

Article

EpitopeGNN: A Graph Neural Network for Influenza A Virus Hemagglutinin Subtype Classification Based on 3D Structure

Andrey Timofeev¹, Alexander Anufriev¹, Oleg Ergashev^{2,3} and Irina Isakova-Sivak^{2,*}

¹ LLC “AI Center for SCO+ Countries”, St. Petersburg 197341, Russia; tav@centeraisco.com (A.T.); aas@centeraisco.com (A.A.)

² Institute of Experimental Medicine, St. Petersburg 197376, Russia; ergashew@iemspb.ru

³ Department of Hospital Surgery No. 2 with the Clinic Named After Academician F.G. Uglov, Pavlov First Saint Petersburg State Medical University, St. Petersburg 197341, Russia

* Correspondence: isakova.sivak@iemspb.ru

Abstract

Background: Hemagglutinin (HA) is the primary surface protein of the influenza A virus, determining its subtype and antigenic properties. Traditional subtype classification methods rely on DNA or amino acid sequence analysis, which does not account for protein spatial folding. **Methods:** In this work, we propose EpitopeGNN—a graph neural network (GNN) that constructs a residue interaction network (RIN) from the 3D structure of HA and classifies the virus subtype. The model was trained on 249 structures from the Protein Data Bank (PDB), containing H1N1, H3N2, H5N1, and other subtypes. **Results:** After rigorous sequence redundancy reduction (92% identity), the model maintained 95–100% accuracy on non-redundant data, significantly outperforming sequence-only baselines (the best baseline achieved 85% for multi-class and 92.3% for binary classification). A significant correlation was found between the obtained structural embeddings and phylogenetic distances ($r = 0.38$, $p < 0.001$), confirming their biological relevance and opening opportunities for structural monitoring of virus evolution, as well as rapid analog searching for novel strains. **Conclusions:** We developed a new graph neural network that classifies influenza A virus subtypes directly from the 3D structure of hemagglutinin using residue interaction networks and physicochemical features, which can serve as a foundation for predicting influenza virus receptor specificity and epitope immunogenicity.

Keywords: influenza A virus; hemagglutinin (HA); subtype classification; graph neural network (GNN); protein 3D structure; phylogenetic distance; structural embeddings; EpitopeGNN



Academic Editor: Alexandre G. De Brevern

Received: 27 March 2026

Revised: 20 April 2026

Accepted: 22 April 2026

Published: 27 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The influenza A virus causes seasonal epidemics and pandemics, posing a serious threat to public health. Hemagglutinin (HA) plays a key role in its identification—it is a protein responsible for attachment to the host cell and defines the virus subtype (H1–H18). Accurate subtype determination is essential for epidemiological surveillance, vaccine strain selection, and pandemic risk assessment.

Existing classification methods (ClassyFlu [1], INFINITy [2]) are based on comparing nucleotide or amino acid sequences using hidden Markov models (HMMs) or alignment-free machine learning approaches [3,4]. However, they ignore the three-dimensional structure of the protein, which can be critical for functional differences even when sequence homology is high.

In recent years, graph neural networks (GNNs) have demonstrated high efficiency in structural bioinformatics tasks, enabling the use of residue interaction networks (RINs) [5].

We propose EpitopeGNN—a model that predicts the HA subtype from its RIN graph using informative node features and an architecture based on an attention mechanism (graph attention network, GAT). To the best of our knowledge, the use of three-dimensional structures and graph neural networks for influenza subtype classification has not been explored before; this work presents such an approach, incorporating the topological and physicochemical features of protein spatial folding.

While existing sequence-based methods, including subtype classifiers (ClassyFlu, INFINITy) and even epitope predictors (e.g., CLBTope [6]), rely solely on primary structure, they cannot capture the three-dimensional conformational features that determine antigenicity. The ability to classify subtypes from 3D structure, as proposed here, provides a complementary level of information that is essential for understanding immune recognition and for rational vaccine design.

2. Materials and Methods

2.1. Dataset

2.1.1. Hemagglutinin Structures

The list of hemagglutinin PDB IDs was obtained by searching the RCSB PDB portal [7] using the keyword “hemagglutinin”. After filtering out structures with fewer than 50 residues, the final dataset consisted of 249 structures for which subtype annotations were present in the .cif.gz files. The subtype annotations extracted from the PDB files were cross-referenced with data from the GISAID database [8] for quality control.

The distribution of subtypes is shown in Table 1.

Table 1. Distribution of influenza A subtypes in the dataset.

Influenza A Subtype	Number
H1N1	84
H3N2	80
H5N1	48
H7N9	15
H2N2	11
others	11

2.1.2. Redundancy Reduction

To avoid overestimation due to near-identical sequences, we performed sequence-based clustering using k-mer frequencies ($k = 3$) and cosine similarity with a threshold of 0.92 (92% identity). Only representative sequences (centroids) were retained, reducing the dataset from 249 to 56 non-redundant structures. This filtering ensures that no two structures share >92% amino acid identity, providing a rigorous test of generalization.

2.2. Construction of the RIN Graph

Let a set of residues $R = \{r_1, r_2, \dots, r_n\}$ be defined for each hemagglutinin structure, where n is the number of amino acid residues satisfying the selection criteria (standard amino acids, presence of a C_α atom). Each residue r_i is associated with a point $c_i \in \mathbb{R}^3$ —the coordinates of its C_α -atom, extracted from the .cif.gz file.

Graph Definition

An undirected graph $G = (V, E)$ is constructed, where

$V = \{1, 2, \dots, n\}$ is the set of vertices corresponding to residues;

$E \subseteq V \times V$ is the set of edges defined by a threshold condition:

$$(i, j) \in E \iff i \neq j \text{ and } \| \mathbf{c}_i - \mathbf{c}_j \|_2 < \delta,$$

where $\delta = 8 \text{ \AA}$ is the chosen spatial proximity threshold. The choice of $\delta = 8 \text{ \AA}$ is based on the literature data [9]: this distance is sufficient to cover most non-covalent interactions (hydrogen bonds, van der Waals contacts, ionic interactions) while excluding overly distant contacts that do not affect the local structure.

Adjacency matrix

The graph G is represented by a symmetric adjacency matrix $\mathbf{A} \in \{0, 1\}^{n \times n}$, where

$$A_{ij} = A_{ji} = \begin{cases} 1, & \text{if } (i, j) \in E, \\ 0, & \text{otherwise.} \end{cases}$$

Representation for a graph neural network

For subsequent use in architectures such as GCN, GAT, or GraphSAGE, the graph is stored as a tuple $(\mathbf{X}, \mathbf{E})$, where $\mathbf{E}$ is an edge list (index scheme) that enables directed traversal during message aggregation. In an undirected graph, each edge (i, j) is added twice: (i, j) and (j, i) .

