**1. Benchmark features for peptide embeddings**

1.1 Amino acid composition (AAC)

AAC is the percentage of standard amino acids with a fixed length of 20 features where each feature represents the percentage of one standard amino acid in the given peptide.

1.2 Dipeptide composition (DPC)

DPC is the rate of dipeptides normalized by all possible dipeptide combinations with a fixed length of 400 features where each feature represents the percentage of one dipeptide part occurring in the given peptide.

1.3 Pseudo amino acid composition (PseAAC)

The PseAAC considers both the sequence-ordering information within a local range and the global sequence-ordering effects (Lertampaiporn et al., 2022). Peptide sequence can be encoded to a (20+d)-dimension vector of numeric observations for d-tier correlation structure. In this study, 21 dimension features of PseAAC (lambda=1 and weight=0.05) were adopted since some ACE inhibitory peptides are dipeptides.

1.4 Amino acid Index (AAI)

The AAI database contains a total of 566 numerical values of amino acids including physicochemical properties and biochemical properties of amino acids and pairs of amino acids (Kawashima et al., 2008). Eight high-quality AAIs were identified by clustering, whose accession codes in the AAI database are BIOV880101, BLAM930101, CEDJ970104, LIFS790101, MAXF760101, NAKH920108, TSAJ990101 and MIYS990104 (Manavalan et al., 2019). Since the smallest peptides in our ACE inhibitory peptide dataset is dipeptides, only the features of the N terminus and the C terminus were extracted for peptide embeddings. each peptide will be represented by a fixed length of 16 vectors.

1.5 Overlapping property features (OPF)

The physicochemical properties of amino acids are clustered into ten groups. However, there is still a chance of overlapping a different group. To show the relationship of varied properties, we computed a 10-bit vector comprised of 0/1 to depict each amino acid of a given peptide. If a residue of the peptide belongs to each property, the parameter will be set to 1, else 0. Since the smallest peptides in our ACE inhibitory peptide dataset is dipeptides, only the features of the N terminus and the C terminus were extracted for peptide embeddings. each peptide will be represented by a fixed length of 20 vectors.

1.6 Binary profile/one-hot encoding (BP)

For a given length peptide with k residues, the i-th (i < k) residue has 20 different variations in standard amino acids, which can be converted into a 20-dimensional vector for the i-th peptide. Then, a peptide with k residues will be represented by a 20\*k matrix. Considering peptides in our dataset have different lengths (maximum length is 28 residues), we set an empty 20\*30 matrix for peptide representation. The i-th row represents i-th residue in the given peptide. For example, the second residue arginine can be represented by the second row (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) of the 20\*30 matrix, where the location of the 1 represents the arginine. Finally, the matrix is reshaped as a 600\*1 vector for model development.

1.7 Adaptive skip dipeptide composition (ASDC)

The ASDC is a modified dinucleotide composition, which sufficiently considers the correlation information present not only between adjacent residues but also between intervening residues (Chen et al., 2022). It is a fixed length of 400 features. For given a sequence, the feature vector for ASDC is represented by:

where *fvi* is calculated by

where *fvi* denotes the occurrence frequency of all possible dipeptide with ≤ *L*-1 intervening nucleotides (Chen et al., 2022).

1.8 Composition-transition-distribution (CTD)

CTD is employed to delineate the global composition of amino acid properties for a given protein or peptide sequence (Manavalan et al., 2019). Standard amino acids are divided into thirteen different attributes, including hydrophobicity, normalized Van der Waals Volume, polarity, polarizability, charge, secondary structures, and solvent accessibility (below Table 1) (Chen et al., 2022).

**Table 1.** Amino acid physicochemical attributes and the division of the amino acids into three groups according to each attribute.

