Matthew Poncini
J.B. Speed School of Engineering
University of Louisville

21 November 2025

—

Project 3:
Unsupervised Clustering and Pattern Recognition of a 9-mer Peptide Dataset

Abstract

Peptides are short chains of amino acids joined by peptide bonds. They represent fundamental units of biological activity, serving as signaling molecules, antigens, and binding ligands in a handful of biochemical contexts [1].A 9-mer peptide is a sequence of nine amino acids. Each amino acid occupies a specific position in the peptide chemical space. There are 20^n possible combinations of an n-mer peptide. Systematic exploration of peptide space has driven advances in vaccine design, therapeutic discovery, and molecular screening. Applying unsupervised data mining techniques reveal underlying patterns among peptides and provide insight to to characterize their associated biochemical motifs.

In this study, a 9-mer peptide dataset is drawn from four biological sources with equal representation: humans, viruses, bacteria, and cancer cells. The primary objective was to apply unsupervised clustering and pattern-recognition methods to identify latent biochemical and positional motifs within this diverse collection of peptides. A secondary objective was to assess whether specific feature representations could reveal the biological origin of a given 9-mer.

The results support the first objective:9-mer peptides exhibit clear structural clusters when characterized both through their physicochemical properties and their amino acid composition. The results failed to support the second objective. Among the 229 engineered features examined, none of the subsets of features examined provided sufficient discriminatory power to reliably infer the biological source of a peptide. HDBSCAN and UMAP served as the core Python libraries for clustering, pattern recognition, and multidimensional visualization.

Data Mining Project: Unsupervised Clustering and Pattern Recognition of a 9-mer Peptide Dataset

1. Problem Description

The goal of this project is to investigate whether unsupervised learning techniques can reveal meaningful biochemical structure within peptide sequence space and whether these learned structures can provide insight into the biological origin of a peptide. Specifically, the study focuses on 9-mer peptides, short sequences composed of nine amino acids. These peptides are biologically relevant because 9-mers often represent the core binding length for major histocompatibility complex (MHC) class I molecules, making them central to antigen presentation, immune recognition, and many areas of biomedical research. As a result, 9-mer peptides are the standard analytical unit in immunopeptidomics and computational epitope-prediction research.

The raw dataset consists of proteins from four biological sources: humans, viruses (Influenza B), bacteria (E. coli), and human cancer cells. Each protein is simply a long chain of amino acids, with each sequence ranging from 50 to 2,000 amino acids. Human and cancer cell generated proteins typically 100 to 800 amino acidslong. Viral proteins are typically between 70 and 300 amino acids long, while bacterial proteins are between 200 and 400 amino acids [2]. To construct the 9-mer peptide dataset, each protein sequence was processed using a sliding window of length 9, shifting one amino acid at a time to generate all possible overlapping 9-mer fragments. From these generated peptides, a stratified sample of 250,000 9-mers per source (human, viral, bacterial, and cancer) was selected to assemble the final dataset.

The source of each 9-mer peptide was kept hidden from the unsupervised clustering algorithms. Part of the goal was to see which sourced peptides behaved similarly to each other and which sourced peptides were unidentifiable from each other. If distinct peptide sources formed clearly separable groups, this would suggest that their biochemical and positional properties differ in systematic ways. Conversely, if sources largely overlap, it would indicate that many peptides, particularly human and cancer-derived sequences, share similar amino-acid features and therefore cannot be distinguished without explicit mutation or context information.

2. Previous Analyses on Unsupervised Learning with 9-mer Peptides

Prior studies have applied unsupervised learning techniques to peptide sequence analysis. An example is the article *GibbsCluster: unsupervised clustering and alignment of peptide sequences*, which introduced a probabilistic framework for clustering and aligning peptide sequences without prior knowledge of their grouping [3]. Their method revealed that short peptides—especially 9-mers—contain distinct biochemical signatures that can be uncovered through feature-driven unsupervised analysis. Similarly, immunopeptidome mining efforts were done in *Unsupervised Mining of HLA-I Peptidomes Reveals New Binding Motifs and Potential False Positives in the Community Database* [4]. This study employed clustering and motif-discovery workflows to identify patterns within HLA-I bound peptides derived from endogenous cellular proteins. These studies collectively demonstrate that peptide clustering is a powerful tool for uncovering biochemical structure, motif patterns, and subpopulation heterogeneity within biologically diverse peptide datasets.

