IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 April 2025, accepted 4 May 2025, date of publication 12 May 2025, date of current version 27 May 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3569255

RESEARCH ARTICLE

Generating Anticancer Peptides Sequences Using Seq2Seq Modeling and Machine Learning Methods

MUHAMMAD SOHAIL IBRAHIM[1,2], SAHEED ADEMOLA BELLO[3], YUNSANG KWAK[2], MINSEOK KIM[2,4], AND SHUJAAT KHAN[3,5]

[1]On-Sensor AI Semiconductor Center, Kumoh National Institute of Technology, Gumi-si 39177, Republic of Korea
[2]School of Mechanical System Engineering, Kumoh National Institute of Technology, Gumi-si 39177, Republic of Korea
[3]Department of Computer Engineering, College of Computing and Mathematics, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia
[4]Department of Aeronautics, Mechanical and Electronic Convergence Engineering, Kumoh National Institute of Technology, Gumi-si 39177, Republic of Korea
[5]SDAIA-KFUPM Joint Research Center for Artificial Intelligence, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia

Corresponding authors: Minseok Kim (mkim@kumoh.ac.kr) and Shujaat Khan (shujaat.khan@kfupm.edu.sa)

This work was supported by the Ministry of Science and ICT (MSIT), South Korea, under the Information Technology Research Center (ITRC) Support Program Supervised by the Institute for Information & Communications Technology Planning & Evaluation (IITP) under Grant IITP-2025-RS-2024-00438288.

ABSTRACT Cancer remains a major health threat with rising incidence and mortality rates. Despite the efficacy of chemotherapy, its lack of selectivity and associated severe side effects highlight the need for new, targeted anticancer therapies. Anticancer peptides (ACPs) have emerged as a promising alternative due to their biocompatibility, broad-spectrum anticancer activity, and unique mechanisms of action. This study presents a novel computational approach to design and identify ACPs using a multi-tier filtration system. Our method begins with peptide sequence generation via a recurrent neural network (RNN) trained on the acp740 dataset. The generated sequences undergo rigorous filtration: Tier-1 employs three deep learning-based classifiers (ACP-DL, ACP-MHCNN, ACP-LSE) to identify potential ACPs; Tier-2 uses a nearest centroid classifier to filter out statistically less relevant sequences; Tier-3 involves a final filtration using unsupervised nearest neighbor learning based on fused feature encoding schemes (CKSAAP, k-Mer, and BPF). Experimental results demonstrate a significant improvement in identifying viable ACP candidates, with the proposed method showing a 2.21-fold higher hit-rate compared to random sequence generation. Further analysis using t-SNE, PCA, and antimicrobial peptide (AMP) prediction tools confirms the robustness and effectiveness of the selected ACPs. Furthermore, performance comparisons using the proposed sequence filtering technique reveal that it surpasses the baseline LSTM and RNN-based sequence generation models by 2.95% and 14.11%, respectively. Complementary reverse analyses further validate the robustness and effectiveness of proposed sequence generation framework. The proposed computational approach offers a streamlined and economical alternative to traditional experimental methods, expediting the discovery of new ACPs and enhancing the accuracy of anticancer peptide predictions. The relevant models, codes, and results are also available on the authors github page at (https://github.com/mhdshl/ACP-Seq2Seq).

INDEX TERMS Anticancer peptides (ACPs), multi-tier filtration, unsupervised nearest neighbor learning, fused feature encoding, peptide prediction.

I. INTRODUCTION

Cancer has become a significant threat to human health, with rising incidence and mortality rates [1]. Among current cancer treatment strategies, chemotherapy remains a top priority due to its non-invasiveness and anti-metastatic properties [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Rajeswari Sundararajan.

However, conventional chemotherapeutic drugs often lack selectivity for cancer cells, leading to severe adverse effects and potential therapy discontinuation [3]. Additionally, the development of drug resistance during chemotherapy limits the effectiveness of existing antineoplastic agents [4]. Consequently, there is a continuous need to develop new, selective, and more effective anticancer drugs for tumor therapy.

In recent years, the discovery of numerous peptides with medicinal properties has significantly advanced the field of human treatment, opening up new horizons for therapeutic development [5], [6], [7], [8]. Among such peptides, Anticancer peptides (ACPs) have emerged as a promising alternative to conventional anticancer drugs due to their excellent biocompatibility, broad-spectrum anticancer activity, and unique mechanisms of action [9]. ACPs are typically small peptides, ranging from 5 to 50 amino acids, and are primarily derived from antimicrobial peptides (AMPs) [10]. They generally possess amphipathic structures with more than two net positive charges and a high content of hydrophobic residues—key structural properties that confer their biological activity. ACPs can rapidly disrupt the cell membrane, inducing tumor cell death through electrostatic adsorption and hydrophobic interactions with the negatively charged cancer cell membrane [11]. This selective membrane damage mechanism is advantageous because it is less influenced by tumor heterogeneity and is less likely to lead to drug resistance, offering a significant edge over other chemotherapeutic drugs [12]. Additionally, ACPs can inhibit the growth of various cancer cells through apoptosis or other mechanisms of action [13].

The design and identification of ACPs through experimental methods are often time-consuming and costly, making computational approaches increasingly vital in this field. Machine learning and deep learning techniques have demonstrated significant potential in efficiently predicting and characterizing ACPs, offering a more streamlined and economical alternative to traditional laboratory methods [14]. These computational models can analyze peptide sequences, predict their anticancer properties, and identify patterns within the amino acid sequences that correlate with anticancer activity. This efficiency not only accelerates the discovery process but also enhances the accuracy and specificity of ACP predictions, thereby facilitating the development of new therapeutic agents [15].

In recent years, a surge in scientific exploration has seen an increasing reliance on machine learning models for peptide/protein classification and prediction. For instance, Yi et al. [16] harnessed *Long Short-Term Memory* (LSTM) alongside k-mer sparse matrix and binary profile features to propose an ACP classification method. Another study presented in [17] have leveraged the *composition of K-spaced amino acid pairs* (CKSAAP) for feature extraction, employing a kernel sparse representation classification approach for ACP detection. Another method, termed ACP-MHCNN used a novel multi-headed convolution neural network for the combination and extraction of various discriminating features for accurate detection of ACPs [18]. ACP-LSE [19] employed deep representation learning and latent space encoding for ACP classification. Researchers also employed various feature encoding and feature selection schemes aided by classic machine learning algorithms such as SVM [20], fusion of various classic machine learning algorithms by means of majority voting and genetic algorithms [21]. Other popular deep learning-based techniques include cACP-DeepGram [22], pACP-HybDeep [23], mACPpred2.0 [24], ACP-LSTM-NFR v1 [25] and ACP-LSTM-NFR v2 [26], etc.

