


GRU4ACE: Enhancing ACE inhibitory peptide prediction by integrating gated recurrent unit with multi-source feature embeddings

Saeed Ahmed^{1,2} | Nalini Schaduangrat¹ | Pramote Chumnannpuen^{3,4} |
Watshara Shoombuatong¹ 

¹Center for Research Innovation and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand

²Department of Computer Science, University of Swabi, Swabis, Pakistan

³Department of Zoology, Faculty of Science, Kasetsart University, Bangkok, Thailand

⁴Kasetsart University International College (KUIC), Kasetsart University, Bangkok, Thailand

Correspondence

Watshara Shoombuatong, Center for Research Innovation and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand.

Email: watshara.sho@mahidol.ac.th

Funding information

Mahidol University; National Research Council of Thailand

Review Editor: Nir Ben-Tal

Abstract

Accurate identification of angiotensin-I-converting enzyme (ACE) inhibitory peptides is essential for understanding the primary factor regulating the renin-angiotensin system and guiding the development of new drug candidates. Given the inherent challenges in experimental processes, computational methods for in silico peptide identification can be invaluable for enabling high-throughput characterization of ACE inhibitory peptides. This study introduces GRU4ACE, an innovative deep learning framework based on multi-view information for identifying ACE inhibitory peptides. First, GRU4ACE utilizes multi-source feature encoding methods to capture the information embedded in ACE inhibitory peptides, including sequential information, graphical information, semantic information, and contextual information. Specifically, the feature representations used herein are derived from conventional feature descriptors, natural language processing (NLP)-based embeddings, and pre-trained protein language model (PLM)-based embeddings. Next, multiple feature embeddings were fused, and the elastic net was employed for feature optimization. Finally, the optimal feature subset with strong feature representation was input into a gated recurrent unit (GRU). The proposed GRU4ACE approach demonstrated superior performance over existing methods in terms of the independent test. To be specific, the balanced accuracy, sensitivity, and MCC scores of GRU4ACE reached 0.948, 0.934, and 0.895, which were 6.46%, 8.92%, and 12.51% higher than those of the compared methods, respectively. In addition, when comparing well-regarded feature descriptors, we found that the proposed multi-view features effectively captured crucial information, leading

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2025 The Author(s). *Protein Science* published by Wiley Periodicals LLC on behalf of The Protein Society.

to improved ACE inhibitory peptide prediction performance. These comprehensive results highlight that GRU4ACE enhances prediction accuracy and significantly narrows down the search for new potential antihypertensive drugs.

KEYWORDS

ACE inhibitory peptide, bioinformatics, deep learning, feature representation, feature selection, machine learning

1 | INTRODUCTION

Hypertension, a prevalent chronic condition, is closely associated with an elevated risk of cerebrovascular and cardiovascular diseases. The angiotensin I-converting enzyme (ACE, EC 3.4.15.1) plays a crucial role in blood pressure regulation, which can increase blood pressure via the kallikrein-kinin system (KKS) and the renin-angiotensin system (RAS) (Xue et al., 2021). ACE can act on angiotensin I to produce angiotensin II within the RAS and inactivate bradykinin in the KKS, ultimately leading to elevated blood pressure (Lee & Hur, 2017; Xue et al., 2021). Consequently, ACE has become a prime target for hypertension treatment. In recent years, substantial research has focused on ACE inhibitors, with ACE inhibitory peptides attracting significant interest due to their health and safety benefits compared to manufactured inhibitors like captopril and lisinopril, which may have more severe side effects (Duan et al., 2023; Lee & Hur, 2017). Through the process of fermentation or enzymatic hydrolysis, peptides exhibiting ACE inhibitory activity are derived from natural proteins. These peptides are screened and prepared from various protein sources, including plants like soybeans, rice bran, and zein; animal sources such as meat, milk, eggs, and fish; as well as microorganisms (Ding et al., 2023; Xue et al., 2021). There is growing interest in identifying peptides from plant proteins due to the increasing demand for balanced, healthy diets and their high sustainability (Rizzello et al., 2016).

Currently, a vast number of novel proteins are being generated through next-generation sequencing techniques. Among these proteins, there may be several peptide candidates with ACE inhibitory activity. However, in vitro or in vivo experimental methods are cost-ineffective, time-consuming, and labor-intensive (Basith et al., 2020; Du et al., 2024; Kalyan et al., 2021; Kumar, Chaudhary, Sharma, et al., 2015; Lertampaiporn et al., 2022; Manavalan et al., 2019; Win et al., 2018; Wu et al., 2019; Zhuang et al., 2021). To alleviate these limitations, several computational tools have been developed to identify ACE inhibitory peptides (Kalyan et al., 2021; Kumar, Chaudhary, Sharma, et al., 2015; Lertampaiporn

et al., 2022; Manavalan et al., 2019; Win et al., 2018; Wu et al., 2019; Zhuang et al., 2021). According to this study, Kumar et al. developed the first computational model (AHTpin) to identify ACE inhibitory peptides using only sequence information (Kumar, Chaudhary, Singh Chauhan, et al., 2015). AHTpin was developed based on the first benchmark dataset, which consisted of 386 ACE inhibitory peptides and 386 non-ACE inhibitory peptides. To date, this benchmark dataset remains the most widely used for developing computational models, such as PAAP (Win et al., 2018), mAHTPred (Manavalan et al., 2019), ACHP (Xu et al., 2021), UniDL4BioPep (Du et al., 2023), and Ensemble-AHTPpred (Lertampaiporn et al., 2022). Additional details regarding computational tools designed for identifying ACE inhibitory peptides can be found in the study by (Basith et al., 2020). However, the aforementioned computational models were developed based on incorrect negative samples, which may limit their predictive capability in some cases. Recently, Du et al. (2024) established a new dataset containing true positive and true negative samples. Using this high-quality benchmark dataset, pLM4ACE was developed with machine learning algorithms, including random forest (RF), support vector machine (SVM), k-nearest neighboring (KNN), logistic regression (LR), and multilayer perceptron (MLP). The pre-trained protein language model (PLM) was used as a feature extractor to generate a high-dimensional feature vector. The pLM4ACE model provided reasonable prediction results, with a balanced accuracy (BACC) of 0.883, sensitivity (SN) of 0.845, and Matthew's Correlation Coefficient (MCC) of 0.770. However, this method still has potential limitations: (i) Since pLM4ACE was developed based using a single feature descriptor, it has limited discriminative ability to capture the diverse information of ACE inhibitory peptides. (ii) The overall predictive performance of pLM4ACE is still not sufficient for practical implementation in identifying ACE inhibitory peptides.

In this study, we present a novel deep-learning framework, called GRU4ACE, to enhance the accurate identification of ACE inhibitory peptides by leveraging multi-view information (Figure 1). The major contributions of our proposed approach can be summarized as follows:

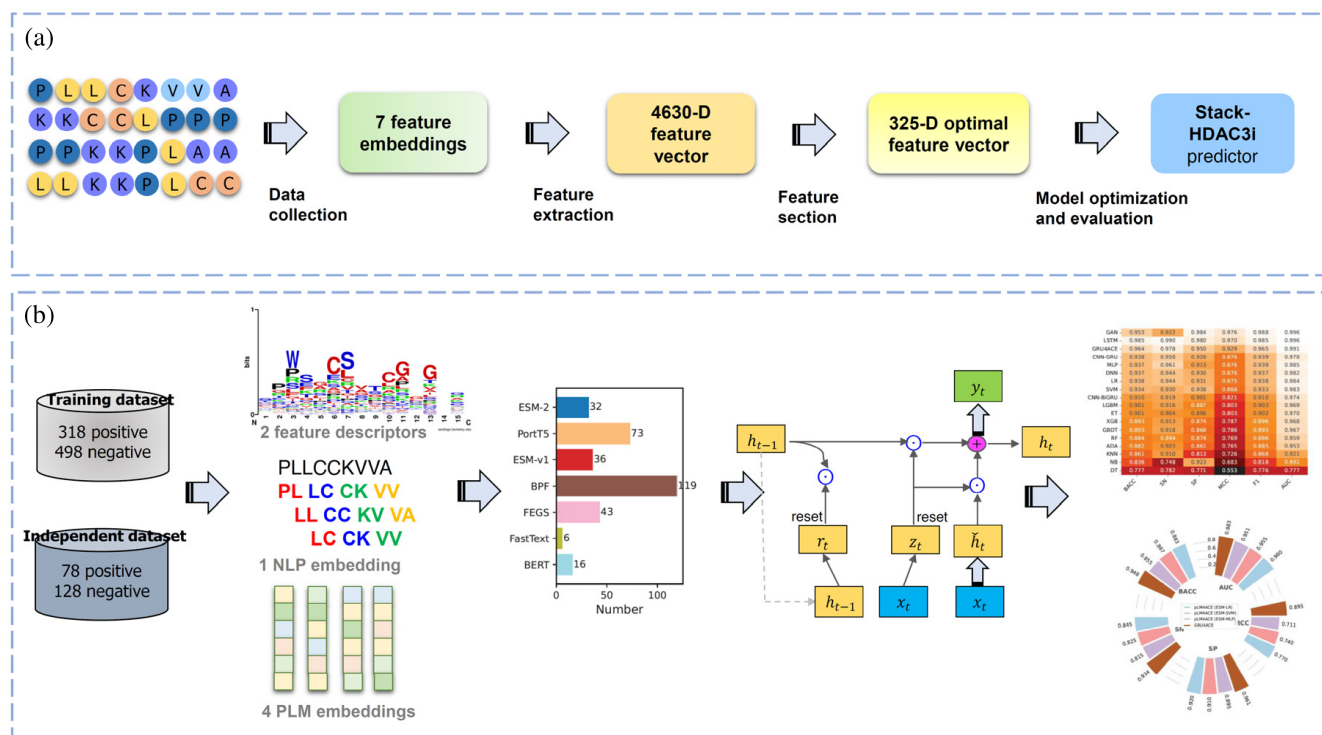


FIGURE 1 An overview of GRU4ACE framework for identifying ACE inhibitory peptides. It involves the following steps: (i) dataset preparation, (ii) feature extraction, (iii) feature importance selection, and (iv) model optimization and evaluation.

