Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A systematic framework for identifying amphipathic cationic motifs in the human proteome reveals a distinct non-random physicochemical class

Taiki Kojima

koji3744@hotmail.com

Kojima Family Clinic https://orcid.org/0000-0003-2069-6079

Research Article

Keywords: antimicrobial peptides, amphipathic helix, human proteome, hydrophobic moment, computational analysis

Posted Date: March 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9206301/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Short amphipathic helices are widely recognized as functional motifs involved in membrane interaction and antimicrobial activity. However, their global distribution and structural characteristics within the human proteome remain incompletely understood.

In this study, we systematically analyzed the human proteome (~ 20,000 proteins) using a fixed sliding window of 12 residues to identify candidate amphipathic motifs. Each segment was evaluated based on hydrophobic moment, mean hydrophobicity, and net charge, followed by stringent physicochemical filtering.

This approach reduced the search space to 475 candidate motifs exhibiting strong amphipathic properties, characterized by high positive charge (≥ + 6) and moderate hydrophobic moment (~ 0.45–0.55). Functional annotation revealed enrichment in membrane-associated proteins, while a substantial fraction of candidates remained unclassified, suggesting the presence of previously unrecognized functional elements.

Positional analysis demonstrated a clear bias toward internal (middle) regions of proteins, supporting a role as structural or interaction modules rather than signal peptides. Comparison with randomly sampled peptide segments confirmed that the identified motifs occupy a distinct physicochemical space, indicating that their properties are not attributable to random sequence composition.

These findings demonstrate that amphipathic cationic motifs with antimicrobial-like physicochemical signatures are widely embedded within human proteins. This systematic framework provides a basis for identifying cryptic functional elements and may inform the design of antimicrobial peptides and membrane-interacting biomolecules.

Introduction

Antimicrobial peptides (AMPs) are fundamental components of innate immunity and represent a promising class of therapeutic agents due to their broad-spectrum activity and membrane-targeting mechanisms (1, 2). A key structural feature underlying their function is the formation of amphipathic α-helices, which enable selective interaction with lipid membranes (3, 4). These physicochemical properties—particularly cationicity and amphipathicity—are central to peptide-mediated membrane disruption and host defense functions (5).

Interestingly, antimicrobial activity is not restricted to classical AMPs. Several studies have demonstrated that functional peptide fragments with antimicrobial or membrane-interacting properties can be derived from larger human proteins. A notable example is apolipoprotein E (ApoE), from which cationic amphipathic peptides exhibiting antibacterial and immunomodulatory activity have been identified (6–8). These observations suggest that short functional motifs embedded within larger proteins may represent a previously underappreciated source of bioactive peptides.

The amphipathic character of peptide segments can be quantitatively described using the hydrophobic moment, a metric that captures the periodic distribution of hydrophobic residues along an α-helix (9). This concept has been widely applied to identify membrane-interacting domains in proteins, including receptor-binding regions in apolipoproteins (10).

Despite these advances, a systematic, proteome-wide identification of short amphipathic motifs with defined physicochemical properties remains lacking. In particular, it is unclear whether such motifs are rare, protein-specific features or represent a broadly distributed structural module embedded across the human proteome.

In this study, we systematically scanned the human proteome (~ 20,000 proteins) using a fixed-length sliding window approach to identify short amphipathic motifs. By integrating hydrophobic moment, mean hydrophobicity, and net charge filtering, we reduced the search space to 475 candidate motifs with highly constrained physicochemical characteristics. This large-scale reduction highlights the selective enrichment of such motifs.

Our results demonstrate that amphipathic, cationic peptide segments with antimicrobial-like physicochemical signatures are not isolated features but are widely embedded across diverse human proteins. These findings support the concept that ApoE-derived peptides represent one example of a broader class of cryptic functional motifs, and suggest that such motifs may constitute a generalizable structural module relevant to membrane interaction and peptide design.

Methods

Dataset

The human proteome dataset was obtained from UniProt (https://www.uniprot.org), comprising all reviewed human protein sequences (Homo sapiens, Swiss-Prot). In total, approximately 20,000 protein sequences were included. Protein sequences were downloaded in FASTA/TSV format and used as input for subsequent analysis.

Sliding window extraction

Each protein sequence was scanned using a fixed-length sliding window of 12 amino acids. The window was shifted by one residue at a time across the full length of each protein sequence, generating overlapping peptide segments. This procedure produced a comprehensive set of candidate segments covering the entire proteome.

Feature calculation

For each 12-residue segment, the following physicochemical properties were calculated:

1. Hydrophobic moment (HM)

The hydrophobic moment was calculated according to Eisenberg et al. (1984) (9), assuming an α-helical structure with a periodicity of 100° per residue. This metric quantifies the amphipathic nature of the peptide.

