Computers in Biology and Medicine 176 (2024) 108538

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

CAPTURE: Comprehensive anti-cancer peptide predictor with a unique amino acid sequence encoder

Hina Ghafoor [a,b,1], Muhammad Nabeel Asim [b,*,1], Muhammad Ali Ibrahim [a,b,1], Sheraz Ahmed [b], Andreas Dengel [a,b]

[a] Department of Computer Science, Rhineland-Palatinate Technical University of Kaiserslautern-Landau, Kaiserslautern, 67663, Germany
[b] German Research Center for Artificial Intelligence GmbH, Kaiserslautern, 67663, Germany

ARTICLE INFO

Dataset link: https://sds_genetic_analysis.opendfki.de/CAPTURE

Keywords:
Anti-cancer peptides
Machine learning
Binary classification
Multi-label classification
Anti-cancer peptides functional types
Anti-cancer peptides tissue types
Anti-microbial peptides
Sequence analysis
Sequence representation learning

ABSTRACT

Anticancer peptides (ACPs) key properties including bioactivity, high efficacy, low toxicity, and lack of drug resistance make them ideal candidates for cancer therapies. To deeply explore the potential of ACPs and accelerate development of cancer therapies, although 53 Artificial Intelligence supported computational predictors have been developed for ACPs and non ACPs classification but only one predictor has been developed for ACPs functional types annotations. Moreover, these predictors extract amino acids distribution patterns to transform peptides sequences into statistical vectors that are further fed to classifiers for discriminating peptides sequences and annotating peptides functional classes. Overall, these predictors remain fail in extracting diverse types of amino acids distribution patterns from peptide sequences. The paper in hand presents a unique CARE encoder that transforms peptides sequences into statistical vectors by extracting 4 different types of distribution patterns including correlation, distribution, composition, and transition. Across public benchmark dataset, proposed encoder potential is explored under two different evaluation settings namely; intrinsic and extrinsic. Extrinsic evaluation indicates that 12 different machine learning classifiers achieve superior performance with the proposed encoder as compared to 55 existing encoders. Furthermore, an intrinsic evaluation reveals that, unlike existing encoders, the proposed encoder generates more discriminative clusters for ACPs and non-ACPs classes. Across 8 public benchmark ACPs and non-ACPs classification datasets, proposed encoder and Adaboost classifier based CAPTURE predictor outperforms existing predictors with an average accuracy, recall and MCC score of 1%, 4%, and 2% respectively. In generalizeability evaluation case study, across 7 benchmark anti-microbial peptides classification datasets, CAPTURE surpasses existing predictors by an average AU-ROC of 2%. CAPTURE predictive pipeline along with label powerset method outperforms state-of-the-art ACPs functional types predictor by 5%, 5%, 5%, 6%, and 3% in terms of average accuracy, subset accuracy, precision, recall, and F1 respectively. CAPTURE web application is available at https://sds_genetic_analysis.opendfki.de/CAPTURE.

1. Introduction

From the period of 2000 to 2023, millions of people have died from just seven different types of cancers [1] including colorectal cancer, lung cancer, liver cancer, breast cancer, stomach cancer, skin cancer, and prostate cancer.[2] According to the World Health Organization (WHO) report, cancer is responsible for one out of every six deaths in the ongoing year [2]. Cancer induces uncontrolled cell growth and possesses the capability to swiftly spread to other parts of the body [3,4]. To mitigate the rapid spread of cancer, numerous drugs and therapies have been developed [3,4]. However, traditional treatments such as chemotherapy inadvertently target healthy cells along with fast-growing cancer cells [5]. Moreover, few treatment methods including radiation and surgery are painful and also cause adverse side effects such as cardiac toxicity, myelosuppression, and gastrointestinal damage [6]. Hence, there is an urgent demand for the development of alternative anti-cancer therapies that demonstrate enhanced effectiveness.

Anticancer peptides (ACPs) have opened new horizons for early detection and treatment of cancer [1]. ACPs are 5-to-50 amino acids [1] based molecules that target cancer cells through various mechanisms.

* Corresponding author.
E-mail address: Muhammad_Nabeel.Asim@dfki.de (M.N. Asim).
[1] These authors contributed equally to this work.
[2] https://www.who.int/news-room/fact-sheets/detail/cancer.

https://doi.org/10.1016/j.compbiomed.2024.108538
Received 8 January 2024; Received in revised form 26 April 2024; Accepted 28 April 2024
Available online 3 May 2024

ACPs can disrupt cellular membranes of cancer cells [7], deliver therapeutic drugs across physiological barriers [8], and induce programmed death of cancer cells [7]. ACPs prevent the growth of cancer cell by inhibiting angiogenesis through which new blood vessels are formed to supply key nutrients to cancer cells [7]. ACPs also have the ability to reverse Epithelial-Mesenchymal Transition process through which cancer cells gain invasive and migratory properties [7]. ACPs also have the potential to enhance body's natural immune response against cancer cells as they can block pro-inflammatory responses which assist in cancer progression [9]. Some cyclic nature ACPs such as ADH-1, apicidin, and chlamydocin have demonstrated potential therapeutic activity against cancer [10]. To accelerate and expedite research for unlocking more ACPs functionalities and therapeutic potential for cancer treatment, an accurate classification of ACPs followed by the categorization of their target function types is an active area of research [3,4,11].

Researchers have employed diverse kind of experimental approaches to identify ACPs and their functional types [12,13]. These approaches include phage display [14], high-throughput screening [13–15], and mass spectrometric analysis [16]. However, the identification of ACPs and their functional types through wet-lab experimental approaches proves time consuming, labor intensive and costly. The substantial investment of time as well as resources required for the identification of ACPs along with their functional types in wet-lab experimental approaches hinder the widespread research into discovery of potential candidates for anti-cancer therapies and further unlocking their other hidden potential.

Following the constraints of experimental methods and to enable widespread identification of anticancer peptides (ACPs) along with the annotation of their functional types, to date, 53 Artificial Intelligence (AI) based predictors have been developed. The working paradigm of these predictors can be divided into two different stages namely sequence encoding and classification. First stage makes use of sequence encoding methods that extract amino acids distributional information from peptides sequences and transform them into statistical vectors. At second stage, classifiers learn discriminative patterns from statistical vectors of training sequences and use this learning to accurately detect anti-cancer sequences during inference. The more informative and discriminative statistical vectors are the better the classifiers perform. Hence, even a simple machine learning classifier can achieve promising predictive performance with discriminative statistical vectors [17]. Whereas, a sophisticated classifier is bound to lack predictive performance on account of less discriminative statistical vectors of peptide sequences.

A brief summary of existing 53 predictors in terms of encoding methods and classifiers is given in Supplementary Table 1. A close analysis of Supplementary Table-1 reveals that among 53 predictors, only 11 predictors make use of standalone sequence encoding methods while 42 predictors reaped the benefits of multiple sequence encoding methods. Main motivation behind the utilization of multiple sequence encoding method was to fed the classifiers with statistical vectors having 4 different types of amino acids information namely correlational, distribution, compositional, and transitional. However, statistical vectors generated by the integration of multiple sequence encoding methods contain some irrelevant and redundant features that hamper the classifiers performance. To remove such features, in 16 different predictors, researchers have used 10 feature selection methods of 3 different types (filter, wrapper, embedded). However, it is difficult to design a generic predictor by integrating feature selection methods in predictive pipeline.

At classification stage, among 53 predictors, 32 predictors have utilized traditional machine learning classifiers and 21 predictors have utilized CNN, RNN, LSTM and Dense-Net architectures based predictors. Prime reason behind the pre-dominant utilization of traditional classifiers is the availability of limited annotated data. While developing computational predictors, main focus of researchers was to generate more discriminative statistical vectors of peptides sequences that can help the classifiers to learn discriminative patterns and accurately classify peptides into ACP and non-ACP classes. However, statistical vectors of existing sequence encoders lack discriminative patterns which is why even sophisticated classifiers fail to accurately discriminate ACPs from non-ACPs, indicating a lot of room for the development of new predictors. Furthermore, for functional types annotations, there exist only one predictor [11]. Following the need of a robust and precise computational predictor for ACPs classification and their functional types annotation, contributions of this manuscript are manifold:

(I) It presents a powerful sequence encoder CARE that transforms peptides sequences into statistical vectors by extracting amino acids four different types of information including correlation, distribution, composition, and transition, **(II)** It compares proposed encoder performance with 55 existing encoders performance under two different evaluation settings namely intrinsic and extrinsic. An intrinsic evaluation objective is to determine which sequence encoder captures amino acids discriminative distribution in ACPs and non-ACPs sequences and generate highly non-overlapping clusters for both classes. Whereas, an extrinsic evaluation objective is to assess and compare the predictive performance of 12 different machine learning classifiers by feeding them with statistical vectors generated through proposed and 55 existing sequence encoders **(III)** In order to demonstrate the predictive power and generalizeability of proposed encoder and Adaboost classifier based CAPTURE predictor, its performance is compared with most recent 36 ACPs and Non-ACPs classification predictors **(IV)** It compares proposed CARE encoder, Label Powerset, and AdaBoost Classifier based Predictive Pipeline Performance with State-of-the-art ACPs functional types Predictor **(V)** To speed up the process of discovering new ACPs along with their functional types, a web application is developed (https://sds_genetic_analysis.opendfki.de/CAPTURE).

Materials and methods

This section provides a concise overview of the 3 different modules of proposed predictor CAPTURE, graphical representation of which is provided in Fig. 1. In Fig. 1, first module describes benchmark ACPs datasets used for experimentation, necessary details of which are given in Section 1.4. Second module demonstrates proposed peptide sequence encoder, necessary details of which are facilitated in Section 1.1. Third module describes classifiers and evaluation measures, necessary details of which are given in Sections 1.3 and 1.5 respectively.

1.1. Proposed sequence encoder

Proposed sequence encoder Comprehensive Amino Acid Relations Explorer (CARE) is an extension of Quasi Sequence Order (QSO) [18]. QSO encoder transforms raw peptides sequences into statistical vectors by capturing two different types of information namely amino acids correlation and distribution. A comprehensive set of discriminative patterns that enable the classifier to more precisely discriminative anti-cancer peptides from non anti-cancer peptides, cannot be obtained by using only correlational and distributional information of amino acids. To generate highly discriminative statistical vectors offering comprehensive patterns that more often exist in anti-cancer peptides sequences or in non anti-cancer peptide sequences, we extend QSO encoder to extract 4 different types of information including correlational, distributional, compositional and transitional.

Correlational information is captured on the basis of physico-chemical properties based distance between amino acids. To compute the distance between amino acids, we use two pre-computed matrices in which values of four different physico-chemical properties including hydrophilicity, hydrophobicity, side chain volume, and polarity are averaged on the basis of Manhattan distance. Distributional information is captured by utilizing 20 unique amino acids occurrence frequencies along with sequence correlational information. Compositional information is acquired by using 20 unique amino acids consecutive two or

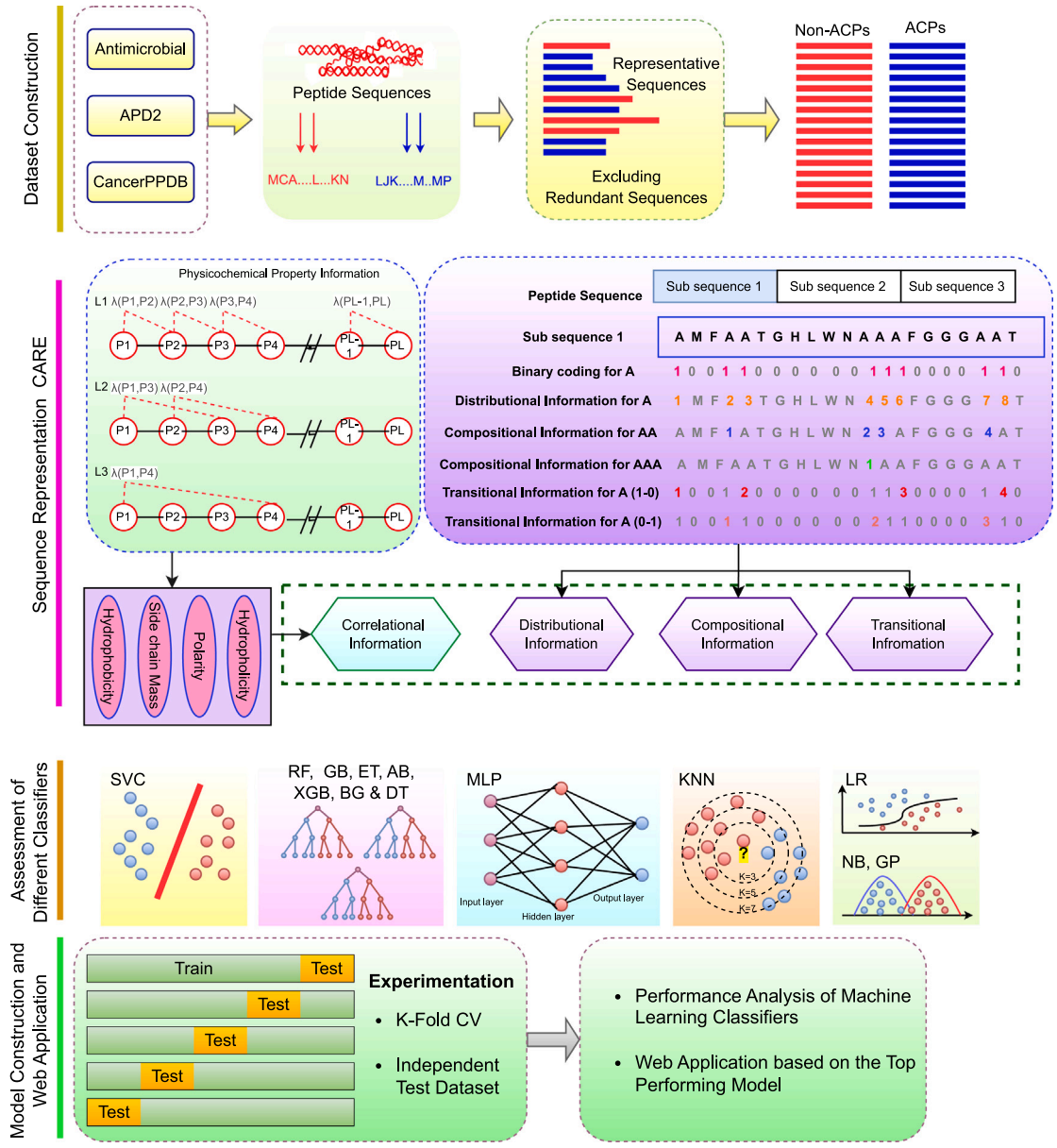


Fig. 1. Three fundamental modules of proposed anti-cancer peptide prediction pipeline, (a) Experimental datasets collection, (B) Feature representation using novel encoder CARE. (C) Assessing the performance of Adaboost classifier under the hood of 10-fold cross validation and multiple independent test sets.

three times occurrence frequencies along with sequence correlational information. Transitional information is captured by computing 20 amino acids number of transitions from one to another amino acid along with sequence correlational information. Furthermore, to more deeply extract all 4 types of information, we divide peptide sequences into same length sub-sequences, extract all 4 types of information from each sub-sequence, and then concatenate extracted information to represent peptide sequences into statistical vectors.

