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

Andrey Timofeev , Alexander Anufriev , Oleg Ergashev , Irina Isakova-Sivak *

Posted Date: 7 April 2026

doi: 10.20944/preprints202604.0291.v1

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

 doi:10.20944/preprints202604.0291.v1

Article

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

Andrey Timofeev [1], Alexander Anufriev [1], Oleg Ergashev [2,3] and Irina Isakova-Sivak [2,*]

[1] LLC "AI Center for SCO+ Countries", St. Petersburg, Russia

[2] Institute of Experimental Medicine, St. Petersburg, Russia

[3] Department of Hospital Surgery No. 2 with the Clinic named after Academician F.G. Uglov of the Pavlov First Saint Petersburg State Medical University, St. Petersburg, Russia

* Correspondence: isakova.sivak@iemspb.ru

Abstract

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. 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. By utilizing physicochemical properties of amino acids and topological centrality measures, we achieved 100% classification accuracy on the test set and 97.6% with five-fold cross-validation. A significant correlation was found between the obtained structural embeddings and phylogenetic distances ($r = 0.48$, $p < 0.001$), confirming their biological relevance and opening opportunities for structural monitoring of virus evolution, as well as rapid analog searching for novel strains.

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

1. Introduction

Influenza A virus causes seasonal epidemics and pandemics, posing a serious threat to public health. Hemagglutinin (HA) plays a key role in its identification—a protein responsible for attachment to the host cell and defining 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). This work is the first to address influenza subtype classification using three-dimensional structures and GNNs, allowing the incorporation of 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

 doi:10.20944/preprints202604.0291.v1

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

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

2.2. Construction of the RIN Graph

Let a set of residues R={r_1, r_2, …, 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 $\mathbf{c}_i \in \mathbb{R}^3$ — 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 $\delta = 8$ Å – is the chosen spatial proximity threshold. The choice of $\delta = 8$ Å is based on 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.

 doi:10.20944/preprints202604.0291.v1

2. **Physicochemical properties** (4 dimensions): charge, hydrophobicity (according to the Kyte–Doolittle scale [10]), volume, 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 Å³). For example, Gly ~48 Å³, Trp ~163 Å³.
- **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

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

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 a global mean pooling operation 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 – nonlinear activation function (ELU is used in this work).

The attention coefficients are computed using an additive attention mechanism:

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

Here:

- $\|$ denotes vector concatenation;
- $\mathbf{a}^{(l)} \in \mathbb{R}^{2F'}$ – 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:

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2026 doi:10.20944/preprints202604.0291.v1

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

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, a 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 = 5e-4, batch size = 4, and the number of epochs was 100. To assess stability, five-fold cross-validation was performed.

3. Results

3.1. Classification Performance

On the test set, the model achieved 100% accuracy for all subtypes. Five-fold cross-validation yielded an average accuracy of 97.6% (σ=2.3%). Table 2 presents the classification report for the test set.

Table 2. Classification metrics on the test set.

Subtype	Precision	Recall	f1-score	Support
H1N1	1.00	1.00	1.00	25
H2N2	1.00	1.00	1.00	3
H3N2	1.00	1.00	1.00	24
H5N1	1.00	1.00	1.00	15
H7N9	1.00	1.00	1.00	5
other	1.00	1.00	1.00	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-SNE algorithm (Figure 1). The projection includes all 249 structures belonging to eight hemagglutinin subtypes: H1N1 (84), H2N2 (11), H3N2 (80), H5N1 (48), H7N3 (4), H7N7 (4), H7N9 (15), and H9N2 (3).

Visualization. In Figure 1, seven compact clusters are clearly visible. Five of them correspond to the abundant subtypes H1N1, H2N2, H3N2, H5N1, and H7N9. The clusters of the rare subtypes

 doi:10.20944/preprints202604.0291.v1

H7N3 and H7N7 (4 structures each) almost completely overlap with the H7N9 cluster on the plane, which is not unexpected since all these subtypes share H7 HA molecules (distances between them in the t-SNE space are only 0.7–1.5). Despite the small number of samples (3 structures), the H9N2 subtype appears as a separate cluster located near H2N2 (distance between centroids 2.21), which is consistent with the phylogenetic relationship of these subtypes.