2. Mean hydrophobicity

The average hydrophobicity of each segment was computed using the Eisenberg hydrophobicity scale.

3. Net charge

Net charge was calculated as the total number of positively charged residues (Arg, Lys) minus negatively charged residues (Asp, Glu) within each segment.

Filtering criteria

Candidate motifs were selected using the following thresholds:

- Hydrophobic moment ≥ 0.45

- Mean hydrophobicity ≤ −0.40

- Net charge ≥ +6

These criteria were chosen to enrich for amphipathic, cationic peptide segments with potential membrane-interacting or antimicrobial-like properties.

Motif selection

For each protein, candidate segments satisfying the filtering criteria were identified. When multiple segments were detected within a single protein, a representative segment was selected based on the highest hydrophobic moment, with net charge used as a secondary criterion.

Functional annotation

Proteins containing candidate motifs were annotated using UniProt metadata, including:

- Presence of signal peptides

- Subcellular localization

- Transmembrane regions

- Functional annotations (receptor, enzyme)

Proteins not assigned to predefined categories were classified as “unclassified”.

Cluster classification

Each candidate motif was assigned to a functional cluster based on the annotation of its parent protein. The distribution of motifs across clusters was analyzed to assess enrichment patterns.

Positional analysis

The relative position of each motif within its parent protein was normalized as:

position ratio = motif start position / protein length

Motifs were categorized into three regions:

- N-terminal (< 0.2)

- Middle (0.2–0.8)

- C-terminal (> 0.8)

This classification enabled assessment of positional bias.

Comparison with randomly sampled peptide segments

To determine whether the identified motifs represent a non-random subset of peptide sequences, we performed a comparative analysis using randomly sampled segments from the same proteome. An equal number of peptide segments (length = 12) were randomly extracted from the human proteome, ensuring identical length and source distribution.

For each random segment, physicochemical properties—including net charge and hydrophobic moment—were calculated using the same computational procedures applied to the identified motifs.

The resulting distributions were directly compared to those of the identified motifs to evaluate enrichment and deviation from background sequence properties.

Statistical analysis

All computations were performed in Python using pandas and numpy. Visualization was performed using matplotlib. Distributions were compared qualitatively to assess enrichment patterns.

Reproducibility

All analyses were implemented using custom Python scripts, including data preprocessing, sliding window extraction, feature calculation, filtering, annotation, and visualization. All intermediate and final results were saved as CSV and PNG files. The analysis pipeline is fully reproducible and can be executed using the provided scripts and parameters.

Results

Identification of amphipathic motifs in the human proteome

Using a fixed-length sliding window of 12 residues (Figure 1), we systematically scanned the human proteome (~20,000 proteins) to identify peptide segments with amphipathic and cationic properties. Each segment was evaluated based on hydrophobic moment, mean hydrophobicity, and net charge. After applying stringent filtering criteria, a total of 475 candidate motifs were identified, representing a substantial reduction of the initial search space.

Functional distribution of candidate proteins

Functional annotation of proteins containing candidate motifs revealed a broad distribution across categories (Figure 2). While a substantial fraction of motifs was found in proteins without clear functional classification ("other"), a notable enrichment was observed in membrane-associated proteins. Additional categories included enzyme-like, signal peptide-containing, secreted, and receptor-like proteins.

The large proportion of motifs classified as "other" suggests that many candidate motifs reside in proteins not currently annotated with membrane-interacting or antimicrobial functions. This finding indicates the potential presence of previously uncharacterized functional elements embedded within diverse protein contexts.

Physicochemical characteristics of identified motifs

All candidate motifs were extracted using a fixed window size of 12 residues (Figure 3A), ensuring consistent comparison across segments. Analysis of net charge distribution demonstrated that the majority of motifs were highly cationic, with strong enrichment at +6 or higher (Figure 3B).

Hydrophobic moment analysis revealed a relatively narrow distribution centered around approximately 0.45–0.55 (Figure 3C), consistent with the expected range for amphipathic α-helices. These results indicate that the identified motifs share a highly constrained physicochemical profile characterized by strong positive charge and moderate amphipathicity.

Positional bias of motifs within proteins

To investigate the localization of motifs within protein sequences, positional analysis was performed by normalizing motif positions relative to full-length protein length (Figure 4). A strong bias toward internal (middle) regions was observed, whereas fewer motifs were located near the N-terminal or C-terminal ends.

This distribution suggests that the identified motifs are more likely to function as internal structural or interaction modules rather than classical signal peptides or terminal targeting sequences.

Representative motif analysis

To further assess the biological relevance of the identified motifs, we examined representative high-scoring candidates in detail. Several motifs, including those derived from PKDREJ, DVL1, and SPEM2, exhibited strongly cationic and amphipathic sequences (e.g., RLIRKHKRRRRK), consistent with known antimicrobial and membrane-interacting peptides.

