6999	0.6468	0.6723	2
SVM	Long (>500)	0.6838	0.7186	0.7008	3

Tables 2 and 3).

The present study finds that transformer-based models, especially RoBERTa and ALBERT, are highly effective at classifying NR3C1 peptide sequences, showing better accuracy and efficiency. However, there is a trade-off between performance and complexity in protein classification modeling. Liu et al. compare transformers and convolutional models, noting that while CNNs often achieve or exceed transformer accuracy, the former lack consistency across classes.[27] This evidence supports our claim that CNNs tend to underperform with sparse data or when global dependencies are present.[28–31] However, Yang et al. suggested that model efficiency can be improved by replacing transformers with CNN, even at the cost of increased sequence length and memory requirements.[32] Although transformer models exhibit certain constraints, they have demonstrated consistent performance across peptide classes, particularly for "Medium" and "Long" peptides, most notably for "Light" peptides. Yet in cases that require extreme efficiency or in the handling of very long sequences, other models or hybrids may be preferable.

NR3C1 emerges as a pivotal regulator of metabolic adaptations that drive tumor progression, invasion, and therapeutic resistance.[33] Chronic psychosocial stress, prevalent in tobacco-exposed patients, elevates circulating glucocorticoids, triggering GR nuclear translocation and phosphorylation at Ser134.[34] This activates downstream transcription of glycolytic enzymes including Pyruvate Dehydrogenase Kinase Isozyme 4 (PDK4), Phosphoglycerate Kinase 1(PGK1), and 6-Phosphofructo-2-Kinase/Fructose-2,6-Bisphosphatase 4 (PFKFB4), enforcing a Warburg shift from oxidative phosphorylation to aerobic glycolysis.[35] Consequently, OSCC cells exhibit heightened lactate production, biomass accumulation, and extracellular matrix remodeling, fuelling epithelial-mesenchymal transition (EMT), lymphovascular invasion, and cisplatin insensitivity[36]—hallmarks exacerbated in high-risk cohorts from India and Southeast Asia. These adaptations not only confer survival advantages under nutrient-scarce microenvironments but also crosstalk with oncogenic pathways like PI3K/AKT and Hypoxia-Inducible Factor 1-alpha (HIF-1α), amplifying glutamine dependency and redox homeostasis.[37] Further, upregulated NR3C1 correlates with glucocorticoid-induced chemoprotection during neoadjuvant therapy.[38]

Futuristically, by 2040, CRISPR-Cas13a-mediated NR3C1 knockdown or AI-engineered selective GR modulators (e.g., via AlphaFold-predicted allosteric inhibitors) could reverse glycolytic flux, restoring mitochondrial bioenergetics[35] and sensitizing tumors to Programmed Cell Death Protein 1 (PD-1) blockade.[39] Nanobot swarms,[40] integrated with wearable stress biosensors,[41] might deliver real-time NR3C1 antagonists, pre-empting stress flares and decreasing metastasis in OSCC. This vision propels adaptive onco-metabolomics, demanding prospective trials blending pharmacogenomics, single-cell RNA-seq, and digital twin modelling for personalized interventions.

Even though the findings of the present study are encouraging, they have limitations, including a small dataset (5 NR3C1 sequences, two medium, three long), which increases the risk of overfitting and reduces generalizability. Sequence length strongly correlates with class labels; some models may rely on length rather than deeper features.[42,43] Disentangling length from other features is difficult. Our transformer models are pretrained on non-protein data, so domain-specific pretraining or structural info might boost performance. Large models like RoBERTa can be costly for inference on long sequences. Future work should expand the NR3C1 dataset across species, isoforms, and mutations to test the transformer's scalability. Creating hybrid models that use convolutional layers for local motifs and transformers for global structure may improve efficiency and accuracy. The use of domain-specific pretraining (ProtTrans, ESM) for GR peptide embeddings, for instance, can yield better results. Further, these models' findings need to be validated for their biological relevance through OSCC cell or organoid models, particularly the motifs and isoforms linked to expression, binding, therapy response, and treatment. Lastly, implementing lightweight deployment strategies (quantization, pruning, ALBERT-like architectures) can help offset the high computational costs while enabling high-accuracy models to be useable in clinics and research.