3. Software tools

This project was completed using Python, using UMAP (Uniform Manifold Approximation and Projection) and HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) libraries.

UMAP works by modeling the structure of high-dimensional data and then creating a lower-dimensional representation that preserves that structure as closely as possible. First, it identifies each point's nearest neighbors and builds a graph that reflects how strongly points are connected based on local distances. This graph represents the underlying manifold—the shape the data is assumed to lie on. UMAP then attempts to place the same points in a low-dimensional space (such as 2D) so that points that were close in the original data stay close, and points that were far apart remain separated. It uses an optimization process to make these two representations match. The result is a clear, interpretable visualization that preserves both local clusters and broader global patterns [5].

UMAP was chosen to be the software tool used because it was capable to discover nonlinear patterns. PCA (principal component analysis) was originally used, but failed to reveal patterns. Where clusters may blend in PCA, UMAP produces well-separated clusters because it emphasizes local relationships and density. Unlike PCA, which prioritizes global variance, UMAP preserves both local neighborhood relationships and global manifold structure. This is essential for peptide feature data, where biochemical similarity often depends on subtle, localized feature patterns that PCA cannot separate. Peptide features interact in highly nonlinear ways. UMAP is specifically designed to capture these nonlinear interactions.

HDBSCAN is a density-based clustering algorithm that extends DBSCAN by allowing clusters of varying density and automatically identifying noise points. It begins by transforming the dataset into a "mutual reachability" space, which adjusts distances based on the local density around each point. Using these distances, the algorithm constructs a minimum spanning tree (MST) and then progressively removes edges to build a hierarchy of clusters ordered by density. As sparse connections break, dense regions separate into distinct cluster groups. HDBSCAN then analyzes the resulting cluster tree and extracts the most stable clusters—those that persist across a wide range of density thresholds. Points that never belong to any stable region are labeled as noise. This approach makes HDBSCAN highly effective at uncovering natural structures in complex, irregular data [6].

HDBSCAN was chosen because it excels at finding meaningful clusters in datasets where the underlying structure is irregular, noisy, or variable in density—properties common in biochemical datasets such as peptide feature embeddings. Traditional clustering methods, including k-means, require specifying the number of clusters and assume spherical, equally sized groups, which is unrealistic for biological feature spaces. In contrast, HDBSCAN automatically determines the number of clusters and can detect both large and small groups without forcing points into poorly fitting assignments. It is especially effective when paired with UMAP, which produces a low-dimensional layout with well-separated regions of differing densities. HDBSCAN can then interpret these regions naturally, identifying genuine peptide subgroups while classifying ambiguous or non-patterned data as noise.

4. Preprocessing

Each protein was sourced from four separate queries on the UniProt database, based on source: human, cancer, bacteria (E. Coli) and virus (Influenza B). The protein was then sliced into all possible 9-mers. From each of the four datasets, 250k 9-mer peptides were randomly selected to be in the dataset used for data mining. Table 1 shows the preview of this dataset.

Table 1: Rows from Peptide Dataset (1 million rows × 6 columns)

peptide	protein_id	gene	start_pos	end_pos	source
TLGINLPKL	sp|Q53EV4|LRC23_HUMAN	sp	193	201	human
APDSVEGLL	sp|Q9H0Q0|CYRIA_HUMAN	sp	290	298	human
ERPDQTNPS	sp|P17482|HXB9_HUMAN	sp	167	175	human
GRRKCKPQM	sp|P41162|ETV3_HUMAN	sp	76	84	human
ASVIGVFIT	sp|Q8NFF2|NCKX4_HUMAN	sp	155	163	human
	...				
SASGVSECR	tr|A0A6G9JXH8|A0A6G9JXH8_9INFB	tr	243	251	viral
IADYRYTYK	tr|A0A6M4ZXM1|A0A6M4ZXM1_9INFB	tr	550	558	viral
YTGEHAKAI	tr|A0A6C0S8R8|A0A6C0S8R8_9INFB	tr	321	329	viral
TSHGAHGVA	tr|A0A8K1SRK9|A0A8K1SRK9_9INFB	tr	383	391	viral
SHIRGSTII	tr|A0A346A7L2|A0A346A7L2_9INFB	tr	14	22	viral

From the data set shown in Table 1, a new dataset was created hiding the protein ID, gene, starting and ending positions, and source. Each peptid was then split by amino acid with its position notated as shown in Table 2.