To more precisely understand the working paradigm of proposed encoder, consider a generic sequence $S = AA_1, AA_2, \ldots \ldots AA_n$ where AA_i represents a particular amino acid from 20 unique amino acids. First of all, to compute correlational information, proposed CARE encoder generates sub-sequences of given sequence and represents each sub-sequence in terms of bi-mers separated by different Lag values. Here, number of sub-sequences and Lag values are two hyperparameters of proposed encoder. We tweak number of sub-sequence value from 1 to 5. Here sub-sequence value 1 means that proposed encoder takes complete peptide sequence and extracts all 4 types of information. Whereas, sub-sequence value 2 means that proposed encoder segregates the peptide sequence into 2 equal length sub-sequences and extracts all 4 types of information from both sequences separately. Sub-sequence value 3 means that proposed encoder divides the peptide sequence into 3 equal length sub-sequences and extracts all 4 types of information from three sub-sequences separately. Similarly, all 4 types of information is extracted from sub-sequences generated by sub-sequence value 4 and 5 respectively. As proposed encoder captures correlation information at sub-sequence level, so it manages to capture importance of different regions of peptide sequence.

With an aim to capture different level of details, we tweak the Lag value from 1 to 5 for each sub-sequence value. For instance, for sub-sequence value 1, tweaking the Lag value from 1 to 5 produces five different chains of bi-mers, showing bi-mers at different distances

within the sequence. This leads to produce five different sets of unique bi-mers, shown in Eq. (1).

$$\begin{cases} AA_1AA_2, AA_2AA_3, AA_3AA_4, \ldots\ldots AA_{L-1}AA_L \text{ with Lag 1} \\ AA_1AA_3, AA_2AA_4, AA_3AA_5, \ldots\ldots AA_{L-2}AA_L \text{ with Lag 2} \\ AA_1AA_4, AA_2AA_5, AA_3AA_6, \ldots\ldots AA_{L-3}AA_L \text{ with Lag 3} \\ AA_1AA_5, AA_2AA_6, AA_3AA_7, \ldots\ldots AA_{L-4}AA_L \text{ with Lag 4} \\ AA_1AA_6, AA_2AA_7, AA_3AA_8, \ldots\ldots AA_{L-5}AA_L \text{ with Lag 5} \end{cases} \tag{1}$$

Then for each set of bi-mers, coupling factor which aims to capture the distance between two amino acids is computed for every bi-mer in order to extract correlation information. Following Chou et al. [18], we have used two pre-computed amino acid distance matrices (dimensions $\rightarrow 20 \times 20 = 400$) provided by Schneider et al. [19] and Grantham et al. [20]. In these matrices, values of four different physico-chemical properties including hydrophilicity, hydrophobicity, side chain volume, and polarity are averaged on the basis of Manhattan distance. The computation of coupling factor for every bi-mer_k based on two amino acids AA_k, AA_j using Schneider and Grantham's amino acid distance matrices can be mathematically expressed as:

$$\begin{gathered} Coupling\ \ Factor[bimer_k] = Distance\ \ Matrix_i\ (AA_k, AA_j)^2 \\ Distance\ \ Matrix_i \in \{Schneider, Grantham\} \end{gathered} \tag{2}$$

Afterward, for each bi-mers sequence, bi-mers correlational values are summed up and divided by the length of the sequence to get normalized correlation information. As correlational values are acquired from two different amino acid distance matrices, hence, five normalized correlation values are obtained for five bi-mers sequences using Schneider et al. [19] matrix and five normalized correlation values are obtained using Grantham et al. [20] matrix using Eq. (3).

$$\text{Normalized Correlation}[Distance\ Matrix_i][Lag_k] = \frac{\sum_{m=1}^{Sequence\ Length-1}(Coupling\ Factor[bimer]_k)}{Sequence\ Length - 1}. \tag{3}$$

Here, Lag_k denotes specific chain of bi-mers having certain sequence length. Afterward, five Schneider and five Grantham matrix based floating point values are separately summed up to get two different values using Eq. (4). In this way, we capture the overall correlation of bi-mers at five distinct Lag values using two different average values of 4 unique physico-chemical properties.

$$\text{Overall Correlation}[Distance\ Matrix_i] = \sum_{k=1}^{5} \text{Normalized Correlation}[Distance\ Matrix_i][Lag_k]. \tag{4}$$

Afterward, these estimated normalized correlation values are optimized by taking a weight factor of 0.25 using Eq. (5).

$$\text{Optimized Normalized Correlation}[Distance\ Matrix_i][Lag_k] = \frac{weight \times \text{Normalized Correlation}[Distance\ Matrix_i][Lag_k]}{1 + weight \times \text{Overall Correlation}[Distance\ Matrix_i]}. \tag{5}$$

We integrate correlational information with distributional information of amino acids. To achieve this, in second step, we count the occurrence frequency of every amino acid inside a sequence using Eq. (6) and normalized the resulting value with sequence overall correlation values computed using Eq. (4).

$$\text{Distribution}[AA_j] = \frac{AA_j \text{ occurrence frequency in sequence}}{weight \times \text{Overall Correlation}[Distance\ Matrix_i] + 1}. \tag{6}$$

In third step, proposed CARE encoder captures compositional information by considering consecutive two times as well as three times occurrence frequencies of amino acids within sequences using Eqs. (7) and (8). To illustrate better, Fig. 1 describes the process of estimating composition information of amino acids. More specifically, consecutive two times occurrence frequency of amino acid A in a given hypothetical sequence is 4 and consecutive three times occurrence frequency of amino acid A is 1. Proposed CARE encoder incorporates composition information of amino acids present in given sequences using a 40-dimensional vector, where initial 20-dimensions denote consecutive two times occurrence frequency and other 20-dimensions represent the consecutive three times occurrence frequency of a distinct amino acid.

$$\text{Composition}[AA_j^2] = \frac{\text{Consecutive Two Times } AA_j \text{ occurrence}}{weight \times \text{Overall Correlation}[Distance\ Matrix_i] + 1}. \tag{7}$$

$$\text{Composition}[AA_j^3] = \frac{\text{Consecutive Three Times } AA_j \text{ occurrence}}{weight \times \text{Overall Correlation}[Distance\ Matrix_i] + 1}. \tag{8}$$

In fourth step, CARE encodes transition information by characterizing the 20 unique amino acids shift from one to another amino acid within sequences using Eqs. (9) and (10). To illustrate better, Fig. 1 describes the process of estimating transition information of a specific amino acid A with respect to 19 other amino acids. More specifically, count of amino acid A for the case where other 19 amino acids are coming after this amino acid is shown by 1-0 transition. Whereas, count of amino acid A for the case where 19 distinct amino acids are coming before this amino acid is shown by 0–1 transition.

$$\text{Transition}[!AA_j][AA_j] = \frac{0-1 \text{ Transition Count}}{weight \times \text{Overall Correlation}[Distance\ Matrix_i] + 1}. \tag{9}$$

$$\text{Transition}[AA_j][!AA_j] = \frac{1-0 \text{ Transition Count}}{weight \times \text{Overall Correlation}[Distance\ Matrix_i] + 1}. \tag{10}$$

Using Eqs. (9) and (10), proposed CARE encoder incorporates transition information of a amino acid present in given sequences using a 40-dimensional vector. In this vector, 20-dimensions denote the 1-0 transition information and other 20-dimensions represent the 0–1 transitions of a distinct amino acid.

In fifth step, all four different kinds of amino acid features are concatenated, which can be mathematically expressed using Eq. (11).

$$\text{CARE} = \begin{cases} Optimized\ Normalized \\ Correlation[Distance Matrix_i][Lag_k] \oplus \\ \text{Distribution}[AA_j] \oplus \\ \text{Composition}[AA_j^2] \oplus \\ \text{Composition}[AA_j^3] \oplus \\ \text{Transition}[!AA_j][AA_j] \oplus \\ \text{Transition}[AA_j][!AA_j] \end{cases} \tag{11}$$

With an aim to extract more comprehensive information about amino acids distribution, composition, and transition, we divide peptide sequences into same length sub-sequences denoted by l. To better illustrate, let us consider an imaginary peptide sequence $AA_1, AA_2, AA_3, \ldots, AA_n$ and $l = 3$. As shown in Fig. 1, given peptide sequence will be segregated into 3 equal length sub-sequences, where statistical vectors based on distribution, composition, and transition information will be generated for each sub-sequence separately. In this way, instead of getting 20-dimensional distributional information vector, 40-dimensional composition, and 40-dimensional transition vectors, $20 \times l$ dimensional distributional information vector and $40 \times l$ compositional and transitional information vectors will be generated. These statistical vectors are concatenated to generate final statistical vectors. These statistical vectors will have the dimension of $[(20 \times l) \times 5 + lag] \times n$, where 20 denotes unique amino acids, 'l' represents number of sub-sequences, 5 denotes number of ways through which different features are captured, and n represents the number of physico-chemical properties.

1.2. An overview of existing encoders

According to our best of knowledge, 55 unique protein sequence encoders have been developed that transform amino acid sequences into statistical vectors. These sequence encoders can be categorized into 14 different types based on the information they capture, such as amino acids distribution, gap based amino acid distribution, amino acids groups distribution, autocorrelation, co-variance, local–global context-aware, sequence order, binary, physico-chemical properties, traditional networks, pre-trained deep neural network, optimize physico-chemical properties, substitution matrix, and Fourier transformation based encoders.

Amino acid distribution encoders such as Kmer [21,22], DPC [21, 22], TPC [21,22], ANF [23], EAAC [24,25], EGAAC, DDE [22], are type of protein sequence encoders that capture the frequency or proportion of each individual or group of amino acids called k-mers in a protein sequence. This type of encoder reflects the overall composition of a single amino acid or k-mers in a protein sequence, providing information about the relative abundance or scarcity of specific amino acids or k-mers. Gap based amino acid distribution encoders such as CKSAPP [26,27], Adaptive skip Dipeptide composition (ASDC) [28], and CKSAAGP [25,29] segregate the protein sequences into bi-mers with distinct gap values and capture the distribution of unique bi-mers. The gap value determines the distance between the amino acids that are considered as a pair, impacting the local context-aware representation of the protein sequence. Different gap values can provide different insights into the arrangement and distribution of amino acids pairs in the protein sequence. A smaller gap value focuses on capturing short-range interactions, while a larger gap value allows the encoders to capture long-range associations within the sequence.

Amino acid group distribution encoders such as CTDC [30–34], CTDD [30–34], CTDT [30–34], GAAC [25,29], GDPC [25,29], GTPC [25,29], KSCTriad [25,29], CTriad [35] categorize the amino acids into different groups based on specific physico-chemical properties such as hydrophobicity, charge, or polarity. These encoders capture the distribution of amino acid groups in the protein sequence, providing insight into the overall physicochemical properties of the sequence. Autocorrelation encoders such as Geary [36], Moran [37,38], NMBroto [39] capture the relatedness between amino acids or k-mers in a protein sequence. These encoders compute the correlation coefficients between different amino acids or k-mers based on their physico-chemical properties such as hydrophobicity or charge. These encoders provide information on the pairwise interactions and dependencies of amino acids within the protein sequence, allowing for the identification of specific functional motifs. Covariance encoders such as auto-covariance [40–42], auto-crosscovariance [40–42], bi-autocovariance [40–42] measure the joint variability of two amino acids or k-mers in a protein sequence.

Covariance encoders provide information about how two amino acids or k-mers vary together. If the covariance is positive, it means that when one amino acid or k-mer tends to be above its mean, the other amino acid or k-mer is also likely to be above its mean. On the contrary, a negative covariance indicates that when one amino acid or k-mer is above its mean, the other one is likely to be below its mean. Unlike correlation encoders, which measure the strength and direction of a relationship between two amino acids or k-mers, covariance encoders do not indicate the strength of the relationship, only the direction of the relationship between the two amino acids or k-mers.

Local-Global context aware protein encoders such as WSRC-local [17], WDRC-global [17], WSRC-local–global [17], consider composition and transition of amino acids, proving key information about distribution as well as changes of amino acids in different segments of protein sequences. Sequence order category encoders such as PAAC [43], APAAC [44], QSOrder [45–47], SOCNumber [45–47], consider the distribution as well as order or arrangement of amino acids in a protein sequence on the basis of different distances. Different distances encompass different levels of local or global interactions between specifically arranged amino acids. Binary encoders [28,48–53] working paradigm typically involve converting the amino acid sequences into statistical vectors having 0s and 1s.

Physico-chemical properties and network based encoders such as AAIndex [54] and AESNN3 [48,49] respectively substitute amino acids with pre-computed numerical values. Optimized physico-chemical properties based encoder such as ZScale [55] utilizes physicochemical properties to characterize amino acids. It makes use of different strategies like principal component analysis (PCA), Partial least squares (PLS), and Multiple Linear regression to eliminate less informative properties and retain only highly informative properties of amino acids. Traditional network based encoders such as complex network, enhanced complex-network use network-based approaches to characterize protein sequences by representing them as graphs or networks, where nodes represent amino acids and edges represent interactions or relationships between them.

The functional paradigm of Fourier transformation-based sequence encoders, such as MappingClass-eiip-fourier and MappingClass-integer-fourier, involves the utilization of Electron-Ion Interaction Potential values or integer values to replace individual amino acids. By applying Fourier transformation, these encoders aim to enhance the encoding of hidden patterns and trends, including frequent components, in protein sequences. On the other hand, substitution matrix-based encoders like BLOSUM62 [56] generate matrices that assign scores to amino acid substitutions based on their observed frequencies in related protein sequences. This scoring system provides a measure of the similarity between different amino acids. Higher scores are assigned to more similar amino acid substitutions, indicating a greater likelihood of their occurrence in related proteins.