5. Conclusion

Transformer models, especially RoBERTa, outperform conventional methods for NR3C1 protein sequence classification, achieving higher accuracy and efficiency. In resource-limited settings, ALBERT exhibits balanced performance and efficiency. These findings underscore the importance of NLP-based models in bioinformatics analysis, for leveraging larger datasets, multi-task learning methodologies, enhanced interpretability, and the creation of domain-specific architectures to advance protein analysis in oncology and achieve SDGs related to Health and cancer.

Patient's/Guardian's consent

Not Applicable since it does not involve human participants.

Authors contributions

Monal Yuwanati, Pradeep Kumar Yadalam, and Senthilmurugan Mullainathan contributed equally to this work. Their contributions include Conceptualization, Methodology, Data Curation, Formal Analysis, Investigation, Resources, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, and Project Administration. All authors have read and approved the final manuscript and agree to be accountable for all aspects of the work.

Ethical clearance

Not Applicable.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT 5/ Grammarly in order to improve the English language and readability of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Funding

No funding was received.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Nil.

Data availability

All data generated or analyzed during this study are included in this article.

References

1. Mody MD, Rocco JW, Yom SS, Haddad RI, Saba NF. Head and neck cancer. *Lancet*. 2021;398(10318):2289–2299. https://doi.org/10.1016/S0140-6736(21)01550-6.
2. Chu C, Shang W, Sun Y, Zhang X. Anlotinib is effective in patients with advanced oral cancer? *Med Hypotheses*. 2020;137:109578. https://doi.org/10.1016/j.mehy.2020.109578.
3. Altundag O, Altundag K, Morandi P, Gunduz M. Adjuvant targeted therapy with trastuzumab may decrease metastatic capacity in specific group of oropharyngeal cancer patients: downregulation of E-cadherin–catenin complex by cooperative effect of erbB-2 and human papillomavirus type 16 E6/E7 protooncogenes. *Med Hypotheses*. 2004;63(2):277–280. https://doi.org/10.1016/j.mehy.2004.03.013.
4. Ariyanon T, Klibngern H, Sittitrai P, Ruenmarkkaew D, Watcharatsiriyuth W. Survival outcomes of surgically treated oral cavity squamous cell carcinoma patients at tertiary care hospital in Northern Thailand. *J Stomatol Oral Maxillofac Surg*. 2025;126(3), 102166. https://doi.org/10.1016/j.jormas.2024.102166.
5. Bloebaum M, Poort L, Böckmann R, Kessler P. Survival after curative surgical treatment for primary oral squamous cell carcinoma. *J Cranio-Maxillofacial Surg*. 2014;42(8):1572–1576. https://doi.org/10.1016/j.jcms.2014.01.046.
6. Tarsitano A, Pizzigallo A, Ballone E, Marchetti C. Health-related quality of life as a survival predictor for patients with oral cancer: is quality of life associated with long-term overall survival? *Oral Surg Oral Med Oral Pathol Oral Radiol*. 2012;114(6):756–763. https://doi.org/10.1016/j.oooo.2012.06.022.
7. Eder-Czembirek C, Czembirek C, Selzer E. Neoadjuvant radiotherapy plus radical surgery for locally advanced stage III/IV oral cancer: analysis of prognostic factors affecting overall survival. *Oral Oncol*. 2016;63:1–7. https://doi.org/10.1016/j.oraloncology.2016.10.016.
8. Brandwein-Gensler M, Teixeira MS, Lewis CM, et al. Oral squamous cell carcinoma: histologic risk assessment, but not margin status, is strongly predictive of local disease-free and overall survival. *Am J Surg Pathol*. 2005;29(2):167–178. https://doi.org/10.1097/01.pas.0000149687.90710.21.
9. Riders A, Oberste M, Abbaspour B, Beule A, Rudack C. [Prognostic factors for overall survival in oropharyngeal carcinoma depending on HPV status]. *HNO*. June 2021:1–8. https://doi.org/10.1007/s00106-021-01076-3. Published online.
10. Farhadieh RD, Smee R, Ow K, et al. Down-regulation of KAI1/CD82 protein expression in oral cancer correlates with reduced disease free survival and overall patient survival. *Cancer Lett*. 2004;213(1):91–98. https://doi.org/10.1016/j.canlet.2004.03.004.
11. Ravindran S, Ranganathan S, R K, et al. The role of molecular biomarkers in the diagnosis, prognosis, and treatment stratification of oral squamous cell carcinoma: a comprehensive review. *J Liq Biopsy*. 2025;7, 100285. https://doi.org/10.1016/j.jlb.2025.100285.
12. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144(5):646–674. https://doi.org/10.1016/j.cell.2011.02.013.
13. Chen S, Ke Y, Xie Y, et al. Identification of the key immune gene NR3C1 as a diagnostic biomarker in differentiating ovarian borderline tumors from benign tumors. *Front Cell Dev Biol*. 2025;13. https://doi.org/10.3389/fcell.2025.1602192.
14. Ramamoorthy S, Cidlowski JA. Corticosteroids. *Rheum Dis Clin N Am*. 2016;42(1):15–31. https://doi.org/10.1016/j.rdc.2015.08.002.