Drug discovery represents a complex challenge, consuming substantial time and financial resources [27]. The process can be delineated into four key stages: (1) target selection and validation, (2) compound screening and optimization, (3) preclinical studies, and (4) clinical trials. Following exhaustive *in-vitro* and *in-vivo* assessments, the drug candidate undergoes FDA scrutiny before commercialization [28]. This conventional workflow typically spans over 12 years, with costs estimated at approximately $2.6 billion [29]. Therefore, there is a shared interest in mitigating costs and expediting candidate development.

In tandem with technological progress and the proliferation of digital pharmaceutical data, AI has emerged as a potent tool for managing vast datasets and finding diverse applications in the pharmaceutical domain [30]. In chemical-based drug development, AI facilitates primary and secondary drug screening and predicts drug–target interactions [31]. Moreover, computational approaches enable the prediction of pharmacological properties, potential efficacy, and *in-silico* absorption, distribution, metabolism, excretion, and toxicity (ADMET) profiles of drug candidates [32]. These active engagements with AI hold promise for accelerating and cost-reducing the process of drug discovery. In this context, Grisoni et al. [33] designed an LSTM-based constructive machine learning model trained on α-helical cationic amphipathic peptide sequences, and fined-tuned with known ACP sequences to generate novel ACP sequences that were tested and verified againts MCF7 cancer cells. Similarly in [34], the authors used a generative RNN model to generate novel ACP sequences followed by another RNN classification model for activity and homolysis. Their experimental evaluation resulted in eleven active ACP sequences. Yue et al. [35] presented, CNBT-ACPred, a three-channel deep learning-based ACP prediction technique augmented by in-vivo and in-vitro testing of anticancer activity in candidate ACPs.

The current methods of ACP design and discovery rely heavily on the conventional methods to test the novelty, activity, helicity, and amphiphilicity of the designed sequences, which can act as the bottleneck for the design process. Therefore, in this study, we present lightweight computational methods to filter/select statistically viable ACP candidates for experimentation and evaluation to expedite the peptide discovery pipeline. The key contributions of this study are as follows:

- Development of a GRU-based ACP sequence generation model that eliminates the need for complex feature encoding during the peptide synthesis process.
- Introduction of a novel multi-tier filtration framework for identifying statistically viable ACP candidates.
- Comprehensive analysis of candidate sequences at each stage of the filtration process.
- Evaluation of antimicrobial activity using a state-of-the-art AMP prediction tool to validate functional relevance.
- Structural modeling and 3D visualization of top candidate sequences using AlphaFold3 and ChimeraX.
- Performance benchmarking against baseline LSTM and RNN-based sequence generation models.
- Reverse validation using classical machine learning techniques to assess the robustness and generalizability of the proposed generation and filtration pipeline.

The rest of the paper is organized as; Section II presents the proposed method and its workflow, followed by the results in Section III. Finally the paper concludes in Section IV.

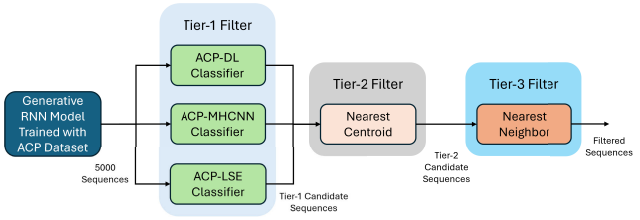


FIGURE 1. Workflow of the proposed method.

II. MATERIALS AND METHODS

The workflow of the proposed method, as illustrated in Fig. 1, involves peptide sequence generation using a RNN-based generative model trained on the acp740 dataset [36], [37]. The generated sequences are then forwarded to a rigorous filtration and statistical evaluation process where the filtration and evaluation is carried out in three tiers; 1) filter using known ACP classifiers, 2) filtering out statistically less significant sequences by a clustering-based centroid filter, and 3) finer filtration on the basis of the statistical similarity between the known and generated sequences using unsupervised nearest neighbor learning.

A. DATASET

The dataset used in this study is titled *acp740*, as presented in [36] and [37]. It comprises 740 peptide sequences, with 376 classified as ACP and 364 as non-ACP. A refined version of this dataset is available in [16]. The *acp740* dataset is employed for training both the sequence generation model and the classifiers and machine learning models used in the various filtration steps. Fig. 3 in Sec. III illustrates the distribution of sequence lengths in the *acp740* database. It was found that the ACP sequence length within the dataset ranges from 10 amino acids to 97 amino acids, with an average sequence length of 30.1. The minimum and maximum sequence lengths of a known benchmark dataset are used when designing a model for ACP sequence generation.

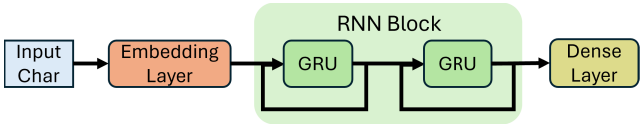


FIGURE 2. The proposed ACP sequence generation model.

B. ACP SEQUENCE GENERATION RNN MODEL

The sequence generation model is presented in Fig. 2. The sequence generation model has a simple architecture where the input character batch is forwarded to an embedding layer with its embedding dimensions as 256. Each character in the peptide sequence is transformed into a dense vector representation by the embedding layer, enabling the model to understand the relationships between different characters. The embedding layer is followed by a block of 2 RNN layers composed of gated recurrent unit (GRU) layers. The GRU layers are capable of capturing long-term dependencies within the sequences thanks to their recurrent connections and gating mechanisms. Each of the GRU layers have 1024 RNN units. The RNN block is followed by a dense layer with the number of units as the number of unique characters in the training dataset. During the inference phase, the output of the model is fed to the input to generate the sequences autonomously, with the seed input (\n) as the sequence termination code.

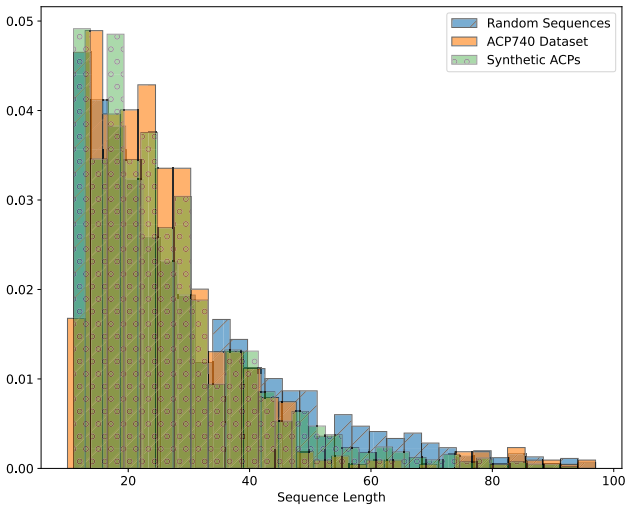


FIGURE 3. Distribution of sequences from *acp740* dataset, randomly generated sequences, and sequences generated using the trained sequence generation model.