These motifs were located in structurally exposed or functionally relevant regions, suggesting that they may represent previously unrecognized functional elements embedded within larger proteins.

Comparison with random peptide segments

To determine whether the observed physicochemical properties were specific to the identified motifs, we compared their distributions with those of randomly generated peptide segments of identical length.

Random peptides exhibited broad distributions of net charge centered around neutral values, whereas the identified motifs showed a strong enrichment of highly cationic sequences, predominantly ≥ +6 (Figure 5A).

Similarly, hydrophobic moment values of random peptides were widely distributed, while the identified motifs formed a distinct cluster within a narrow range (~0.45–0.55), consistent with amphipathic α-helical structures (Figure 5B).

These results demonstrate that the identified motifs occupy a well-defined physicochemical space that is not recapitulated by random peptide sequences, supporting their non-random occurrence and potential biological relevance.

Results

Identification of amphipathic motifs in the human proteome

Using a fixed-length sliding window of 12 residues (Fig. 1), we systematically scanned the human proteome (~ 20,000 proteins) to identify peptide segments with amphipathic and cationic properties. Each segment was evaluated based on hydrophobic moment, mean hydrophobicity, and net charge. After applying stringent filtering criteria, a total of 475 candidate motifs were identified, representing a substantial reduction of the initial search space.

Functional distribution of candidate proteins

Functional annotation of proteins containing candidate motifs revealed a broad distribution across categories (Fig. 2). While a substantial fraction of motifs was found in proteins without clear functional classification ("other"), a notable enrichment was observed in membrane-associated proteins. Additional categories included enzyme-like, signal peptide-containing, secreted, and receptor-like proteins.

The large proportion of motifs classified as "other" suggests that many candidate motifs reside in proteins not currently annotated with membrane-interacting or antimicrobial functions. This finding

indicates the potential presence of previously uncharacterized functional elements embedded within diverse protein contexts.

Physicochemical characteristics of identified motifs

All candidate motifs were extracted using a fixed window size of 12 residues (Fig. 3A), ensuring consistent comparison across segments. Analysis of net charge distribution demonstrated that the majority of motifs were highly cationic, with strong enrichment at + 6 or higher (Fig. 3B).

Hydrophobic moment analysis revealed a relatively narrow distribution centered around approximately 0.45–0.55 (Fig. 3C), consistent with the expected range for amphipathic α-helices. These results indicate that the identified motifs share a highly constrained physicochemical profile characterized by strong positive charge and moderate amphipathicity.

Positional bias of motifs within proteins

To investigate the localization of motifs within protein sequences, positional analysis was performed by normalizing motif positions relative to full-length protein length (Fig. 4). A strong bias toward internal (middle) regions was observed, whereas fewer motifs were located near the N-terminal or C-terminal ends.

This distribution suggests that the identified motifs are more likely to function as internal structural or interaction modules rather than classical signal peptides or terminal targeting sequences.

Representative motif analysis

To further assess the biological relevance of the identified motifs, we examined representative high-scoring candidates in detail. Several motifs, including those derived from PKDREJ, DVL1, and SPEM2, exhibited strongly cationic and amphipathic sequences (e.g., RLIRKHKRRRRK), consistent with known antimicrobial and membrane-interacting peptides.

These motifs were located in structurally exposed or functionally relevant regions, suggesting that they may represent previously unrecognized functional elements embedded within larger proteins.

Comparison with random peptide segments

To determine whether the observed physicochemical properties were specific to the identified motifs, we compared their distributions with those of randomly generated peptide segments of identical length.

Random peptides exhibited broad distributions of net charge centered around neutral values, whereas the identified motifs showed a strong enrichment of highly cationic sequences, predominantly ≥ + 6 (Fig. 5A).

Similarly, hydrophobic moment values of random peptides were widely distributed, while the identified motifs formed a distinct cluster within a narrow range (~ 0.45–0.55), consistent with amphipathic α-helical structures (Fig. 5B).

These results demonstrate that the identified motifs occupy a well-defined physicochemical space that is not recapitulated by random peptide sequences, supporting their non-random occurrence and potential biological relevance.

Discussion

In this study, we performed a systematic analysis of the human proteome (~ 20,000 proteins) to identify short peptide segments with amphipathic and cationic properties. By combining sliding window extraction with physicochemical filtering, we identified 475 candidate motifs that share highly constrained structural features.

Importantly, the use of a fixed window size of 12 residues (Fig. 1) defines the analytical framework of this study. Therefore, the observed uniform motif length reflects the design of the screening approach rather than an intrinsic biological constraint. Nevertheless, the convergence of multiple independent properties—including high positive charge, moderate hydrophobic moment, and consistent positional bias—strongly suggests that these motifs represent a coherent structural class.