Table 2: Rows Showing Peptides and Their Positional Amino Acids

peptide	pos1	pos2	pos3	pos4	pos5	pos6	pos7	pos8	pos9
TLGINLPKL	T	L	G	I	N	L	P	K	L
APDSVEGLL	A	P	D	S	V	E	G	L	L
ERPDQTNPS	E	R	P	D	Q	T	N	P	S
GRRKCKPQM	G	R	R	K	C	K	P	Q	M
ASVIGVFIT	A	S	V	I	G	V	F	I	T
				...					
SASGVSECR	S	A	S	G	V	S	E	C	R
IADYRYTYK	I	A	D	Y	R	Y	T	Y	K
YTGEHAKAI	Y	T	G	E	H	A	K	A	I
TSHGAHGVA	T	S	H	G	A	H	G	V	A
SHIRGSTII	S	H	I	R	G	S	T	I	I

Each amino acid has various biochemical properties as shown in Table 3. These properties can be aggregated to the peptide, providing numerical data describing the physicochemical behavior of the peptide. Hydrophobicity (Kyte–Doolittle) measures how strongly a peptide avoids water, influencing peptide folding, membrane affinity, and MHC binding. Polarity (Grantham) reflects how unevenly charge is distributed within the peptide, negatively affecting hydrogen bonding and interaction with hydogen associated environments. Charge indicates whether an amino acid is positively charged, negatively charged, or neutral at physiological pH. This plays a role in electrostatic interactions and peptide stability. Volume (Zamyatnin) gives the size of each peptide, contributing to constraints and packing within protein or binding pockets. Aromaticity identifies residues with aromatic rings (F, Y, W), which participate in hydrophobic interactions. Flexibility describes how easily the backbone can bend around that peptide. Helix propensity and turn propensity indicate how likely a residue is to appear in α-helices or β-turns. Solvent-accessible surface area (SASA) measures how exposed a residue is to solvent. Disorder propensity reflects how likely the residue is to appear in disordered regions, capturing instability or flexibility of the peptide chain [7].

Table 3: Amino Acid Biochemical Properties Used in Feature Engineering

AA	Name	Hydro	Polarity	Charge	Volume	Arom	Flex	Helix	Turn	SASA	Disorder
A	Alanine	1.8	8.1	0	88.6	0	0.360	1.42	0.66	115	0.06
R	Arginine	-4.5	10.5	+1	173.4	0	0.530	0.98	0.95	225	0.39
N	Asparagine	-3.5	11.6	0	114.1	0	0.460	0.67	1.56	160	0.41
D	Aspartic acid	-3.5	13.0	-1	111.1	0	0.510	1.01	1.46	150	0.54
C	Cysteine	2.5	5.5	0	108.5	0	0.380	0.70	1.19	135	0.17
E	Glutamic acid	-3.5	12.3	-1	138.4	0	0.500	1.51	0.74	190	0.51
Q	Glutamine	-3.5	10.5	0	143.8	0	0.490	1.11	0.98	180	0.42
G	Glycine	-0.4	9.0	0	60.1	0	0.540	0.57	1.64	75	0.54
H	Histidine	-3.2	10.4	+1	153.2	0	0.510	1.00	0.95	195	0.35
I	Isoleucine	4.5	5.2	0	166.7	0	0.460	1.08	0.47	175	0.22
L	Leucine	3.8	4.9	0	166.7	0	0.370	1.21	0.59	170	0.25
K	Lysine	-3.9	11.3	+1	168.6	0	0.470	1.16	1.01	215	0.39
M	Methionine	1.9	5.7	0	162.9	0	0.300	1.45	0.60	185	0.23
F	Phenylalanine	2.8	5.2	0	189.9	1	0.310	1.13	0.60	210	0.29
P	Proline	-1.6	8.0	0	112.7	0	0.510	0.59	1.52	145	0.43
S	Serine	-0.8	9.2	0	89.0	0	0.510	0.77	1.43	115	0.46
T	Threonine	-0.7	8.6	0	116.1	0	0.440	0.83	1.19	140	0.40
W	Tryptophan	-0.9	5.4	0	227.8	1	0.310	1.08	0.96	255	0.32
Y	Tyrosine	-1.3	6.2	0	193.6	1	0.420	0.69	1.14	230	0.28
V	Valine	4.2	5.9	0	140.0	0	0.390	1.06	0.50	155	0.21

After applying aggregate calculations based on the biochemical properties, features were added to the dataset. Of the 1 million peptides, 884 peptides were removed from the dataset due to missing amino acids leaving a a dataset of 999,116 rows. The numeric data was then normalized using a z-score transformation. This standardization places all physicochemical features on a comparable scale, ensuring that no single property dominates the analysis due to differences in magnitude.