C. FILTRATION MECHANISM

The filtration mechanism is executed in three stages: tier-1 filter, tier-2 filter, and tier-3 filter.

1) TIER-1 FILTER

In tier-1, three contemporary deep learning-based classifiers are employed: ACP-DL [16], ACP-MHCNN [18], and ACP-LSE [19]. The generated sequences are evaluated by these three classifiers, and only those sequences that are positively classified by all three are selected to proceed to the next filtration step.

ACP-DL [16] is a deep learning model using a long short-term memory (LSTM) neural network to predict novel anticancer peptides. It integrates binary profile features and k-mer sparse matrices of the reduced amino acid alphabet for efficient feature representation, enabling automatic identification of anticancer and non-anticancer peptides. ACP-MHCNN [18] is a multi-headed deep convolutional neural network model designed to extract and combine discriminative features from various information sources interactively. It identifies ACPs by extracting sequence, physicochemical, and evolutionary features using different numerical peptide representations while minimizing parameter overhead. Finally, ACP-LSE [19] presents an intuitive classification strategy based on representation learning, specifically employing a deep latent-space encoding scheme. It excels in scenarios with limited sample sizes and abundant features by embedding high-dimensional features, such as g-spaced amino acid pair compositions, into a compressed latent space using an auto-encoder-inspired network. Unlike conventional auto-encoders, ACP-LSE ensures that the learned feature set is both compact and effective for classification, providing a transparent alternative to typical closed-box approaches.

The three deep learning-based classifier networks explained earlier are used in the first step of filtration where the generated sequences are classified using the three classifiers. The result of the classification from each network are compared and the sequences that qualify as ACPs by all three classifiers are then selected for further filtration and analysis in the Tier-2 filtration.

2) TIER-2 FILTER

In the tier-2 of the filtration process, we employ the nearest centroid classifier to filter-out the statistically less relevant sequences that might have been misclassified as ACPs in the tier-1 filter. The nearest centroid [38] method is a simple and efficient classification algorithm that assigns a data point to the class whose centroid (mean position of all points in the class) is closest to it. During the training phase, the centroids for each class are calculated by averaging the feature vectors of all points in that class. In the classification phase, the distance from the new data point to each centroid is computed, and the point is assigned to the class with the smallest distance. This method is particularly effective for well-separated classes and is computationally inexpensive, making it suitable for large datasets with relatively simple structures.

For the tier-2 filtration analysis, t-SNE features of the combined BPF and k-Mer features for the *acp740* database are used to train the nearest centroid model. After training, the t-SNE feature embeddings of the combined BPF and k-Mer features for the tier-1 filtered ACP candidates are analyzed by the nearest centroid classifier and the resulting sequences are then forwarded to the final tier-3 filtration step.

3) TIER-3 FILTER

The tier-3 filter involves a statistical analysis and evaluation of the tier-2 filtered candidates using an unsupervised nearest neighbor model [39]. The unsupervised nearest neighbor method is a type of clustering algorithm that groups data points based on their proximity to each other without using predefined labels. This method involves calculating the distance between all pairs of points in the dataset and then forming clusters by linking each point to its nearest neighbors. These clusters are formed iteratively, with each point being added to the cluster containing its nearest neighbor or forming a new cluster if it does not closely match existing clusters. This approach is useful for discovering natural groupings in data and is commonly used in applications like anomaly detection, data compression, and pattern recognition.

For the analysis at this stage, a fusion of three feature encoding schemes i.e., CKSAAP, k-Mer, and BPF [16], [17] are used. For the nearest neighbor analysis, only the positive sequences from the *acp740* dataset have been selected and forwarded to the feature encoding process where the three features are extracted and fused together. The candidate sequences from the tier-2 filtered follow the same feature encoding process. The nearest neighbor model is trained using the fused features from the positive sequences of the benchmark dataset and the trained model is used to evaluate the closest distance from the training set. This results in a distance vector and a vector correspoding to the closest known ACP sequences from the benchmark dataset. The distances are analyzed and the sequences with the minimum distances are selected for further visualization and analysis using AlphaFold3 [40] and ChimeraX [41].

III. EXPERIMENTAL RESULTS

This section presents the experimental evaluation and the analysis performed in this study. As explained in the previous section, the sequence generation model is trained with *acp740* dataset. As mentioned in Sec. II-A, the ACP sequences in the training dataset have the sequence lengths with the ranges (10,97), the minimum and maximum threshold for sequence generation is kept within this range. The sequence generation model is trained for 100 epochs, keeping a batch size of 64, and an early stopping criterion with patience value of 25 epochs is used to avoid model overfitting. The model is optimized using the Adam optimizer for a sparse categorical cross-entropy loss, keeping the optimizer learning rate at 10^{-3}. To evaluate the performance of the sequence generation model, random sequences were also generated using the untrained model to compare the distributions of the sequences generated from the trained model and that

TABLE 1. Tier-1 filtration analysis of random sequences compared to sequences generated using the trained model.

Sequences	Hit-Rate (%)			Combined Hit-Rate (%)
	ACP-DL	ACP-MHCNN	ACP-LSE	
Random Sequences	36.31	43.07	38.32	11.31
Trained Sequences	43.07	41.22	41.34	36.25

of the randomly generated sequences with the presence of the known ACP dataset. Fig. 3 presents a comparison of the distribution of the length of the known ACPs in *acp740* dataset with randomly generated sequences and sequences generated using the trained RNN model. It can be observed from the figure that the distribution of synthetic ACPs (sequences generated using the trained model) closely follow the distribution of the *acp740* dataset. The relevant models, codes, and results are also available on the authors github page at (https://github.com/mhdshl/ACP-Seq2Seq).

A. TIER-1 FILTRATION ANALYSIS

For the experiment and analysis, we generated 5000 random sequences and 5000 sequences using the trained sequence generation model. These random and predicted sequences were evaluated using ACP-DL, ACP-MHCNN, and ACP-LSE benchmark classifiers as stated in previous section. The experiment in tier-1 filter is evaluated on the basis of the individual hit-rate of the classifiers as well as the combined hit-rate of the benchmark classifiers. Table 1 presents the comparison of the hit-rate of the individual classifiers and the combined hit-rate, where a candidate sequence is classified as a positive ACP sequence from all three classifiers. It can be observed from the combined hit-rate that the sequences generated using the proposed model have 2.21× higher hit-rate compared to the randomly generated sequences.