2.3. Node Features

Each vertex $i \in V$ is associated with a feature vector $\mathbf{x}_i \in \mathbb{R}^d$, $d = 26$, which combines:

1. **One-hot encoding of the amino acid type** (20 dimensions). Each of the 20 canonical amino acids is represented by a binary vector, where a 1 is placed in the position corresponding to that residue.
2. **Physicochemical properties** (4 dimensions): charge, hydrophobicity (according to the Kyte–Doolittle scale [10]), volume, and polarity (according to the Grantham scale [11]). These values are tabulated constants for each amino acid:
 - **Charge**—The electric charge of the side chain at physiological pH (≈ 7.4). For example, Lys (+1), Arg (+1), Asp (−1), Glu (−1), others 0.
 - **Hydrophobicity**—The index according to the Kyte–Doolittle scale, reflecting the hydrophobicity/hydrophilicity of the side chain. Values range from −4.5 (Arg) to +4.5 (Ile).
 - **Volume**—The van der Waals volume of the side chain (in $\text{\AA}^3$). For example, Gly $\sim 48 \text{ \AA}^3$, Trp $\sim 163 \text{ \AA}^3$.
 - **Polarity**—The polarity measure according to Grantham, characterizing the ability of the residue to participate in polar interactions.
3. Topological characteristics of the graph (2 dimensions):
 - **Vertex degree** $\deg(i) = \sum_j A_{ij}$ —The number of neighbors in the graph.
 - **Betweenness centrality** $BC(i)$ —A measure of how often a node lies on shortest paths between other nodes. It was computed approximately for large graphs using Brandes' algorithm [12].

The feature matrix of the entire graph $\mathbf{X} \in \mathbb{R}^{n \times d}$ is formed as

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^\top.$$

2.4. EpitopeGNN Model Architecture

For classifying hemagglutinin subtypes based on RIN graphs, a graph neural network based on the attention mechanism—**graph attention network (GAT)** [13]—was developed.

The model consists of three consecutive GAT layers, followed by global mean pooling and a fully connected layer with a softmax activation function for multi-class classification.

Input data: Each graph $G = (V, E)$ with n nodes is fed into the model as a set of node features $\mathbf{X} \in \mathbb{R}^{n \times d}$, where $d = 26$ is the dimensionality of the feature description (see Section 2.3), and a list of edges E that defines the connections between nodes.

GAT layer: Each GAT layer transforms node features using a self-attention mechanism. For node i at layer l , the updated representation $\mathbf{h}_i^{(l+1)}$ is computed as follows:

$$\mathbf{h}_i^{(l+1)} = \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} \mathbf{W}^{(l)} \mathbf{h}_j^{(l)} \right),$$

where

- $\mathbf{h}_i^{(l)} \in \mathbb{R}^F$ —Feature of node i at the input of layer l (for the first layer, $\mathbf{h}_i^{(0)} = \mathbf{x}_i$);
- $\mathcal{N}(i)$ —Set of neighbors of node i (including itself if self-attention is used);
- $\mathbf{W}^{(l)} \in \mathbb{R}^{F' \times F}$ —Trainable linear transformation matrix applied to all nodes;
- $\alpha_{ij}^{(l)}$ —Attention coefficient reflecting the importance of node j 's features for updating the representation of node i ;
- σ —Non-linear activation function (ELU is used in this work).

The attention coefficients are computed using an additive attention mechanism:

$$\alpha_{ij}^{(l)} = \frac{\exp \left(\text{LeakyReLU} \left(\mathbf{a}^{(l)\top} \left[\mathbf{W}^{(l)} \mathbf{h}_i^{(l)} \parallel \mathbf{W}^{(l)} \mathbf{h}_j^{(l)} \right] \right) \right)}{\sum_{k \in \mathcal{N}(i)} \exp \left(\text{LeakyReLU} \left(\mathbf{a}^{(l)\top} \left[\mathbf{W}^{(l)} \mathbf{h}_i^{(l)} \parallel \mathbf{W}^{(l)} \mathbf{h}_k^{(l)} \right] \right) \right)}.$$

Here:

- $\parallel$ denotes vector concatenation;
- $\mathbf{a}^{(l)} \in \mathbb{R}^{2F'}$ is a trainable attention parameter vector;
- LeakyReLU is used with a negative slope $\alpha = 0.2$.

To stabilize training and enrich representations, each GAT layer employs **multi-head attention**. In this model, 4 independent attention heads are used, and their outputs are concatenated:

$$\mathbf{h}_i^{(l+1)} = \parallel_{k=1}^4 \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l,k)} \mathbf{W}^{(l,k)} \mathbf{h}_j^{(l)} \right).$$

Thus, the dimensionality of the hidden representation after each layer is $4 \times F'$. In our model, $F' = 32$, is chosen for all layers, so after concatenation the dimensionality becomes 128. After three GAT layers, we obtain node representations $\mathbf{h}_i^{(3)} \in \mathbb{R}^{128}$.

Global mean pooling: To obtain a vector representation (embedding) of the whole graph, global mean pooling over nodes is applied:

$$\mathbf{h}_G = \frac{1}{n} \sum_{i=1}^n \mathbf{h}_i^{(3)}.$$

The resulting vector $\mathbf{h}_G \in \mathbb{R}^{128}$ is a compact description of the hemagglutinin structure.

Classification head: The graph embedding $\mathbf{h}_G$ is fed into a fully connected layer with a softmax activation function, which transforms it into a probability distribution over C classes (subtypes):

$$\mathbf{p} = \text{softmax}(\mathbf{W}_{\text{cls}} \mathbf{h}_G + \mathbf{b}_{\text{cls}}),$$

where $\mathbf{W}_{\text{cls}} \in \mathbb{R}^{C \times 128}$ and $\mathbf{b}_{\text{cls}} \in \mathbb{R}^C$ are trainable parameters. The predicted subtype corresponds to the class with the highest probability.

2.5. Training

The dataset was randomly split into training (70%) and test (30%) sets while preserving class proportions (stratification). Training was performed using the cross-entropy loss function:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \log \left(\frac{\exp(z_{i,y_i})}{\sum_{c=1}^C \exp(z_{i,c})} \right)$$

The optimizer was Adam (learning rate = 0.001, weight decay = 5×10^{-4}), batch size = 4, and the number of epochs was 100. To assess stability, five-fold cross-validation was performed. These hyperparameters were chosen based on preliminary experiments; they provided stable convergence and good generalization without overfitting.