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Division** | | |
| Hydrophobicity\_PRAM900101 | Polar: RKEDQN | Neutral: GASTPHY | Hydrophobicity: CLVIMFW |
| Hydrophobicity\_ARGP820101 | Polar: QSTNGDE | Neutral: RAHCKMV | Hydrophobicity: LYPFIW |
| Hydrophobicity\_ZIMJ680101 | Polar: QNGSWTDERA | Neutral: HMCKV | Hydrophobicity: LPFYI |
| Hydrophobicity\_PONP930101 | Polar: KPDESNQT | Neutral: GRHA | Hydrophobicity: YMFWLCVI |
| Hydrophobicity\_CASG920101 | Polar: KDEQPSRNTG | Neutral: AHYMLV | Hydrophobicity: FIWC |
| Hydrophobicity\_ENGD860101 | Polar: RDKENQHYP | Neutral :SGTAW | Hydrophobicity: CVLIMF |
| Hydrophobicity\_FASG890101 | Polar: KERSQD | Neutral: NTPG | Hydrophobicity: AYHWVMFLIC |
| Normalized van der Waals volume | Volume range: 0-2.78  GASTPD | Volume range: 2.95-94.0  NVEQIL | Volume range: 4.03-8.08  MHKFRYW |
| Polarity | Polarity value: 4.9-6.2  LIFWCMVY | Polarity value: 8.0-9.2  PATGS | Polarity value: 10.4-13.0  HQRKNED |
| Polarizability | Polarizability value: 0-1.08  GASDT | Polarizability value: 0.128-120.186  GPNVEQIL | Polarizability value: 0.219-0.409  KMHFRYW |
| Charge | Positive: KR | Neutral: ANCQGHILMFPSTWYV | Negative: DE |
| Secondary structure | Helix: EALMQKRH | Strand: VIYCWFT | Coil: GNPSD |
| Solvent accessibility | Buried: ALFCGIVW | Exposed: PKQEND | Intermediate: MPSTHY |

Note: the table information is from the supplementary of the study of Chen et al., 2022.

1.8.1 Composition-transition-distribution composition (CTDC)

CTDC computes the percentage composition values of the above thirteen different attributes from a given peptide sequence, which will generate a fixed length of 39 vectors (Chen et al., 2022).

1.8.2 Composition-transition-distribution transition (CTDT)

CTDT computes the percentage composition values of the transition from one cluster to another cluster among the above thirteen different attributes from a given peptide sequence, which will generate a fixed length of 39 vectors. For example, it will generate three feature groups: polar and neural, neural and hydrophobicity, and hydrophobicity and polar in terms of the attribute “Hydrophobicity\_PRAM900101”. Then the given peptide will check its dipeptide compositions with consideration of their feature group for the attribute “Hydrophobicity\_PRAM900101”

the features of, count the total numbers, and calculate the percentage as its features regarding to “Hydrophobicity\_PRAM900101” (Chen et al., 2022).

1.8.3 Composition-transition-distribution distribution (CTDD)

CTDD also computes the percentage composition values of the above thirteen different attributes from a given peptide sequence. The given peptide sequence is used to generate the fraction of the entire sequence, where 25, 50, 75, and 100% of occurrences are contained. Then the CTDC values of each fraction are calculated and concatenated as the peptide representation with a fixed length of 195 vectors (Chen et al., 2022).

1.9 Grouped amino acid composition (GAAC)

GAAC computes the percentage composition the residues belonging to five groups, including the aliphatic group (*g1*: GAVLMI), aromatic group (*g2*: FYW), positive charge group (*g3*: KRH), negative charged group (*g4*: DE) and uncharged group (*g5*: STCPNQ), where the given peptide is represented by a fixed length of 5 vectors (Chen et al., 2022).

1.10 Grouped dipeptide composition (GDPC)

GDPC combines the GAAC and DPC for peptide representation, where the dipeptide composition type encrypted in the given peptide was counted. There are 25 different combinations for the five groups in GAAC and generate a fixed length of 25 vectors (Chen et al., 2022).