Table 4: Peptide Features Used in Analysis

peptide	pos1	pos2	pos3	pos4	pos5	pos6
pos7	pos8	pos9	hydro_mean	polarity_mean	charge_sum	volume_mean
aromatic_count	flexibility_mean	helix_mean	turn_mean	sasa_mean	disorder_mean	

UMAP served as the dimensionality-reduction step in the analysis. UMAP reduces dimensionality by modeling the high-dimensional structure of the data and then reconstructing it in a lower-dimensional space. It begins by identifying the nearest neighbors for each point and building a weighted graph that represents how strongly points are connected in the original feature space. This graph captures the local manifold structure, which is the underlying shape the data is assumed to lie on. UMAP then constructs a second graph in a low-dimensional space and optimizes the positions of the points so that the relationships in this lower-dimensional graph match those of the original one as closely as possible. In doing so, UMAP preserves both local neighborhoods and global patterns while compressing the dataset features into two dimensions suitable for visualization and further analysis. The source remained hidden during dimensionality-reduction.

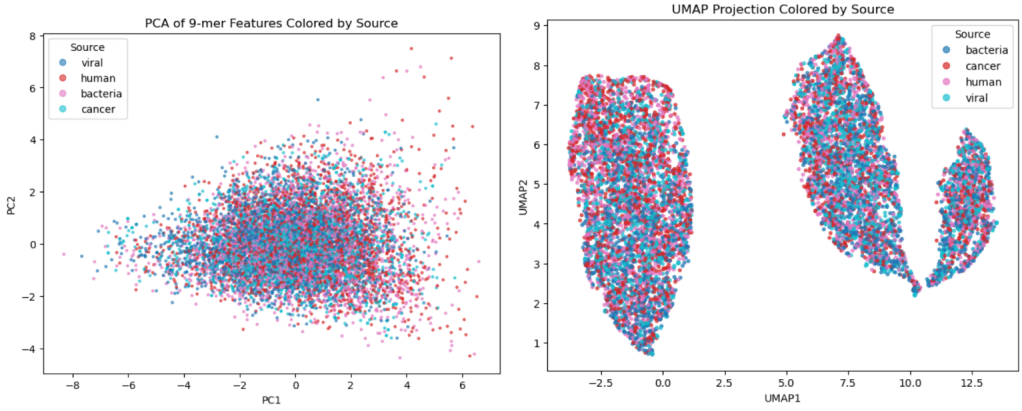

Figure 1: Dimension reduction through PCA (left) versus UMAP (right).

Figure 1 shows the PCA and UMAP projections. The PCA projection shows that the 9-mer peptide features form a large, continuous cloud with substantial overlap among all four sources—human, viral, bacterial, and cancer. This is expected because PCA is a linear dimensionality-reduction method. It captures only the directions of greatest variance, not nonlinear interactions between biochemical properties. As a result, no meaningful group structure emerges. The UMAP projection (right) reveals well-defined nonlinear manifolds where the points organize into distinct clusters.

5. Data Mining Process

To better understand how individual biochemical features contribute to the organization of the peptide manifold, the UMAP projection was colored according to each aggregated property (Figure 2). Clear gradients are visible across the UMAP structure, indicating that the manifold learned by UMAP reflects underlying physicochemical patterns rather than random spatial arrangement. Most features vary smoothly across the three major UMAP clusters, demonstrating that the embedding preserves meaningful biochemical continuity. The gradients shown in the hydrophobicity and polarity clusters appear to be inverses of each other. This is consistent with their biochemical definitions. Hydrophobicity reflects the tendency of a peptide to avoid water, while polarity measures the degree of uneven charge distribution within the amino acids. In general, amino acids that are strongly hydrophobic exhibit low polarity, whereas highly polar residues exhibit low hydrophobicity.