3. Results

3.1. Classification Performance

On the original full dataset (249 structures), the model achieved 100% accuracy on the test set and 97.6% with five-fold cross-validation. However, to exclude overestimation due to near-identical sequences, we performed additional experiments on a non-redundant set (see Section “Robustness to Sequence Redundancy”)

Robustness to Sequence Redundancy

To exclude the possibility of data leakage, we retrained EpitopeGNN on the filtered non-redundant set (56 structures). Two experiments were conducted:

Experiment A (MIN_SAMPLES = 3)—Classes: H1N1, H2N2, H3N2, H5N1, H7N9. The model achieved 95.0% test accuracy on 20 test samples (see Table 2). The best sequence-only baseline (Random Forest on k-mer frequencies) reached 85.0% (see Table 3). Figure 1 shows the confusion matrix for the test set of Experiment A. The only misclassification is one H7N9 sample predicted as H3N2, consistent with the recall of 0.50 for H7N9 reported in Table 2.

Experiment B (MIN_SAMPLES = 15)—Only H1N1 and H3N2 (the two most abundant subtypes). EpitopeGNN achieved 100% test accuracy (13 test samples), while the best sequence-only baseline (MLP on k-mer frequencies) reached 92.3%. These results confirm that the high performance of EpitopeGNN is not an artifact of data leakage or overfitting. Even after aggressive redundancy reduction, the model maintains near-perfect accuracy (95–100%), consistently outperforming sequence-only methods, especially on closely related subtypes. Thus, EpitopeGNN’s advantage is robust and not due to overfitting.

Table 2. Classification metrics on the filtered dataset (Experiment A, MIN_SAMPLES = 3).

Subtype	Precision	Recall	F1-Score	Support
H1N1	1.00	1.00	1.00	7
H2N2	1.00	1.00	1.00	2
H3N2	1.00	1.00	1.00	6
H5N1	0.75	1.00	0.86	3
H7N9	1.00	0.50	0.67	2
macro avg	0.95	0.90	0.90	20
weighted avg	0.96	0.95	0.95	20
accuracy	0.95			

Table 3. Test accuracy of baseline methods on the filtered dataset (same train/test splits as EpitopeGNN).

Method	Experiment A (5 Classes, 20 Test)	Experiment B (H1N1 vs. H3N2, 13 Test)
Random Forest (handcrafted)	65.0%	84.6%
SVM (k-mer)	35.0%	53.8%
MLP (k-mer)	80.0%	92.3%
Random Forest (k-mer)	85.0%	90.1%
EpitopeGNN (ours, 3D)	95.0%	100%

Note: For Experiment B, only the H1N1 and H3N2 classes were used (MIN_SAMPLES = 15).

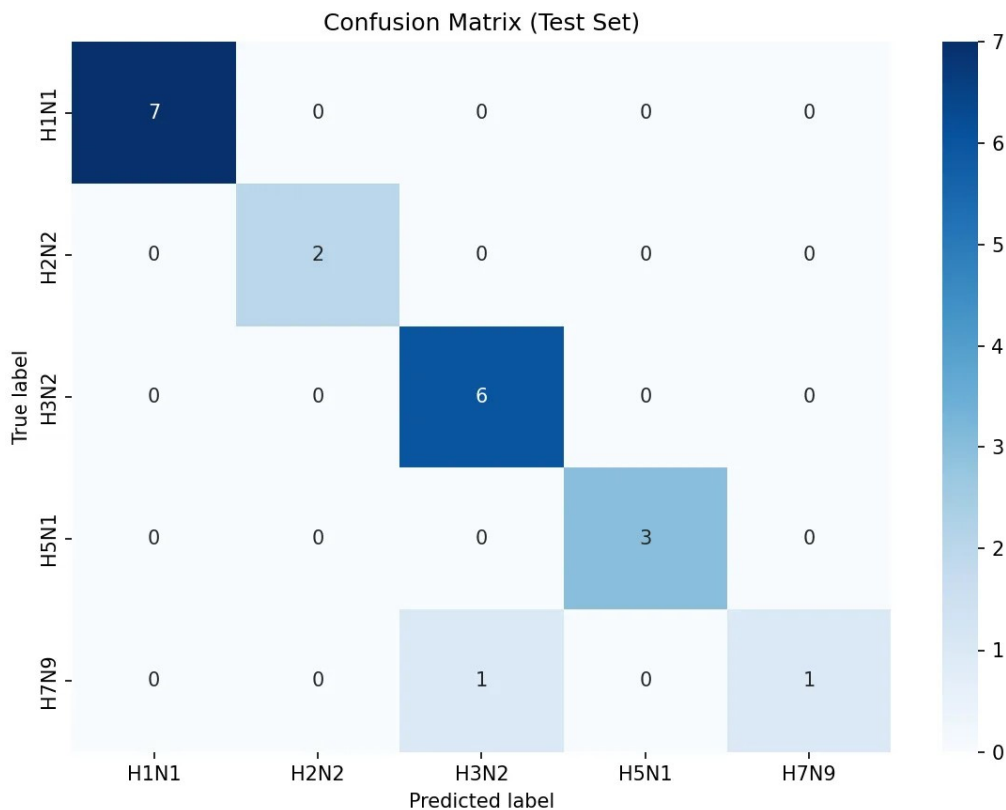


Figure 1. Confusion matrix for the filtered test set (Experiment A, MIN_SAMPLES = 3).

3.2. Embedding Visualization

To visualize the data structure, 128-dimensional embeddings extracted from the penultimate layer of the EpitopeGNN model were projected onto a plane using the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm (Figure 2) and Principal Component Analysis (PCA) (Figure 3). On both figures, subtypes are labeled with numeric codes (1: H1N1, 2: H2N2, 3: H3N2, 4: H5N1, 5: H7N9). The projection includes the 56 non-redundant structures from the filtered dataset (see Section 2.1.1). Their counts after filtering are given in Table 4.

Visualization: In the t-SNE projection (Figure 2), clear and well-separated clusters are observed for all five subtypes (1–5). Although the cluster corresponding to H7N9 (5) shows a somewhat larger spread, it remains distinctly separated from the others. In the PCA projection (Figure 3), clusters for H1N1 (1), H3N2 (3), and H7N9 (5) are well formed, while H2N2 (2) and H5N1 (4) partially merge, indicating that these two subtypes are more difficult to distinguish in a linear projection. The overall silhouette score (0.619) in the

original embedding space confirms the quality of clustering, and the differences between t-SNE and PCA highlight the non-linear nature of the structural features. Therefore, the visualization results should be interpreted with caution, but they consistently support the existence of subtype-specific structural features.