One of the most striking findings is the enrichment of motifs within membrane-associated proteins (Fig. 2), supporting the hypothesis that these sequences function in membrane interaction. This is consistent with previous studies demonstrating that amphipathic helices play a central role in membrane binding and disruption (3, 4).

At the same time, a large fraction of motifs was found in proteins classified as "other," indicating that many of these sequences are not currently annotated as membrane-interacting or antimicrobial. This observation suggests that the human proteome may contain a substantial number of cryptic functional motifs that have not yet been characterized.

The comparison with randomly generated peptide segments further strengthens this interpretation. The identified motifs occupy a distinct and narrowly defined physicochemical space characterized by high positive charge and moderate hydrophobic moment, which is not recapitulated by random sequences (Fig. 5). This finding indicates that the observed motif properties are unlikely to arise by chance and instead reflect a specific structural and functional constraint within protein sequences.

The relevance of such motifs is further supported by prior work on apolipoprotein E (ApoE), in which short cationic amphipathic peptides derived from the protein exhibit antimicrobial and immunomodulatory activity (6–8). In our analysis, ApoE itself was identified among the 475 candidate proteins, and its corresponding 12-residue segment satisfied the same physicochemical criteria, including high net positive charge and amphipathic character.

This observation indicates that ApoE is not an isolated example, but rather represents one instance of a broader class of embedded cationic amphipathic motifs distributed throughout the human proteome.

Our findings therefore extend previous observations by demonstrating that ApoE-like functional segments are part of a widespread structural motif shared across diverse proteins.

Positional analysis revealed a strong bias toward internal regions of proteins (Fig. 4), suggesting that these motifs may function as embedded structural modules rather than independent secreted peptides. This distinguishes them from classical antimicrobial peptides and points toward a broader functional role, potentially involving transient membrane interactions or protein–lipid interfaces.

Taken together, our results support the existence of a widespread class of amphipathic, cationic motifs embedded within human proteins. These motifs may represent a generalizable membrane-interacting module with potential applications in peptide engineering, antimicrobial design, and functional annotation of proteins.

Several limitations of this study should be acknowledged. First, the analysis is based on computational filtering and does not directly demonstrate biological activity. Second, the use of a fixed window size may exclude motifs of different lengths. Third, functional annotations are incomplete, particularly within the "other" category, which may include both false positives and undiscovered functional elements.

Future studies should focus on experimental validation of selected motifs, structural characterization, and investigation of their biological roles in cellular contexts.

References

1. Hancock REW, Sahl HG (2006) Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. Nat Biotechnol 24(12):1551–1557. 10.1038/nbt1267
2. Zasloff M (2002) Antimicrobial peptides of multicellular organisms. Nature 415(6870):389–395. 10.1038/415389a
3. Wimley WC (2010) Describing the mechanism of antimicrobial peptide action with the interfacial activity model. ACS Chem Biol 5(10):905–917. 10.1021/cb1001558
4. Fjell CD, Hiss JA, Hancock REW, Schneider G (2012) Designing antimicrobial peptides: form follows function. Nat Rev Drug Discov 11(1):37–51. 10.1038/nrd3591
5. Wang G (2014) Human antimicrobial peptides and proteins. Pharmaceuticals (Basel) 7(5):545–594. 10.3390/ph7050545
6. Azuma M, Kojima T, Yokoyama I, Tajiri H, Yoshikawa K, Saga S, Del Carpio CA (2000) A synthetic peptide of human apoprotein E with antibacterial activity. Peptides 21(3):327–330. 10.1016/S0196-9781(00)00165-0
7. Pane K, Sgambati V, Zanfardino A et al (2016) A new cryptic cationic antimicrobial peptide from human apolipoprotein E with antibacterial activity and immunomodulatory effects on human cells. FEBS J 283(11):2115–2131. 10.1111/febs.13725

8. Zhao YY, Ding C, Shi T et al (2023) ApoE mimetic peptide COG1410 exhibits strong additive efficacy against Mycobacterium smegmatis with first-line antituberculosis drugs. Eur J Med Chem 247:115012 PMID:37013167
9. Eisenberg D, Weiss RM, Terwilliger TC (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. Proc Natl Acad Sci U S A 81(1):140–144 PMID:6582470
10. De Loof H, Rosseneu M, Brasseur R, Ruysschaert JM (1986) Use of hydrophobicity profiles to predict receptor binding domains on apolipoprotein E and the low density lipoprotein apolipoprotein B-E receptor. Proc Natl Acad Sci U S A 83(8):2295–2299. 10.1073/pnas.83.8.2295