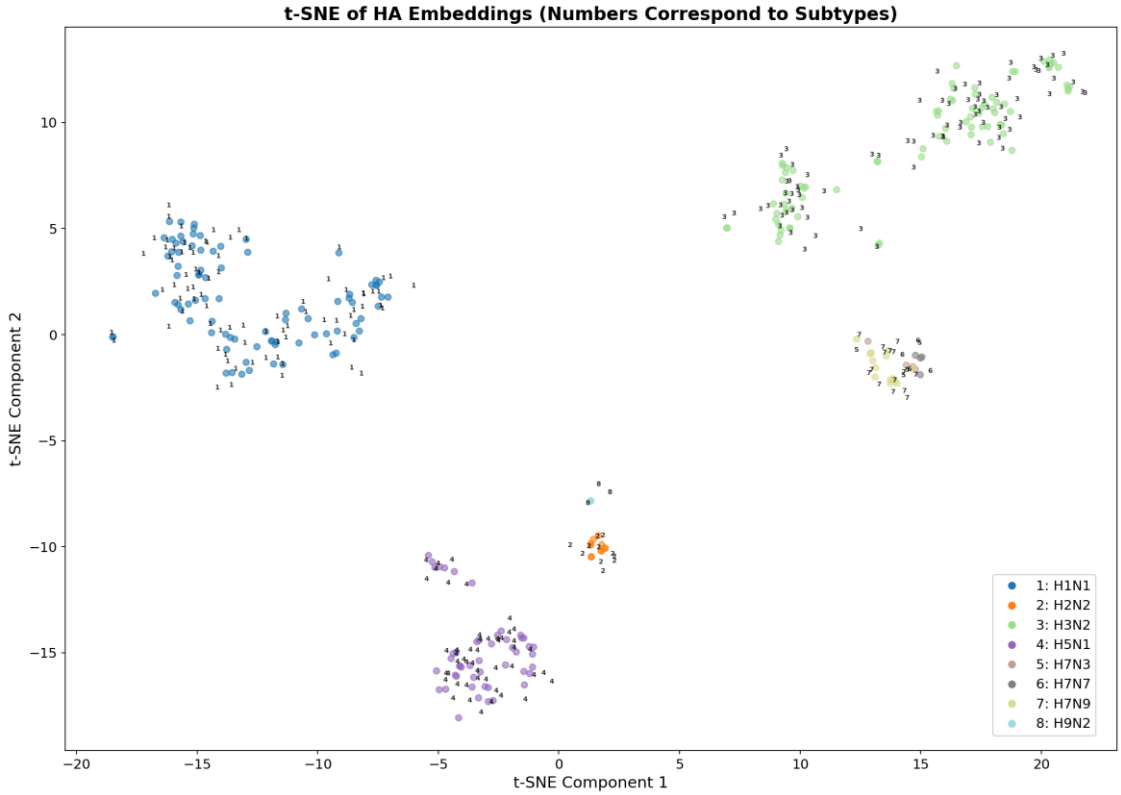

Figure 1. T-SNE clustering of embeddings.

Quantitative analysis. To objectively assess cluster separability, centroid coordinates in the t-SNE space (Table 3) and pairwise Euclidean distances between them (Table 4) were calculated. In addition, the average silhouette coefficient was computed for each subtype, characterizing cluster compactness and separation from neighboring clusters (Table 5). The overall silhouette score for all 249 structures was 0.495, indicating good separation quality (values above 0.5 are considered excellent). It should be noted that for subtypes with a small number of samples (H7N3, H7N7, H9N2), silhouette estimates may be less stable due to high sensitivity to outliers.

Table 3. Cluster characteristics in t-SNE space.

Subtype	Centroid coordinates (x, y)	Number of dots
H1N1	(-12.74, 1.67)	84
H2N2	(1.58, -10.02)	11
H3N2	(14.81, 9.07)	80
H5N1	(-3.22, -14.91)	48
H7N3	(14.15, -1.21)	4
H7N7	(14.93, -1.25)	4
H7N9	(13.48, -1.54)	15
H9N2	(1.31, -7.82)	3

 doi:10.20944/preprints202604.0291.v1

Table 4. Pairwise Euclidean distances between cluster centroids.