Furthermore, the candidate sequences obtained after tier-1 filtration are visualized using t-stochastic neighbor embedding (t-SNE) of their feature space. The t-SNE visualization of the tier-1 candidate sequences compared to the *acp740* dataset is presented in Fig. 4. We used a combination of BPF and k-mer feature encoding schemes as utilized in [16] and evaluated the 2-component t-SNE embeddings for both sets of sequences. In the figure, the green dots represent the tier-1 candidate sequences, demonstrating that they overlap nicely with the positive ACP sequences (blue dots) from *acp740* while maintaining good separation from the non-ACP sequences. After the tier-1 filtration, 1909 candidate sequences are forwarded for further filtration and analysis to the tier-2 filter. The corresponding tier-1 candidate sequences can be found on the authors github page in the directory titled *synthetic_ACPs* under the file name *ACP_seq2seq_20240605_first_tier_filter.xlsx*.

B. TIER-2 FILTRATION ANALYSIS

For tier-2 filtration, the t-SNE of the combined k-Mer and BPF features of the *acp740* dataset is used to train a nearest

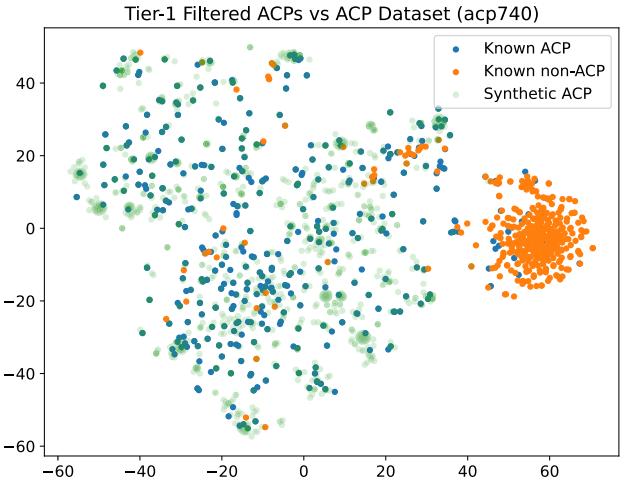


FIGURE 4. t-SNE of the Tier-1 filter candidate sequences vs. *acp740* dataset.

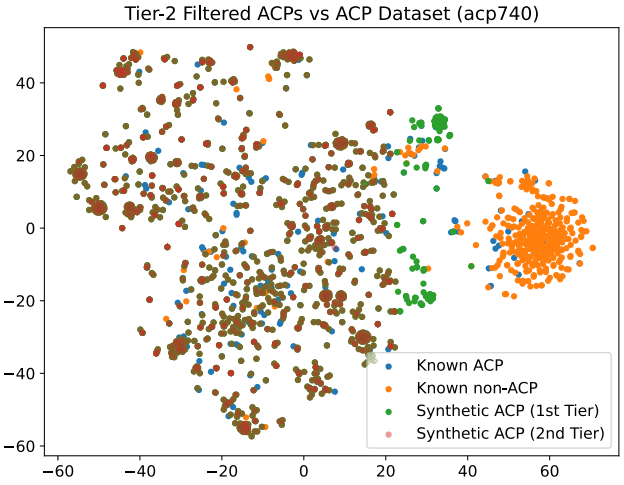


FIGURE 5. t-SNE of the tier-2 candidate sequences compared to tier-1 candidate sequences in the presence of *acp740* dataset.

centroid classifier. The trained nearest centroid classifier is then used to evaluate the t-SNE features of the tier-1 candidate sequences. The nearest centroid classifier aims to restrict the spread of the tier-1 candidate sequences and removes any outlier sequences. A visualization of the effect of the tier-2 filtration is also visualized using the t-SNE. A visualization of the effect of the tier-2 filtration is presented in Fig. 5. It can be observed that the tier-2 candidates are more restricted within the vicinity of the majority of the known *acp740* positive sequences. The filtered tier-2 candidate sequences can be found on the authors github page in the directory titled *synthetic_ACPs* under the file name *ACP_seq2seq_20240605_second_tier_filter.xlsx*.

C. TIER-3 FILTRATION ANALYSIS

In tier-3 filtration, an unsupervised nearest neighbor model is trained on only the positive ACP sequences from *acp740*

database. For training the unsupervised nearest neighbor model, we used a fusion of BPF, k-Mer, and CKSAAP features to enable the search in a high-dimensional space. After training the nearest neighbor model, we performed the inference on the tier-2 candidate sequences features and analyzed the distances of each tier-2 candidate with its nearest matching positive ACP sequence from the *acp740* database. For further analysis, the tier-2 candidate sequences with the minimum distance values were selected as tier-3 candidate sequences. As an initial analysis for the tier-3 candidate sequences, the 2-component PCA embeddings of the tier-3 candidate sequences were visualized in the presence of their neighboring sequences, from *acp740* database, obtained using the minimum distance. Fig. 6 presents the PCA embeddings of the tier-3 candidate sequences compared to their neighboring known ACP sequences from the dataset. It can be observed in the figure that the PCA embeddings of the tier-3 candidate sequences closely match with their corresponding known ACP sequences. After tier-3 filtration, a total of 187 candidate sequences were obtained for further analysis. The tier-3 candidate sequences, their closely matching positive ACP sequences from the benchmark dataset, the corresponding distance values, and the results of further analysis are maintained in spreadsheets on the authors github page mentioned earlier in this section. The tier-3 candidate sequences can be found in the directory titled *synthetic_ACPs* under the file name *ACP_seq2seq_20240605_third_tier_filter_second_method.xlsx*.

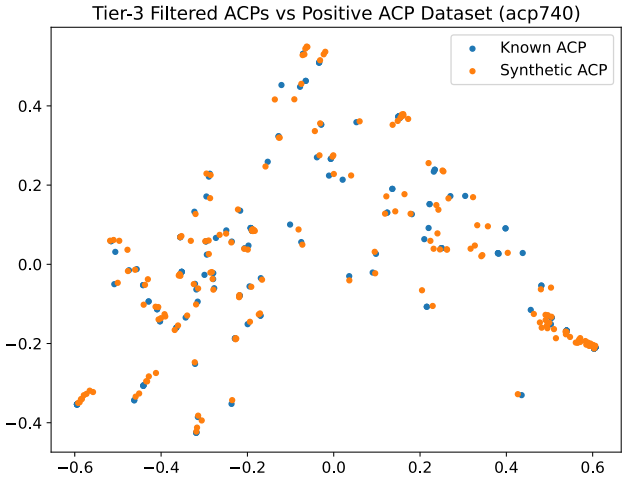


FIGURE 6. PCA emdeddings of the Tier-3 filter candidate sequences vs. positive sequences from *acp740* dataset.