- GRU4ACE is the first approach based on a gated recurrent unit (GRU) for the identification of ACE inhibitory peptides.
- Recently, an increasing number of prediction models have been developed using pre-trained protein language model (PLM)-based embeddings. Instead of relying solely on PLM embeddings, GRU4ACE combines multi-source feature embeddings to capture diverse information from different perspectives, including sequential, graphical, semantic, and contextual information. Specifically, GRU4ACE utilizes three types of embeddings: PLM embeddings, conventional feature descriptors, and natural language processing (NLP)-based embeddings.
- The elastic net (EN) method was assigned to determine the optimal feature subset with strong feature representation, while the synthetic minority over-sampling technique (SMOTE) was used to address the imbalanced dataset.
- Experimental results demonstrated that the proposed multi-view features exhibit remarkably superior predictive performance compared to conventional feature descriptors.
- Through the independent test, GRU4ACE significantly outperformed the existing method in terms of BACC, SN, and MCC. To be specific, GRU4ACE achieved BACC, SN, and MCC scores of 0.948, 0.934,

and 0.895, respectively, representing improvements of 6.46%, 8.92%, and 12.51% over the existing method.

2 | MATERIALS AND METHODS

2.1 | Benchmark dataset

Recently, Du et al. (2024) established a high-quality benchmark dataset to develop their proposed model (i.e., pLM4ACE). Specifically, this benchmark dataset was created by manually curating up-to-date ACE inhibitory peptides (referred to as the positive dataset) from multiple sources, such as AHTPDB (Kumar, Chaudhary, Sharma, et al., 2015), BIOPEP-UWM (Minkiewicz et al., 2019), and a research article (Kalyan et al., 2021). All ACE inhibitory peptides curated by Du et al. were experimentally verified with half maximal inhibitory concentration (IC_{50}) values ranging from 0 to 50 μ M. Peptides with low or non-ACE inhibitory activity ($IC_{50} > 50 \mu$ M) were considered as the negative dataset. After the screening process, 394 positive samples and 626 negative samples were obtained. Finally, the benchmark dataset was split in an 8:2 ratio to construct the training (318 positives and 498 negatives) and

independent test (76 positives and 128 negatives) datasets. ACE inhibitory peptides typically range from 2 to 12 amino acids, with some extending up to 21 amino acids (Li et al., 2004). Our dataset includes peptides of 2–21 amino acids (Table S1), prioritizing experimentally validated sequences. Over 90% of peptides in both the positive and negative datasets are shorter than 10 amino acids. To maintain consistency, the negative dataset was limited to a similar length distribution, focusing on the most relevant ACE inhibitory peptides.

2.2 | Peptide feature representation

2.2.1 | Conventional feature encoding

The binary profile (BPF) encodes each amino acid into a 20-D binary vector (Hasan et al., 2020). Specifically, a binary one-hot vector is used to represent each amino acid, ensuring a consistent feature dimension. For example, Ala is represented as $f(A) = [1, 0, 0, 0, \dots, 0, 0]$. Thus, for a given protein sequence P , it can be mathematically defined as follows:

$$BPF(P) = (f(A), f(C), \dots, f(Y)). \quad (1)$$

In addition, we employed the recent feature encoding (FEGS) developed by Mu et al. to provide graphical information embedding in peptide sequences (Mu et al., 2021). FEGS exploits two kinds of information in peptide sequences, including graphical information and statistical information. For a given protein sequence P , FEGS provides a 578-D feature vector.

2.2.2 | FastText encodings

In the field of NLP, vector representation of words (i.e., word embeddings) plays an important role in enabling machines to understand natural language (Gasparetto et al., 2022; Le & Mikolov, 2014). Additionally, this approach can effectively address the limitations of manual feature encoding methods. Since protein sequences and residues are considered as sentences and words, respectively, NLP-based embedding methods can be applied to generate sequence embeddings. FastText, a well-known word representation library, is one such embedding method that has been applied in bioinformatics and computational biology (Charoenkwan et al., 2021; Jin & Yang, 2022; Raza et al., 2023). This method is considered efficient for vector representation as it employs morphological cues to identify challenging words, enhancing its generalizability. In particular, FastText

employs single words and custom n -grams as features. By using n -grams, this method can generate useful representations for handling unknown words (Umer et al., 2023). For a given protein sequence P , FastText provides a 100-D feature vector.

2.2.3 | Protein language model

Extracting features from small-scale datasets may not ensure a high-accuracy and robust prediction model. To address this limitation, a promising strategy is to leverage pre-trained PLMs, which have been trained on over a million protein sequences. Remarkably, these approaches allow us to capture crucial patterns that might be missed by small-scale datasets. Recently, many research teams have worked to develop several versions of pre-trained PLMs, each based on different architectures and fine-tuned with various databases. Numerous studies have indicated that PLMs can reveal information embedded in protein sequences and effectively generate high-dimensional representations for downstream tasks. Therefore, we decided to employ four powerful pre-trained PLMs to generate high-dimensional and crucial feature representations embedded in ACE inhibitory peptides, including BERT (Devlin et al., 2018), ProtTrans (Elnaggar et al., 2020; Raffel et al., 2020), and ESM (Charoenkwan, Schaduengrat, et al., 2023; Hayes et al., 2024; Rives et al., 2021).

The first PLM is BERT, developed by Devlin et al. (2018). BERT is based on a bidirectional transformer pre-trained with a novel masked language model (MLM) (Devlin et al., 2018). This design allows the model's encoding layer to process all words in a sentence simultaneously, handling left and right contexts concurrently through a multi-head self-attention mechanism. The attention mechanism relies on three key concepts: Value, Key, and Query. Specifically, Value and Key represent the original value and key vector representations of each word in the context, respectively, while Query represents the target word or the annotated word to be generated. Because these transformer-based models were trained on large-scale protein sequences, they can effectively capture diverse information embedded in protein sequences (Ghazikhani & Butler, 2024; Villegas-Morcillo et al., 2022). Consequently, we utilized two types of transformer-based models, ProtTrans (Raffel et al., 2020) and ESM (Charoenkwan, Schaduengrat, et al., 2023; Rives et al., 2021). ProtTrans are encoder-decoder models built on several architectures (Raffel et al., 2020). We selected ProtTrans employing the T5 architecture (referred to as ProtT5), which was trained on the Big Fantastic Database (BFD) with 3B parameters and

subsequently fine-tuned using Uniref50. For the ESM models, we utilized two popular versions, ESM-1b (Rives et al., 2021) and ESM-2 (Charoenkwan, Schaduengrat, et al., 2023), both developed by the Meta Fundamental AI Research Team in 2019. The ESM-1b model is a transformer-based, 33-layer PLM with 650M parameters, while the ESM-2 model, selected for its efficiency, is an 8-layer PLM with 8M parameters. Both ESM-1b and ESM-2 were trained on millions of protein sequences from UniRef50.

2.3 | Feature selection method

In supervised learning, feature selection aims to identify an optimal feature subset comprising m out of n features, where $m \ll n$. Several studies have shown that using an optimal feature subset can enhance prediction performance and reduce time complexity (Akbar, Zou, et al., 2024; Raza et al., 2024; Shoombuatong, Homdee, et al., 2024; Shoombuatong, Meewan, et al., 2024; Zou et al., 2016; Zou & Hastie, 2005a). In 1996, Tibshirani introduced a powerful feature selection method called LASSO (Tibshirani, 1996), which incorporates the L_1 norm into a simple linear model to automatically select important features. However, LASSO has limitations in predicting correlated variables. To address this, Zou and Hastie proposed the elastic net (EN) method by incorporating both the L_1 and L_2 norms into a simple linear model (Zou & Hastie, 2005b). The simple linear model and EN method can be defined as

$$\min_{\omega} \sum_{i=1}^m (y_i - \omega x_i)^2, \quad (2)$$

$$\min_{\omega} \frac{1}{2 * n} \|y - X\omega\|_2^2 + \alpha * \beta \|\omega\|_1 + \frac{1}{2} (1 - \beta) \|\omega\|_2^2, \quad (3)$$

where X and y are the feature vector and the class label (i.e., positives and negatives), respectively, while n and ω indicate the number of samples and the regression coefficient, respectively.