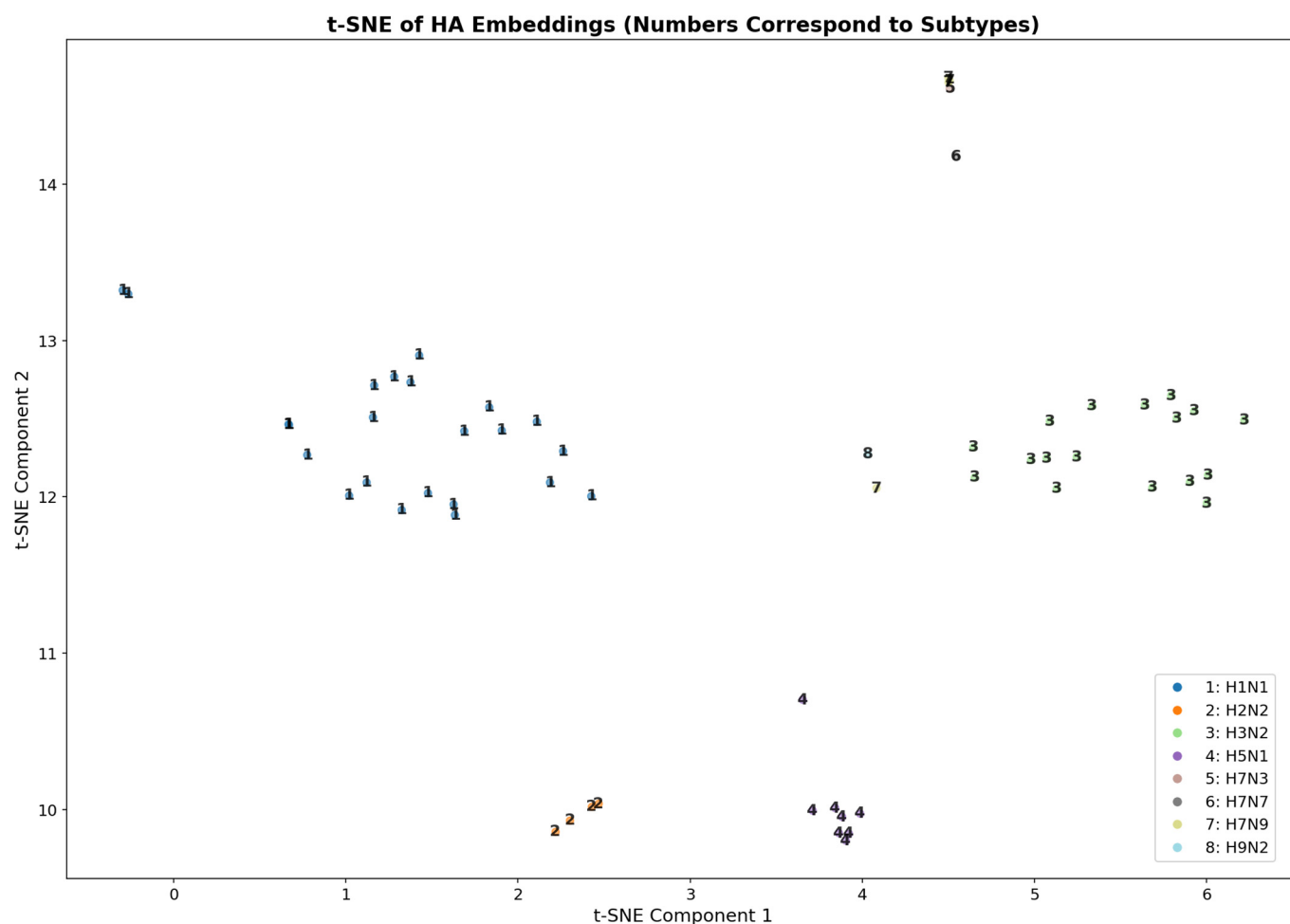


Figure 2. t-SNE projection of HA embeddings for the filtered dataset (56 non-redundant structures, numbers correspond to subtypes). For a complementary PCA view, see Figure 3.

Quantitative analysis: To objectively assess cluster separability, centroid coordinates in the t-SNE space (Table 4) and pairwise Euclidean distances between them (Table 5) were calculated. In addition, the average silhouette coefficient was computed for each subtype (Table 6). The overall silhouette score for the filtered dataset (56 structures) was 0.619, indicating good separation quality (values above 0.5 are considered excellent). All five subtypes show positive silhouette values, with H2N2 achieving the highest (0.698), followed by H5N1 (0.647), H3N2 (0.635), H1N1 (0.613), and H7N9 (0.449). The lower value for H7N9 may be attributed to its smaller sample size (four structures). The t-SNE projection shows well-separated clusters for all subtypes, while PCA reveals partial merging of H2N2 and H5N1, highlighting the non-linear nature of the structural features. Despite the known tendency of t-SNE to amplify separation, the agreement with PCA and the good silhouette score indicate that the observed clusters reflect genuine structural differences. The visualization results therefore support the effectiveness of the proposed approach, while acknowledging the limitations of non-linear dimensionality reduction.