1) EVALUATING ANTIMICROBIAL PROPERTIES

To support the visual and statistical analysis, we analyzed the tier-3 candidate sequences for antimicrobial peptide (AMP) properties using the widely known CAMPSign [42] tool for identification of AMP family signatures. This analysis is conducted due to the reason that ACPs can be considered a sub-family of the AMPs due to their common properties such as selective toxicity, broad spectrum activity, cationicity (positive net charge), high hydrophobicity, and amphipathic structure, giving them an increased affinity for cell membranes [10]. In the AMP analysis using the CAMPSign tool, we used the three well-known classifiers namely support vector machine (SVM), random forest (RF), and artificial neural network (ANN) available in the tool. A comparison of the AMP prediction analysis of the tier-3 candidate sequences with their neighboring ACP sequences from *acp740* dataset is presented in Table 2. The AMP analysis presented in the Table also confirm the findings of the analysis conducted in tier-3 filtration stage that the tier-3 candidate sequence contain similar characteristics as their closely matching known ACP sequences. The results of AMP analysis, tier-3 candidate sequences, their corresponding ACP sequences from dataset, and other analysis can be found in the spreadsheet titled *ACP_seq2seq_Filtration_Analysis.xlsx* in the authors github page.

TABLE 2. AMP analysis of the tier-3 candidate sequences compared to their closest matching positive ACP sequences.

Dataset	Hit-Rate (%)			Mean Hit-Rate (%)
	SVM	RF	ANN	
ACPs (acp740)	90.91	95.18	90.91	92.33
Tier-3 Candidate Sequences	86.09	86.09	86.09	86.09

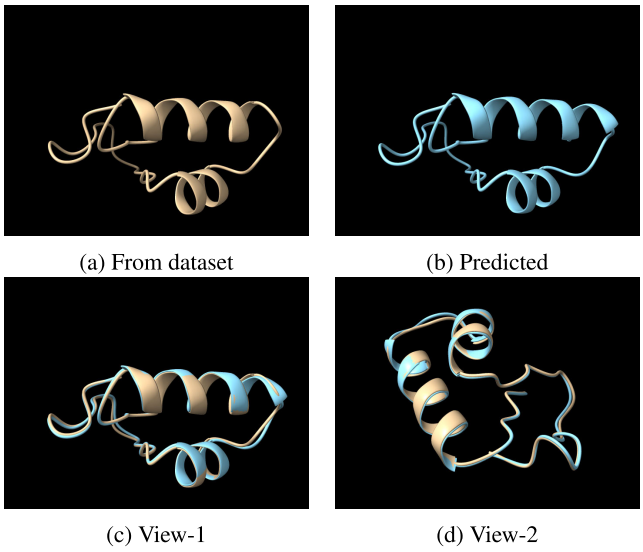


FIGURE 7. 3D structure comparison of a tier-3 candidate sequence and its closely matching known ACP sequence from the *acp740* dataset using AlphaFold3 and ChimeraX.(a) Known ACP sequence (ACP-67) from *acp740* dataset,(b) Tier-3 candidate sequence (ACP-39) corresponding to the sequence structure in (a),(c) Superimposed view of the known and its corresponding sequence (view-1), and(d) Another superimposed view of the known and its corresponding sequence (view-2).

2) EVALUATING STRUCTURAL SIMILARITIES

For structure analysis, we selected a sequence from the tier-3 candidate sequences and its closely matching known ACP sequence. We utilized Alphafold3 [40] for

the prediction of the 3D structure of the sequence pair and used ChimeraX [41] for the visualization of these sequences individually and as a superimposed pair to visualize the structure similarity of both the known and synthesized sequences. The selected sequence pair for visualization can be observed in the Fig. 7. For the 3D visualization of the sequences, we selected ACP-39 (KSCCPNTTGRNIYNACRLTGAPRPTCAKLSGCKIISGSTCPSDYPK) from the tier-3 candidate sequences and its closely matching sequence ACP-67 (KSCCPNTTGRNIYNTCRFGGGSREVCARISGCKIISASTCPSDYPK) from the positive ACP sequences in the *acp740* dataset. The visual comparison of the structures shows a high degree of structural similarity, further supporting the validity of the generated candidate sequences as potential ACPs. The candidate sequences at each filtration stage, their parent sequences, the results of the analysis conducted during each tier, etc., have been maintained in various spreadsheets and made available on the authors github page.

D. ABLATION STUDIES

1) PERFORMANCE EVALUATION ON A LARGER DATASET

To further assess the generalizability and robustness of the proposed sequence generation model, we conducted additional training using the recently introduced mACPpred2.0 dataset [24]. This dataset contains an equal number of positive and negative ACP sequences, with 1175 samples in each class. The proposed sequence generation model was trained on mACPpred2.0 dataset using the same architecture and training hyperparameters as described previously. For the filtration process, the first-tier filter was adapted to utilize the mACPpred2.0 online ACP prediction tool, while the second and third-tier filters followed the same methodology as outlined in the earlier sections. Upon evaluating 4000 sequences generated by the trained model, the first-tier classifier achieved an accuracy of 96.37%. After completing all three filtration stages, 240 high-confidence ACP candidates were identified.

These final candidates were further subjected to antimicrobial peptide (AMP) analysis using the same set of classifiers—Support Vector Machine (SVM), Random Forest (RF), and Artificial Neural Network (ANN)—as used in the primary experiments. The AMP evaluation yielded a mean hit-rate of 92.74%, with individual hit-rates of 92.6% (SVM), 94.94% (RF), and 90.66% (ANN). These results further validate the effectiveness of the proposed sequence-to-sequence generation model and its associated three-tier filtration mechanism in identifying statistically viable ACP candidates from larger and more diverse datasets.

2) PERFORMANCE COMPARISON WITH BASELINE MODELS

To evaluate the effectiveness of the proposed sequence generation framework, we designed two baseline models based on standard RNN and LSTM architectures. These baseline models adopt the same architecture and training hyperparameters as the proposed GRU-based model, as described in Sec. II-B, with the only difference being the replacement of GRU layers with standard RNN and LSTM layers, respectively.

TABLE 3. Tier-1 filtration analysis and comparison of the sequence generation ability of the proposed model compared to the baseline RNN models.