1.3. Machine learning classifiers

At classification stage, we utilize traditional machine learning classifiers to design two different predictive pipelines namely binary classification and multi-label classification. In binary classification predictive pipeline for ACPs and non-ACPs sequences classification, we assess the performance impact of proposed sequence encoder CARE using 12 most widely used machine learning classifiers including Naive Bayes (NB) [57], Gaussian Process (GP) [58], Extra-Tree (ET) [59], Random Forest (RF) [59], Bagging (BG) classifier [59], Decision-Tree (DT) [59], Adaboost (AB) [59], Gradient Boost (GB) [59], Extreme Gradient Boost (XGB) [59], Logistic Regression (LR) [60], Support Vector Machine (SVM) [61], and K-Nearest Neighbour (KNN) [62].

On the other hand, in multi-label classification predictive pipeline, for ACPs functional types annotation, we utilize proposed encoder along with label powerset method and Adaboost classifier. As binary classifiers cannot handle multi-label peptide sequences, hence Label Powerset [63] serves as a data transformation approach that treats each unique combination of functional types as a separate class. Discriminative statistical vectors of raw sequences are generated through CARE encoder, multi-label sequences are transformed to unique class sequences through label powerset, and functional types annotation is performed using transformed data and AdaBoost classifier.

1.4. Benchmark datasets

Selection of comprehensive datasets is an important task for the development and evaluation of a computational predictor. Supplementary Table-2 illustrates the details of unique datasets across which existing predictors are evaluated. Overall 40 different datasets have been developed, and several datasets are quite similar as they vary in terms of only few sequences. A brief description about these datasets is given in Supplementary file. Overall, datasets belong to two different categories namely extracted and derived. In extracted category,

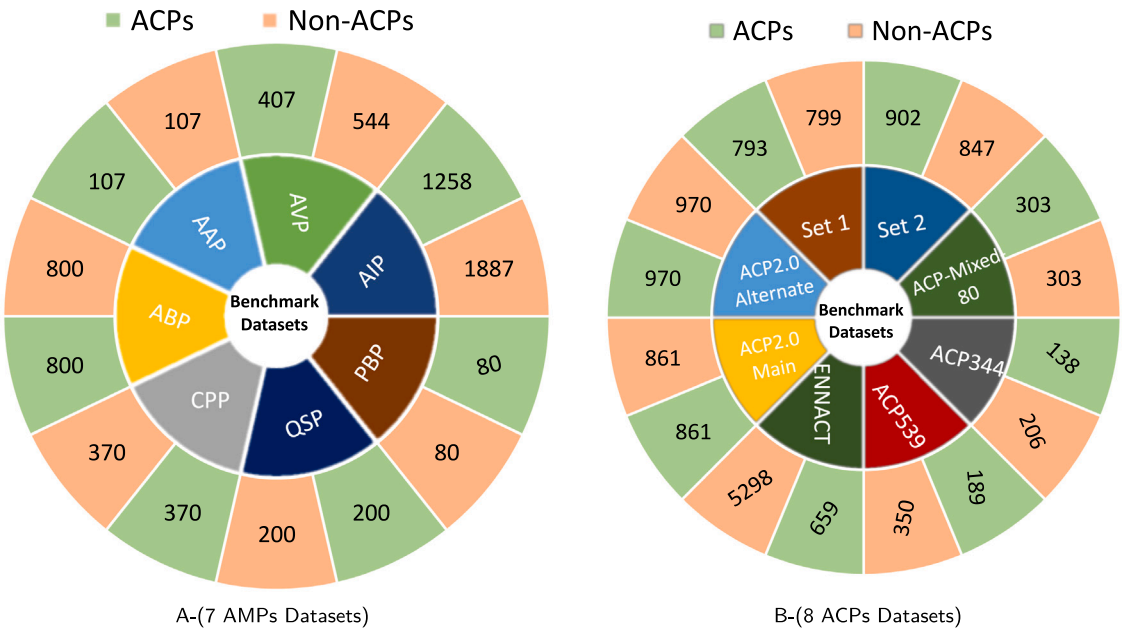


Fig. 2. Statistics of 15 benchmark binary ACPs and AMPs datasets where green color denotes positive sequences and orange color denotes negative sequences.

datasets are developed by acquiring sequences from public databases, while in derived category, datasets are developed by combining existing datasets. A thorough analysis of Supplementary Table-2 indicates that from 2020-to-2023, 27 ACPs predictors are evaluated across different combinations of following datasets: ACP-Mixed-80 [11,64], $ACP2.0\_Main$, $ACP2.0\_Alterante$ [65], Set-1 and Set-2 [66], ENNACT core [67], ACP539 [68], and ACP344 [69]. From these 8 datasets, 5 datasets including ACP-Mixed-80 [11,64], $ACP2.0\_Main$, $ACP2.0\_Alterante$ [65], Set-1 and Set-2 [66] belong to derived datasets category. These 5 derived datasets have comprehensively solved the issues of short length sequences, repeated sequences, and annotation conflicts among different datasets. The remaining 3 datasets including ENNACT core [67], ACP539 [68], and ACP344 [69] belong to a category where datasets are developed by acquiring sequences from databases. These datasets are reliable and authoritative as they do not contain redundant sequences and annotation conflicts.

In the same time span, 17 predictors are evaluated across different combinations of following datasets: ACP240, ACP740 [70], Tyagi et al. [71] Main and Alternate, Vijayakumar [72] ACPMain and ACPIndependent. ACP240 and ACP740 datasets contain annotation conflicts [64], other datasets such as Tyagi et al. [71] Main and Alternate datasets, Vijayakumar [72] ACPMain and ACPIndependent are already well accommodated by 5 aforementioned derived datasets, so they do not need to be used separately. Considering, experimentation criteria of most recent studies [11], we have used 8 highly reliable and most widely used benchmark ACPs datasets to comprehensively evaluate the potential of proposed CAPTURE predictor.

The prime reason of using these 8 benchmark datasets is manifold. First, it covers almost all the datasets used in the literature so far and used by 27 predictors since 2020. Second, as mixed datasets such as $ACP2.0\_Main$, $ACP2.0\_Alterante$ [65], ACP-Mixed-80 [11,64], Set-1 and Set-2 [66] were prepared by the comprehensive processing of more than 20 different datasets, hence, these datasets contain decent number of correct peptide sequences. Third, these datasets have correct length (less than 50 amino acids) peptide sequences and contain no sequences of very huge lengths that may become outliers for machine learning classifier. Fourth, these datasets have no redundancy at all which is crucial to avoid homology bias as well high similarity between sequences. Fifth, these datasets have been developed using a balanced CD-HIT similarity threshold which assist to prevent over-estimation of classifier performance. Statistics of all 8 ACPs datasets are described in Fig. 2-B.

Apart from type specific anti-cancer peptides classification datasets, we have used 7 different anti-microbial peptides (AMPs) datasets including PBP, QSP, CPP, ABP, AAP, AVP, and AIP [73–79] to assess the generalizeability of proposed CAPTURE predictor for identifying anti-cancer charateristics from generic AMPs sequences. Statistics of all 7 AMPs datasets are described in Fig. 2-A.

To facilitate the development of anti-cancer functional types predictors, there exist only one public benchmark dataset developed by Deng et al. [11]. Authors prepared this dataset by acquiring anti cancers peptides sequences and their associated functional types annotations from CancerPPD database [80]. The label space of this dataset involves 7 functional types including blood, breast, colon, cervix, lung, skin, and prostate. Statistics and sequence-to-functional type distribution of ACP-Functional dataset are shown in Fig. 3.

1.5. Evaluation metrics

Following the evaluation criterion of existing ACPs classification predictors [11,65,68], proposed CAPTURE predictor performance is assessed using 7 distinct evaluation metrics namely Accuracy (ACC), Precision (PRE), Recall (REC) or Sensitivity (SEN), F1-score, Specificity (SPE), Matthews correlation coefficient (MCC), and area under the receiver operating characteristic (AU-ROC). These evaluation metrics are discussed extensively in literature [11,65,68], so here we only provide their mathematical expressions.

$$f(x)=\begin{cases}\text{Accuracy (ACC)} = (T_P+T_N)/(T_P+T_N+F_P+F_N)\\ \text{Precision (PRE)} = T_P/(T_P+F_P)\\ \text{Recall or Sensitivity (SEN)} = T_P/(T_P+F_N)\\ \text{Specificity (SPE)} = T_N/(T_N+F_P)\\ \text{F1-Score} = 2\times\dfrac{PRE\times REC}{PRE+REC}\\ \text{MCC} = T_P\times T_N - F_P\times F_N/Q\\ \text{Q} = \sqrt{(T_P+F_N)(T_P+F_P)(T_N+F_P)(T_N+F_N)}\end{cases} \tag{12}$$

Here, TN and TP denote the number of accurately predicted non-ACPs and ACPs respectively. On the other hand, FN and FP denote the number of in-correctly predicted non-ACPs and ACPs respectively. It is

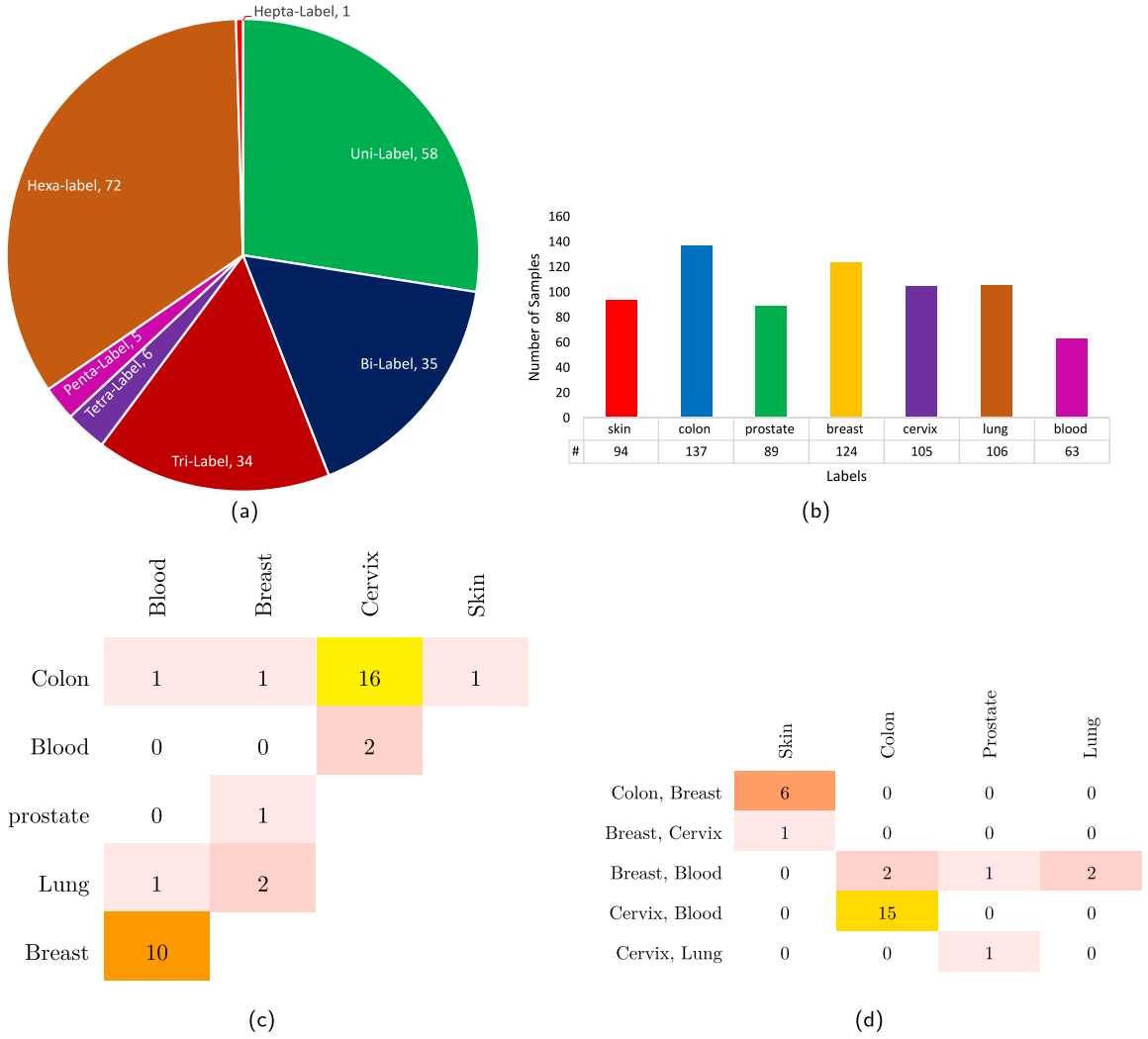


Fig. 3. (a) Descriptive statistic of benchmark acpfunctional dataset with division of sequences in terms of label cardinality (b) Count of sequences in each functional type (c) Dense bi-functional type confusion matrix (d) Dense tri-functional type confusion matrix.

important to mention that the higher the values of these 7 evaluation metrics are, the greater the classifier performance is.

ACPs functional types annotation is a multi-label classification task. ACPs sequences may belong to multiple functional types at the same time and multi-label classification predictor cannot be evaluated through binary evaluation measures. Following the evaluation criterion of existing ACPs functional types annotation predictor [11], we use both example based and label based evaluation metrics. The example based evaluation metrics such as accuracy, subset accuracy, precision, F1-score, recall, and Hamming loss are first estimated for every sequence and later averaged to get the final performance value.

$$\begin{cases} \text{Accuracy} = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \\ \text{Precision} = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \cap Z_i|}{|Z_i|} \\ \text{Recall} = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \cap Z_i|}{|Y_i|} \\ \text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\ \text{Hamming loss} = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \Delta Z_i|}{|L|} \\ \text{Subset accuracy} = \frac{1}{|D|}\sum_{i=1}^{|D|}[\![Y_i = Z_i]\!] \end{cases} \tag{13}$$

Here, $|D|$ denotes the number of sequences present in the dataset and L denotes number of functional types. Y_i denotes set of predicted functional types for sequence i, and Z_i denotes true set of functional types for sequence i. The mathematical operator $\cup$ and $\cap$ denote the union and intersection of two sets respectively. The Δ symbol represents the symmetric difference among predicted and actual functional types. If Y_i and Z_i are same then [Y_i == Z_i] = 1, otherwise 0.