2.4 | Synthetic minority over-sampling technique

As mentioned above, the training dataset was imbalanced between positive and negative samples, which could degrade the model performance. To mitigate the effect of data imbalance, we applied the SMOTE method to create new minority samples (positive samples)

(Chawla et al., 2002). To date, this method has been successfully applied to various biological and chemical classification problems (Hassan et al., 2023; Jiang et al., 2024; Kabir et al., 2018; Schaduengrat et al., 2019; Zhou et al., 2022). In SMOTE, Euclidean distance is used to compute the distances between x and the minority samples to generate new samples according to the following formula:

$$x_{\text{new}} = x + \text{rand}(0, 1) \times |x - x_n|, i = 1, \dots, n, \quad (4)$$

where x_{new} and x are the SMOTE-based and original positive samples, respectively. After applying the SMOTE method, the balanced training dataset contained 498 positive and 498 negative samples.

2.5 | Gated recurrent unit

Deep learning (DL) models have so far achieved satisfactory prediction results in bioinformatics and related tasks. One well-regarded DL method is the GRU, which is a type of long short-term memory (LSTM) and an improved version of the recurrent neural network (RNN). The RNN has a limitation known as the vanishing gradient problem in its network structure. To address this, GRU was developed with a specialized network architecture that uses a gating mechanism. As a result, GRU can naturally learn and retain long-term dependency information. In GRU, there are two important gates (i.e., the update gate (z_t) and the reset gate (r_t)). After obtaining feature embeddings, each layer of GRU can be computed as follows:

(i) Update gate: This gate enables the model to decide the extent of past information (from prior time steps) that should be retained and carried forward. The formula used to compute z_t at the current time t is defined as follows:

$$z_t = \sigma(c_z \cdot [h_{t-1}, x_t] + b_z), \quad (5)$$

where x_t and h_{t-1} represent the input at time t and the hidden state at time $t - 1$, respectively. Meanwhile, W is the weight, and b is the bias. Consequently, z_t and r_t determine the information that can be employed as the final output.

(ii) Reset gate: This gate calculates how much of the past information is forgotten:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r), \quad (6)$$

(iii) Employing r_t to store relevant information from the past:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \check{h}_t, \quad (7)$$

where $\check{h}_t = \tanh(W \cdot [r_t \odot h_{t-1}, x_t] + b)$ and \odot represents the Hadamard product. Finally, the sigmoid activation function is applied in the final output layer, formulated as follows:

$$\text{Sigmoid}(x) = \frac{1}{1 + \exp(-x)}. \quad (8)$$

2.6 | Performance evaluation

This study used six commonly employed performance metrics for binary classification tasks, including F1, BACC, MCC, SN, specificity (SP), and the area under the receiver operating curve (AUC) (Akbar et al., 2022; Akbar, Raza, & Zou, 2024; Azadpour et al., 2014), to evaluate the performance of the proposed models. SN, SP, BACC, MCC, and F1 are defined as follows:

$$\text{SN} = \frac{\text{TP}}{(\text{TP} + \text{FN})}, \quad (9)$$

$$\text{SP} = \frac{\text{TN}}{(\text{TN} + \text{FP})}, \quad (10)$$

$$\text{BACC} = \frac{\text{SN} + \text{SP}}{2}, \quad (11)$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}, \quad (12)$$

$$\text{F1} = 2 \times \frac{\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \quad (13)$$

Here, TP and TN refer to the numbers of correctly predicted positive and negative samples, respectively, while FP and FN denote the numbers of falsely predicted positive and negative samples (Amjad et al., 2024; Arif et al., 2024; Arshad et al., 2024; Kanwal et al., 2024; Kurata et al., 2022; Manavalan et al., 2019).

3 | RESULTS AND DISCUSSIONS

3.1 | Amino acid composition analysis

In this section, we performed amino acid composition analysis to highlight significant trends in amino acid

preferences. As can be seen in Figure 2, proline (P) appears more frequently in peptides with higher activity (16.51%) compared to those with lower activity (14.32%), suggesting its potential role in enhancing ACE inhibition. Similarly, glutamic acid (E) shows a marked contrast, being present in 5.35% of active peptides versus only 1.15% in inactive ones, underscoring its beneficial effect on inhibitory activity (Sun et al., 2019; Wang et al., 2021). Additionally, amino acids such as glycine (G) and serine (S) are more abundant in active peptides, further supporting their significance in structural configurations favorable for ACE inhibition (He et al., 2021; Liu et al., 2021). On the other hand, amino acids like tyrosine (Y) and cysteine (C) are more common in less active peptides, suggesting a potential negative impact on ACE inhibition (Sarbon, Howell, & Ahmad, 2019; Wang et al., 2021). The importance of hydrophobic residues at the N- and C-termini of peptides is also emphasized, with research indicating that terminal residues such as leucine (L), isoleucine (I), and proline (P) enhance ACE binding and inhibitory capacity (Sun et al., 2019; Wang et al., 2021). Experimentally validated datasets are essential, as they ensure that predictive models are based on biological data rather than solely on theoretical or computational assumptions. This makes these datasets indispensable for the rational design of novel ACE inhibitors for therapeutic applications.

3.2 | Performance of different feature encodings

As mentioned in Section 2.2, ESM-2, PortT5, ESM-v1, BPF, FECS, FastText, and BERT were used to extract 320-D, 1024-D, 1280-D, 560-D, 578-D, 100-D, and 768-D feature vectors, respectively. To assess the discriminative ability of these feature encoding methods in identifying ACE inhibitory peptides, we constructed and evaluated the performance of GRU models using these feature vectors in a fivefold cross-validation test on the training dataset. Specifically, we applied seven commonly used performance measures (i.e., ACC, BACC, SN, SP, MCC, AUC, and F1) to analyze the performance of each feature encoding. The detailed prediction results of these encodings based on two standard evaluation strategies are recorded in Table 1. From Table 1, the MCC scores of ESM-2, PortT5, ESM-v1, BPF, FECS, FastText, and BERT are 0.877, 0.898, 0.875, 0.799, 0.775, 0.505, and 0.864, respectively. It is notable that the top-three features, including PortT5, ESM-2, and ESM-v1, were derived from PLM-based feature encoding methods. In terms of AUC scores, as depicted in Figure 3, PortT5, ESM-2, and ESM-v1 also outperformed other feature encodings.

FIGURE 2 A bar graph to represent percentage amino acid composition of ACE inhibitory (positive) and non-ACE inhibitory (negative) peptides.

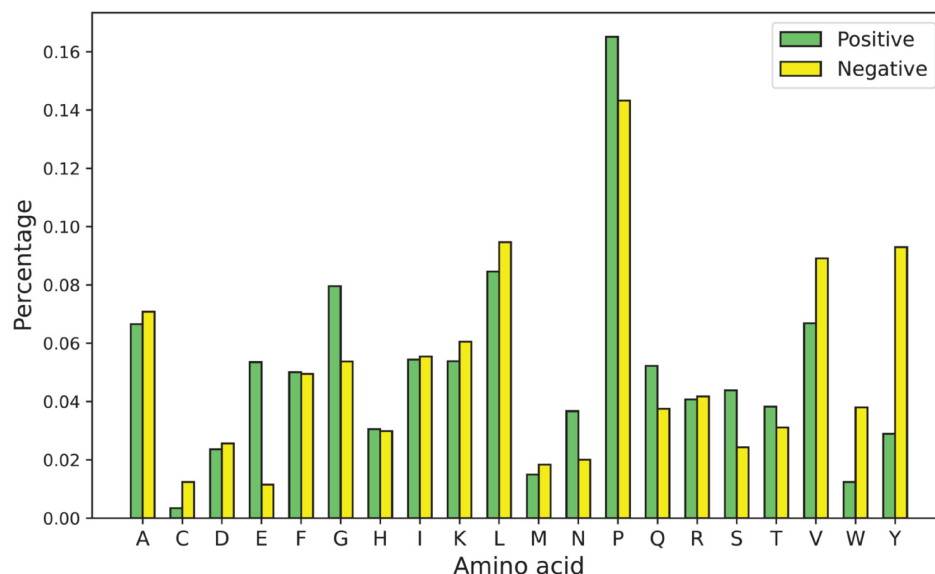


TABLE 1 Performance of seven feature encoding schemes in identifying ACE inhibitory peptides on the training dataset.