Model	Hit-Rate (%)			Combined Hit-Rate (%)
	ACP-DL	ACP-MHCNN	ACP-LSE	
RNN	42.34	38.69	35.35	21.25 ± 3.59
LSTM	53.90	48.58	48.25	32.41 ± 0.50
Proposed	54.03	47.75	49.92	35.36 ± 1.91

To ensure fair evaluation, each baseline model was trained over five independent trials. In each trial, 4000 sequences were generated and subjected to the first-tier filtration using three benchmark ACP classifiers (ACP-DL, ACP-MHCNN, and ACP-LSE). The classifier hit-rates for each model were then compared with those of the proposed method. Table 3 presents the mean hit-rate of individual classifiers, along with the combined hit-rate (mean $\pm$ standard deviation) where all three classifiers simultaneously identify a sequence as an ACP. Interestingly, while the LSTM-based model achieved comparable or slightly better individual hit-rates, the proposed GRU-based model consistently outperformed both RNN and LSTM baselines in terms of the combined hit-rate. Specifically, it achieved a performance gain of 2.95% over the LSTM-based model and 14.11% over the RNN-based model. These results demonstrate the superior generalization ability of the GRU-based model for generating high-quality ACP candidates.

We also compared the computational complexity of the proposed model with the RNN and LSTM baselines, as summarized in Table 4. The proposed sequence-to-sequence model contains approximately 10.26 million parameters (39.15 MB), which is lower than the LSTM-based model (13.66 million parameters, 52.13 MB), but higher than the simpler RNN model (3.43 million parameters, 13.11 MB). Despite having more parameters than the RNN, the proposed GRU model offers a balanced trade-off between model size and performance, achieving better generalization and sequence generation quality while maintaining a relatively lightweight computational footprint compared to the LSTM counterpart.

TABLE 4. Comparison of model parameters and memory size for the proposed, LSTM, and RNN sequence generation models.

Model	RNN	LSTM	Proposed
Parameters (Million)	3.43	13.66	10.26
Model Size (MB)	13.11	52.13	39.15

3) REVERSE ANALYSIS

To further assess the robustness and practical utility of the proposed sequence generation framework, we conducted

TABLE 5. Tier-wise reverse analysis and comparison of the proposed sequence generation method and random sequence generation.

Filtration Tier	Accuracy		F1-Score		Sensitivity		Specificity		MCC	
	Random	Proposed	Random	Proposed	Random	Proposed	Random	Proposed	Random	Proposed
Tier-1	67.70	79.86	0.66	0.80	0.45	0.78	0.90	0.81	0.40	0.59
Tier-2	69.32	80.13	0.68	0.80	0.47	0.78	0.92	0.82	0.43	0.60
Tier-3	-	74.86	-	0.75	-	0.75	-	0.75	-	0.50

a reverse analysis using a classic machine learning classifier. Specifically, a linear Support Vector Machine (SVM) was trained on the synthetic ACP sequences produced and filtered through the three-tier pipeline and then tested on the benchmark *acp740* dataset. For comparison, a similar analysis was conducted using sequences generated by an untrained (random) model. To ensure fairness, the training data at each filtration tier was balanced by incorporating non-ACP sequences from the independent *mACPpred2.0* dataset [24]. The performance of the reverse classifier, evaluated in terms of Accuracy, F1-Score, Sensitivity, Specificity, and MCC, is summarized in Table 5. The results clearly show that sequences generated by the proposed method consistently outperform those from the random model across all metrics and filtration tiers. Notably, after the third filtration stage, all random sequences were excluded, further demonstrating the effectiveness of the proposed three-tier filtration strategy. These findings highlight the ability of the generated sequences to generalize and retain discriminative features, reinforcing the value of our sequence generation and filtration framework in ACP discovery.

IV. CONCLUSION

In this study, we presented a comprehensive computational approach to generate, filter, and evaluate potential ACPs using a combination of deep learning models and statistical analysis. The proposed method leverages a GRU-based generative model for sequence generation, followed by a rigorous three-tier filtration process involving state-of-the-art classifiers, nearest centroid classification, and unsupervised nearest neighbor analysis. The results demonstrate that the proposed approach is capable of generating candidate sequences that closely match known ACPs in terms of sequence length distribution, feature space, and structural properties. The final tier-3 candidate sequences exhibit strong antimicrobial properties, as verified through external tools, and show significant structural similarity to known ACPs, suggesting their potential efficacy in anticancer therapy. The statistical evaluation, structural analysis, baseline comparisons, and reverse analysis collectively validate the robustness and effectiveness of the proposed sequence generation and three-tier filtration method. The framework holds practical relevance for accelerating early-stage anticancer peptide discovery, offering a cost-efficient alternative to conventional screening processes.

However, this study also has certain limitations. Firstly, it is restricted to computational evaluations; the predicted ACP candidates have not undergone experimental validation through in-vitro or in-vivo assays, which is essential to confirm their biological activity. Secondly, the model training is based on publicly available datasets, which may not comprehensively capture the sequence diversity found in real-world biological systems. These limitations may impact the generalizability of the findings in clinical settings. Future work may involve experimental validation of the predicted peptides through in-vitro and in-vivo assays to confirm their biological activity. In addition, future directions will explore training on larger and more diverse ACP datasets, and adopting advanced generative models such as transformer-based architectures to further improve sequence diversity, accuracy, and clinical relevance. Furthermore, the integration of safety profiling and pharmacokinetic modeling will be essential for advancing promising candidates towards preclinical development. These enhancements are expected to facilitate smoother clinical translation and align the discovery pipeline with regulatory requirements for peptide-based therapeutics.

REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, A Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021.

[2] J. Chen, Y. Song, W. Yang, J. Guo, S. Zhang, D. Wan, Y. Liu, and J. Pan, "Enzyme and reduction dual-responsive peptide micelles as nanocarriers for smart drug delivery," *ACS Appl. Nano Mater.*, vol. 6, no. 18, pp. 16179–16188, Sep. 2023.

[3] J.-X. Su, S.-J. Li, X.-F. Zhou, Z.-J. Zhang, Y. Yan, S.-L. Liu, and Q. Qi, "Chemotherapy-induced metastasis: Molecular mechanisms and clinical therapies," *Acta Pharmacol. Sinica*, vol. 44, no. 9, pp. 1725–1736, Sep. 2023.

[4] H. F. Cabanos and A. N. Hata, "Emerging insights into targeted therapy-tolerant persister cells in cancer," *Cancers*, vol. 13, no. 11, p. 2666, May 2021.

[5] S. Park, A. Wahab, M. Usman, I. Naseem, and S. Khan, "Editorial: Artificial intelligence in bioimaging and signal processing," *Frontiers Physiol.*, vol. 14, Aug. 2023, Art. no. 1267632.

[6] M. Usman, S. Khan, S. Park, and J.-A. Lee, "AoP-LSE: Antioxidant proteins classification using deep latent space encoding of sequence features," *Current Issues Mol. Biol.*, vol. 43, no. 3, pp. 1489–1501, Oct. 2021.