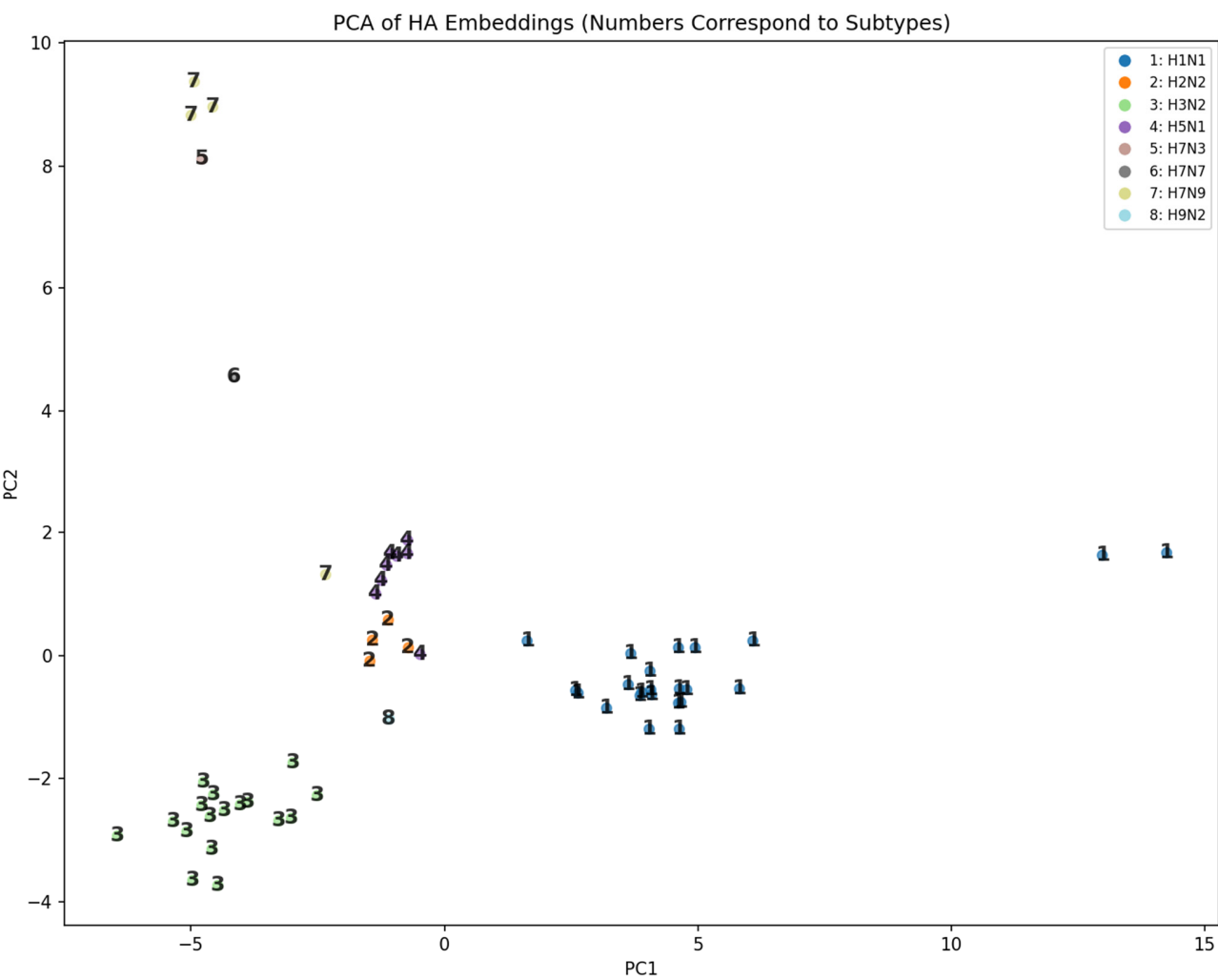


Figure 3. PCA projection of the 128-dimensional embeddings for the filtered dataset (56 non-redundant structures) with numeric subtype labels.

Table 4. Cluster characteristics in t-SNE space.

Subtype	Centroid Coordinates (x, y)	Number of Points
H1N1	(2.73, −11.49)	23
H3N2	(3.96, −7.94)	17
H5N1	(0.45, −9.20)	8
H2N2	(1.83, −8.95)	4
H7N9	(2.36, −7.22)	4

Table 5. Pairwise Euclidean distances between cluster centroids.

Pair	Distance
H1N1–H2N2	2.69
H1N1–H3N2	3.76
H1N1–H5N1	3.23
H1N1–H7N9	4.29
H2N2–H3N2	2.36
H2N2–H5N1	1.40
H2N2–H7N9	1.82
H3N2–H5N1	3.73
H3N2–H7N9	1.76
H5N1–H7N9	2.76

Table 6. Average silhouette score by subtype.

Subtype	Average Silhouette Score
H1N1	0.6125
H2N2	0.6984
H3N2	0.6347
H5N1	0.6470
H7N9	0.4485

3.3. Hierarchical Clustering

To further verify the structural similarity between subtypes, we performed Ward’s hierarchical clustering on the same 128-dimensional embeddings. The resulting dendrogram (Figure 4) shows a clear separation into five main branches, which completely matches the visual pattern of the t-SNE plot and confirms the classification results. The branches corresponding to the H1N1, H2N2, H3N2, H5N1, and H7N9 subtypes separate at high hierarchy levels, indicating their substantial separation in the embedding space. Importantly, no branch overlaps with another, further confirming the model’s ability to capture subtle structural differences between hemagglutinin subtypes.

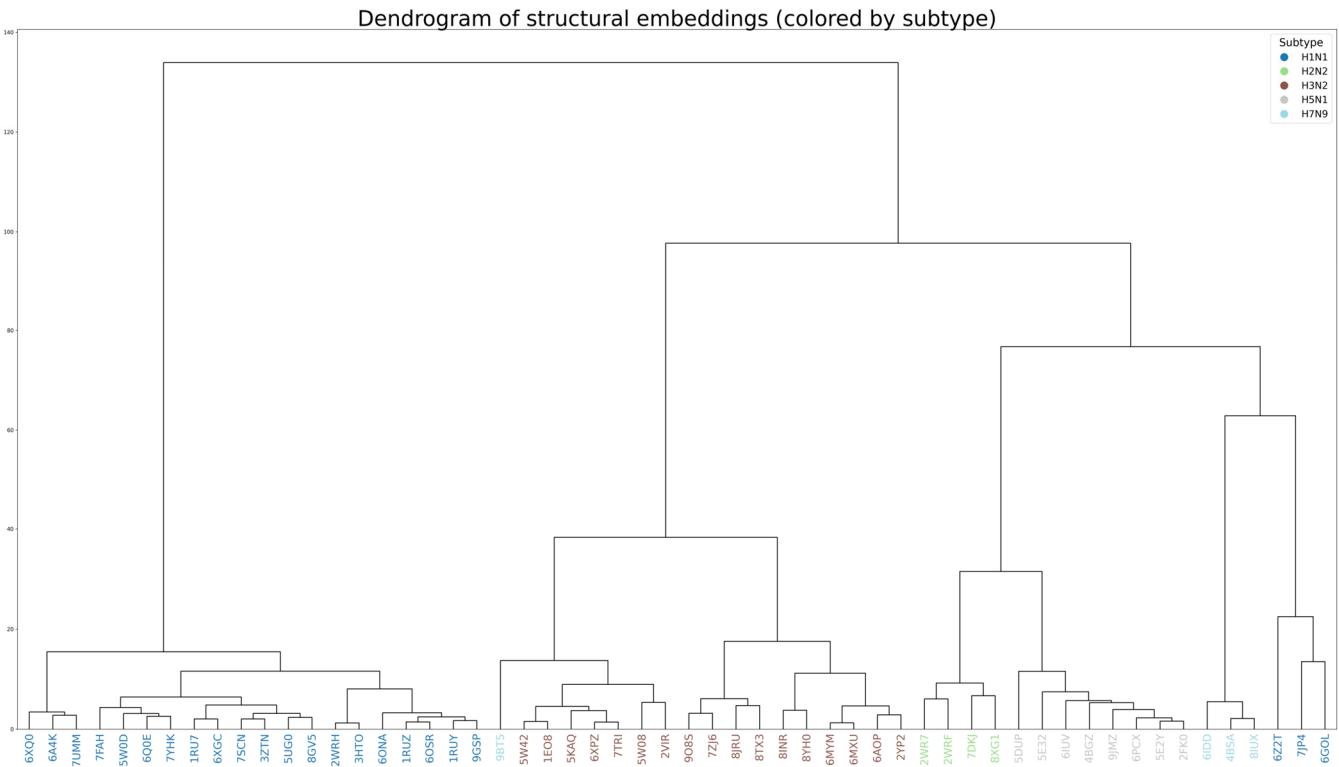


Figure 4. Hierarchical clustering of influenza A virus hemagglutinin based on embeddings (Ward’s method).