Descriptor	BACC	SN	SP	MCC	F1	AUC
ESM-2	0.938	0.924	0.952	0.877	0.924	0.977
PortT5	0.947	0.922	0.972	0.898	0.935	0.976
ESM-v1	0.937	0.925	0.950	0.875	0.924	0.972
BPF	0.899	0.871	0.926	0.799	0.873	0.955
FEGS	0.888	0.862	0.914	0.775	0.861	0.942
FastText	0.750	0.701	0.799	0.505	0.688	0.837
BERT	0.932	0.919	0.946	0.864	0.917	0.959
Fusion	0.951	0.944	0.958	0.901	0.940	0.980

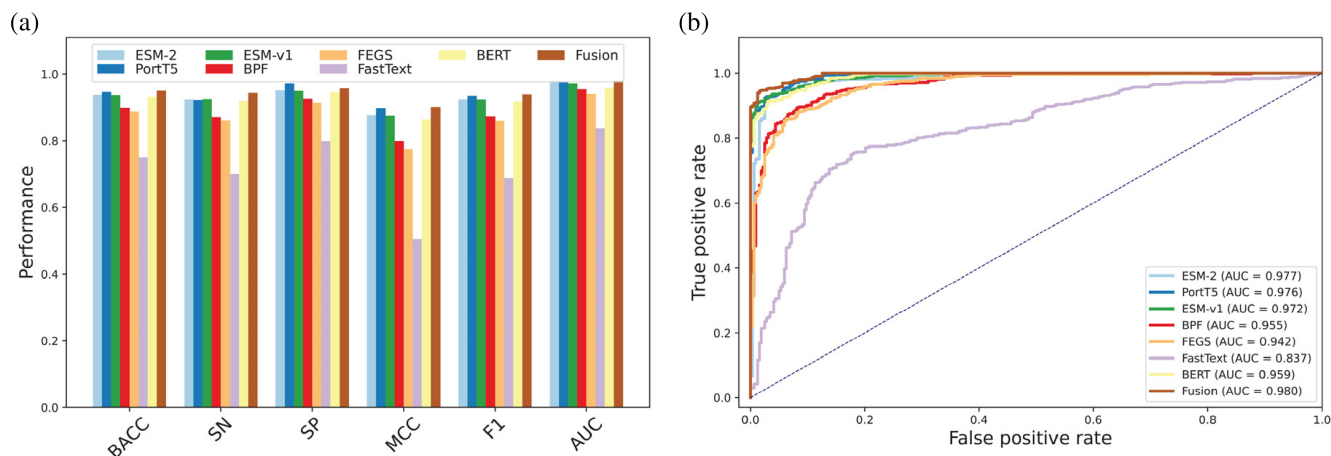


FIGURE 3 Performance comparison of seven feature encoding schemes on the training (a,b) and independent test (c,d) datasets.

Interestingly, the best-performing feature was obtained from PortT5, achieving a BACC of 0.947, SN of 0.922, SP of 0.972, MCC of 0.898, F1 of 0.935, and AUC of 0.976. This indicates that PLM-based feature embeddings might

be beneficial for capturing crucial information in ACE inhibitory peptides. Although PortT5 provides reasonable results, its generalizability remains unsatisfactory. Rather than selecting a single best feature encoding to develop

the model, we combined multiple feature representations to provide more comprehensive information. Herein, we combined ESM-2, PortT5, ESM-v1, BPF, FECS, FastText, and BERT to generate fused features with a 4630-D feature vector (Fusion). As can be seen from Table 1, fusion outperformed other feature representations in terms of BACC, SN, MCC, F1, and AUC scores.

3.3 | Analyzing the effect of the data balancing technique on predictive performance

Since the number of positive samples in the training dataset is less than that of negative samples, models trained with this dataset tend to be biased toward the negative class (Bourel et al., 2021; Mahmud et al., 2021). Herein, the SMOTE technique was thus applied to generate positive samples that were equivalent in number to the negative samples. In addition, we applied an under-sampling technique to address the imbalance problem. The training datasets (positive and negative) generated using the SMOTE and under-sampling techniques consisted of (498 and 498) and (318 and 318) samples, respectively. The performance of models trained with three different training datasets is provided in Table 2, where the control refers to the original training dataset. Table 2 shows that the model trained with the SMOTE-based balanced training dataset achieved the highest MCC scores across both evaluation strategies. On the independent test dataset, the SN, SP, MCC, F1, and AUC scores of this model were 0.816, 0.906, 0.726, 0.827, and 0.918, respectively. Altogether, we employed the SMOTE-based balanced training dataset to develop the new model proposed in this study.

3.4 | Analyzing the impact of the feature selection method on predictive performance

Although the SMOTE technique improved the SN score on the training dataset, its SN score on the independent test dataset was lower than that of the under-sampling technique. The possible reason is that the number of features was larger than the size of the training dataset, potentially leading to dimensionality issues and overfitting (Chowdhury et al., 2017; Mahmud et al., 2021). To solve this problem while improving prediction performance, we applied five commonly used feature selection methods to determine the optimal feature subsets, including LASSO (Mahmud et al., 2024; Tibshirani, 1996), EN (Arif et al., 2021), mRMR (Zou et al., 2016), PCA (Yan

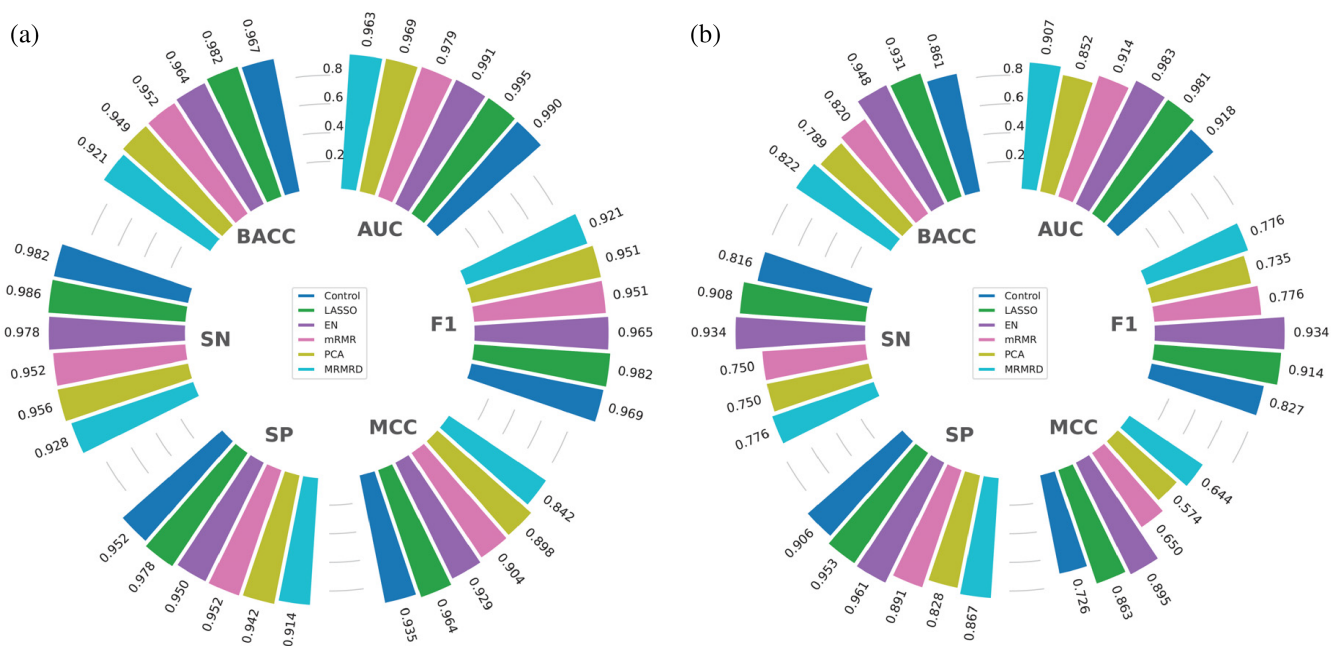
et al., 2024), and MRMRD (Zou et al., 2016). The numbers of selected important features obtained by LASSO, EN, mRMR, PCA, and MRMRD were 264, 325, 200, 1000, and 100, respectively. We then input these optimal feature subsets into GRU models and evaluated their performance in terms of F1, AUC, BACC, MCC, SN, and SP across both evaluation strategies, as summarized in Figure 4 and Table 3. As can be seen, the cross-validation MCC scores of LASSO, EN, mRMR, PCA, and MRMRD are 0.964, 0.929, 0.904, 0.898, and 0.842, respectively. This suggests that LASSO achieved the best performance, with EN and the control obtaining the second and third highest MCC scores, respectively. Although EN's cross-validation MCC score was slightly lower than that of LASSO (0.964 versus 0.929) and the control (0.935 versus 0.929), its prediction results on the independent test dataset outperformed those of the other methods across all performance metrics. Remarkably, the BACC, SN, SP, MCC, F1, and AUC scores of EN were 0.948, 0.934, 0.961, 0.895, 0.934, and 0.983, which were 7.84%, 8.66%, 11.84%, 5.47%, 16.91%, 10.75%, and 6.50% higher than the control, respectively, highlighting the contribution of the optimal feature subset obtained by EN in precisely identifying ACE inhibitory peptides on both the training and independent test datasets. Therefore, we applied the optimal feature subset obtained by EN to construct our proposed model. As shown in Figure 5, this optimal feature set consists of 325 features derived from 32 ESM-2, 73 PortT5, 36 ESM-v1, 119 BPF, 43 FECS, 6 FastText, and 16 BERT features. This indicates that the top four important feature descriptors are from BPF, PortT5, FECS, and ESM-v1, accounting for 36.62%, 22.46%, 13.23%, and 11.08%, respectively. This finding reaffirms the value of PortT5, FECS, and ESM-v1 in capturing critical information about ACE inhibitory peptides. These results are consistent with previous work (Du et al., 2024), where the LR classifier coupled with ESM-2 feature embedding demonstrated superior performance.