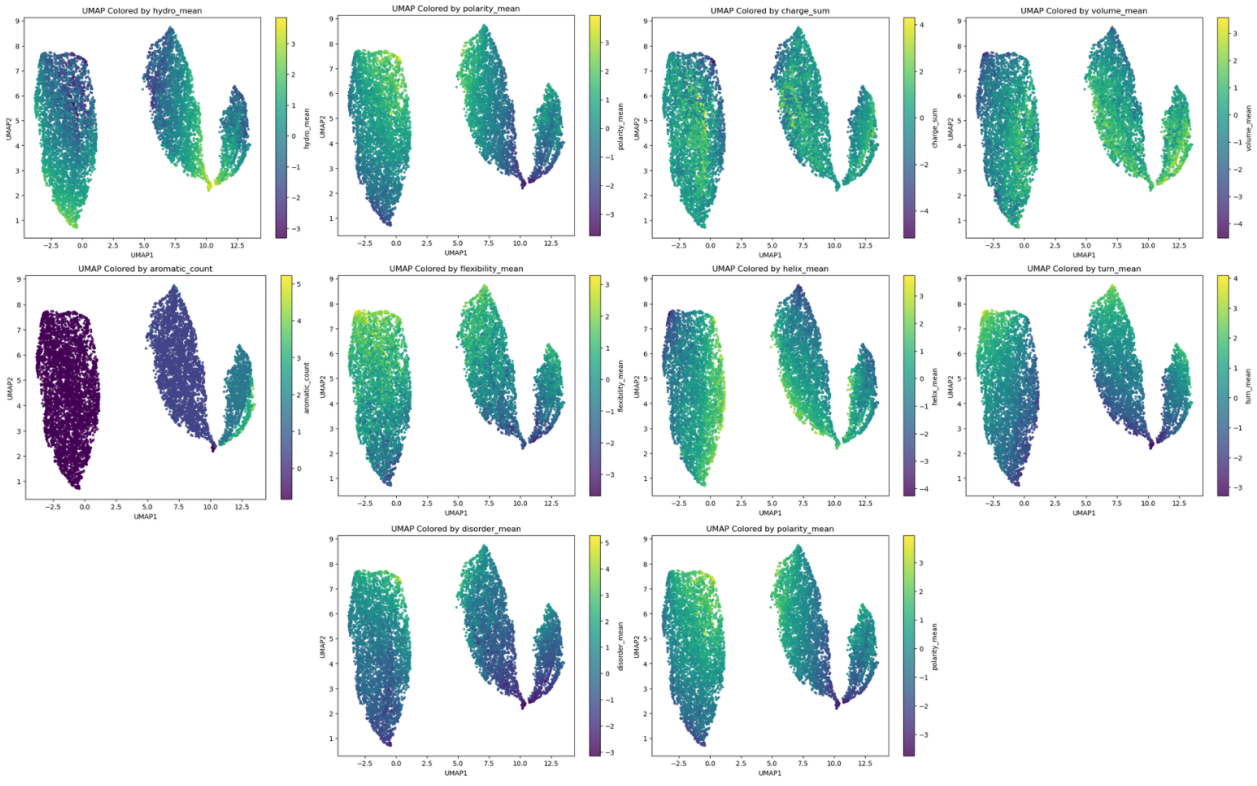

Figure 2: UMAP Projections, colored by feature values

Coloring each UMAP projection by its biochemical features revealed that aromaticity was the strongest indicator of cluster separation. The aromaticity feature produced a sharp contrast between the 3, arguably 4 clusters. Peptides with even a single aromatic residue (F, Y, or W) tended to localize toward specific regions of the UMAP branches, whereas peptides lacking aromatic content clustered distinctly elsewhere. Aromatic amino acids have large, rigid ring structures that contribute disproportionately to several physicochemical properties, including hydrophobicity, volume, and electronic interactions. Because these residues strongly influence peptide behavior, their presence introduces discrete structural shifts that UMAP captures as dis-

tinct submanifolds. This suggests that aromatic content acts as a key organizing factor in the underlying feature space, driving the formation of the most pronounced clusters observed in the dimensionality-reduced representation.