[7] I. Naseem, S. Khan, R. Togneri, and M. Bennamoun, "ECMSRC: A sparse learning approach for the prediction of extracellular matrix proteins," *Current Bioinf.*, vol. 12, no. 4, pp. 361–368, Jul. 2017.

[8] U. M. Al-Saggaf, M. Usman, I. Naseem, M. Moinuddin, A. A. Jiman, M. U. Alsaggaf, H. K. Alshoubaki, and S. Khan, "ECM-LSE: Prediction of extracellular matrix proteins using deep latent space encoding of k-Spaced amino acid pairs," *Frontiers Bioeng. Biotechnol.*, vol. 9, Oct. 2021, Art. no. 752658.

[9] Z. Dong, X. Zhang, Q. Zhang, J. Tangthianchaichana, M. Guo, S. Du, and Y. Lu, "Anticancer mechanisms and potential anticancer applications of antimicrobial peptides and their nano agents," *Int. J. Nanomedicine*, vol. Volume 19, pp. 1017–1039, Feb. 2024.

[10] A. Jafari, A. Babajani, R. Sarrami Forooshani, M. Yazdani, and M. Rezaei-Tavirani, "Clinical applications and anticancer effects of antimicrobial peptides: From bench to bedside," *Frontiers Oncol.*, vol. 12, Feb. 2022, Art. no. 819563.

[11] Y. Liscano, J. Oñate-Garzón, and J. P. Delgado, "Peptides with dual antimicrobial–anticancer activity: Strategies to overcome peptide limitations and rational design of anticancer peptides," *Molecules*, vol. 25, no. 18, p. 4245, Sep. 2020.

[12] N. Shao, L. Yuan, P. Ma, M. Zhou, X. Xiao, Z. Cong, Y. Wu, G. Xiao, J. Fei, and R. Liu, "Heterochiral β-Peptide polymers combating multidrug-resistant cancers effectively without inducing drug resistance," *J. Amer. Chem. Soc.*, vol. 144, no. 16, pp. 7283–7294, Apr. 2022.

[13] A. Lath, A. R. Santal, N. Kaur, P. Kumari, and N. P. Singh, "Anti-cancer peptides: Their current trends in the development of peptide-based therapy and anti-tumor drugs," *Biotechnol. Genetic Eng. Rev.*, vol. 39, no. 1, pp. 45–84, Jan. 2023.

[14] X. Xu, C. Li, X. Yuan, Q. Zhang, Y. Liu, Y. Zhu, and T. Chen, "ACP-DRL: An anticancer peptides recognition method based on deep representation learning," *Frontiers Genet.*, vol. 15, Apr. 2024, Art. no. 1376486.

[15] M. Liu, T. Wu, X. Li, Y. Zhu, S. Chen, J. Huang, F. Zhou, and H. Liu, "ACPPfel: Explainable deep ensemble learning for anticancer peptides prediction based on feature optimization," *Frontiers Genet.*, vol. 15, Feb. 2024, Art. no. 1352504.

[16] H.-C. Yi, Z.-H. You, X. Zhou, L. Cheng, X. Li, T.-H. Jiang, and Z.-H. Chen, "ACP-DL: A deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation," *Mol. Therapy - Nucleic Acids*, vol. 17, pp. 1–9, Sep. 2019.

[17] E. Fazal, M. S. Ibrahim, S. Park, I. Naseem, and A. Wahab, "Anticancer peptides classification using kernel sparse representation classifier," *IEEE Access*, vol. 11, pp. 17626–17637, 2023.

[18] S. Ahmed, R. Muhammod, Z. H. Khan, S. Adilina, A. Sharma, S. Shatabda, and A. Dehzangi, "ACP-MHCNN: An accurate multi-headed deep-convolutional neural network to predict anticancer peptides," *Sci. Rep.*, vol. 11, no. 1, p. 23676, Dec. 2021.

[19] S. Khan, "Deep-representation-learning-based classification strategy for anticancer peptides," *Mathematics*, vol. 12, no. 9, p. 1330, Apr. 2024.

[20] S. Akbar, M. Hayat, M. Tahir, and K. T. Chong, "CACP-2LFS: Classification of anticancer peptides using sequential discriminative model of KSAAP and two-level feature selection approach," *IEEE Access*, vol. 8, pp. 131939–131948, 2020.

[21] S. Akbar, M. Hayat, M. Iqbal, and M. A. Jan, "IACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space," *Artif. Intell. Med.*, vol. 79, pp. 62–70, Jun. 2017.

[22] S. Akbar, M. Hayat, M. Tahir, S. Khan, and F. K. Alarfaj, "CACP-DeepGram: Classification of anticancer peptides via deep neural network and skip-gram-based word embedding model," *Artif. Intell. Med.*, vol. 131, Sep. 2022, Art. no. 102349.

[23] Shahid, M. Hayat, W. Alghamdi, S. Akbar, A. Raza, R. A. Kadir, and M. R. Sarker, "PACP-HybDeep: Predicting anticancer peptides using binary tree growth based transformer and structural feature encoding with deep-hybrid learning," *Sci. Rep.*, vol. 15, no. 1, p. 565, Jan. 2025.

[24] V. K. Sangaraju, N. T. Pham, L. Wei, X. Yu, and B. Manavalan, "MACPpred 2.0: Stacked deep learning for anticancer peptide prediction with integrated spatial and probabilistic feature representations," *J. Mol. Biol.*, vol. 436, no. 17, Sep. 2024, Art. no. 168687.

[25] N. Al Tahifah, M. Sohail Ibrahim, and S. Khan, "Anticancer peptides classification with high-accuracy feature representation using long short-term memory," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2024, pp. 5707–5711.

[26] N. Al Tahifah, M. Sohail Ibrahim, E. Rehman, N. Ahmed, A. Wahab, and S. Khan, "Anticancer peptides classification using long-short-term memory with novel feature representation," *IEEE Access*, vol. 13, pp. 67–79, 2025.

[27] W. Cui, A. Aouidate, S. Wang, Q. Yu, Y. Li, and S. Yuan, "Discovering anti-cancer drugs via computational methods," *Frontiers Pharmacol.*, vol. 11, p. 733, May 2020.

[28] H. C. S. Chan, H. Shan, T. Dahoun, H. Vogel, and S. Yuan, "Advancing drug discovery via artificial intelligence," *Trends Pharmacol. Sci.*, vol. 40, no. 10, p. 801, Oct. 2019.

[29] R. C. Mohs and N. H. Greig, "Drug discovery and development: Role of basic biological research," *Alzheimer's Dementia: Translational Res. Clin. Interventions*, vol. 3, no. 4, pp. 651–657, Nov. 2017.