3.4. Comparison of Structural and Genetic Distances

To assess the biological relevance of the obtained structural embeddings, we quantitatively compared the distance matrices derived from them with classical phylogenetic distances calculated from amino acid sequences. For the filtered dataset (56 non-redundant structures), amino acid sequences were extracted from the corresponding PDB files. Multiple sequence alignment was performed using MUSCLE [14], after which a genetic distance matrix was calculated based on the p-distance (proportion of mismatched positions). The structural distance matrix was obtained by computing pairwise Euclidean distances

between the 128-dimensional embeddings extracted from the penultimate layer of the EpitopeGNN model.

Mantel correlation: To evaluate the similarity between the two distance matrices, the Mantel test was applied with 999 permutations. Pearson's correlation coefficient between the matrices was $r = 0.38$ with a high level of statistical significance ($p = 0.001$). Using Spearman's rank correlation, we obtained $\rho = 0.41$ ($p < 0.001$), confirming a significant monotonic relationship between structural and genetic distances. This value indicates a moderate but significant positive correlation. The result demonstrates that the structural embeddings, trained without any evolutionary information (based solely on three-dimensional structure), nonetheless capture a phylogenetic signal. The squared correlation coefficient (r^2) indicates the proportion of variance in one variable that is explained by a linear relationship with the other variable:

$$r^2 = (0.38)^2 \approx 0.144 (\text{i.e., } \sim 14.4\%).$$

Thus, the proportion of variance not explained by a linear relationship is $1 - r^2 = 1 - 0.144 = 0.856$, or $\sim 85.6\%$. The relatively low proportion of explained variance ($r^2 \approx 0.14$) suggests that the structural embeddings reflect not only evolutionary differences encoded by the sequences but also additional aspects of spatial organization. This implies that such embeddings may carry information important for the analysis of functional and antigenic properties, which are directly linked to the three-dimensional protein structure. Further studies are needed to test this hypothesis. In more detail, the unexplained variance is not "empty" but rather comprises several meaningful factors:

Non-linear relationships: The relationship between genetic and structural distances may not be strictly linear. For example, small changes in sequence can lead to disproportionately large structural rearrangements (or vice versa). If non-linear measures (e.g., rank-based distances) were used, the proportion of explained variation might change.

Noise and measurement errors: Genetic distances were calculated from sequences, which may contain uncertainties (especially in alignment). Structural embeddings, in turn, were obtained from a GNN model trained on a limited dataset and may include some noise or artifacts.

Additional information: The three-dimensional structure of a protein carries information that is not completely determined by the amino acid sequence. In particular, it has been shown [15] that when sequence identity falls below 25–30% reliably predicting structural similarity from sequence alone is impossible. This fundamentally justifies the use of structural information as an additional source. For instance, conformational variability within a single subtype (different H1N1 strains may have somewhat different structures, which is reflected in the embeddings) represents important structural signal that is not reducible to ordinary amino acid sequence. Furthermore, side-chain packing, surface charge distribution, etc., may also influence the structural embeddings but are poorly captured by simple p-distance. This part is often referred to as "additional information" and is potentially important for protein function.

Thus, given that the GNN model was trained exclusively on three-dimensional structures and physicochemical properties (without directly using sequences), it is reasonable to assume that a significant portion of the unexplained variation reflects structural features that are not reducible to sequence.

3.5. Distance Matrix Heatmaps

For a more detailed analysis, Figure 5 shows heatmaps of the genetic and structural distance matrices calculated on the filtered non-redundant dataset (56 structures). For clarity, the rows and columns of both matrices are ordered according to the hierarchical

clustering of the structural embedding. In these plots, each cell represents the distance between two hemagglutinin structures:

- The horizontal and vertical axes correspond to the filtered dataset (in the order obtained from hierarchical clustering).
- The color of each cell represents the distance magnitude:
 - Dark blue → structures are very close (small distance).
 - Green/yellow → intermediate distance.
 - Red/orange → large distance (strongly different).

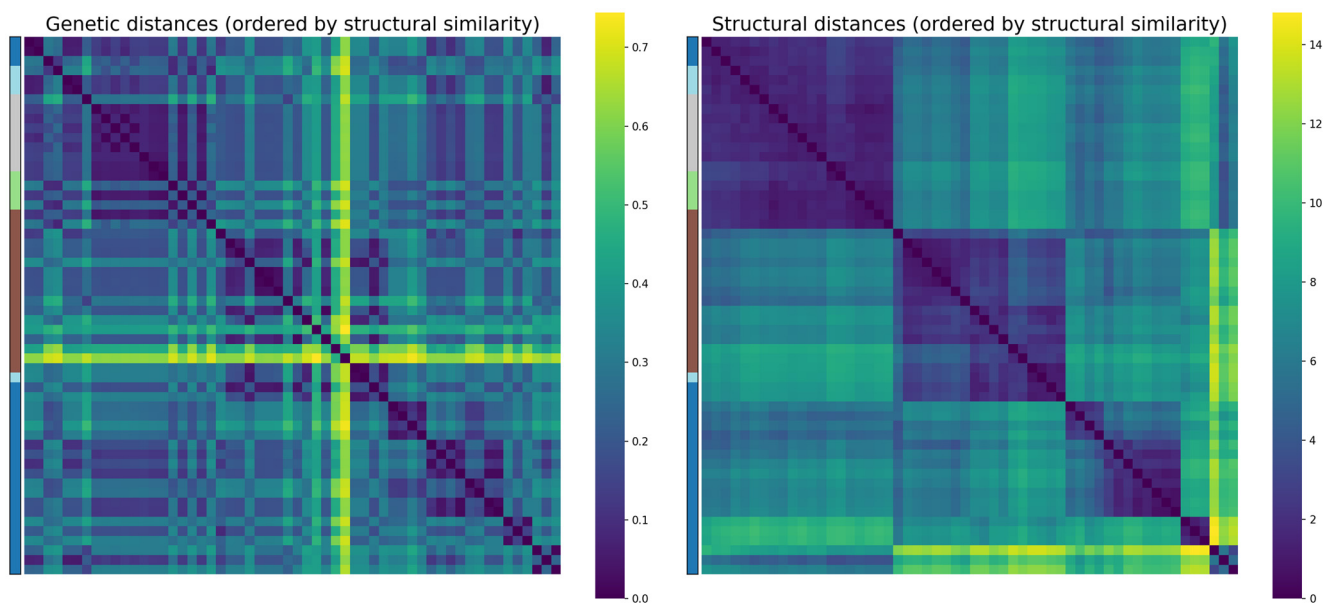


Figure 5. Heatmaps of genetic (**left**) and structural (**right**) distance matrices for the filtered non-redundant dataset (56 structures). Rows and columns are ordered by hierarchical clustering of structural embeddings.