Figures

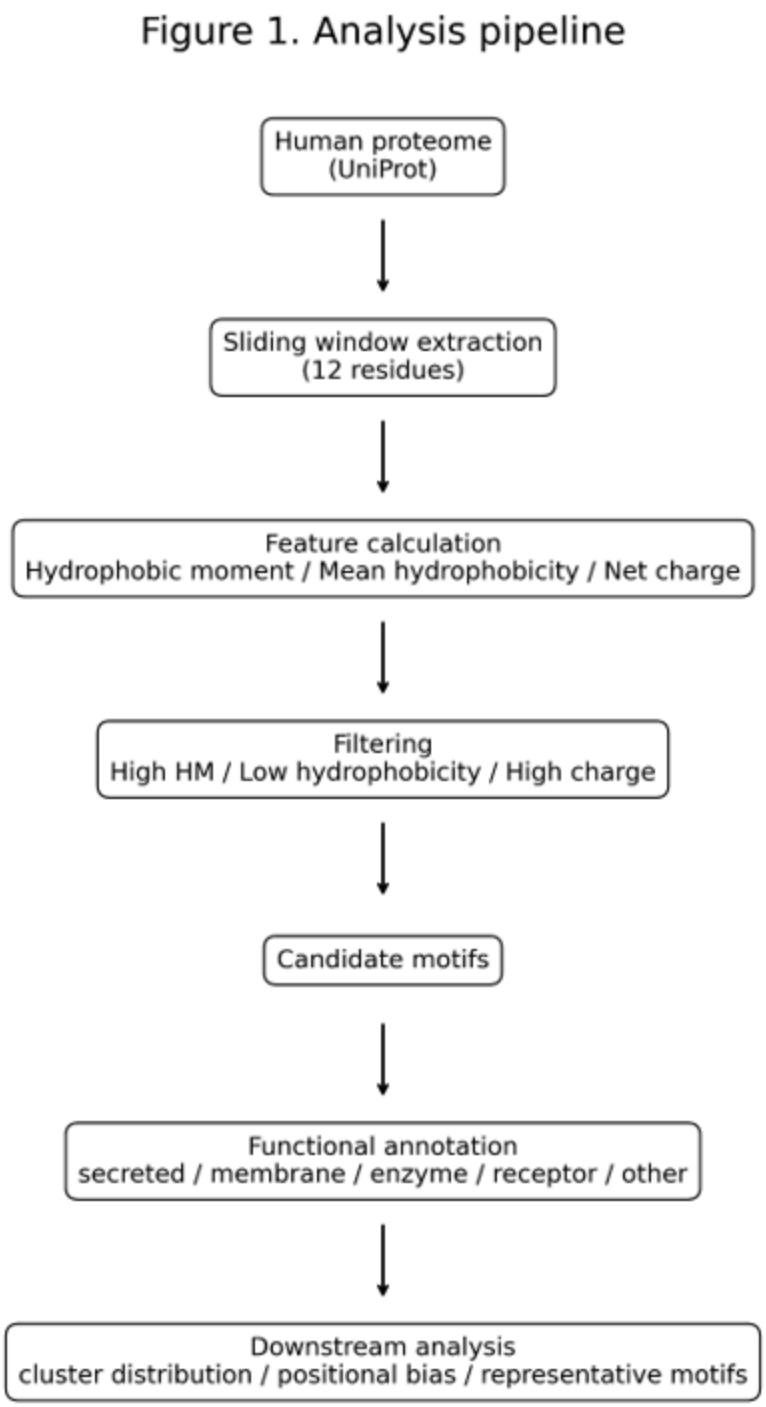


Figure 1

Overview of the analysis pipeline used to identify amphipathic motifs in the human proteome. Protein sequences from the human proteome (UniProt) were systematically scanned using a sliding window approach with a fixed length of 12 residues. For each segment, physicochemical properties including hydrophobic moment, mean hydrophobicity, and net charge were calculated.
Candidate motifs were selected based on predefined thresholds for high hydrophobic moment, low hydrophobicity, and high net positive charge. The resulting candidate set was further annotated using

available functional classifications (secreted, membrane-associated, enzyme-like, receptor-like, or other). Downstream analyses included functional category distribution, positional bias within protein sequences, and identification of representative motifs.