3.5 | Multi-view features contribute to performance improvement

As aforementioned, we combined different sources of information to generate multi-view features, enabling the capture of critical information about ACE inhibitory peptides. To demonstrate the impact of combining different perspectives, we first compared the performance of our proposed features with that of their constituent feature descriptors (i.e., ESM-2, PortT5, ESM-v1, BPF, FECS, FastText, and BERT). Table S2 summarizes the prediction results of our proposed features and their constituent feature descriptors. From Table S1, our proposed features

TABLE 2 The influence of different data balancing methods over the cross-validation and independent tests.

Evaluation strategy	Method	BACC	SN	SP	MCC	F1	AUC
Cross-validation	Control	0.951	0.944	0.958	0.901	0.940	0.980
	Under-sampling	0.942	0.940	0.944	0.884	0.942	0.978
	SMOTE	0.967	0.982	0.952	0.935	0.969	0.990
Independent test	Control	0.851	0.803	0.898	0.705	0.813	0.913
	Under-sampling	0.842	0.855	0.828	0.668	0.798	0.919
	SMOTE	0.861	0.816	0.906	0.726	0.827	0.918

**FIGURE 4** Performance comparison of different feature selection methods on the training (a) and independent test (b) datasets.**TABLE 3** Comparison of different feature selection methods over the cross-validation and independent tests.

Evaluation strategy	Method	BACC	SN	SP	MCC	F1	AUC
Cross-validation	Control	0.967	0.982	0.952	0.935	0.969	0.990
	LASSO	0.982	0.986	0.978	0.964	0.982	0.995
	EN	0.964	0.978	0.950	0.929	0.965	0.991
	mRMR	0.952	0.952	0.952	0.904	0.952	0.979
	PCA	0.949	0.956	0.942	0.898	0.951	0.969
	MRMRD	0.921	0.928	0.914	0.842	0.921	0.963
Independent test	Control	0.861	0.816	0.906	0.726	0.827	0.918
	LASSO	0.931	0.908	0.953	0.863	0.914	0.981
	EN	0.948	0.934	0.961	0.895	0.934	0.983
	mRMR	0.820	0.750	0.891	0.650	0.776	0.914
	PCA	0.789	0.750	0.828	0.574	0.735	0.852
	MRMRD	0.822	0.776	0.867	0.644	0.776	0.907

achieved the best prediction results in terms of BACC, SN, MCC, AUC, and F1, as indicated by the cross-validation strategy. Furthermore, when comparing the best-performing individual feature descriptor (PortT5) on the independent test, our proposed features demonstrated significantly superior predictive performance, with 9.46%

higher BACC, 15.79% higher SN, 17.11% higher MCC, and 11.48% higher F1. Additionally, to intuitively analyze the discriminative ability of our multi-view features, we projected both our proposed features and the constituent feature descriptors into a 2-D feature space using the well-known feature reduction method (t-Distributed Stochastic Neighbor Embedding (t-SNE)) (Boschin et al., 2014). Figure 6 displays the t-SNE visualization of ESM-2, PortT5, ESM-v1, BPF, FECS, FastText, BERT, and our proposed features. As shown in Figure 6I, the two classes are clearly distinguishable through the multi-view features.

Secondly, we conducted ablation experiments to investigate how different views of features contribute to performance improvement. The ablation experiments were performed by designing and testing several variants of GRU4ACE. Taking ESM-2 as an example, GRU4ACE trained solely on 32 ESM-2 features was defined as GRU4ACE_ESM-2. Herein, the performance of seven variants of GRU4ACE was assessed using two standard evaluation strategies, with the prediction results recorded in Table S3. It could be observed that GRU4ACE significantly outperformed its variants in both cross-validation and independent tests. Impressively, the BACC and AUC

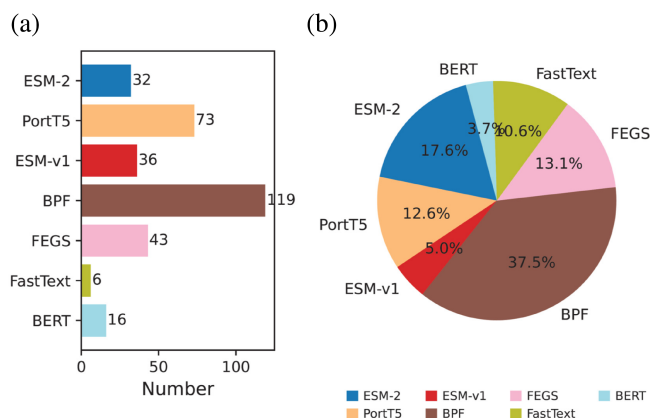


FIGURE 5 Analysis of the optimal feature set. The number (a) and proportion (b) of each type feature embedding selected from the optimal feature set.

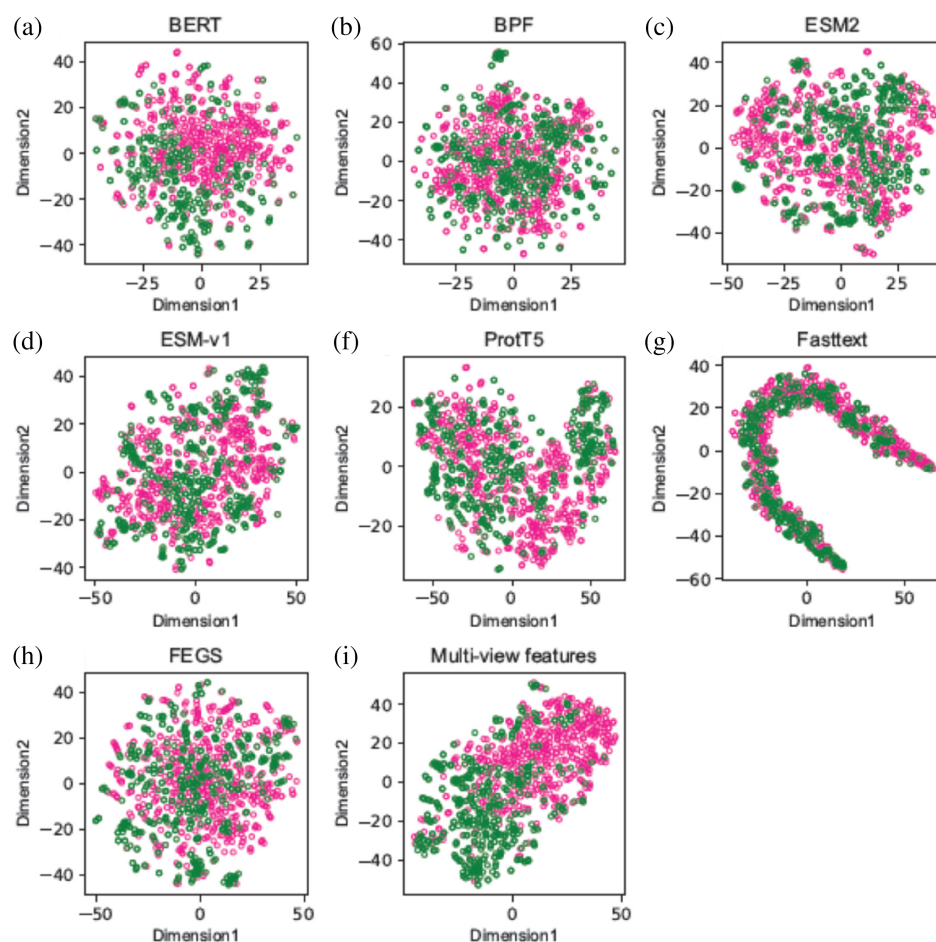


FIGURE 6 Analysis of feature ability of different feature embedding by t-distributed stochastic neighbor embedding (t-SNE).

achieved by GRU4ACE were 20.19%–44.76% and 15.84%–43.17% higher, respectively, than those of all the variants in the independent test. Altogether, the combination of different views of information enables the model to capture critical and diverse information about ACE inhibitory peptides, thereby contributing to the improved performance of ACE inhibitory peptide prediction.

3.6 | Comparison of GRU4ACE with well-regarded machine learning and deep learning methods

Selecting a suitable classifier is crucial for obtaining reliable and stable and prediction models (Basith et al., 2020; Charoenkwan et al., 2021; Charoenkwan et al., 2022; Manavalan et al., 2019; Shoombuatong, Homdee, et al., 2024; Shoombuatong, Meewan, et al., 2024). To confirm the effectiveness of GRU, we compared its performance with popular ML and DL classifiers using two standard evaluation strategies. In this study, the classifiers used for comparative analysis included 5 DL classifiers (i.e., CNN-BiGRU, DNN, CNN-GRU, LSTM, and GAN) (Table S4) and 12 ML classifiers (i.e., MLP, RF, LR, SVM, light gradient boosting machine (LGBM), partial least squares (PLS), k-nearest neighboring (KNN), extremely randomized trees (ET), naive Bayes (NB), decision tree (DT), AdaBoost (ADA), and extreme gradient boosting (XGB)) (Table S5). For a fair performance comparison, all ML and DL models were constructed based on the selected 325 features, and their prediction performance was evaluated using five-fold cross-validation and independent tests on the same benchmark training

and independent test datasets, respectively (as described in Figure 7 and Table 4, along with Table S6). Among the 18 classifiers, the top-five were GAN, LSTM, GRU, CNN-GRU, and MLP, as indicated by cross-validation MCC scores. This demonstrates that DL methods are well-suited for identifying ACE inhibitory peptides. The MCC scores of the top-five classifiers were 0.976, 0.970, 0.929, 0.876, and 0.876, respectively (Figure 7a and Table S6). On the independent test dataset, the highest MCC scores of 0.895, 0.874, 0.872, 0.864, and 0.853 were obtained by GRU, CNN-GRU, CNN-BiGRU, MLP, and LR, respectively (Figure 7b and Table 4). By considering both evaluation strategies, GRU, CNN-GRU, and MLP exhibited stable performance on both the training and independent test datasets. Furthermore, compared with other classifiers, GRU achieved the best BACC, MCC, and AUC across both standard evaluation strategies, highlighting that GRU not only has high prediction accuracy but also excellent generalization ability. Therefore, we selected GRU in conjunction with the selected 325 features to construct the proposed GRU4ACE model.