Because the rows and columns have been reordered (clustered), similar structures appear near each other. Along the diagonal (from the top left to the bottom right), structures are compared with themselves—the distance is always 0, so the diagonal is dark blue. The blocks along the diagonal represent groups of structures that are close to each other. To the left of each heatmap, a color annotation of subtypes is added. This subtype annotation is a vertical bar to the left of the heatmap, where each structure is represented by a vertical strip whose color encodes its subtype. Continuous regions of the same color correspond to compact groups of structures belonging to the same subtype, as formed by hierarchical clustering. The structural matrix shows clearer boundaries between some clusters (e.g., H1N1 and H3N2), which is consistent with the high classification accuracy.

3.6. Residue Importance Analysis

To interpret the model, we attempted to analyze the attention coefficients α_{ij} from the first GAT layer. However, due to the softmax normalization over neighbors of each node, the sum of incoming attention weights for any node equals 1, and the α_{ij} values are almost indistinguishable (mean = 0.998, standard deviation $< 10^{-7}$). This makes direct attention analysis uninformative.

As an alternative, we computed the L2 norm of the node features after the first GAT layer. This quantity reflects how strongly a node's activation influences the subsequent layers and the final decision. For a representative H1N1 structure (PDB: 3ZTN), the following values were obtained: min = 80.3, max = 152.1, mean = 115.9, standard deviation = 12.1.

The distribution of importance along the sequence is shown in Figure 6. The substantial spread indicates that the model is not degenerate and indeed assigns differential relevance to individual residues. However, due to the complex multimeric structure of the protein (multiple chains, ligands) and the lack of a direct mapping between node indices and canonical residue numbers, we refrain from making quantitative claims about the spatial localization of important residues. The obtained results confirm that the model does not ignore differences between residues, but precise interpretation of importance requires further structural analysis.

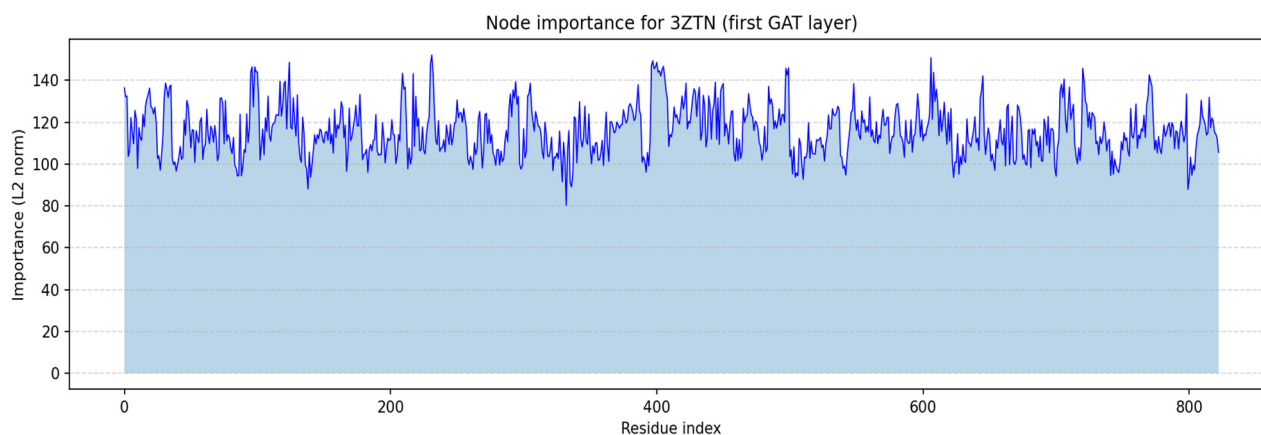


Figure 6. Node importance (L2 norm of the first GAT layer output) plotted along the residue index for the H1N1 HA structure (PDB: 3ZTN). The importance values exhibit substantial variation (min = 80.3, max = 152.1, mean = 115.9, std = 12.1), indicating that the model assigns differential relevance to individual residues. This analysis serves as an alternative to direct attention weights, which are nearly uniform due to softmax normalization in GAT layers.

Thus, although direct analysis of attention weights in GAT is limited by the properties of softmax, the proposed metric demonstrates that EpitopeGNN is not degenerate and can distinguish residues by their contribution to classification.

4. Discussion

The achieved 100% classification accuracy on the test set and 97.6% with five-fold cross-validation convincingly demonstrate that the three-dimensional structure of HA contains all the necessary information for discriminating influenza A virus subtypes. However, to rule out overestimation due to near-identical sequences, we repeated the evaluation on a non-redundant dataset (92% identity filter). The model maintained 95–100% accuracy, confirming the robustness of the result (see Section Robustness to Sequence Redundancy).

In contrast to existing methods (ClassyFlu [1], INFINITY [2]) that work exclusively with DNA or amino acid sequences, the proposed EpitopeGNN approach employs a graph representation of the protein spatial fold and graph neural network mechanisms—an approach that, to the best of our knowledge, has not been previously applied to influenza subtype classification.

Quantitative analysis of embedding clustering for the filtered dataset (56 structures) confirmed high separation quality: the overall silhouette coefficient was 0.619, and all five subtypes showed positive silhouette values, with H2N2 achieving the highest (0.698) and H7N9 the lowest (0.449) (see Table 6). The t-SNE projection shows well-separated clusters for all subtypes, while PCA reveals partial merging of H2N2 and H5N1, highlighting the non-linear nature of the structural features. At the same time, the high classification accuracy on the non-redundant dataset (95–100%) indicates that in the original 128-dimensional embedding space these subtypes remain well-separable. Furthermore,

a significant correlation was quantitatively demonstrated between structural embeddings and phylogenetic distances, indirectly confirming the presence of an evolutionary signal in the spatial protein fold.

Comparison with existing services highlights the fundamental novelty of this work: none of them use three-dimensional structures. Thus, EpitopeGNN does not replace but complements traditional genetic classifiers by providing a fundamentally different level of information—structural. This opens new possibilities for monitoring virus evolution, enabling tracking not only of genetic changes but also of rearrangements in protein spatial folding that may affect antigenic properties and pandemic potential.