On the other hand, label based evaluation metrics are estimated for every functional type instead of every sequence. It mainly includes two different methods in order to aggregate the values of the functional types: macro-average, and micro-average. While the macro-average approach computes each evaluation metric independently for every functional type and later perform the average to obtain the final performance value. Micro-average approach first estimates true negatives, true positives, false negatives, and false positives for all functional types and later computes all evaluation metrics.

$$\begin{cases} \text{MacroMetric} = \frac{1}{|L|}\sum_{l=1}^{|L|}\text{evalMetric}(TP_l, FP_l, TN_l, FN_l) \\ \text{MicroMetric} = \text{evalMetric}\begin{pmatrix} \sum_{l=1}^{|L|}\text{TP}_l, \sum_{l=1}^{|L|}\text{FP}_l, \\ \sum_{l=1}^{|L|}\text{TN}_l, \sum_{l=1}^{|L|}\text{FN}_l \end{pmatrix} \end{cases} \tag{14}$$

Here, evalMetric includes accuracy, recall, precision, and F1-score.

2. Experimental setup

Proposed novel sequence encoder CARE is implemented using Python programming language. The existing 55 encoders implementations are taken from iLearnPlus [81]. All 12 distinct classifiers implementations are taken from Scikit-Learn [82] library. Proposed CAPTURE predictor web server is implemented using Django framework [83]. We have optimized different hyperparameters of encoders and 12 different classifiers. To find optimal values of different hyperparameters, we have splitted the datasets into training and test sets. Using training data, we search the optimal values of hyperparameters from range of values through Grid search. Using optimal value of hyperparameters, we compute the results on test sets of all datasets. To ensure the reproduceability of the results, proposed CARE encoder and AdaBoost classifier initial and optimal hyperparameters values across all 16 datasets are provided in Supplementary Table-3.

3. Results

This section performs an extensive performance comparison of proposed CARE encoder with 55 existing encoders using 12 different machine learning classifiers. It also assess the potential of proposed and top performing existing encoders for generating highly non-overlapping clusters for distinct classes. It performs a detailed performance comparison of proposed CAPTURE predictor with 37 existing ACPs and non-ACPs classification predictors across 8 benchmark datasets. It evaluates the generalizeability of proposed CAPTURE predictor for identifying anti-cancer characteristics from generic AMPs sequences, and compares it with 3 existing predictors. It evaluates and compares the performance potential of proposed CAPTURE predictor for functional types annotation classification with state-of-the-art predictor.

3.1. Extrinsic performance comparison of the proposed peptide sequence encoder with existing sequence encoders

In order to truly illustrate the effectiveness of proposed CARE encoder, we compare the performance of proposed CARE encoder with 55 existing encoders of 14 different categories using benchmark ACP-Mixed-80 dataset and 12 different machine learning classifiers. From 55 existing encoders, 7 encoders belong to amino acids distribution, 4 belong to gap based amino acid distribution, 8 belong to amino acids groups distribution, 3 belong to correlation, 3 belong to co-variance, 3 belong to local–global context-aware, 5 belong to sequence order, 15 belong to binary, 1 belong to physico-chemical properties, 1 belong to optimized physico-chemical properties, 3 belong to network, 1 belong to substitution matrix, and 2 belong to Fourier transformation based paradigms. Results of proposed and all existing encoders in terms of different evaluation metrics across 12 different classifiers on benchmark ACP-Mixed-80 dataset are given in Supplementary Table-4.

A thorough performance analysis of Supplementary Table-4 indicates that from amino acids distribution, gap based amino acid distribution, and amino acids groups distribution encoders, Kmer, Adaptive skip Dipeptide composition, and CTDC achieve better performance respectively. From correlation and co-variance, NMBroto and auto-covariance mark top performance, whereas, from local–global context-aware, sequence order, and binary encoders, WSRC-local–global, QSOrder, OPF-10 bit achieve top performance. From physico-chemical properties based encoders, AAINDEX produces better performance than ZScale. From network, and Fourier based encoders, AESNN3, complex-network, and MappingClass-integer-fourier achieve best performance respectively. To assist the readers, from 55 encoders of 14 different categories, here in Fig. 4, we have picked top performing encoders from each category and compared their accuracy values with the performance of proposed CARE encoder.

It is evident in Fig. 4 from all top performing existing encoders, QSOrder encoder achieves the top average accuracy of 75% due to its aptitude to capture sequential and distributional information. Whereas, from 14 top performing existing encoders, co-variance based encoder auto-covariance marks lowest performance across most classifiers. Co-variance based encoders only capture linear associations and neglect non-linear relationships like spatial arrangement of the amino acids, that are important for comprehending how relative positions of specific amino acids largely contribute to the peptide function. Second worst performance is achieved by Fourier-transformation based encoder MappingClass-integer-fourier. Fourier-transformation based encoders fail to capture the distinctive patterns and abrupt changes in amino acids within peptide sequences. These encoders make the assumption that peptide sequences adhere to a consistent pattern. However, this assumption overlooks the complexity and variability present in peptide sequences.

Apart from these two categories, other types of existing sequence encoders also have some disadvantages. Amino acid distribution and gap-based amino acid distribution encoders, do not effectively capture the sequential arrangement of amino acids. Additionally, group-based amino acid distribution encoders tend to oversimplify peptide sequences due to their shorter lengths, potentially missing important functional characteristics of peptides. Moreover, these encoders heavily depend on the criteria used for grouping amino acids, making it challenging to find a general criteria that works well for different datasets. Physico-chemical property-based encoders like AAIndex and network-based encoders like AESNN3 use pre-computed values, which limits their ability to capture comprehensive relationships and interactions between amino acids within peptide sequences. Traditional network-based encoders, such as complex network, also fail to capture comprehensive hierarchical and non-linear relationships. Sequence order-based encoders like QSOrder struggles to capture long-range interactions between amino acids. On the other hand, contextual information-aware encoders focus on frequent local and global features but lack to capture discriminative local and global features. In addition, these encoders have a deficiency in capturing the comprehensive correlation and distribution information of amino acids.

Binary encoders only consider the presence or absence of individual amino acids or groups within peptide sequences, and thus fail to capture the diverse properties of individual amino acids. Optimized physico-chemical properties and substitution matrix-based encoders such as ZScale and BLOSUM62 struggle to accurately capture the diversity present in peptide sequences.

In a nutshell, existing sequence encoders lack to capture comprehensive discriminative amino acids relations that can distinguish ACPs sequences from and non-ACPs sequences. By precisely capturing 4 different types of amino acids relations such as correlation, distribution, composition, and transition, proposed CARE encoder outperforms all 55 encoders of 14 different categories across most machine learning classifiers in terms of most evaluation metrics. Overall, CARE encoder achieves best performance with Gradient Boosting classifier and it outperforms top performing existing predictor QSOrder by an average accuracy margin of 3%. A similar performance increment trend for proposed CARE encoder in terms of other evaluation metrics is evident in Supplementary Table-4.

3.2. Intrinsic performance comparison of the proposed and existing sequence encoders

The main objective of intrinsic performance comparison between proposed CARE encoder and 14 existing top-performing encoders is to determine which encoder can create highly disjoint clusters for ACPs and non-ACPs classes. Precisely, the more qualitative statistical representations are, the more non-overlapping clusters will be generated. To visualize the statistical representations of both the proposed and existing encoders, we utilize the t-distributed stochastic neighbor embedding (TSNE) method to reduce the dimensions of all encoders to two, as depicted in Fig. 5.

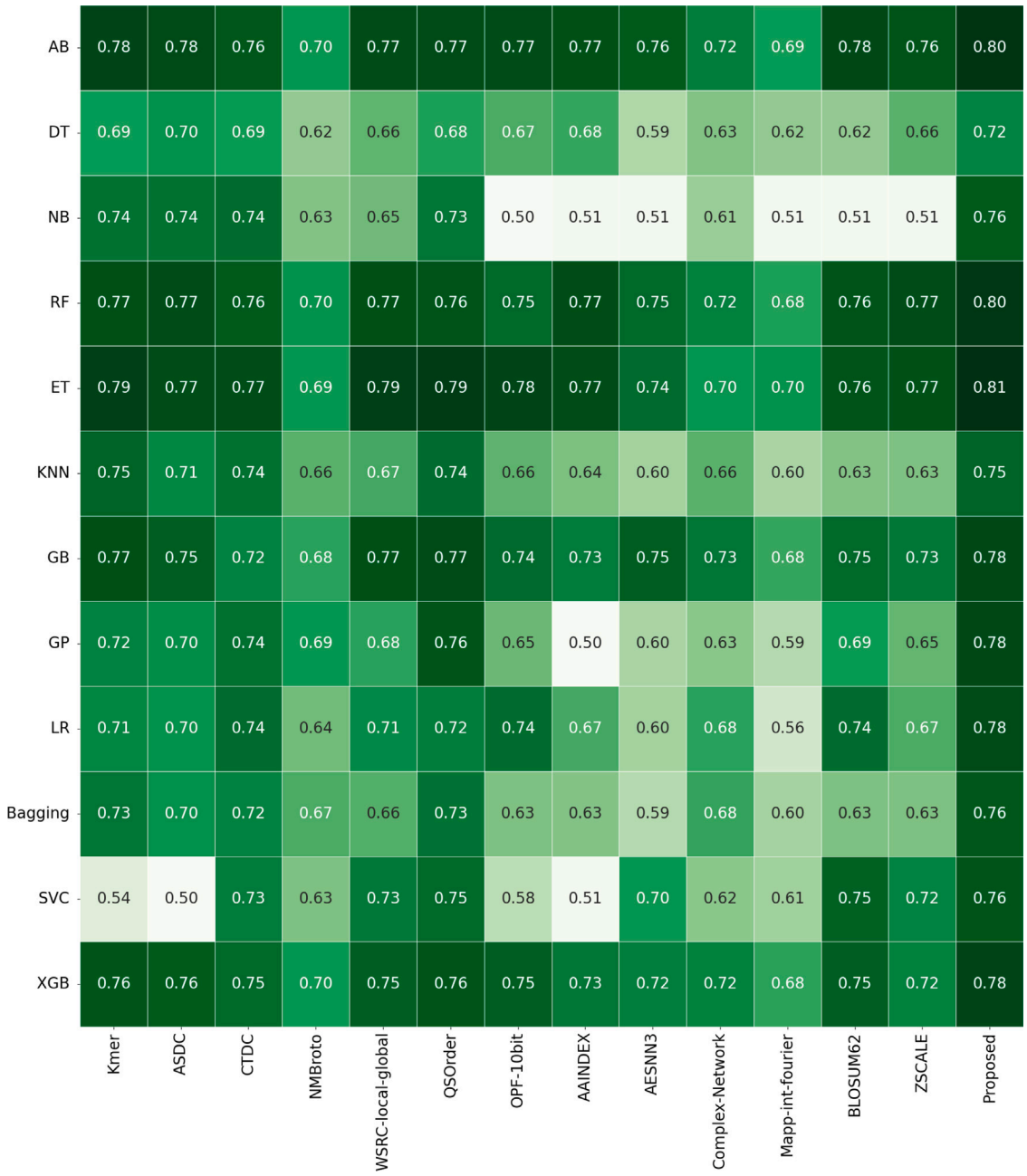


Fig. 4. Accuracy comparison of proposed care encoder and top performing existing encoders of 14 different categories across 12 different classifiers. Here encoders are shown on X-axis and classifiers are shown on Y-axis.

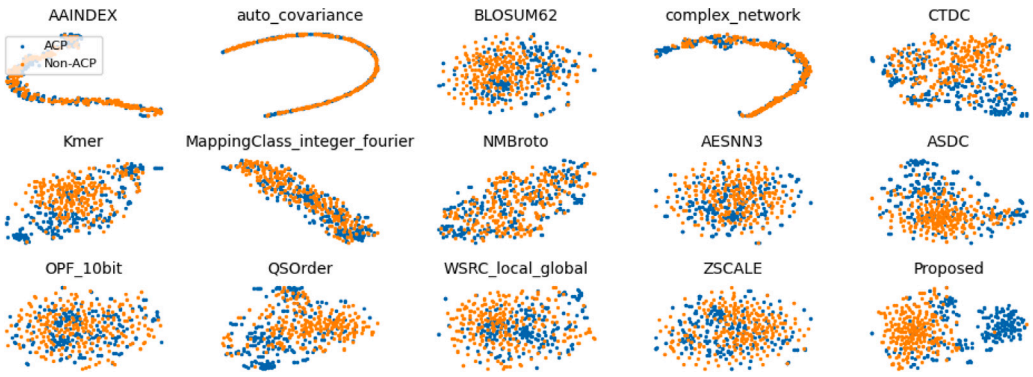


Fig. 5. Intrinsic performance comparison of proposed CARE encoder with 14 top performing existing encoders of different categories using benchmark ACP-Mixed-80 dataset. Peptide sequences of ACP class are represented with orange circles and peptide sequences of Non-ACP class are represented with blue circles.

Fig. 5 clearly demonstrates that all 14 existing top-performing encoders exhibit significant overlap in the clusters produced for ACPs and non-ACPs classes. None of these encoders manage to extract and encode comprehensive discriminative information about the distribution of amino acids for both classes which leads to the generation of low-quality statistical representations and overlapping clusters. Since the quality of the statistical representation greatly affects the performance of classifiers, these representations cannot be utilized to achieve great predictive performance. Whereas, graphical analysis of statistical representation produced by the proposed CARE encoder indicates a clear difference because it formulates highly disjoint clusters for ACPs and non-ACPs classes. The high-quality clusters demonstrate the CARE encoder's ability to capture comprehensive discriminative information about the distribution of amino acids in both classes, which proves effective in distinguishing ACPs class sequences from non-ACPs class sequences.

3.3. Proposed and existing predictors performance comparison for ACPs and non-ACPs classification

We compare the performance of proposed CAPTURE predictor with 37 existing predictors across 8 benchmark ACPs and non-ACPs classification datasets. Performance values of proposed and existing predictors across 8 datasets in terms of 7 different evaluation measures are provided in Supplementary Excel File. To illustrate better, only accuracy, MCC, and F1 score of proposed predictor and existing predictors are shown in Table 1.