3.7 | Comparison of GRU2ACE with the existing methods

To demonstrate the superiority of the proposed GRU4ACE model, we compared its performance with state-of-the-art methods. Although many computational models have been developed to predict ACE inhibitory peptides, these models are often trained on datasets without a clear bioactivity threshold (Du et al., 2024). Recently, Du et al. (2024) established a reliable and

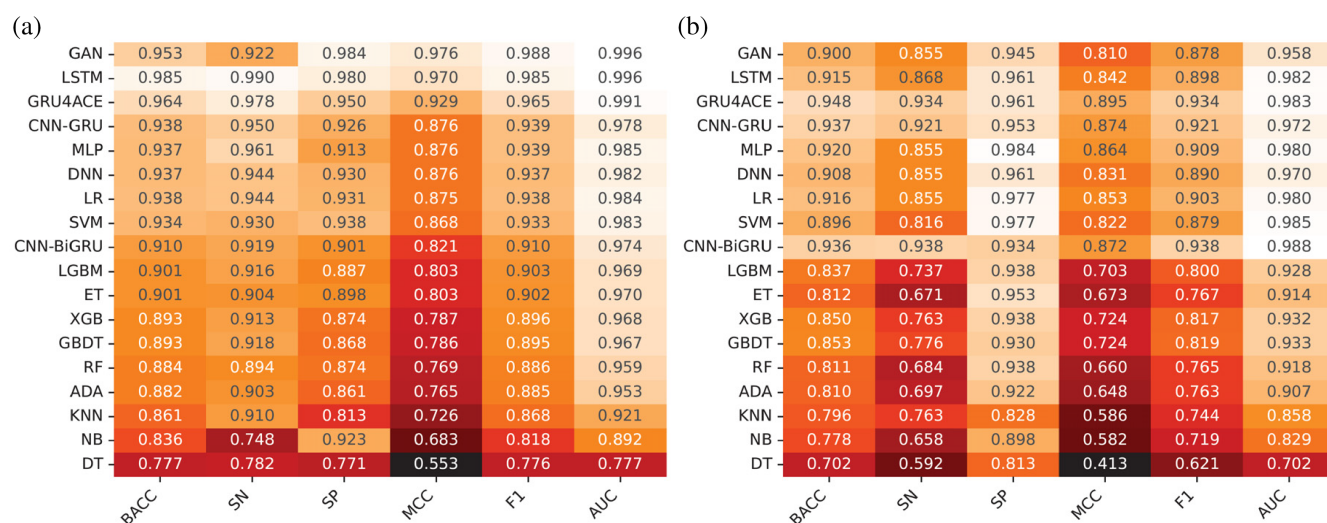


FIGURE 7 Heat-map of the prediction performance of different ML and DL classifiers in terms of the cross-validation (a) and independent tests (b).

Method	BACC	SN	SP	MCC	F1	AUC
DT	0.702	0.592	0.813	0.413	0.621	0.702
NB	0.778	0.658	0.898	0.582	0.719	0.829
KNN	0.796	0.763	0.828	0.586	0.744	0.858
ADA	0.810	0.697	0.922	0.648	0.763	0.907
RF	0.811	0.684	0.938	0.660	0.765	0.918
GBDT	0.853	0.776	0.930	0.724	0.819	0.933
XGB	0.850	0.763	0.938	0.724	0.817	0.932
ET	0.812	0.671	0.953	0.673	0.767	0.914
LGBM	0.837	0.737	0.938	0.703	0.800	0.928
CNN-BiGRU	0.936	0.938	0.934	0.872	0.938	0.988
SVM	0.896	0.816	0.977	0.822	0.879	0.985
LR	0.916	0.855	0.977	0.853	0.903	0.980
DNN	0.908	0.855	0.961	0.831	0.890	0.970
MLP	0.920	0.855	0.984	0.864	0.909	0.980
CNN-GRU	0.937	0.921	0.953	0.874	0.921	0.972
LSTM	0.915	0.868	0.961	0.842	0.898	0.982
GAN	0.900	0.855	0.945	0.810	0.878	0.958
GRU (GRU4ACE)	0.948	0.934	0.961	0.895	0.934	0.983

TABLE 4 Performance comparison of GRU4ACE with conventional ML and DL methods over the independent test.

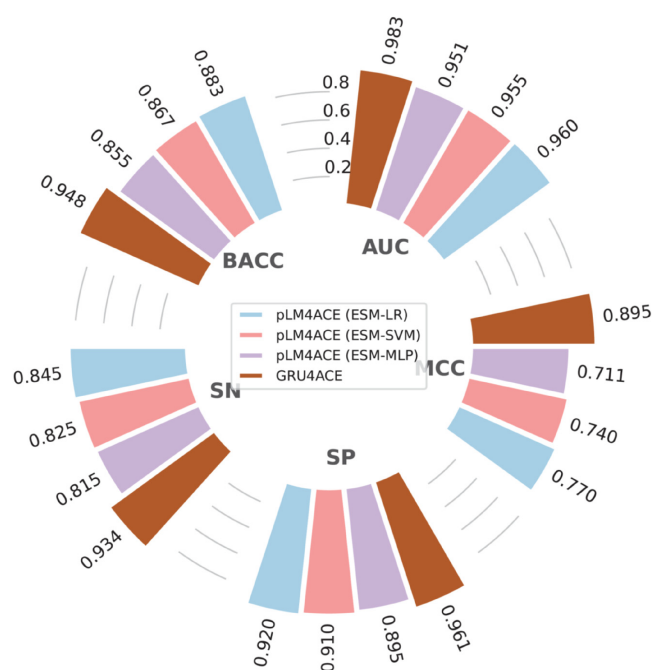


FIGURE 8 Performance comparison of GRU4ACE with the existing method over the independent test.

up-to-date dataset for constructing pLM4ACE. To make a fair comparison, we compared the performance of our model against pLM4ACE. Since three models were developed in pLM4ACE (i.e., pLM4ACE (ESM-LR), pLM4ACE (ESM-SVM), and pLM4ACE (ESM-MLP)), we compared

GRU4ACE with these three pLM4ACE models. The performance of GRU4ACE and three pLM4ACE models was evaluated based on the independent test. As shown in Figure 8 and Table 5, GRU4ACE outperformed the three pLM4ACE models in terms of BACC, SN, SP, MCC, and AUC. To be specific, the MCC score of our model was 0.895, which was 18.41%, 15.51%, and 12.51% higher than pLM4ACE (ESM-MLP), pLM4ACE (ESM-SVM), and pLM4ACE (ESM-LR), respectively. In addition, the BACC, SN, and SP scores of our model were higher than those of the three compared models by ranges of 6.46%–9.26%, 8.92%–11.92%, and 4.09%–6.59%, respectively. Taken together, the comparative results showed that GRU4ACE outperformed existing methods and achieved significantly better performance, especially in terms of SN and MCC. This evidence is sufficient to indicate that GRU4ACE can effectively reduce the number of false negatives, which is beneficial for the prioritizing candidate ACE inhibitory peptides.

4 | CONCLUSIONS

Accurate and cost-effective identification of peptides with ACE inhibition ability plays a significant role in expediting the discovery of new potential antihypertensive drugs. Therefore, an innovative GRU-based computational framework, GRU4ACE, was developed to identify ACE inhibitory peptides through the

TABLE 5 Performance comparison of GRU4ACE with the baseline models and existing predictors on the independent dataset.