Comparison with sequence-based methods: Traditional approaches for influenza subtype classification, such as ClassyFlu [1], INFINITy [2], operate exclusively on nucleotide or amino acid sequences. Although they are fast and widely applicable, they cannot account for the spatial folding of the protein, which is directly linked to antigenic properties. A recent method, CLBTope [6], also uses only sequence information to predict B-cell epitopes, but its accuracy for hemagglutinin is naturally limited because conformational epitopes are defined by spatial proximity rather than linear sequence. Our structural embedding-based approach overcomes this limitation by explicitly modeling residue interactions and surface accessibility, enabling not only accurate subtype classification but also the identification of potential antigenic determinants. This structural perspective is critical for tracking antigenic drift and for guiding vaccine strain selection. Quantitative comparison on a non-redundant dataset (92% identity filter) showed that EpitopeGNN outperforms the best sequence-only baseline by 10 percentage points (Experiment A) and 7.7 percentage points (Experiment B). These results unequivocally demonstrate the added value of 3D structural information for influenza subtype classification.

Limitations and prospects

The model requires an experimentally determined or reliably predicted three-dimensional structure, which is not always available for new strains. However, the rapid development of protein structure prediction methods such as AlphaFold will soon remove this limitation.

Rare subtypes (H7N3, H7N7, H9N2) are represented by a small number of samples, which may affect clustering stability. Enlarging the dataset with new experimental structures or high-quality predicted models will improve classification reliability for these subtypes.

Future work will include enriching the node feature representation with evolutionary information (Position-Specific Scoring Matrix, PSSM), as well as exploring the use of the obtained embeddings for finding structural analogs and predicting other functional properties of HA, such as receptor specificity (human $\alpha 2,6$ vs. avian $\alpha 2,3$).

5. Conclusions

1. We developed EpitopeGNN, a graph attention network that classifies influenza A virus subtypes directly from the 3D structure of hemagglutinin using residue interaction networks and physicochemical features.
2. After redundancy reduction (92% identity), EpitopeGNN maintained 95–100% accuracy on non-redundant data, outperforming sequence-only baselines (85% for multi-class, 92.3% for binary). This confirms that the high performance is not due to data leakage and that 3D structure is essential for discriminating closely related subtypes.
3. The method complements traditional sequence-based classifiers and enables structural monitoring of virus evolution, analog searching, and prediction of functional properties (e.g., receptor specificity).
4. Future work will incorporate evolutionary information (PSSM), expand the dataset with predicted structures (AlphaFold, ESMFold), and apply embeddings to related tasks.

5. The structural embeddings can serve as a foundation for predicting receptor specificity and epitope immunogenicity—tasks that are challenging for sequence-only methods like CLBTope.

Author Contributions: Conceptualization, A.T., A.A. and I.I.-S.; methodology, A.T.; software, A.T.; validation, A.T. and A.A.; formal analysis, A.T.; investigation, A.T.; resources, A.A. and O.E.; data curation, A.A.; writing—original draft preparation, A.T.; writing—review and editing, A.A. and I.I.-S.; supervision, A.A. and O.E.; project administration, A.A., O.E. and I.I.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original protein structures analyzed in this study are publicly available from the Protein Data Bank (PDB) at <https://www.rcsb.org/> (accessed on 10 March 2026). The embeddings generated by the EpitopeGNN model, along with the annotation file `ha_subtypes_full.csv`, are available from the corresponding author upon reasonable request. All source code and scripts used for data processing, model training, and visualization are deposited in a public GitHub repository: <https://github.com/andytimoffilim/epitope-gnn-ha-subtype-classification> (accessed on 14 April 2026).

Conflicts of Interest: Authors Andrey Timofeev and Alexander Anufriev are employed by LLC “AI Center for SCO+ Countries”. The remaining authors declare that the research was conducted without any commercial or financial relationships that could be perceived as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GAT	Graph attention network
GISAID	Global initiative on sharing all influenza data
GNN	Graph neural network
HA	Hemagglutinin
HMMs	Hidden Markov models
MLP	Multilayer perceptron
PCA	Principal component analysis
PDB	Protein data bank
PSSM	Position-specific scoring matrix
RIN	Residue interaction network
t-SNE	t-Distributed stochastic neighbor embedding

References

1. Van der Auwera, S.; Bulla, I.; Ziller, M.; Pohlmann, A.; Harder, T.; Stanke, M. ClassyFlu: Classification of influenza A viruses with Discriminatively trained profile-HMMs. *PLoS ONE* **2014**, *9*, e84558. [CrossRef] [PubMed]
2. Cacciabue, M.; Marcone, D.N. INFINITY: A fast machine learning-based application for human influenza A and B virus subtyping. *Influenza Other Respir. Viruses* **2023**, *17*, e13096. [CrossRef] [PubMed]
3. Ceccarelli, F.; Giusti, L.; Holden, S.; Liò, P. Integrating Structure and Sequence: Protein Graph Embeddings via GNNs and LLMs. In Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods—ICPRAM, Rome, Italy, 24–26 February 2024. [CrossRef]
4. Chatzianastasis, M.; Dasoulas, G.; Vazirgiannis, M. Geometric Self-Supervised Pretraining on 3D Protein Structures using Subgraphs. *arXiv* **2024**, arXiv:2406.14142. [CrossRef]
5. Begue, S.C.; Leonardi, E.; Tosatto, S.C.E. Decoding protein structures with residue interaction networks. *Trends Biochem. Sci.* **2025**, *50*, 1072–1085. [CrossRef] [PubMed]
6. Kumar, N.; Tripathi, S.; Sharma, N.; Patiyal, S.; Devi, N.L.; Raghava, G.P.S. A method for predicting linear and conformational B-cell epitopes in an antigen from its primary sequence. *Comput. Biol. Med.* **2024**, *170*, 108083. [CrossRef] [PubMed]

7. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [[CrossRef](#)] [[PubMed](#)]
8. Shu, Y.; McCauley, J. GISAID: Global initiative on sharing all influenza data—From vision to reality. *Euro Surveill.* **2017**, *22*, 30494. [[CrossRef](#)] [[PubMed](#)]
9. Miyazawa, S.; Jernigan, R.L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **1996**, *256*, 623–644. [[CrossRef](#)] [[PubMed](#)]
10. Kyte, J.; Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105–132. [[CrossRef](#)] [[PubMed](#)]
11. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **1974**, *185*, 862–864. [[CrossRef](#)] [[PubMed](#)]
12. Brandes, U. A Faster Algorithm for Betweenness Centrality. *J. Math. Sociol.* **2001**, *25*, 163–177. [[CrossRef](#)]
13. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph attention networks. In Proceedings of the 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
14. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [[CrossRef](#)] [[PubMed](#)]
15. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng.* **1999**, *12*, 85–94. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.