A first look on Table 1 indicates that proposed CAPTURE is the first predictor that is comprehensively evaluated on 8 different datasets. Whereas, 37 existing predictors are evaluated only on few benchmark datasets. Among 8 datasets, 3 datasets namely ACP-Mixed-80, ACP2.0_Main, ACP2.0_Alternate are balanced as they have same number of ACPs and non-ACPs sequences. Across these datasets, 22 existing predictors have been evaluated and 3 predictors manage to produce decent performances including ACPred-LAF [64], iACP-FSCM [94], and ACPred-BMF [104] respectively. Proposed CAPTURE predictor outperforms previous best ACPred-LAF [64] predictor on ACP-Mixed-80 dataset by a significant recall margin of 12%, precision and MCC margin of 6%, AU-ROC and accuracy margin of 3%. Similarly, on ACP2.0_Main dataset, CAPTURE predictor outperforms iACP-FSCM [94] by recall margin of 14%. On ACP2.0_Alternate dataset, proposed CAPTURE predictor outperforms ACPred-BMF [104] predictor by accuracy and MCC margin of 1%. Unlike CAPTURE predictor, there is a significant gap of 18% in the sensitivity and specificity figures of iACP-FSCM predictor [94] that shows its biaseness towards type II error because it lacks to accurately predict non-ACP sequences on balanced dataset. On 3 balanced datasets, proposed predictor consistently shows best performance due to the comprehensive diverse feature extraction paradigm of underlay novel CARE encoder.

On the other hand, among 8 datasets, 5 datasets namely ENNACT, ACP539, ACP344, SET-1, and SET-2 belong to unbalanced datasets category because in these datasets non-ACPs sequences are higher than ACPs sequences. Researchers have used different strategies such as under-sampling, over-sampling, ensemble learning, and cost-sensitive learning to prioritize correct predictions for the minority class. On 5 imbalanced datasets, proposed CAPTURE predictor outperforms existing predictors by a decent margin without using any aforementioned strategy. On unbalanced ACP344 dataset where non-ACP to ACP class difference is 68 sequences, 9 existing predictors have been evaluated and proposed CAPTURE predictor outperforms previous best Kabir et al. predictor [111] by an accuracy margin of 4% and recall margin of 2%. On ACP539 dataset where non-ACP to ACP class difference is 161 sequences, 5 predictors have been evaluated and CL-ACP [68] predictor achieves good performance. Proposed CAPTURE predictor outperforms state-of-the-art CL-ACP [68] predictor by the recall of 16%, precision of 10%, specificity and MCC of 5%, and accuracy of 3%. On SET-1 dataset having non-ACP to ACP class difference of 6 sequences and SET-2 dataset having ACP to non-ACP class difference of 55 sequences, 10 predictors have been evaluated. On both datasets, proposed CAPTURE predictor outshines previous best Yao et al. [66] predictor by an accuracy and F1 score of 1%. Also, it achieves a recall increment of 2% on SET-1 and 8% on SET-2 dataset. Furthermore, on ENNAACT dataset where non-ACP to ACP sequence difference is very huge (n = 4639), 6 predictors have been evaluated and LGBM-ACP [109] achieves top performance. Proposed CAPTURE predictor outperforms LGBM-ACP predictor by specificity of 2%, accuracy, recall, and AU-ROC of 1%

Instead of using any additional strategy to address the challenges posted by imbalance datasets, proposed CAPTURE predictor only focuses on learning and using effective statistical representation of peptides sequences. Despite the short lengths of peptides sequences, novel CARE encoder does not solely rely on features that are prevalent in the majority class, but also consider features specific to the minority class. This is why even the increase in negative to positive class sequence difference does not hamper the top performance of proposed CAPTURE predictor at all across different datasets.

3.3.1. Proposed predictor generalizeability evaluation for identifying anti-microbial peptides sequences

With an aim to evaluate the generalizeability of ACPs predictors, few researchers have explored the potential of their ACPs predictors for classifying Anti-microbial peptides (AMPs) sequences [68,112,113]. Unlike ACPs that are special class of AMPs and only target cancer cells, AMPs have the potential to treat different microbial infections caused by bacteria, fungi, and viruses [68,112,113]. Following generalizeability evaluation paradigm of existing studies [68,112,113], we compare the generalizeability potential of proposed predictor with 3 existing AMPs classification predictors across 7 most widely used benchmark datasets.

Fig. 6 illustrates the AU-ROC values of proposed and 3 existing predictors across 7 benchmark datasets. A bird's eye view of Fig. 6 reveals that, among 3 existing predictors, Yi et al. predictor [70] remains least performer on 3 datasets (AAP, AVP, PBP) and manages to imitate performance of Wu et al. [90] predictor on 2 datasets (ABP, CCP). Across 2 datasets namely AIP and QSP, Wu et al. [90] predictor remains least performer, however, across all 7 datasets, Wang et al. [68] predictor remains the top performer.

On the other hand, proposed predictor outperforms existing best performing Wang et al. [68] predictor by 1% on 3 datasets (AAP,AIP, QSP), and manages to imitate the top performance of Wang et al. [68] predictor on CCP dataset. Furthermore, it achieves the performance increment of 6% on AVP, 5% on PBP, and 3% on ABP dataset.

Prime reason behind the supreme performance of proposed predictor is incorporation of powerful sequence encoding method CARE in its predictive pipeline. Wang et al. [68] predictor first processes AMPs sequences using amino acid one-hot encoding and structure information. They concatenated the features extracted by self-attention mechanism, convolutional, and long-short term neural network from processed AMPs sequences. While, one-hot encoding fails to capture the order of amino acids because it treats all amino acids independently and neglects the dependency of amino acids on the position and distribution of neighboring amino acids. Multi-head self-attention mechanism only manages to extract few important amino acids patterns in peptides sequences due to attention collapse issue [114] that forces different heads to extract very similar attentive features. On the other hand proposed CARE encoder transforms peptides sequences into statistical vectors by extracting 4 different kinds of amino acid features including correlation, distribution, composition, and transition. CARE encoder paradigm of introducing feature diversity leads to more informative and discriminative statistical representations that largely help machine learning classifier to precisely discriminate AMPs from non-AMPs.

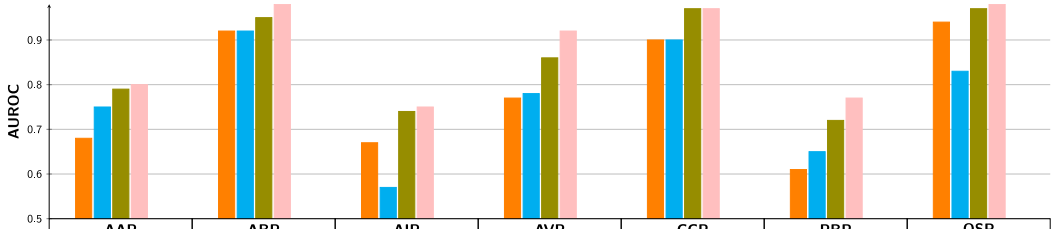
Table 1
Performance comparison of proposed predictor CAPTURE with existing predictors on 8 different benchmark binary ACPs datasets.

Predictor	Benchmark datasets																
	ACP_Main		ACP_Alternate		ENNAACT_main		ACP_539		ACP Mixed 80		ACP_344			Set 1		Set 2	
	Acc	MCC	Acc	MCC	Acc	MCC	Acc	MCC	Acc	MCC	Acc	MCC	F1	Acc	F1	Acc	F1
Tyagi et al. (2013) [71]	–	–	–	–	–	–	–	–	–	–	–	–	–	53.3	67.4	87.9	86.9
iACP (2016) [84]	55.1	0.11	77.6	0.55	0.95	0.76	–	–	–	–	–	–	–	–	–	–	–
MLACP (2017) [85]	–	–	–	–	0.94	0.72	–	–	–	–	–	–	–	–	–	–	–
AntiCP (2017) [86]	50.6	0.07	90	0.8	–	–	–	–	–	–	–	–	–	–	–	–	–
ACPred-FL (2018) [28]	44.8	0.12	43.8	0.15	–	–	–	–	–	–	–	–	–	–	–	–	–
SAP (2018) [87]	–	–	–	–	–	–	–	–	–	–	0.92	0.83	0.89	–	–	–	–
ACPred (2019) [88]	53.5	0.09	85.3	0.71	0.94	0.65	–	–	–	–	–	–	–	54.9	66.0	88.5	88.9
PEPred-Suite (2019) [89]	53.5	0.08	57.5	0.16	–	–	–	–	–	–	–	–		–	–	–	–
ACP-DL (2019) [70]	71.4	0.43	–	–	–	–	0.72	0.41	–	–	–	–		–	–	–	–
PTPD (2019) [90]	–	–	–	–	–	–	0.75	0.43	–	–	–	–		–	–	–	–
ACPred-Fuse (2020) [91]	68.9	0.38	78.9	0.6	–	–	–	–	–	–	–	–	–	–	–	–	–
AMPfun (2020) [92]	–	–	–	–	–	–	–	–	–	–	–	–	–	68.7	69.9	77.3	74.6
DeepACP (2020) [93]	–	–	–	–	–	–	–	–	–	–	–	–	–	58.0	68.7	90.7	90.7
ACP-LDF (2020) [64]	–	–	–	–	–	–	–	–	–	–	0.92	0.84	0.92	–	–	–	–
iACP-FSCM (2021) [94]	**82.5**	**0.64**	88.9	0.77	–	–	–	–	–	–	–	–		–	–	–	–
ACP-MHCNN (2021) [95]	68.4	0.37	–	–	–	–	–	–	–	–	–	–	–	57.1	68.9	91.6	91.6
AntiCP 2.0 (2021) [65]	72.3	0.45	–	–	0.91	0.56	0.82	0.6	–	–	–	–	–	70.2	67.1	91.6	91.9
iACP-DRLF (2021) [96]	74.3	0.49	0.93	0.86	–	–	0.83	0.61	–	–	–	–	–	–	–	–	–
ENNAACT (2021) [67]	–	–	–	–	0.97	**0.9**	–	–	–	–	–	–		–	–	–	–
iACP-FSCM (2021) [94]	–	–	–	–	–	–	0.84	0.66	–	–	–	–	–	–	–	–	–
ACPred-LAF (2021) [64]	–	–	–	–	–	–	–	–	0.81	0.63	–	–	–	–	–	–	–
PreTP-EL (2021) [97]	–	–	–	–	–	–	–	–	0.58	0.17	–	–	–	–	–	–	–
dbAMP2.0 (2022) [98]	–	–	–	–	–	–	–	–	–	–	–	–	–	67.4	71.7	49.7	54.2
iACP-GE (2022) [99]	–	–	–	–	–	–	–	–	–	–	–	–	–	75.9	76.5	89.6	89.8
StackACPred (2022) [100]	–	–	–	–	–	–	–	–	–	–	–	–	–	73.0	72.1	93	93.1
PreTP-Stack (2022) [101]	–	–	–	–	–	–	–	–	0.49	0.02	–	–	–	–	–	–	–
ACPCheck (2022) [102]	78.0	56.0	93.0	86.0	–	–	–	–	–	–	–	–	–	–	–	–	–
ME-ACP (2022) [103]	79.0	58.0	93.5	87.0	–	–	–	–	–	–	–	–	–	–	–	–	–
ACPred-BMF (2022) [104]	81.0	62.0	93.6	87.1	–	–	–	–	–	–	–	–	–	–	–	–	–
AI4ACP (2022) [105]	0.72	0.44	0.89	0.79		–	–	–	0.73	0.48	–	–	–	–	–	–	–
ACP-OPE (2023) [106]	79.0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
iACP-RF (2023) [107]	75.9	0.52	93.1	0.86	–	–	–	–	–	–	–	–	–	–	–	–	–
TriNet (2023) [108]	76.6	0.53	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
LGBM-ACp (2023) [109]	–	–	–	–	0.97	0.87	–	–	–	–	–	–	–	–	–	–	–

(*continued on next page*)

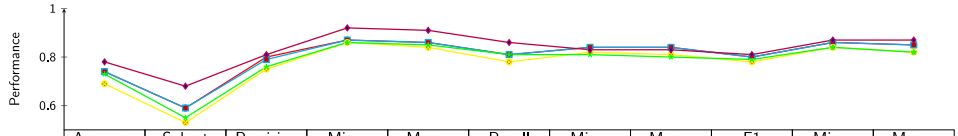
Table 1 (*continued*).

ACP-GRDF (2023) [66]	–	–	–	–	–	–	–	–	–	–	–	–	–	77.1	77.5	94.1	94.2
ACP-MLC (2023) [11]	–	–	–	–	–	–	–	–	0.79	0.57	–	–	–	–	–	–	–
ACP-Kernel-SRC (2023) [110]	–	–	–	–	–	–	–	–	–	–	0.93	0.85	**0.94**	–	–	–	–
Proposed CAPTURE Predictor	76.7	0.54	**94.1**	**0.88**	**0.98**	0.87	**0.87**	**0.71**	**0.84**	**0.69**	**0.93**	**0.86**	0.93	**78.4**	**78.0**	**95.2**	**95.2**



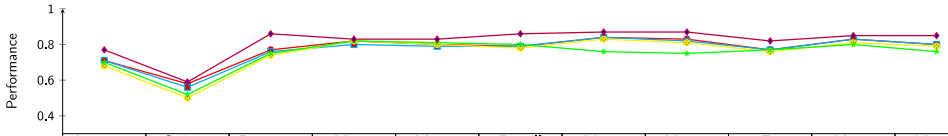
	AAP	ABP	AIP	AVP	CCP	PBP	QSP
Yi et al.	0.68	0.92	0.67	0.77	0.9	0.61	0.94
Wu et al.	0.75	0.92	0.57	0.78	0.9	0.65	0.83
Wang et al.	0.79	0.95	0.74	0.86	0.97	0.72	0.97
Proposed	0.8	0.98	0.75	0.92	0.97	0.77	0.98

Fig. 6. Comparison of generalizeability potential of proposed predictor CAPTURE with 3 existing predictors in terms of AUROC across 7 AMPs datasets.



	Accuracy .	Subset Accuracy	Precision .	Micro Precision	Macro Precision	Recall .	Micro Recall	Macro Recall	F1 .	Micro F1	Macro F1
Deng et al.[BR][1]	0.74	0.59	0.8	0.87	0.86	0.81	0.84	0.84	0.8	0.86	0.85
Deng et al.[CC][1]	0.74	0.59	0.79	0.87	0.86	0.81	0.84	0.84	0.8	0.86	0.85
Deng et al.[CNN][1]	0.69	0.53	0.75	0.86	0.84	0.78	0.82	0.81	0.78	0.84	0.82
Deng et al.[MLKNN][1]	0.73	0.55	0.76	0.86	0.85	0.81	0.81	0.8	0.79	0.84	0.82
Proposed	0.78	0.68	0.81	0.92	0.91	0.86	0.83	0.83	0.81	0.87	0.87

Fig. 7. Five-fold performance comparison of proposed predictor CAPTURE with 4 different classifiers of existing predictor such as Binary Relevance (BR), Classifier Chain (CC), Convolutional Neural Network (CNN), and Algorithm Adaptation (MLKNN) in terms of 11 evaluation metrics.