[30] K.-K. Mak, Y.-H. Wong, and M. R. Pichika, "Artificial intelligence in drug discovery and development," in *Drug Discovery and Evaluation: Safety and Pharmacokinetic Assays*, Springer, 2023, pp. 1–38.

[31] A. C. A. Nascimento, R. B. C. Prudêncio, and I. G. Costa, "A multiple kernel learning algorithm for drug-target interaction prediction," *BMC Bioinf.*, vol. 17, no. 1, pp. 1–16, Jan. 2016.

[32] G. Klopman, S. K. Chakravarti, H. Zhu, J. M. Ivanov, and R. D. Saiakhov, "ESP: A method to predict toxicity and pharmacological properties of chemicals using multiple MCASE databases," *J. Chem. Inf. Comput. Sci.*, vol. 44, no. 2, pp. 704–715, Mar. 2004.

[33] F. Grisoni, C. S. Neuhaus, G. Gabernet, A. T. Müller, J. A. Hiss, and G. Schneider, "Designing anticancer peptides by constructive machine learning," *ChemMedChem*, vol. 13, no. 13, pp. 1300–1302, Jul. 2018.

[34] E. Zakharova, M. Orsi, A. Capecchi, and J.-L. Reymond, "Machine learning guided discovery of non-hemolytic membrane disruptive anticancer peptides," *ChemMedChem*, vol. 17, no. 17, Sep. 2022, Art. no. 202200291.

[35] J. Yue, T. Li, J. Xu, Z. Chen, Y. Li, S. Liang, Z. Liu, and Y. Wang, "Discovery of anticancer peptides from natural and generated sequences using deep learning," *Int. J. Biol. Macromolecules*, vol. 290, Feb. 2025, Art. no. 138880.

[36] W. Chen, H. Ding, P. Feng, H. Lin, and K.-C. Chou, "IACP: A sequence-based tool for identifying anticancer peptides," *Oncotarget*, vol. 7, no. 13, pp. 16895–16909, Mar. 2016.

[37] L. Wei, C. Zhou, H. Chen, J. Song, and R. Su, "ACPred-FL: A sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides," *Bioinformatics*, vol. 34, no. 23, pp. 4007–4016, Dec. 2018.

[38] I. Levner, "Feature selection and nearest centroid classification for protein mass spectrometry," *BMC Bioinf.*, vol. 6, no. 1, pp. 1–14, Mar. 2005.

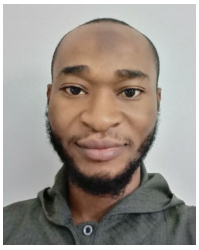
[39] S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood component analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, vol. 17, nos. 513–520, p. 4.

[40] J. Abramson et al., "Accurate structure prediction of biomolecular interactions with AlphaFold 3," *Nature*, vol. 630, no. 8016, pp. 493–500, May 2024.

[41] E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, and T. E. Ferrin, "UCSF ChimeraX: Structure visualization for researchers, educators, and developers," *Protein Sci.*, vol. 30, no. 1, pp. 70–82, Jan. 2021.

[42] F. H. Waghu, R. S. Barai, P. Gurung, and S. Idicula-Thomas, "CAMPR3: A database on sequences, structures and signatures of antimicrobial peptides: Table 1," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1094–D1097, Jan. 2016.

MUHAMMAD SOHAIL IBRAHIM received the B.E. degree in electronic engineering from Iqra University, Pakistan, in 2012, the M.E. degree in telecommunications from the NED University of Engineering and Technology, Pakistan, in 2016, and the Ph.D. degree from Chosun University, Republic of Korea, in 2024. From 2013 to 2019, he was a Lecturer with the Faculty of Engineering, Science, and Technology, Iqra University, Karachi, Pakistan. From 2013 to 2014, he was a Research Assistant with the Embedded Systems Research Group, Karachi Institute of Economics and Technology, Pakistan. He is currently a Postdoctoral Researcher with the Department of Mechanical Systems Engineering, Kumoh National Institute of Technology, Republic of Korea. His research interests include deep learning, computer vision, bioinformatics, and medical image processing. He was honored with the 2020 Highly Cited Review Paper Award from Applied Energy, Elsevier, for his paper titled "Machine Learning Driven Smart Electric Systems: Current Trends and New Perspectives."



SAHEED ADEMOLA BELLO received the master's degree in computer engineering from King Abdulaziz University, Jeddah, Saudi Arabia. He is currently pursuing the Ph.D. degree with the Department of Computer Engineering, King Fahd University of Petroleum & Minerals. His research interests include the development and application of deep learning models in solving critical human problems in the area of biomedical systems, computer vision, and cybersecurity. His work focuses on designing innovative solutions to advance medical diagnostics and secure digital environments. Driven by a passion for leveraging technology to improve lives, he aims to contribute significantly to the fields of artificial intelligence and its applications in addressing global challenges.

YUNSANG KWAK received the B.S. and Ph.D. degrees from Hanyang University, Seoul, South Korea, in 2013 and 2018, respectively. From 2018 to 2019, he conducted research at French National Research Center (CNRS), followed by a postdoctoral position at Purdue University, from 2019 to 2021. Since 2021, he has been a Professor with the Kumoh National Institute of Technology. His research interests include dynamics-based deep learning and integrating vibration-wave system modeling with advanced signal processing techniques.

MINSEOK KIM received the B.S. degree in mechanical system engineering from Chonnam National University, Gwangju, South Korea, in 2010, and the Ph.D. degree in mechanical engineering from Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea, in 2015. Following the Ph.D. degree, he conducted postdoctoral research at Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, from 2015 to 2017. He later became a Senior Researcher with Daegu Convergence Technology Research Center, part of Korea Institute of Machinery and Materials (KIMM), Daegu, South Korea. He is currently a Professor with the Kumoh National Institute of Technology (KIT), Gumi, South Korea. His research interests include microfluidics, biosensors, bio-MEMS, nanofabrication, and the application of deep learning in bio-chemical analysis, driving innovations in these interdisciplinary fields.

SHUJAAT KHAN received the Ph.D. degree from the Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2022. He was a Researcher with the Synergistic Bioinformatics (SynBi) and the Bio Imaging, Signal Processing Learning (BISPL), KAIST. He is currently an Assistant Professor with the Department of Computer Engineering and a fellow with Saudi Data and AI Authority (SDAIA) and the King Fahd University of Petroleum & Minerals (KFUPM), Dhahran, Saudi Arabia, under the SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRC-AI), KFUPM. Prior to joining KFUPM, he was a Senior AI Scientist with the Digital Technology and Innovation, Siemens Medical Solutions USA, Inc. His research interests include machine learning, optimization, inverse problems, and signal processing, with a focus on their applications in various domains.