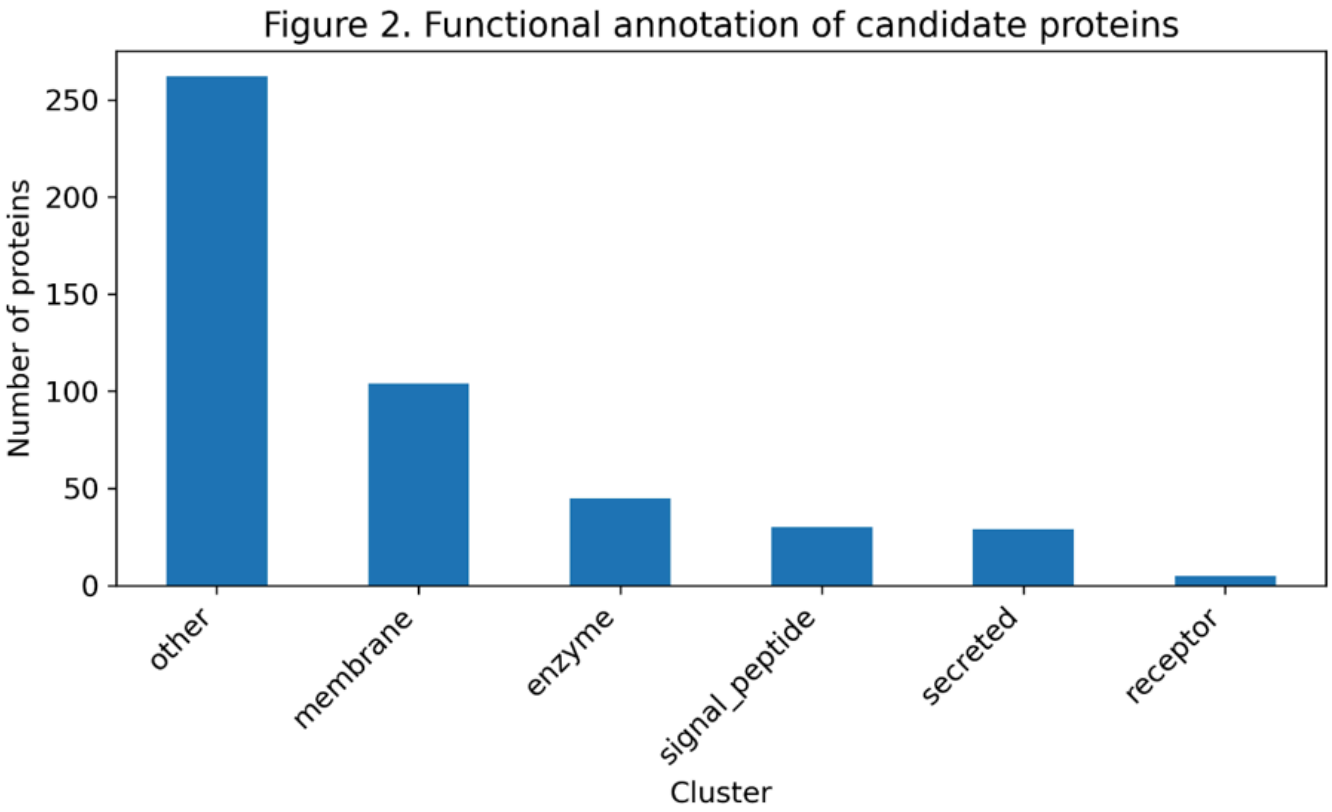


Figure 2

Functional annotation of candidate proteins containing amphipathic motifs.
The distribution of candidate motifs across functional categories is shown. While a substantial fraction of motifs was found in proteins without clear functional classification ("other"), a notable enrichment was observed in membrane-associated proteins.
This distribution suggests that amphipathic motifs are not restricted to well-characterized functional groups and may represent broadly distributed structural or interaction modules.