	Accuracy .	Subset Accuracy	Precision .	Micro Precision	Macro Precision	Recall .	Micro Recall	Macro Recall	F1 .	Micro F1	Macro F1
Deng et al.[BR][1]	0.71	0.58	0.77	0.82	0.80	0.79	0.84	0.83	0.77	0.83	0.80
Deng et al.[CC][1]	0.71	0.56	0.76	0.80	0.79	0.79	0.84	0.82	0.77	0.83	0.80
Deng et al.[CNN][1]	0.68	0.50	0.74	0.82	0.80	0.78	0.83	0.81	0.76	0.81	0.79
Deng et al.[MLKNN][1]	0.70	0.52	0.75	0.82	0.81	0.80	0.76	0.75	0.77	0.80	0.76
Proposed	0.77	0.59	0.86	0.83	0.83	0.86	0.87	0.87	0.82	0.85	0.85

Fig. 8. Independent test based performance comparison of proposed predictor CAPTURE with 4 different classifiers of existing predictor such as Binary Relevance (BR), Classifier Chain (CC), Convolutional Neural Network (CNN), and Algorithm Adaptation (MLKNN) in terms of 11 evaluation metrics.

3.4. Proposed and existing predictors performance comparison for ACPs functional types classification

State-of-the-art ACPs functional types classification predictor [11] explored the potential of two data transformation approaches namely binary relevance and classifier chains to transform the multi-label data into binary classification data and trained a separate machine learning classifier for binary classification data. Additionally they explored the potential of algorithm adaptation approach MLKNN and convolutional neural network for accurate classification of ACPs functional types. To illustrate the potential of proposed CAPTURE predictor for accurate prediction of ACPs functional types by precisely capturing functional types dependencies. We perform 5-fold performance comparison of proposed CAPTURE predictor, 3 baseline approaches [11], top performing binary relevance and tree classifier based approach called ACP-MLC [11] on a benchmark dataset in Fig. 7.

It can be seen Fig. 7 that proposed CAPTURE predictor outshines all 4 approaches [11] across almost all evaluation metrics. It achieves an accuracy increment of 5%, subset accuracy rise of 9%, micro and macro precision and recall rise of 5%, macro F1 rise of 2%, precision, F1, and micro F1 rise of 1%.

Similarly, on independent test, analysis of performance figures shown in Fig. 8 indicates that, previous performance figures of proposed CAPTURE predictor jump to even more promising values. Specifically, proposed predictor CAPTRUE achieves 10% increment in precision, 6% rise in accuracy and recall, 5% increment in F1 and macro

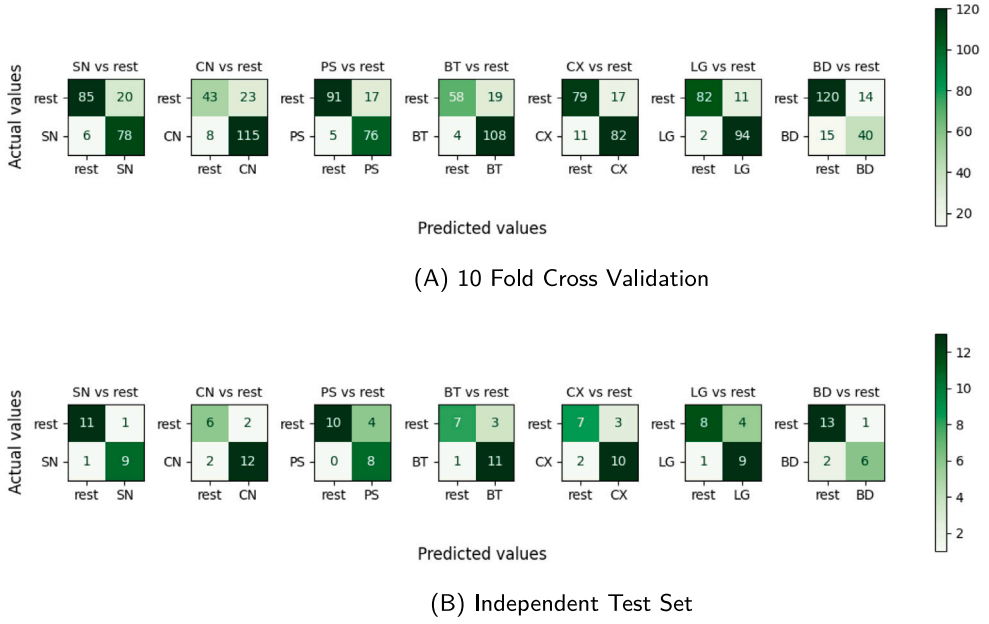


Fig. 9. Accuracy confusion matrices produced by proposed CAPTURE predictor using 10-fold cross validation [A] and independent test set [B].

F1, 4% rise in macro recall, 3% increment in micro recall, 2% rise in macro precision and micro F1, 1% increment in subset accuracy and micro precision. The significant performance rise achieved by proposed predictor CAPTURE on 5-fold and independent test set is due to the ability of novel encoder CARE to most effectively handle heterogeneity of sequences, imbalance distribution of functional types, and their correlations. Whereas existing data transformation approaches such as binary relevance and classifier chains disregard potential dependencies and correlations between functional types by treating each functional type independently. Also, classifier chains is sensitive the order of functional types and struggle to generalize well on new ACPs sequences. Likewise existing algorithm adaptation approach MLKNN is sensitive to imbalance distribution of functional types, hence it shows biaseness towards dominant functional types. CNN lacks to capture key long range dependencies of amino acids present in ACPs sequences and requires a large and balanced training data to achieve decent performance for multi-label prediction.

3.5. Proposed predictor in-depth performance analysis for ACPs functional types classification

In order to truly evaluate the effectiveness of proposed CAPTURE predictor for multi-label classification of ACPs in relevant functional types, we analyze the accuracy confusion matrices (Fig. 9) produced under the hood of two different settings: one-vs-rest and independent test set.

A critical analysis of confusion matrices (Fig. 9A) produced by proposed CAPTURE predictor under the hood of 10-fold cross validation indicates that decent number of peptide sequences are accurately classified into their corresponding 7 different functional types. For three functional types including Lung (LG), Breast (BT), and Prostate (PT), only 5 or less than 5 peptide sequences are miss-classified. For two functional types including Skin (SN) and Colon (CN), less than 10 peptides sequences are miss-classified. Furthermore, despite having least only 55 peptide sequences for Blood (BD)functional type, CAPTURE still manages to accurately identify BD functional type for 30 peptide sequences due to comprehensive discriminative features extracted by proposed CARE encoder. For Cervix functional type, CAPTURE has only miss-classified 11 peptide sequences. Overall, more than 100 peptide sequences belonging to CN and BT, more than 75 peptide sequences belonging to SN, PS, and CX, and more than 90 peptide sequences belonging to LG functional types are correctly predicted by CAPTURE.

A similar performance trend is shown by proposed CAPTURE predictor on test set (Fig. 9B) where every functional type has only 12 or less than 12 peptide sequences. CAPTURE manages to identify 6 functional types for most peptide sequences and only miss-classify 1 or 2 peptide sequences. Whereas, all 8 peptide sequences belonging to Prostrate (PS) functional type are accurately classified mainly due to the powerful statistical representation generated by proposed CARE encoder.

Furthermore, in order to assess up to what degree proposed CAPTURE predictor manages to correctly predict distinct combination of functional types on account of distinct functional type cardinality in benchmark ACPFunctional dataset. We analyze sequence-to-functional type distribution as well as correctly identified functional types out of all functional types highlighted in vertical bar chart and respective confusion matrices shown in Fig. 10.

A critical analysis of Fig. 10[A–B] produced using 10-fold cross validation indicates that from 52 uni-functional type sequences, 37 are correctly predicted by CAPTURE. From 32 bi-functional type sequences, 25 are correctly predicted by CAPTURE. From 29 tri-functional type sequences, 21 are correctly predicted by CAPTURE. Unlike typical multi-label classification approaches whose performance drop with the increase of label cardinality, from 65 hexa-functional type sequences, 61 sequences are correctly predicted by CAPTURE due to the supremely effective statistical representations produced proposed CARE encoder. Although Tetra-functional types and Penta-functional types have only 6 and 4 peptide sequences respectively. However, once again unlike existing traditional multi-label classification approaches whose performance plunge to lowest figures due to limited number of sequences. Proposed CAPTURE predictor still manages to accurately predict all four and five functional types for almost all corresponding peptide sequences. One peptide sequence that has all 7 functional types is miss-classified by CAPTURE as 1 sequence is not sufficient to learn complex underlay distribution of hepta-functional types based sequences.

A similar performance trend is shown by proposed CAPTURE predictor on test set. Analysis of Fig. 10[C–D] indicates that almost all uni, bi, penta, hexa functional type sequences are correctly predicted by CAPTURE. A decent number of tri-functional type sequences are correctly predicted whereas tetra-functional types have no peptide sequence in test set.

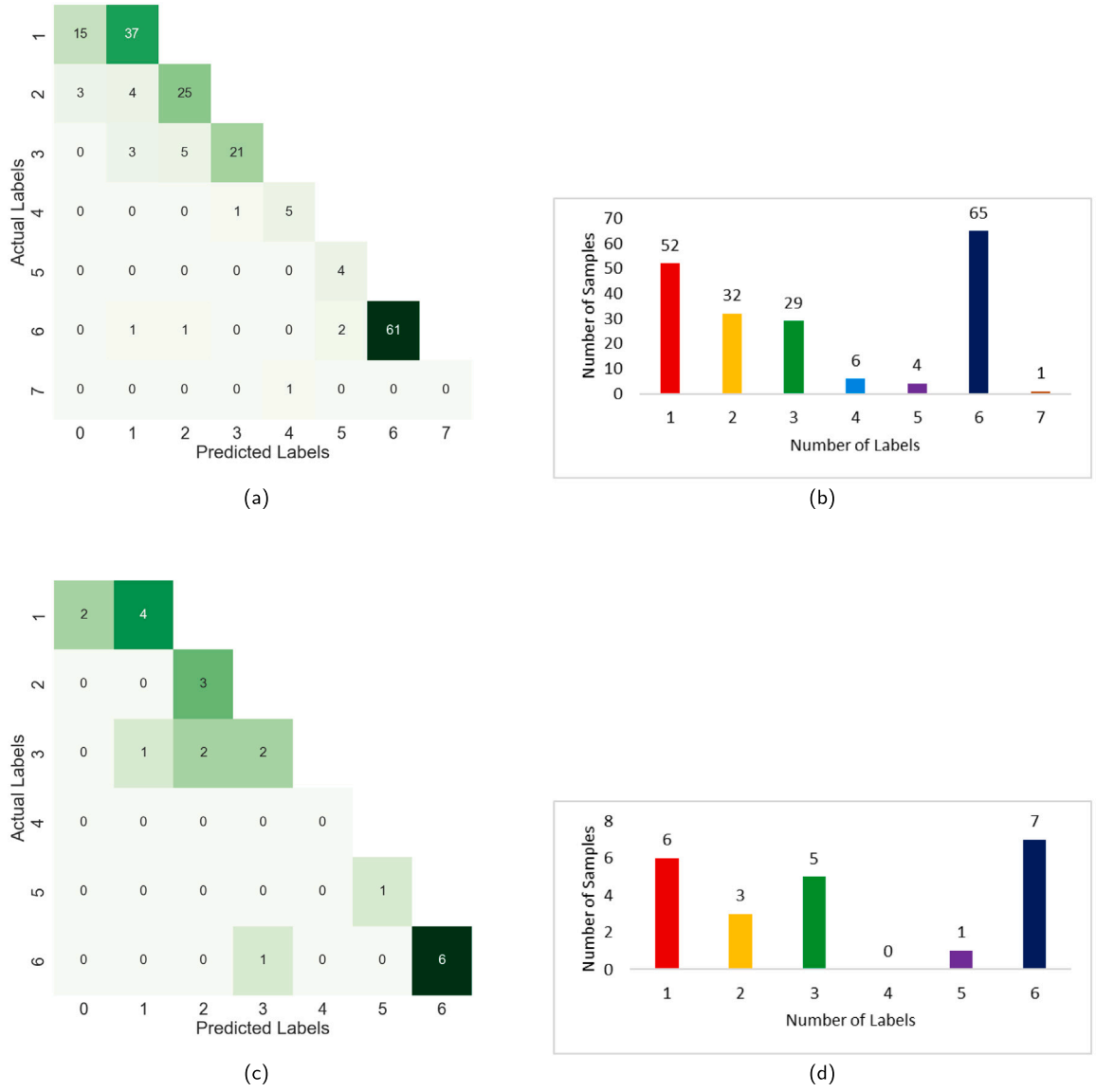


Fig. 10. Multi-label performance values of proposed CAPTURE predictor produced using 10-fold cross validation and independent test set corresponding to unique sequence-to-functional types distribution.

4. Limitation

Section 1 describes two main modules of ACPs classification pipeline: sequence representation or sequence transformation into statistical feature space and classification. The primary emphasis of the current study is to transform ACPs sequences into statistical feature space by extracting diverse types of amino acids distribution patterns. Then, it utilizes traditional machine learning classifiers at classification stage to identify ACPs. Although proposed sequence encoding method along with Adaboost classifier manages to produce state-of-the-art performance but this study does not reap the combine potential of multiple classifiers by designing a meta predictor. Despite the promising performance of recent protein representation learning methods [115,116] and deep learning architectures for diverse proteomics sequence analysis tasks like protein function prediction [117,118]. It neither explores the aptitude of recent protein representation learning methods [115,116] nor the potential of proposed sequence encoder with deep learning architectures [117,118].

Moreover, it is briefly described in Section 1 that ACPs functional types annotation is a multi-label classification task. This study performs functional types annotations by first transforming ACPs sequences into statistical vectors, then transforming multi-label task into binary classification task and finally utilizing a machine learning classifier for prediction. Along with proposed encoder, it does not explore the potential of machine learning and deep learning algorithms [119] that are competent to deal multi-label data directly.

Furthermore, proposed CAPTURE predictor web application may not make correct predictions for cyclic class peptides because proposed predictor is trained on public benchmark datasets that do not contain sequences of cyclic class peptides. The application can only categorize peptides sequences into ACPs and non-ACPs classes and can make functional annotations to ACPs sequences. It is not capable to predict ACPs special characteristics like bio-activity and toxicity [10,120–123].