Method	BACC	SN	SP	MCC	AUC
pLM4ACE (ESM-LR)	0.883	0.845	0.920	0.770	0.960
pLM4ACE (ESM-SVM)	0.867	0.825	0.910	0.740	0.955
pLM4ACE (ESM-MLP)	0.855	0.815	0.895	0.711	0.951
GRU4ACE	0.948	0.934	0.961	0.895	0.983

integration of multi-source feature embeddings, including conventional feature descriptors, NLP-based embeddings, and pre-trained PLM-based embeddings. Through a series of comparative experiments, we found that the proposed multi-view features exhibited remarkably outstanding prediction performance compared to single feature descriptors. When compared with well-regarded ML and DL methods, GRU4ACE attained the best predictive performance in terms of BACC, MCC, and AUC scores across both standard evaluation strategies. In the independent test, GRU4ACE significantly outperformed pLM2ACE in terms of BACC, SN, and MCC scores. Specifically, the BACC, SN, and MCC scores of GRU4ACE are 0.948, 0.934, and 0.895, which are 6.46%, 8.92%, and 12.51% higher than those of pLM2ACE, respectively. Experimental results showed that GRU4ACE was highly effective and outperformed other models in identifying ACE inhibitory peptides, attributed to the following factors: (i) GRU4ACE was developed and optimized using a reliable and high-quality benchmark dataset. (ii) GRU4ACE exploited multi-source feature embedding methods to encode the information embedded in ACE inhibitory peptides. The resulting feature embeddings effectively captured crucial information, including sequential, graphical, semantic, and contextual information. This approach contrasts with the existing method (pLM2ACE), which relied on a single PLM embedding (ESM-2). (iii) Several well-known feature selection methods were applied and tested to determine the optimal feature subset for this prediction task. We anticipate that GRU4ACE will contribute to community-wide efforts for high-efficiency detection and assessment of peptides with ACE inhibition ability, accelerating the discovery of new potential antihypertensive drugs.

Although GRU4ACE significantly improved the identification of ACE inhibitory peptides, there are still some limitations. Firstly, to enhance the interpretability of GRU4ACE, we plan to apply a propensity score representation learning scheme (Charoenkwan, Chumnannuen, et al., 2023; Charoenkwan, Pipattanaboon, et al., 2023) to generate propensity scores of amino acids and dipeptides and integrate these scores into GRU4ACE. Secondly, we

aim to extract structural information by converting FASTA-formatted ACE inhibitory peptides into their corresponding SMILES format (Charoenkwan, Kongsompong, et al., 2023; Shoombuatong, Meewan, et al., 2024; Wei et al., 2021). Then, we will employ popular molecular encoding methods, such as PubChem, AP2D, CDK, KR, FP4C, and MACCS, to generate molecular fingerprints. Finally, we will develop an online web server for GRU4ACE to enable rapid identification and detailed profiling of ACE inhibitory peptides.

AUTHOR CONTRIBUTIONS

Saeed Ahmed: Methodology; software; conceptualization; validation; writing – original draft; visualization. **Nalini Schaduagrat:** Formal analysis; writing – original draft. **Pramote Chumnannuen:** Formal analysis; writing – original draft. **Watshara Shoombuatong:** Conceptualization; methodology; data curation; investigation; validation; formal analysis; supervision; funding acquisition; visualization; project administration; resources; writing – original draft; writing – review and editing.

ACKNOWLEDGMENTS

This project is supported by the National Research Council of Thailand and Mahidol University (N42A660380); Mahidol University Partnering Initiative under the MU-KMUTT Biomedical Engineering & Biomaterials Research Consortium.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no competing interests.

ORCID

Watshara Shoombuatong  <https://orcid.org/0000-0002-3394-8709>

REFERENCES

- Akbar S, Hayat M, Tahir M, Khan S, Alarfaj FK. cACP-DeepGram: classification of anticancer peptides via deep neural network and skip-gram-based word embedding model. *Artif Intell Med*. 2022;131:102349.

- Akbar S, Raza A, Zou Q. Deepstacked-AVPs: predicting antiviral peptides using tri-segment evolutionary profile and word embedding based multi-perspective features with deep stacking model. *BMC Bioinformatics*. 2024;25(1):102.
- Akbar S, Zou Q, Raza A, Alarfaj FK. iAFPs-Mv-BiTCN: predicting antifungal peptides using self-attention transformer embedding and transform evolutionary based multi-view features with bidirectional temporal convolutional networks. *Artif Intell Med*. 2024;151:102860.
- Amjad A, Ahmed S, Kabir M, Arif M, Alam T. A novel deep learning identifier for promoters and their strength using heterogeneous features. *Methods*. 2024;230:119–28.
- Arif M, Kabir M, Ahmed S, Khan A, Ge F, Khelifi A. DeepCPPred: a deep learning framework for the discrimination of cell-penetrating peptides and their uptake efficiencies. *IEEE/ACM Trans Comput Biol Bioinform*. 2021;19(5):2749–59.
- Arif R, Kanwal S, Ahmed S, Kabir M. A computational predictor for accurate identification of tumor homing peptides by integrating sequential and deep BiLSTM features. *Interdiscip Sci: Comput Life Sci*. 2024;16:1–16.
- Arshad F, Ahmed S, Amjad A, Kabir M. An explainable stacking-based approach for accelerating the prediction of antidiabetic peptides. *Anal Biochem*. 2024;691:115546.
- Azadpour M, McKay CM, Smith RL. Estimating confidence intervals for information transfer analysis of confusion matrices. *J Acoust Soc Am*. 2014;135(3):EL140–6.
- Basith S, Manavalan B, Hwan Shin T, Lee G. Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. *Med Res Rev*. 2020;40(4):1276–314.
- Boschin G, Scigliuolo GM, Resta D, Arnoldi A. ACE-inhibitory activity of enzymatic protein hydrolysates from lupin and other legumes. *Food Chem*. 2014;145:34–40.
- Bourel M, Segura AM, Crisci C, López G, Sampognaro L, Vidal V, et al. Machine learning methods for imbalanced data set for prediction of faecal contamination in beach waters. *Water Res*. 2021;202:117450.
- Charoenkwan P, Chumnanpuen P, Schaduagratt N, Oh C, Manavalan B, Shoombuatong W. PSRQSP: an effective approach for the interpretable prediction of quorum sensing peptide using propensity score representation learning. *Comput Biol Med*. 2023;158:106784.
- Charoenkwan P, Kongsompong S, Schaduagratt N, Chumnanpuen P, Shoombuatong W. TIPred: a novel stacked ensemble approach for the accelerated discovery of tyrosinase inhibitory peptides. *BMC Bioinformatics*. 2023;24(1):356.
- Charoenkwan P, Nantasenamat C, Hasan MM, Manavalan B, Shoombuatong W. BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. *Bioinformatics*. 2021;37(17):2556–62.
- Charoenkwan P, Pipattananaboon C, Nantasenamat C, Hasan MM, Moni MA, Shoombuatong W. PSRTTCA: a new approach for improving the prediction and characterization of tumor T cell antigens using propensity score representation learning. *Comput Biol Med*. 2023;152:106368.
- Charoenkwan P, Schaduagratt N, Moni MA, Shoombuatong W, Manavalan B. Computational prediction and interpretation of druggable proteins using a stacked ensemble-learning framework. *Iscience*. 2022;25(9):104883.
- Charoenkwan P, Schaduagratt N, Pham NT, Manavalan B, Shoombuatong W. Pretoria: an effective computational approach for accurate and high-throughput identification of CD8+ t-cell epitopes of eukaryotic pathogens. *Int J Biol Macromol*. 2023;238:124228.
- Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321–57.
- Chowdhury SY, Shatabda S, Dehzangi A. iDNAProt-ES: identification of DNA-binding proteins using evolutionary and structural features. *Sci Rep*. 2017;7(1):14938.
- Devlin J, Chang M-W, Lee K, Toutanova K. Pre-training of deep bidirectional transformers for language understanding. BERT; 2018.
- Ding Q, Sheikh AR, Chen Q, Hu Y, Sun N, Su X, et al. Understanding the mechanism for the structure-activity relationship of food-derived ACEI peptides. *Food Rev Intl*. 2023;39(4):1751–69.
- Du Z, Ding X, Hsu W, Munir A, Xu Y, Li Y. pLM4ACE: a protein language model based predictor for antihypertensive peptide screening. *Food Chem*. 2024;431:137162.
- Du Z, Ding X, Xu Y, Li Y. UniDL4BioPep: a universal deep learning architecture for binary classification in peptide bioactivity. *Brief Bioinform*. 2023;24(3):bbad135.
- Duan X, Dong Y, Zhang M, Li Z, Bu G, Chen F. Identification and molecular interactions of novel ACE inhibitory peptides from rapeseed protein. *Food Chem*. 2023;422:136085.
- Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: towards cracking the language of Life's code through self-supervised deep learning and high performance computing. *arXiv*. 2020; arXiv preprint arXiv:2007.06225, 2007.
- Gasparetto A, Marcuzzo M, Zangari A, Albarelli A. A survey on text classification algorithms: from text to predictions. *ACM Trans Intell Syst Technol*. 2022;13(2):83.
- Ghazikhani H, Butler G. Exploiting protein language models for the precise classification of ion channels and ion transporters. *Proteins Struct Funct Bioinf*. 2024;92(8):998–1055.
- Hasan MM, Schaduagratt N, Basith S, Lee G, Shoombuatong W, Manavalan B. HLPpred-fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. *Bioinformatics*. 2020;36(11):3350–6.
- Hassan MT, Tayara H, Chong KT. Meta-IL4: an ensemble learning approach for IL-4-inducing peptide prediction. *Methods*. 2023;217:49–56.
- Hayes T, Rao R, Akin H, et al. Simulating 500 million years of evolution with a language model. *bioRxiv*. 2024;2024.07.01.600583.
- He Z, Liu G, Qiao Z, Cao Y, Song M. Novel angiotensin-I converting enzyme inhibitory peptides isolated from rice wine lees: purification, characterization, and structure-activity relationship. *Front Nutr*. 2021;8:746113.
- Jiang J, Pei H, Li J, Li M, Zou Q, Lv Z. FEOpti-ACVP: identification of novel anti-coronavirus peptide sequences based on feature engineering and optimization. *Brief Bioinform*. 2024;25(2):bbae037.
- Jin Y, Yang Y. ProtPlat: an efficient pre-training platform for protein classification based on FastText. *BMC Bioinformatics*. 2022;23(1):66.
- Kabir M, Arif M, Ahmad S, Ali Z, Swati ZNK, Yu D-J. Intelligent computational method for discrimination of anticancer