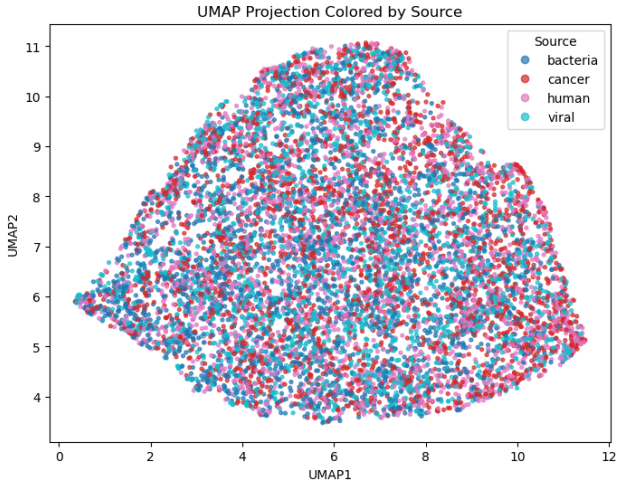

Figure 3: UMAP Projection with Aromaticity removed from the dataset

When aromaticity was removed from the feature set, the structure of the UMAP projection changed dramatically. The distinct multi-branch manifold observed previously collapsed into a single, more amorphous cluster with no clear separation or submanifold structure. This outcome reinforces the earlier observation that aromaticity was one of the strongest drivers of global organization in the peptide space. The resulting projection shows that peptides from human, viral, bacterial, and cancer sources remain broadly intermixed, indicating that source itself does not define separable biochemical patterns. However, the loss of manifold structure after removing aromaticity demonstrates that this single property contributes disproportionately to shaping the geometry of the feature space.

HDBSCAN was applied to the two-dimensional UMAP representation to identify natural groupings within the peptide feature space. As a density-based clustering algorithm, HDBSCAN does not require the number of clusters to be specified in advance. Instead, it identifies regions of high point density as clusters while labeling sparse or irregular points as noise.

The clustering result reveals three major peptide clusters in the UMAP projection, represented by the orange, purple, and green regions. These clusters correspond directly to the large leaf-like branches produced by UMAP, indicating that the structure of the manifold reflects genuine density variations in the high-dimensional biochemical feature space. HDBSCAN successfully isolates these high-density regions and assigns them cluster labels (0, 1, and 2), while points that do not belong to any stable region are marked as noise (shown in dark red). The relatively small number of noise points suggests that the majority of peptides fall into well-defined biochemical neighborhoods.

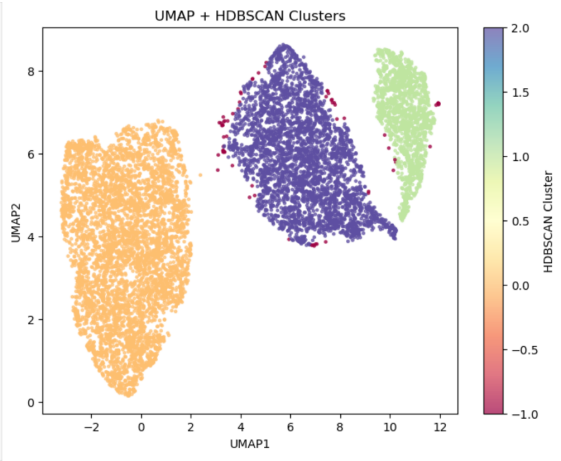

Figure 4: HDBSCAN cluster assignments based on the UMAP projection.

These clusters do not correspond to biological source labels (human, viral, bacterial, cancer), as earlier analyses showed. Instead, they reflect underlying physicochemical organization, particularly features such as aromaticity. Thus, the combined UMAP, HDBSCAN pipeline reveals that while peptide sources are intermixed, the embedded feature space contains distinct biochemical zones that represent different classes of physicochemical character.