5. Conclusion

ACPs ability to block the growth, migration, and invasion of cancer cells through multiple mechanisms make ACPs promising candidates for the development of highly effective cancer treatment. Distinguishing ACPs from non-ACPs and determining the functional types of ACPs are important to gain a deeper understanding of the biological role of ACPs and their potential for cancer therapies. This paper presents

a unique sequence encoding method CARE that has the competency to extract 4 different types of information (correlational, distributional, compositional and transitional) for amino acids present in raw peptides sequences. Across ACPs classification benchmark datasets, a comprehensive experimentation reveals that, unlike existing sequence encoding methods, proposed encoder significantly enhances the performance of various machine learning classifiers. Moreover, an intrinsic analysis proves that proposed encoder extracts more useful patterns of amino acids in comparison to existing sequence encoding methods. Furthermore, proposed encoder along with AdaBoost classifier named as CAPTURE predictor, outperforms existing ACPs and non ACPs classification predictors across 8 public benchmark datasets. A case study based on AMPs classification proves the generalizability of CAPTURE predictor and its potential to use for other types of peptides classification. On the other hand, for ACPs functional types annotations, in comparison to existing predictors, proposed predictor superior performance makes it ideal predictor for the exploration of biological roles of ACPs and their use cases in cancer therapies. A compelling future line of current work would be to design more accurate ACPs classification pipeline by utilizing proposed sequence encoding method and deep learning classifiers. Deep learning classifiers may contain standalone architectures including Convolutional Neural Network (CNN), Long short term memory network (LSTM,), Gated Recurrent Unit (GRU) or Hybrid architectures comprising of different networks. Moreover, for ACPs functional annotations, rather than transforming multi-label data into binary label data, the proposed encoder can be utilized with machine or deep learning predictors that can directly deal with multi-label data. Moreover, proposed CAPTURE predictor web application functional scope can be enhanced by training predictor on large dataset which also contain diverse types of peptides sequences such as cyclic peptides that are not present in current benchmark datasets.

Funding

This research received no funding from public, commercial or non-profit agencies.

Additional information

Additional information is given in supplementary files.

CRediT authorship contribution statement

Hina Ghafoor: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Muhammad Nabeel Asim:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Muhammad Ali Ibrahim:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sheraz Ahmed:** Writing – review & editing, Supervision. **Andreas Dengel:** Writing – review & editing, Supervision.

Declaration of competing interest

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

Data availability

Datasets used in this study are available at CAPTURE web server https://sds_genetic_analysis.opendfki.de/CAPTURE.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.108538.

References

[1] Q. Li, W. Zhou, D. Wang, S. Wang, Q. Li, Prediction of anticancer peptides using a low-dimensional feature model, Front. Bioeng. Biotechnol. 8 (2020).
[2] R.L. Siegel, K.D. Miller, N.S. Wagle, A. Jemal, Cancer statistics, 2023, CA Cancer J. Clin. 73 (1) (2023) 17–48.
[3] D. Hanahan, R.A. Weinberg, Hallmarks of cancer: the next generation, Cell 144 (5) (2011) 646–674.
[4] B. Vogelstein, N. Papadopoulos, V.E. Velculescu, S. Zhou, L.A. Diaz Jr., K.W. Kinzler, Cancer genome landscapes, Science 339 (6127) (2013) 1546–1558.
[5] V.T. DeVita Jr., E. Chu, A history of cancer chemotherapy, Cancer Res. 68 (21) (2008) 8643–8653.
[6] M. Chidambaram, R. Manavalan, K. Kathiresan, Nanotherapeutics to overcome conventional cancer chemotherapy limitations, J. Pharm. Pharm. Sci. 14 (2011) 67.
[7] M.K. Shin, B.-Y. Jang, K.-B. Bu, S.-H. Lee, D.-H. Han, J.W. Oh, J.-S. Sung, De novo design of AC-P19M, a novel anticancer peptide with apoptotic effects on lung cancer cells and anti-angiogenic activity, Int. J. Mol. Sci. 23 (24) (2022) 15594.
[8] M. Karami Fath, K. Babakhaniyan, M. Zokaei, A. Yaghoubian, S. Akbari, M. Khorsandi, A. Soofi, M. Nabi-Afjadi, H. Zalpoor, F. Jalalifar, et al., Anti-cancer peptide-based therapeutic strategies in solid tumors, Cell. Mol. Biol. Lett. 27 (1) (2022) 33.
[9] A.K. Tripathi, J.K. Vishwanatha, Role of anti-cancer peptides as immunomodulatory agents: Potential and design strategy, Pharmaceutics 14 (12) (2022) 2686.
[10] I.W. Hamley, Introduction to Peptide Science, John Wiley & Sons, 2020.
[11] H. Deng, M. Ding, Y. Wang, W. Li, G. Liu, Y. Tang, ACP-MLC: A two-level prediction engine for identification of anticancer peptides and multi-label classification of their functional types, Comput. Biol. Med. 158 (2023) 106844.
[12] F. López-Vallejo, T. Caulfield, K. Martínez-Mayorga, M. A Giulianotti, A. Nefzi, R. A Houghten, J. L Medina-Franco, Integrating virtual screening and combinatorial chemistry for accelerated drug discovery, Comb. Chem. High Throughput Screen. 14 (6) (2011) 475–487.
[13] R. Liu, X. Li, K.S. Lam, Combinatorial chemistry in drug discovery, Curr. Opin. Chem. Biol. 38 (2017) 117–126.
[14] D. Sahin, S.O. Taflan, G. Yartas, H. Ashktorab, D.T. Smoot, Screening and identification of peptides specifically targeted to gastric cancer cells from a phage display peptide library, Asian Pac. J. Cancer Prev.: APJCP 19 (4) (2018) 927.
[15] M. Poreba, P. Kasperkiewicz, W. Rut, M. Drag, Screening combinatorial peptide libraries in protease inhibitor drug discovery, in: Extracellular Targeting of Cell Signaling in Cancer: Strategies Directed at MET and RON Receptor Tyrosine Kinase Pathways, Wiley Online Library, 2018, pp. 307–350.
[16] E.L. Boys, J. Liu, P.J. Robinson, R.R. Reddel, Clinical applications of mass spectrometry-based proteomics in cancer: Where are we? Proteomics 23 (7–8) (2023) 2200238.
[17] M. Nabeel Asim, M. Ali Ibrahim, A. Fazeel, A. Dengel, S. Ahmed, DNA-MP: a generalized DNA modifications predictor for multiple species based on powerful sequence encoding method, Brief. Bioinform. 24 (1) (2023) bbac546.
[18] K.-C. Chou, Prediction of protein subcellular locations by incorporating quasi-sequence-order effect, Biochem. Biophys. Res. Commun. 278 (2) (2000) 477–483.
[19] G. Schneider, P. Wrede, The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site, Biophys. J. 66 (2) (1994) 335–344.
[20] R. Grantham, Amino acid difference formula to help explain protein evolution, Science 185 (4154) (1974) 862–864.
[21] M. Bhasin, G.P.S. Raghava, Classification of nuclear receptors based on amino acid composition and dipeptide composition, J. Biol. Chem. 279 (22) (2004) 23262–23266.
[22] V. Saravanan, N. Gautham, Harnessing computational biology for exact linear B-cell epitope prediction: A novel amino acid composition-based feature descriptor, Omics : J. Integr. Biol. 19 (10) (2015) 648–658.
[23] W. Chen, H. Tran, Z.-Y. Liang, H. Lin, L. Zhang, Identification and analysis of the N(6)-methyladenosine in the Saccharomyces cerevisiae transcriptome, Sci. Rep. 5 (1) (2015) 13859.
[24] C. Zhou, C. Wang, H. Liu, Q. Zhou, Q. Liu, Y. Guo, T. Peng, J. Song, J. Zhang, L.-L. Chen, Y. Zhao, Z. Zeng, D.-X. Zhou, Identification and analysis of adenine N6-methylation sites in the rice genome, Nat. Plants 4 (8) (2018) 554–563.
[25] H. Chen, F. Li, L. Wang, Y. Jin, C.-H. Chi, L. Kurgan, J. Song, J. Shen, Systematic evaluation of machine learning methods for identifying human-pathogen protein-protein interactions, Brief. Bioinform. 22 (3) (2020) 1–NA.

[26] K. Chen, L. Kurgan, J. Ruan, Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs, BMC Struct. Biol. 7 (1) (2007) 25.
[27] K. Chen, Y. Jiang, L. Du, L. Kurgan, Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs, J. Comput. Chem. 30 (1) (2008) 163–172.
[28] L. Wei, C. Zhou, H. Chen, J. Song, R. Su, ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides, Bioinform. (Oxf. Engl.) 34 (23) (2018) 4007–4016.
[29] J. Zhou, C.L. Theesfeld, K. Yao, K.M. Chen, A.K. Wong, O.G. Troyanskaya, Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk, Nature Genet. 50 (8) (2018) 1171–1179.
[30] C.Z. Cai, L. Han, Z. Ji, X. Chen, Y.Z. Chen, SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence, Nucleic Acids Res. 31 (13) (2003) 3692–3697.
[31] C.Z. Cai, L. Han, Z. Ji, Y.Z. Chen, Enzyme family classification by support vector machines, Proteins 55 (1) (2004) 66–76.
[32] I. Dubchak, I. Muchnik, S.R. Holbrook, S.-H. Kim, Prediction of protein folding class using global description of amino acid sequence, Proc. Natl. Acad. Sci. USA 92 (19) (1995) 8700–8704.
[33] I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk, S.-H. Kim, Recognition of a protein fold in the context of the SCOP classification, Proteins 35 (4) (1999) 401–407.
[34] L. Han, C.Z. Cai, S.L. Lo, M.C.M. Chung, Y.Z. Chen, Prediction of RNA-binding proteins from primary sequence by a support vector machine approach, RNA (N. Y. N.Y.) 10 (3) (2004) 355–368.
[35] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, H. Jiang, Predicting protein-protein interactions based only on sequences information, Proc. Natl. Acad. Sci. USA 104 (11) (2007) 4337–4341.
[36] R.R. Sokal, B.A. Thomson, Population structure inferred by local spatial auto-correlation: an example from an Amerindian tribal population, Am. J. Phys. Anthropol. 129 (1) (2005) 121–131.
[37] Z.-P. Feng, C.-T. Zhang, Prediction of membrane protein types based on the hydrophobic index of amino acids, J. Protein Chem. 19 (4) (2000) 269–275.
[38] Z. Lin, X.-M. Pan, Accurate prediction of protein secondary structural content, J. Protein Chem. 20 (3) (2001) 217–220.
[39] D.S. Horne, Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities, Biopolymers 27 (3) (1988) 451–477.
[40] Y. Guo, L. Yu, Z. Wen, M. Li, Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences, Nucleic Acids Res. 36 (9) (2008) 3025–3030.
[41] Q. Dong, S. Zhou, J. Guan, A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation, Bioinform. (Oxf. Engl.) 25 (20) (2009) 2655–2662.
[42] B. Liu, L. Fang, R. Long, X. Lan, K.-C. Chou, iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition, Bioinform. (Oxf. Engl.) 32 (3) (2015) 362–369.
[43] K.-C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins 43 (3) (2001) 246–255.
[44] K.-C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinform. (Oxf. Engl.) 21 (1) (2004) 10–19.
[45] G. Schneider, P. Wrede, The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site, Biophys. J. 66 (2) (1994) 335–344.
[46] K.-C. Chou, Prediction of protein subcellular locations by incorporating quasi-sequence-order effect, Biochem. Biophys. Res. Commun. 278 (2) (2000) 477–483.
[47] K.-C. Chou, Y.-D. Cai, Prediction of protein subcellular locations by GO–FunD–PseAA predictor, Biochem. Biophys. Res. Commun. 320 (4) (2004) 1236–1239.
[48] K. Lin, A.C. May, W.R. Taylor, Amino acid encoding schemes from protein structure alignments: multi-dimensional vectors to describe residue types, J. Theoret. Biol. 216 (3) (2002) 361–365.
[49] B. Liu, X. Gao, H. Zhang, BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches, Nucleic Acids Res. 47 (20) (2019) e127.
[50] X. Chen, J.-D. Qiu, S.-P. Shi, S. Suo, S.-Y. Huang, R.-P. Liang, Incorporating key position and amino acid residue features to identify general and species-specific ubiquitin conjugation sites, Bioinform. (Oxf. Engl.) 29 (13) (2013) 1614–1622.
[51] Z. Chen, Y. Zhou, J. Song, Z. Zhang, hCKSAAP_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties, Biochim. Biophys. Acta (BBA)-Proteins Proteom. 1834 (8) (2013) 1461–1467.
[52] G. White, W. Seffens, Using a neural network to backtranslate amino acid sequences, Electron. J. Biotechnol. 1 (3) (1998) 17–18.
[53] G. White, W. Seffens, Using a neural network to backtranslate amino acid sequences, Electron. J. Biotechnol. 1 (2) (1998) 196–201.
[54] C.W. Tung, S.-Y. Ho, Computational identification of ubiquitylation sites from protein sequences, BMC Bioinform. 9 (1) (2008) 310.
[55] Y.Z. Chen, Z. Chen, Y.A. Gong, G. Ying, SUMOhydro: a novel method for the prediction of sumoylation sites based on hydrophobic properties, PLoS One 7 (6) (2012) e39195–NA.
[56] T.-Y. Lee, S.-A. Chen, H.-Y. Hung, Y.-Y. Ou, Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites, PLoS One 6 (3) (2011) e17331–NA.
[57] L. Jiang, Z. Cai, D. Wang, Improving naive Bayes for classification, Int. J. Comput. Appl. 32 (3) (2010) 328–332.
[58] M.N. Gibbs, D.J. MacKay, Variational Gaussian process classifiers, IEEE Trans. Neural Netw. 11 (6) (2000) 1458–1464.
[59] V. Korde, C.N. Mahender, Text classification and classifiers: A survey, Int. J. Artif. Intell. Appl. 3 (2) (2012) 85.
[60] A. Ng, M. Jordan, On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, Adv. Neural Inf. Process. Syst. 14 (2001).
[61] Y. Zhang, Support vector machine classification algorithm and its application, in: Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings, Part II 3, Springer, 2012, pp. 179–186.
[62] S. Tan, An effective refinement strategy for KNN text classifier, Expert Syst. Appl. 30 (2) (2006) 290–298.
[63] L. Tenenboim-Chekina, L. Rokach, B. Shapira, Identification of label dependencies for multi-label classification, in: Working Notes of the Second International Workshop on Learning from Multi-Label Data, Citeseer, 2010, pp. 53–60.
[64] W. He, Y. Wang, L. Cui, R. Su, L. Wei, Learning embedding features based on multisense-scaled attention architecture to improve the predictive performance of anticancer peptides, Bioinformatics 37 (24) (2021) 4684–4693.
[65] P. Agrawal, D. Bhagat, M. Mahalwal, N. Sharma, G.P. Raghava, AntiCP 2.0: an updated model for predicting anticancer peptides, Brief. Bioinform. 22 (3) (2021) bbaa153.
[66] L. Yao, W. Li, Y. Zhang, J. Deng, Y. Pang, Y. Huang, C.-R. Chung, J. Yu, Y.-C. Chiang, T.-Y. Lee, Accelerating the discovery of anticancer peptides through deep forest architecture with deep graphical representation, Int. J. Mol. Sci. 24 (5) (2023) 4328.
[67] P.B. Timmons, C.M. Hewage, ENNAACT is a novel tool which employs neural networks for anticancer activity classification for therapeutic peptides, Biomed. Pharmacother. 133 (2021) 111051.
[68] H. Wang, J. Zhao, H. Zhao, H. Li, J. Wang, CL-ACP: a parallel combination of CNN and LSTM anticancer peptide recognition model, BMC Bioinform. 22 (2021) 1–22.
[69] Z. Hajisharifi, M. Piryaiee, M.M. Beigi, M. Behbahani, H. Mohabatkar, Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test, J. Theoret. Biol. 341 (2014) 34–40.
[70] H.-C. Yi, Z.-H. You, X. Zhou, L. Cheng, X. Li, T.-H. Jiang, Z.-H. Chen, ACP-DL: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation, Mol. Ther.-Nucleic Acids 17 (2019) 1–9.
[71] A. Tyagi, P. Kapoor, R. Kumar, K. Chaudhary, A. Gautam, G. Raghava, In silico models for designing and discovering novel anticancer peptides, Sci. Rep. 3 (1) (2013) 2984.
[72] S. Vijayakumar, L. Ptv, ACPP: a web server for prediction and design of anti-cancer peptides, Int. J. Pept. Res. Ther. 21 (2015) 99–106.
[73] A.S. Ettayapuram Ramaprasad, S. Singh, R. Gajendra P. S, S. Venkatesan, AntiAngioPred: a server for prediction of anti-angiogenic peptides, PLoS One 10 (9) (2015) e0136990.
[74] S. Lata, B. Sharma, G.P. Raghava, Analysis and prediction of antibacterial peptides, BMC Bioinform. 8 (1) (2007) 1–10.
[75] B. Manavalan, T.H. Shin, M.O. Kim, G. Lee, AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest, Front. Pharmacol. 9 (2018) 276.
[76] N. Thakur, A. Qureshi, M. Kumar, AVPpred: collection and prediction of highly effective antiviral peptides, Nucleic Acids Res. 40 (W1) (2012) W199–W204.
[77] L. Wei, P. Xing, R. Su, G. Shi, Z.S. Ma, Q. Zou, CPPred-RF: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency, J. Proteome Res. 16 (5) (2017) 2044–2053.
[78] A. Rajput, A.K. Gupta, M. Kumar, Prediction and analysis of quorum sensing peptides based on sequence features, PLoS One 10 (3) (2015) e0120066.
[79] N. Li, J. Kang, L. Jiang, B. He, H. Lin, J. Huang, et al., PSBinder: a web service for predicting polystyrene surface-binding peptides, BioMed Res. Int. 2017 (2017).
[80] A. Tyagi, A. Tuknait, P. Anand, S. Gupta, M. Sharma, D. Mathur, A. Joshi, S. Singh, A. Gautam, G.P. Raghava, CancerPPD: a database of anticancer peptides and proteins, Nucleic Acids Res. 43 (D1) (2015) D837–D843.
[81] Z. Chen, P. Zhao, C. Li, F. Li, D. Xiang, Y.-Z. Chen, T. Akutsu, R.J. Daly, G.I. Webb, Q. Zhao, et al., iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization, Nucleic Acids Res. 49 (10) (2021) e60.
[82] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[83] J. Forcier, P. Bissex, W.J. Chun, Python Web Development with Django, Addison-Wesley Professional, 2008.
[84] W. Chen, H. Ding, P. Feng, H. Lin, K.-C. Chou, iACP: a sequence-based tool for identifying anticancer peptides, Oncotarget 7 (13) (2016) 16895.