- peptides by incorporating sequential and evolutionary profiles information. *Chemom Intell Lab Syst.* 2018;182:158–65.
- Kalyan G, Junghare V, Khan MF, Pal S, Bhattacharya S, Guha S, et al. Anti-hypertensive peptide predictor: a machine learning-empowered web server for prediction of food-derived peptides with potential angiotensin-converting enzyme-I inhibitory activity. *J Agric Food Chem.* 2021;69(49):14995–5004.
- Kanwal S, Arif R, Ahmed S, Kabir M. A novel stacking-based predictor for accurate prediction of antimicrobial peptides. *J Biomol Struct Dyn.* 2024;1–12.
- Kumar R, Chaudhary K, Sharma M, Nagpal G, Chauhan JS, Singh S, et al. AHTPDB: a comprehensive platform for analysis and presentation of antihypertensive peptides. *Nucleic Acids Res.* 2015;43(D1):D956–62.
- Kumar R, Chaudhary K, Singh Chauhan J, Nagpal G, Kumar R, Sharma M, et al. An in silico platform for predicting, screening and designing of antihypertensive peptides. *Sci Rep.* 2015;5(1):12512.
- Kurata H, Tsukiyama S, Manavalan B. iACVP: markedly enhanced identification of anti-coronavirus peptides using a dataset-specific word2vec model. *Brief Bioinform.* 2022;23(4):bbac265.
- Le Q, Mikolov T. Distributed representations of sentences and documents. *International conference on machine learning.* New York: PMLR; 2014. p. 1188–96.
- Lee SY, Hur SJ. Antihypertensive peptides from animal products, marine organisms, and plants. *Food Chem.* 2017;228:506–17.
- Lertampaiporn S, Hongsthong A, Wattanapornprom W, Thammarongtham C. Ensemble-AHTPpred: a robust ensemble machine learning model integrated with a new composite feature for identifying antihypertensive peptides. *Front Genet.* 2022;13:883766.
- Li G-H, Le G-W, Shi Y-H, Shrestha S. Angiotensin I-converting enzyme inhibitory peptides derived from food proteins and their physiological and pharmacological effects. *Nutr Res.* 2004;24(7):469–86.
- Liu W-Y, Zhang J-T, Miyakawa T, Li G-M, Gu R-Z, Tanokura M. Antioxidant properties and inhibition of angiotensin-converting enzyme by highly active peptides from wheat gluten. *Sci Rep.* 2021;11(1):5206.
- Mahmud SH, Goh KOM, Hosen MF, Nandi D, Shoombuatong W. Deep-WET: a deep learning-based approach for predicting DNA-binding proteins using word embedding techniques with weighted features. *Sci Rep.* 2024;14(1):2961.
- Mahmud SMH, Chen W, Liu Y, et al. PreDTIs: prediction of drug-target interactions based on multiple feature information using gradient boosting framework with data balancing and feature selection techniques. *Brief Bioinform.* 2021;22(5):bbab046.
- Manavalan B, Basith S, Shin TH, Wei L, Lee G. mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics.* 2019;35(16):2757–65.
- Minkiewicz P, Iwaniak A, Darewicz M. BIOPEP-UWM database of bioactive peptides: current opportunities. *Int J Mol Sci.* 2019;20(23):5978.
- Mu Z, Yu T, Liu X, Zheng H, Wei L, Liu J. FECS: a novel feature extraction model for protein sequences and its applications. *BMC Bioinformatics.* 2021;22:1–15.
- Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res.* 2020;21(140):1–67.
- Raza A, Uddin J, Almuhaimeed A, Akbar S, Zou Q, Ahmad A. AIPs-SnTCN: predicting anti-inflammatory peptides using fastText and transformer encoder-based hybrid word embedding with self-normalized temporal convolutional networks. *J Chem Inf Model.* 2023;63(21):6537–54.
- Raza A, Uddin J, Zou Q, Akbar S, Alghamdi W, Liu R. AIPs-DeepEnC-GA: predicting anti-inflammatory peptides using embedded evolutionary and sequential feature integration with genetic algorithm based deep ensemble model. *Chemom Intell Lab Syst.* 2024;254:105239.
- Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci.* 2021;118(15):e2016239118.
- Rizzello CG, Tagliazucchi D, Babini E, Rutella GS, Saa DLT, Gianotti A. Bioactive peptides from vegetable food matrices: research trends and novel biotechnologies for synthesis and recovery. *J Funct Foods.* 2016;27:549–69.
- Sarbon NM, Howell NK, Ahmad WAN. Angiotensin-I converting enzyme (ACE) inhibitory peptides from chicken skin gelatin hydrolysate and its antihypertensive effect in spontaneously hypertensive rats. *Int Food Res J.* 2019;26(3):903–911.
- Schaduangrat N, Nantasenamat C, Prachayasittikul V, Shoombuatong W. ACPred: a computational tool for the prediction and analysis of anticancer peptides. *Molecules.* 2019;24(10):1973.
- Shoombuatong W, Homdee N, Schaduangrat N, Chumnanpuen P. Leveraging a meta-learning approach to advance the accuracy of Nav blocking peptides prediction. *Sci Rep.* 2024;14(1):4463.
- Shoombuatong W, Meewan I, Mookdarsanit L, Schaduangrat N. Stack-HDAC3i: a high-precision identification of HDAC3 inhibitors by exploiting a stacked ensemble-learning framework. *Methods.* 2024;230:147–57.
- Sun S, Xu X, Sun X, Zhang X, Chen X, Xu N. Preparation and identification of ACE inhibitory peptides from the marine macroalga *Ulva intestinalis*. *Mar Drugs.* 2019;17(3):179.
- Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodology.* 1996;58(1):267–88.
- Umer M, Imtiaz Z, Ahmad M, Nappi M, Medaglia C, Choi GS, et al. Impact of convolutional neural network and FastText embedding on text classification. *Multimed Tools Appl.* 2023;82(4):5569–85.
- Villegas-Morcillo A, Gomez AM, Sanchez V. An analysis of protein language model embeddings for fold prediction. *Brief Bioinform.* 2022;23(3):bbac142.
- Wang J, Wang G, Zhang Y, Zhang R, Zhang Y. Novel angiotensin-converting enzyme inhibitory peptides identified from walnut glutelin-1 hydrolysates: molecular interaction, stability, and antihypertensive effects. *Nutrients.* 2021;14(1):151.
- Wei L, Ye X, Xue Y, Sakurai T, Wei L. ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. *Brief Bioinform.* 2021;22(5):bbab041.
- Win TS, Schaduangrat N, Prachayasittikul V, Nantasenamat C, Shoombuatong W. PAAP: a web server for predicting antihypertensive activity of peptides. *Future Med Chem.* 2018;10(15):1749–67.
- Wu C, Gao R, Zhang Y, De Marinis Y. PTPD: predicting therapeutic peptides by deep learning and word2vec. *BMC Bioinformatics.* 2019;20(1):1–8.

- Xu D, Wu Y, Cheng Z, Yang J, Ding Y. ACHP: a web server for predicting anti-cancer peptide and anti-hypertensive peptide. *Int J Pept Res Ther*. 2021;27(3):1933–44.
- Xue L, Yin R, Howell K, Zhang P. Activity and bioavailability of food protein-derived angiotensin-I-converting enzyme-inhibitory peptides. *Compr Rev Food Sci Food Saf*. 2021;20(2):1150–87.
- Yan Z, Ge F, Liu Y, Zhang Y, Li F, Song J, et al. TransEFVP: a two-stage approach for the prediction of human pathogenic variants based on protein sequence embedding fusion. *J Chem Inf Model*. 2024;64(4):1407–18.
- Zhou T, Rong J, Liu Y, Gong W, Li C. An ensemble approach to predict binding hotspots in protein–RNA interactions based on SMOTE data balancing and random grouping feature selection strategies. *Bioinformatics*. 2022;38(9):2452–8.
- Zhuang Y, Liu X, Zhong Y, Wu L. A deep ensemble predictor for identifying anti-hypertensive peptides using pretrained protein embedding. *IEEE/ACM Trans Comput Biol Bioinform*. 2021;19(4):1986–92.
- Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodology*. 2005a;67(2):301–20.

- Zou Q, Zeng J, Cao L, Ji R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing*. 2016;173:346–54.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ahmed S, Schaduangrat N, Chumnannuen P, Shoombuatong W. GRU4ACE: Enhancing ACE inhibitory peptide prediction by integrating gated recurrent unit with multi-source feature embeddings. *Protein Science*. 2025;34(6):e70026. <https://doi.org/10.1002/pro.70026>