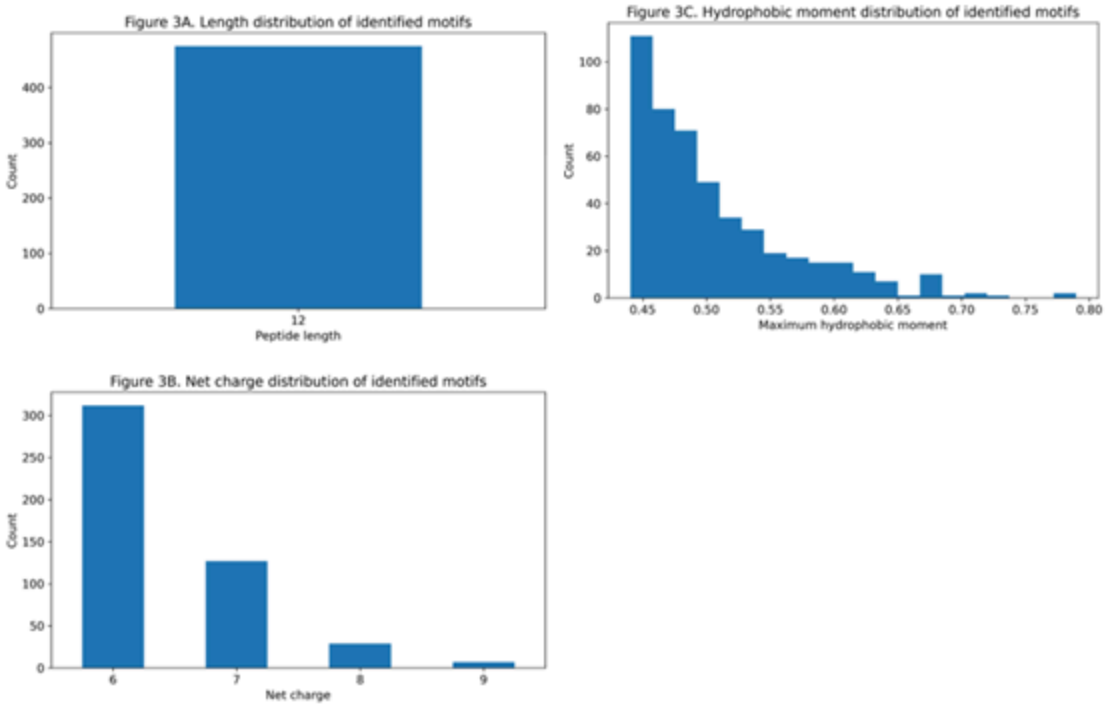


Figure 3

Physicochemical properties

Figure 3A. Length distribution of identified motifs.
All candidate motifs were extracted using a fixed window size of 12 residues. The uniform length reflects the analytical design and ensures consistent comparison of physicochemical properties across motifs.

Figure 3B. Net charge distribution of identified motifs.
The majority of motifs exhibited high net positive charge (≥ +6), indicating a strong enrichment of cationic residues. This property is consistent with known membrane-interacting and antimicrobial peptides.

Figure 3C. Hydrophobic moment distribution of identified motifs.
Candidate motifs showed a narrow distribution of hydrophobic moment values, primarily within the range of approximately 0.45–0.55. This range is characteristic of amphipathic α-helices, supporting the structural consistency of the identified motifs.