15. Zhai P, Zhang H, Li Q, et al. SETBP1 activation upon MDM4-enhanced ubiquitination of NR3C1 triggers dissemination of colorectal cancer cells. *Clin Exp Metastasis*. 2024;41(5):747–764. https://doi.org/10.1007/s10585-024-10294-2.
16. Iftikhar A, Shepherd S, Jones S, Ellis I. Effect of mifepristone on migration and proliferation of oral cancer cells. *Int J Mol Sci*. 2024;25(16):8777. https://doi.org/10.3390/ijms25168777.
17. Yan M, Wang J, Wang H, et al. Knockdown of NR3C1 inhibits the proliferation and migration of clear cell renal cell carcinoma through activating endoplasmic reticulum stress–mitophagy. *J Transl Med*. 2023;21(1):701. https://doi.org/10.1186/s12967-023-04560-2.
18. Lin C, Li S, Yi L, et al. TGM2 regulated by transcription factor NR3C1 drives p38 MAPK-mediated tumor progression and immune evasion in lung squamous cell carcinoma. *Front Immunol*. 2025;16. https://doi.org/10.3389/fimmu.2025.1595907.
19. Huang B, Kong L, Wang C, et al. Protein structure prediction: challenges, advances, and the shift of research paradigms. *Genom Proteom Bioinform*. 2023;21(5):913–925. https://doi.org/10.1016/j.gpb.2022.11.014.
20. Asim MN, Asif T, Hassan F, Dengel A. Protein sequence analysis landscape: a systematic review of task types, databases, datasets, word embeddings methods, and language models. *Database*. 2025;2025. https://doi.org/10.1093/database/baaf027.
21. Ghosh N, Santoni D, Saha I, Felici G. A review on the applications of Transformer-based language models for nucleotide sequence analysis. *Comput Struct Biotechnol J*. 2025;27:1244–1254. https://doi.org/10.1016/j.csbj.2025.03.024.
22. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*. 2022;38(8):2102–2110. https://doi.org/10.1093/bioinformatics/btac020.
23. Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat Commun*. 2022;13(1). https://doi.org/10.1038/s41467-022-32007-7.
24. Schmirler R, Heinzinger M, Rost B. Fine-tuning protein language models boosts predictions across diverse tasks. *Nat Commun*. 2024;15(1):7407. https://doi.org/10.1038/s41467-024-51844-2.
25. Ko YS, Parkinson J, Wang W. Benchmarking text-integrated Protein Language Model Embeddings and Embedding Fusion on Diverse Downstream Tasks. Preprint posted online August 26, 2024. doi:10.1101/2024.08.24.609531.
26. Cao J, Xiong L. Protein sequence classification with improved extreme learning machine algorithms. *BioMed Res Int*. 2014;2014:1–12. https://doi.org/10.1155/2014/103054.
27. Liu Z, Qian S, Xia C, Wang C. Are transformer-based models more robust than CNN-based models? *Neural Netw*. 2024;172, 106091. https://doi.org/10.1016/j.neunet.2023.12.045.
28. Cui F, Zhang Z, Zou Q. Sequence representation approaches for sequence-based protein prediction tasks that use deep learning. *Brief Funct Genom*. 2021;20(1):61–73. https://doi.org/10.1093/bfgp/elaa030.
29. Laine E, Eismann S, Elofsson A, Grudinin S. Protein sequence-to-structure learning: is this the end(-to-end revolution)? *Proteins: Struct, Funct, Bioinf*. 2021;89(12):1770–1786. https://doi.org/10.1002/prot.26235.
30. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583–589. https://doi.org/10.1038/s41586-021-03819-2.
31. Mckenna A, Dubey S. Machine learning based predictive model for the analysis of sequence activity relationships using protein spectra and protein descriptors. *J Biomed Inf*. 2022;128, 104016. https://doi.org/10.1016/j.jbi.2022.104016.
32. Yang KK, Fusi N, Lu AX. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Syst*. 2024;15(3):286–294.e2. https://doi.org/10.1016/j.cels.2024.01.008.
33. Liu B, Deng WZ, Hu WH, et al. Psychological stress-activated NR3C1/NUPR1 axis promotes ovarian tumor metastasis. *Acta Pharm Sin B*. 2025;15(6):3149–3162. https://doi.org/10.1016/j.apsb.2025.04.001.
34. Galliher-Beckley AJ, Williams JG, Cidlowski JA. Ligand-independent phosphorylation of the glucocorticoid receptor integrates cellular stress pathways with nuclear receptor signaling. *Mol Cell Biol*. 2011;31(23):4663–4675. https://doi.org/10.1128/MCB.05866-11.
35. Dwyer AR, Perez Kerkvliet C, Truong TH, et al. Glucocorticoid receptors drive breast cancer cell migration and metabolic reprogramming via PDK4. *Endocrinology*. 2023;164(7). https://doi.org/10.1210/endocr/bqad083.
36. Turlej E, Domaradzka A, Radzka J, Drulis-Fajdasz D, Kulbacka J, Gizak A. Cross-talk between cancer and its cellular environment—A role in cancer progression. *Cells*. 2025;14(6):403. https://doi.org/10.3390/cells14060403.
37. Hoxhaj G, Manning BD. The PI3K–AKT network at the interface of oncogenic signalling and cancer metabolism. *Nat Rev Cancer*. 2020;20(2):74–88. https://doi.org/10.1038/s41568-019-0216-7.
38. Celentano A, McCullough M, Cirillo N. Glucocorticoids reduce chemotherapeutic effectiveness on OSCC cells via glucose-dependent mechanisms. *J Cell Physiol*. 2019;234(3):2013–2020. https://doi.org/10.1002/jcp.27227.
39. Liu H, Li Z, Qiu F, et al. Association between NR3C1 mutations and glucocorticoid resistance in children with acute lymphoblastic leukemia. *Front Pharmacol*. 2021;12. https://doi.org/10.3389/fphar.2021.634956.
40. Kong X, Gao P, Wang J, Fang Y, Hwang KC. Advances of medical nanorobots for future cancer treatments. *J Hematol Oncol*. 2023;16(1):74. https://doi.org/10.1186/s13045-023-01463-z.
41. Singh NK, Chung S, Chang AY, Wang J, Hall DA. A non-invasive wearable stress patch for real-time cortisol monitoring using a pseudoknot-assisted aptamer. *Biosens Bioelectron*. 2023;227, 115097. https://doi.org/10.1016/j.bios.2023.115097.
42. Liu J, Gong X. Attention mechanism enhanced LSTM with residual architecture and its application for protein-protein interaction residue pairs prediction. *BMC Bioinf*. 2019;20(1). https://doi.org/10.1186/s12859-019-3199-1.
43. Lupo U, Sgarbossa D, Bitbol AF. Protein language models trained on multiple sequence alignments learn phylogenetic relationships. *Nat Commun*. 2022;13(1). https://doi.org/10.1038/s41467-022-34032-y.