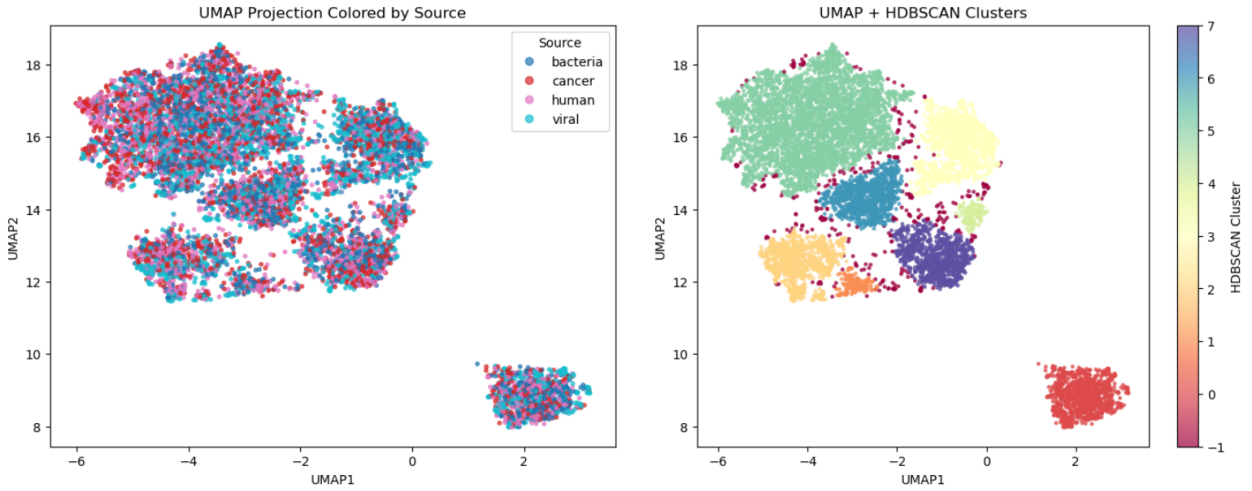

Figure 5: UMAP, HDBSCAN Clusterings from dataset with only amino acid counts

Unsatisfied that a pattern was not revealed to identify each peptide's source, the process was repeated by switching the physicochemical features to a simpler 20-dimensional representation based solely on amino-acid counts per peptide. This changed the structure of the UMAP projection. Each peptide is exactly nine amino acids long, and so the count features capture only composition. As a result, peptides with similar amino-acid proportions cluster together, producing a more compact and discrete set of manifolds compared to the complex leaf-like structures observed when using biochemical aggregates. In the UMAP projection colored by source, peptides from human, viral, bacterial, and cancer origins again show heavy mixing, confirming that source does not dictate amino-acid composition strongly enough to form separable groups.

To further explore the internal structure of the clusters identified from the biochemical feature space (shown in Figure 4), the simpler 20-dimensional amino-acid count feature set was applied separately within each of the three major HDBSCAN clusters from the original UMAP embedding.

This hierarchical approach allows us to test whether the broad biochemical clusters discovered in the first stage contain additional substructure detectable when focusing purely on amino-acid composition.

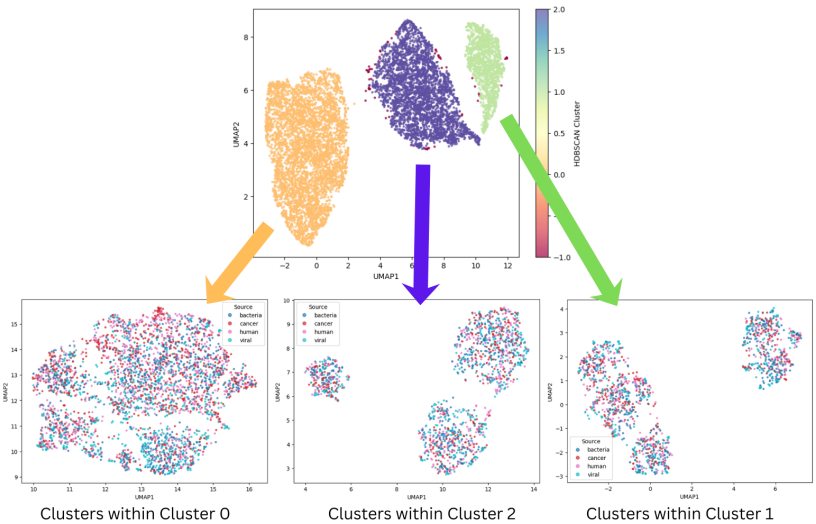

Figure 6: Hierarchical clustering results.