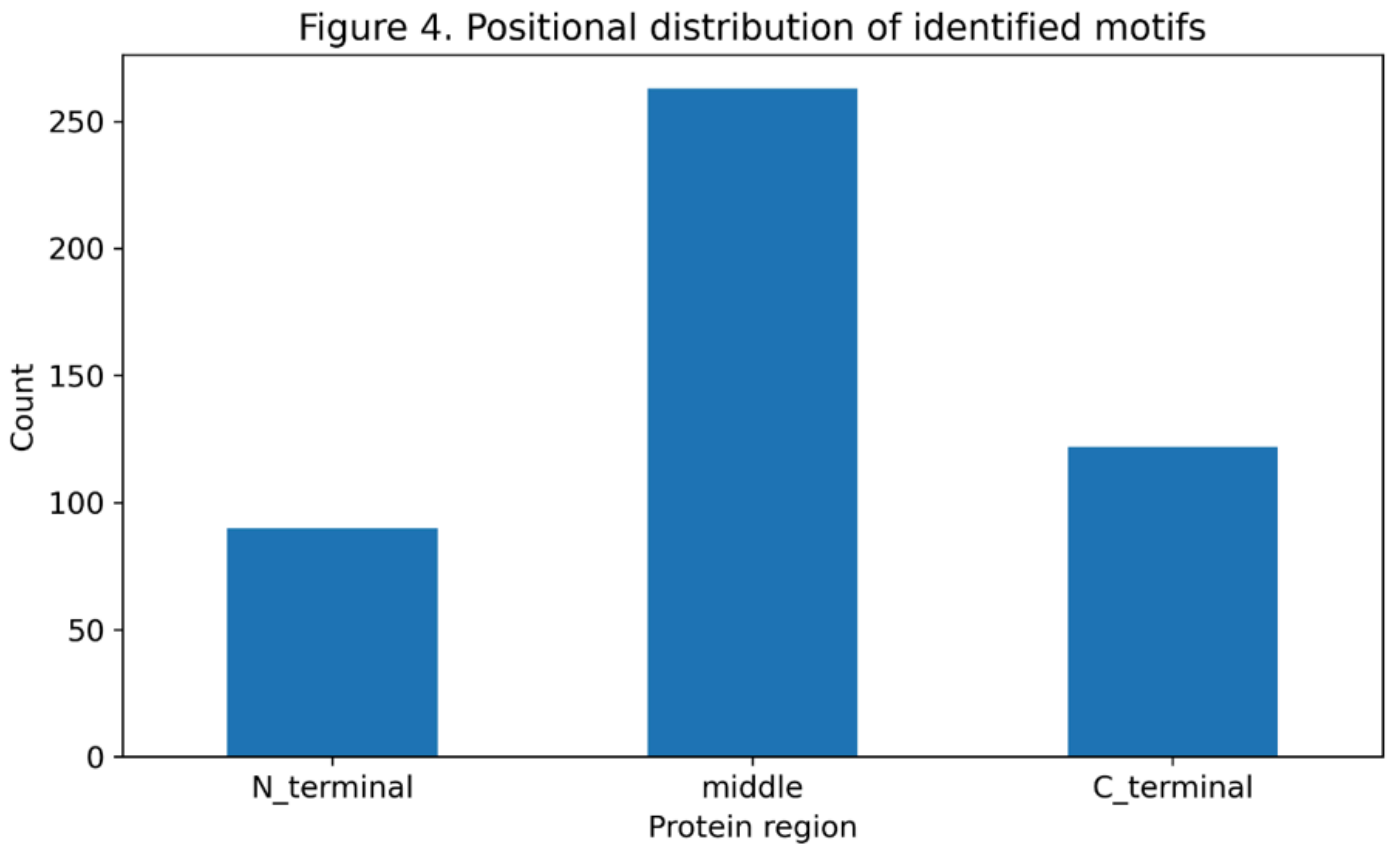


Figure 4

Positional distribution of identified motifs within proteins.
Motifs were categorized based on their relative position within the full-length protein sequence. A strong bias toward internal (middle) regions was observed, while fewer motifs were located near the N-terminal or C-terminal ends.
This distribution suggests that the identified motifs are more likely to function as internal structural or interaction modules rather than signal peptides or terminal targeting sequences.

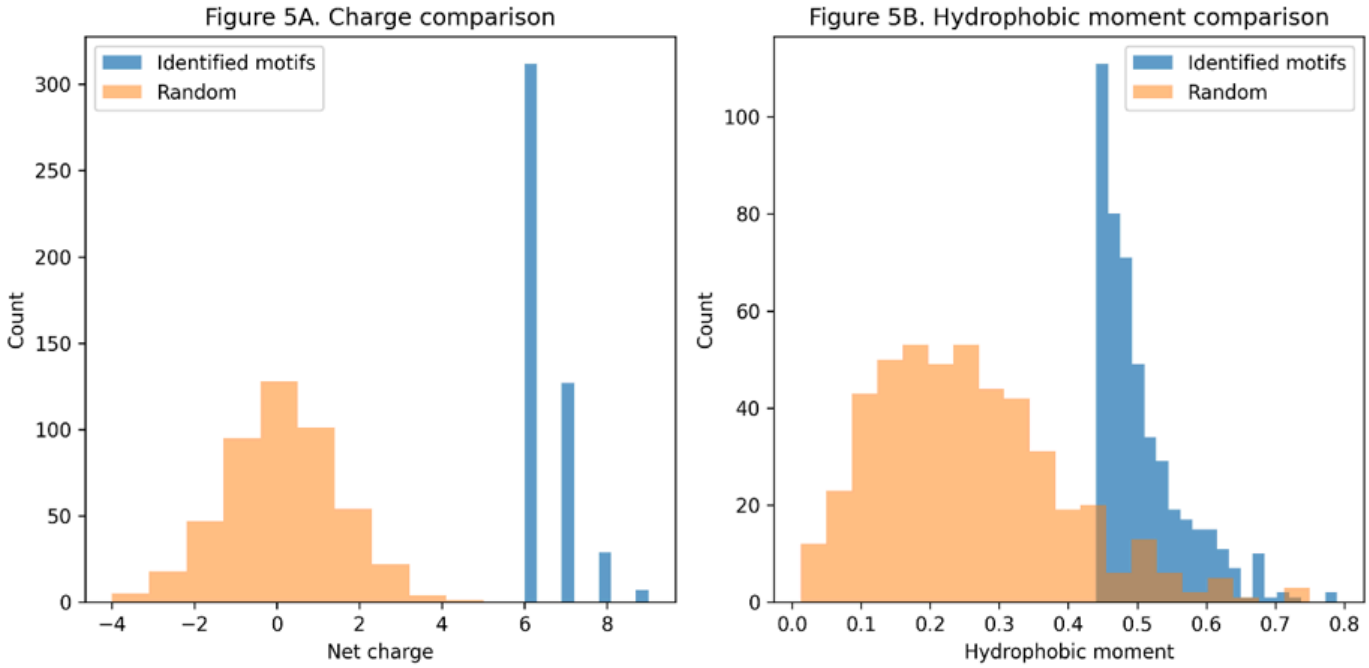


Figure 5

Comparison of identified motifs with random peptide segments.

(A) Net charge distribution. Identified motifs show strong enrichment of highly cationic sequences (≥ +6), whereas random peptides are centered around neutral values.

(B) Hydrophobic moment distribution. Identified motifs form a narrow distribution consistent with amphipathic α-helices, while random peptides exhibit a broader distribution.

These differences indicate that the identified motifs are not randomly distributed within protein sequences.