[85] B. Manavalan, S. Basith, T.H. Shin, S. Choi, M.O. Kim, G. Lee, MLACP: machine-learning-based prediction of anticancer peptides, Oncotarget 8 (44) (2017) 77121.
[86] S. Kumar, H. Li, In silico design of anticancer peptides, in: Proteomics for Drug Discovery: Methods and Protocols, Springer, 2017, pp. 245–254.
[87] L. Xu, G. Liang, L. Wang, C. Liao, A novel hybrid sequence-based model for identifying anticancer peptides, Genes 9 (3) (2018) 158.
[88] N. Schaduangrat, C. Nantasenamat, V. Prachayasittikul, W. Shoombuatong, ACPred: a computational tool for the prediction and analysis of anticancer peptides, Molecules 24 (10) (2019) 1973.
[89] L. Wei, C. Zhou, R. Su, Q. Zou, PEPred-Suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning, Bioinformatics 35 (21) (2019) 4272–4280.
[90] C. Wu, R. Gao, Y. Zhang, Y. De Marinis, PTPD: predicting therapeutic peptides by deep learning and word2vec, BMC Bioinform. 20 (1) (2019) 1–8.
[91] B. Rao, C. Zhou, G. Zhang, R. Su, L. Wei, ACPred-Fuse: fusing multi-view information improves the prediction of anticancer peptides, Brief. Bioinform. 21 (5) (2020) 1846–1855.
[92] C.-R. Chung, T.-R. Kuo, L.-C. Wu, T.-Y. Lee, J.-T. Horng, Characterization and identification of antimicrobial peptides with different functional activities, Brief. Bioinform. 21 (3) (2020) 1098–1114.
[93] L. Yu, R. Jing, F. Liu, J. Luo, Y. Li, DeepACP: a novel computational approach for accurate identification of anticancer peptides by deep learning algorithm, Mol. Ther.-Nucleic Acids 22 (2020) 862–870.
[94] P. Charoenkwan, W. Chiangjong, V.S. Lee, C. Nantasenamat, M.M. Hasan, W. Shoombuatong, Improved prediction and characterization of anticancer activities of peptides using a novel flexible scoring card method, Sci. Rep. 11 (1) (2021) 3017.
[95] S. Ahmed, R. Muhammod, Z.H. Khan, S. Adilina, A. Sharma, S. Shatabda, A. Dehzangi, ACP-MHCNN: An accurate multi-headed deep-convolutional neural network to predict anticancer peptides, Sci. Rep. 11 (1) (2021) 23676.
[96] Z. Lv, F. Cui, Q. Zou, L. Zhang, L. Xu, Anticancer peptides prediction with deep representation learning features, Brief. Bioinform. 22 (5) (2021) bbab008.
[97] Y. Guo, K. Yan, H. Lv, B. Liu, PreTP-EL: prediction of therapeutic peptides based on ensemble learning, Brief. Bioinform. 22 (6) (2021) bbab358.
[98] J.-H. Jhong, L. Yao, Y. Pang, Z. Li, C.-R. Chung, R. Wang, S. Li, W. Li, M. Luo, R. Ma, et al., dbAMP 2.0: updated resource for antimicrobial peptides with an enhanced scanning method for genomic and proteomic data, Nucleic Acids Res. 50 (D1) (2022) D460–D470.
[99] Y. Liang, X. Ma, iACP-GE: accurate identification of anticancer peptides by using gradient boosting decision tree and extra tree, SAR QSAR Environ. Res. 34 (1) (2023) 1–19.
[100] M. Arif, S. Ahmed, F. Ge, M. Kabir, Y.D. Khan, D.-J. Yu, M. Thafar, Stack-ACPred: Prediction of anticancer peptides by integrating optimized multiple feature descriptors with stacked ensemble approach, Chemometr. Intell. Lab. Syst. 220 (2022) 104458.
[101] K. Yan, H. Lv, J. Wen, Y. Guo, Y. Xu, B. Liu, PreTP-Stack: prediction of therapeutic peptides based on the stacked ensemble learing, IEEE/ACM Trans. Comput. Biol. Bioinform. 20 (2) (2022) 1337–1344.
[102] L. Zhu, C. Ye, X. Hu, S. Yang, C. Zhu, ACP-check: An anticancer peptide prediction model based on bidirectional long short-term memory and multi-features fusion strategy, Comput. Biol. Med. 148 (2022) 105868.
[103] G. Feng, H. Yao, C. Li, R. Liu, R. Huang, X. Fan, R. Ge, Q. Miao, ME-ACP: Multi-view neural networks with ensemble model for identification of anticancer peptides, Comput. Biol. Med. 145 (2022) 105459.
[104] B. Han, N. Zhao, C. Zeng, Z. Mu, X. Gong, ACPred-BMF: bidirectional LSTM with multiple feature representations for explainable anticancer peptide prediction, Sci. Rep. 12 (1) (2022) 21915.
[105] Y.-Y. Sun, T.-T. Lin, W.-C. Cheng, I.-H. Lu, C.-Y. Lin, S.-H. Chen, Peptide-based drug predictions for cancer therapy using deep learning, Pharmaceuticals 15 (4) (2022) 422.
[106] Q. Yuan, K. Chen, Y. Yu, N.Q.K. Le, M.C.H. Chua, Prediction of anticancer peptides based on an ensemble model of deep learning and machine learning using ordinal positional encoding, Brief. Bioinform. 24 (1) (2023) bbac630.
[107] S.M. Azim, N.H.N. Sabab, I. Noshadi, H. Alinejad-Rokny, A. Sharma, S. Shatabda, I. Dehzangi, Accurately predicting anticancer peptide using an ensemble of heterogeneously trained classifiers, Inform. Med. Unlocked 42 (2023) 101348.
[108] W. Zhou, Y. Liu, Y. Li, S. Kong, W. Wang, B. Ding, J. Han, C. Mou, X. Gao, J. Liu, TriNet: A tri-fusion neural network for the prediction of anticancer and antimicrobial peptides, Patterns 4 (3) (2023).
[109] S. Garai, J. Thomas, P. Dey, D. Das, LGBM-ACp: an ensemble model for anticancer peptide prediction and in silico screening with potential drug targets, Mol. Divers. (2023) 1–17.
[110] E. Fazal, M.S. Ibrahim, S. Park, I. Naseem, A. Wahab, Anticancer peptides classification using kernel sparse representation classifier, IEEE Access 11 (2023) 17626–17637.
[111] M. Kabir, M. Arif, S. Ahmad, Z. Ali, Z.N.K. Swati, D.-J. Yu, Intelligent computational method for discrimination of anticancer peptides by incorporating sequential and evolutionary profiles information, Chemometr. Intell. Lab. Syst. 182 (2018) 158–165.
[112] Y.P. Zhang, Q. Zou, PPTPP: a novel therapeutic peptide prediction method using physicochemical property encoding and adaptive feature representation learning, Bioinformatics 36 (13) (2020) 3982–3987.
[113] C.H. Rodrigues, A. Garg, D. Keizer, D.E. Pires, D.B. Ascher, CSM-peptides: A computational approach to rapid identification of therapeutic peptides, Prot. Sci. 31 (10) (2022) e4442.
[114] B. An, J. Lyu, Z. Wang, C. Li, C. Hu, F. Tan, R. Zhang, Y. Hu, C. Chen, Repulsive attention: Rethinking multi-head attention as bayesian inference, 2020, arXiv preprint arXiv:2009.09364.
[115] L. Zheng, S. Shi, M. Lu, P. Fang, Z. Pan, H. Zhang, Z. Zhou, H. Zhang, M. Mou, S. Huang, et al., AnnoPRO: a strategy for protein function annotation based on multi-scale protein representation and a hybrid deep learning of dual-path encoding, Genome Biol. 25 (1) (2024) 41.
[116] M. Mou, Z. Pan, Z. Zhou, L. Zheng, H. Zhang, S. Shi, F. Li, X. Sun, F. Zhu, A transformer-based ensemble framework for the prediction of protein–protein interaction sites, Research 6 (2023) 0240.
[117] J. Hong, Y. Luo, Y. Zhang, J. Ying, W. Xue, T. Xie, L. Tao, F. Zhu, Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning, Brief. Bioinform. 21 (4) (2020) 1437–1447.
[118] J. Hong, Y. Luo, M. Mou, J. Fu, Y. Zhang, W. Xue, T. Xie, L. Tao, Y. Lou, F. Zhu, Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery, Brief. Bioinform. 21 (5) (2020) 1825–1836.
[119] W. Xia, L. Zheng, J. Fang, F. Li, Y. Zhou, Z. Zeng, B. Zhang, Z. Li, H. Li, F. Zhu, PFmulDL: a novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods, Comput. Biol. Med. 145 (2022) 105465.
[120] A.V. Singh, A. Romeo, K. Scott, S. Wagener, L. Leibrock, P. Laux, A. Luch, P. Kerkar, S. Balakrishnan, S.P. Dakua, et al., Emerging technologies for in vitro inhalation toxicology, Adv. Healthc. Mater. 10 (18) (2021) 2100633.
[121] A.V. Singh, V. Chandrasekar, P. Laux, A. Luch, S.P. Dakua, P. Zamboni, A. Shelar, Y. Yang, V. Pandit, V. Tisato, et al., Micropatterned neurovascular interface to mimic the blood–brain barrier's neurophysiology and micromechanical function: a BBB-on-CHIP model, Cells 11 (18) (2022) 2801.
[122] V. Chandrasekar, A.V. Singh, R.S. Maharjan, S.P. Dakua, S. Balakrishnan, S. Dash, P. Laux, A. Luch, S. Singh, M. Pradhan, Perspectives on the technological aspects and biomedical applications of virus-like particles/nanoparticles in reproductive biology: Insights on the medicinal and toxicological outlook, Adv. NanoBiomed Res. 2 (8) (2022) 2200010.
[123] A.V. Singh, P. Laux, A. Luch, S. Balkrishnan, S.P. Dakua, Bottom-UP assembly of nanorobots: extending synthetic biology to complex material design, Front. Nanosci. Nanotechnol. 5 (1) (2019).