6. Results and Conclusion

The main goal of the project was to see whether unsupervised learning could reveal structure in the peptide feature space and whether any of that structure matched the biological source of the peptides. Several types of outputs were generated during this process, including dimensionality-reduced visualizations, feature-colored UMAP plots, cluster assignments from HDBSCAN, and a second round of UMAP and clustering using only amino-acid counts. Each output helped explain how peptides organize themselves based on their biochemical properties and overall composition.

When PCA was applied, the dataset formed one large cloud with no meaningful separation. This indicated that linear variance alone was not strong enough to reveal patterns in the 9-mer features. UMAP produced three large, clearly shaped manifolds. These structures showed that the peptide dataset does contain natural groupings, but they were not tied to the peptides' biological source. The groupings came from the biochemical features themselves. Certain features formed smooth gradients across the UMAP clusters, such as hydrophobicity and polarity, which acted as near opposites of each other. This confirmed that UMAP was capturing real biochemical relationships. The strongest and most surprising pattern came from aromaticity. Coloring UMAP by aromaticity showed three sharp, distinct groups. When aromaticity was removed entirely, the structure collapsed into one disorganized cluster. This showed that aromatic residues (F, Y, W) play a major role in defining how peptides arrange themselves in feature space. Aromaticity became the single most important feature driving the global shape of the UMAP manifold.

After dimensionality reduction, HDBSCAN was used to identify clusters in the UMAP space. HDBSCAN found three large clusters that matched the three main branches of the UMAP plot. These clusters were based on density and biochemical similarity, not on biological source. A small number of points were labeled as noise, meaning they didn't clearly belong to any group. The important takeaway was that the clusters reflected how peptides differ chemically, not where they came from.

The results clearly show that peptides do form natural groups when described by biochemical features or amino-acid composition. These groups reflect how peptides differ chemically, but they do not reveal the biological source. One possible reason why the clusters don't separate by source is that different organisms often evolve peptides that are very chemically similar. Viruses frequently use mimicry, shaping their peptides to resemble human sequences so they can evade immune detection. Cancer cells generate mutated peptides, which still preserve the majority of amino acids in normal human proteins. This overlap made the second objective rather ambitious.

The main conclusion is that unsupervised learning does uncover meaningful biochemical patterns in 9-mer peptides. Each feature set revealed different layers of structure: biochemical features gave broad manifolds, while amino-acid counts revealed smaller, tighter clusters. A useful direction for future work would be to try more advanced peptide-embedding models, such as ESM, ProtBERT, or transformer-based embeddings, which can capture higher-order sequence patterns. Motif discovery tools, or supervised models trained to detect subtle source differences, could also be used to see whether faint signals exist that unsupervised methods could not find. Finally, analyzing the biological meaning of each cluster, such as motif enrichment or structural tendencies, could help connect these clusters to real peptide functions.

References

[1] Peptides Guide, "Comprehensive information on peptides, their functions, and uses," May 3, 2025. [Online]. Available: https://peptidesguide.com

[2] UniProt, "UniProtKB." [Online]. Available: https://www.uniprot.org/uniprotkb

[3] M. Andreatta and M. Nielsen, "GibbsCluster: unsupervised clustering and alignment of peptide sequences," Nucleic Acids Research, vol. 45, no. W1, pp. W458–W463, 2017, doi: 10.1093/nar/gkx248.

[4] C. Sricharoensuk, A. Lomsadze, and A. Seetharam, "Unsupervised Mining of HLA-I Peptidomes Reveals New Ligands and Refines Current Understanding," PMC, Article 8977642, 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8977642

[5] McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Journal of Open Source Software, 3(29), 861. https://doi.org/10.21105/joss.00861 JOSS +1 -Documentation: UMAP-learn. (n.d.). Basic Usage. Retrieved from https://umap-learn.readthedocs.io/en/latest/basic_usage.html

[6] McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. Journal of Open Source Software, 2(11), 205. https://doi.org/10.21105/joss.00205 JOSS +1 — Documentation: HDBSCAN. (n.d.). Basic HDBSCAN Usage. Retrieved from https://hdbscan.readthedocs.io/en/latest/basic_hdbscan.html

[7] S. Kawashima and M. Kanehisa, "AAindex: Amino acid index database," Nucleic Acids Research, vol. 28, no. 1, p. 374, 2000, doi: 10.1093/nar/28.1.374.

Submitted by Matthew Poncini on 21 November 2025.