Pair	Distance	Pair	Distance
H1N1 – H2N2	18.48	H2N2 – H3N2	23.23
H1N1 – H3N2	28.53	H2N2 – H5N1	6.85
H1N1 – H5N1	19.11	H2N2 – H7N3	15.35
H1N1 – H7N3	27.05	H2N2 – H7N7	15.98
H1N1 – H7N7	27.83	H2N2 – H7N9	14.62
H1N1 – H7N9	26.42	H2N2 – H9N2	2.21
H1N1 – H9N2	16.95	H3N2 – H5N1	30.00
H3N2 – H7N3	10.30	H3N2 – H7N7	10.32
H3N2 – H7N9	10.69	H3N2 – H9N2	21.63
H5N1 – H7N3	22.12	H5N1 – H7N7	22.72
H5N1 – H7N9	21.40	H5N1 – H9N2	8.41
H7N3 – H7N7	0.78	H7N3 – H7N9	0.74
H7N3 – H9N2	14.45	H7N7 – H7N9	1.48
H7N7 – H9N2	15.13	H7N9 – H9N2	13.70

Table 5. Average silhouette score by subtype.

Subtype	Average silhouette score
H1N1	0.469
H2N2	0.644
H3N2	0.480
H5N1	0.661
H7N3	-0.025
H7N7	0.345
H7N9	0.171
H9N2	0.959

High silhouette values for H1N1, H2N2, H3N2, H5N1, and especially for H9N2 (0.96) indicate exceptional compactness and isolation of these clusters in the original 128-dimensional space. The negative silhouette for H7N3 (–0.025) is explained by the small number of points (4) and their proximity to the larger H7N9 cluster (distance between centroids 0.74). The visual merging of the H7N3, H7N7, and H7N9 clusters in the t-SNE projection does not contradict the classification results: in the high-dimensional space, these subtypes remain distinguishable (100% classification accuracy), but the two-dimensional projection, with its inevitable loss of information, cannot capture the subtle differences between such close clusters. Thus, the visualization results are fully consistent with the performance metrics and confirm the effectiveness of the proposed approach.

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 2) shows a clear separation into seven 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. The rare subtypes H7N3, H7N7, and H9N2 also form distinct branches, albeit at lower heights, which is consistent with their small numbers and proximity to larger clusters (e.g., H7N3 and H7N7 are nested within the H7N9 branch but retain internal clustering). Importantly, no branch overlaps with another, further confirming the model's ability to capture subtle structural differences between hemagglutinin subtypes.

 doi:10.20944/preprints202604.0291.v1

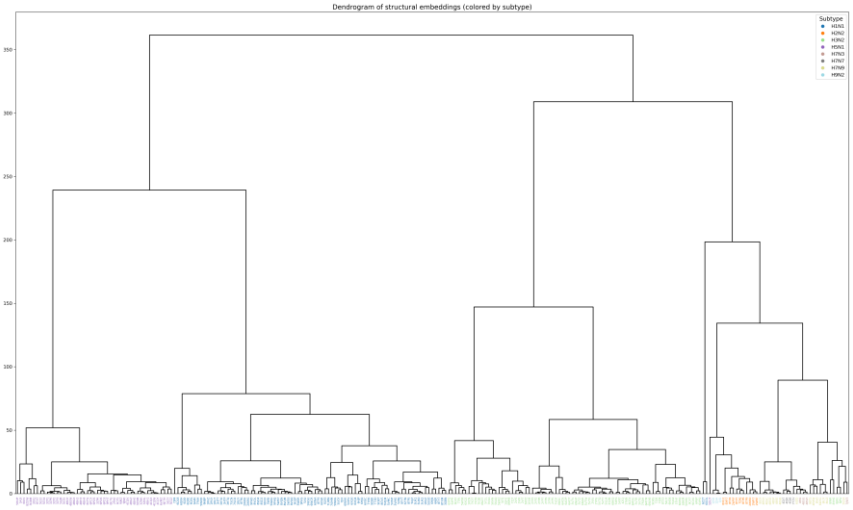

Figure 2. Hierarchical clustering of influenza A virus hemagglutinin based on embeddings (Ward's method).

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 all 249 structures in the dataset, 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 = 0.484** with a high level of statistical significance (**p = 0.001**). 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:

$$r^2 = (0.4842)^2 \approx 0.234 \text{ (e.g. ~23.4\%).}$$

Thus, the proportion of variance not explained by a linear relationship is $1 - r^2 = 1 - 0.234 = 0.766$, or ~76.6%, rounded to ~77%. The relatively low proportion of explained variance ($r^2 \approx 0.23$) 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:

Nonlinear 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 nonlinear 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

 doi:10.20944/preprints202604.0291.v1

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 3 shows heatmaps of the genetic and structural distance matrices. For clarity, the rows and columns of both matrices are ordered according to the hierarchical clustering of the structural embeddings. In these plots, each cell represents the distance between two hemagglutinin structures:

- The horizontal and vertical axes correspond to all 249 structures (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).

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.

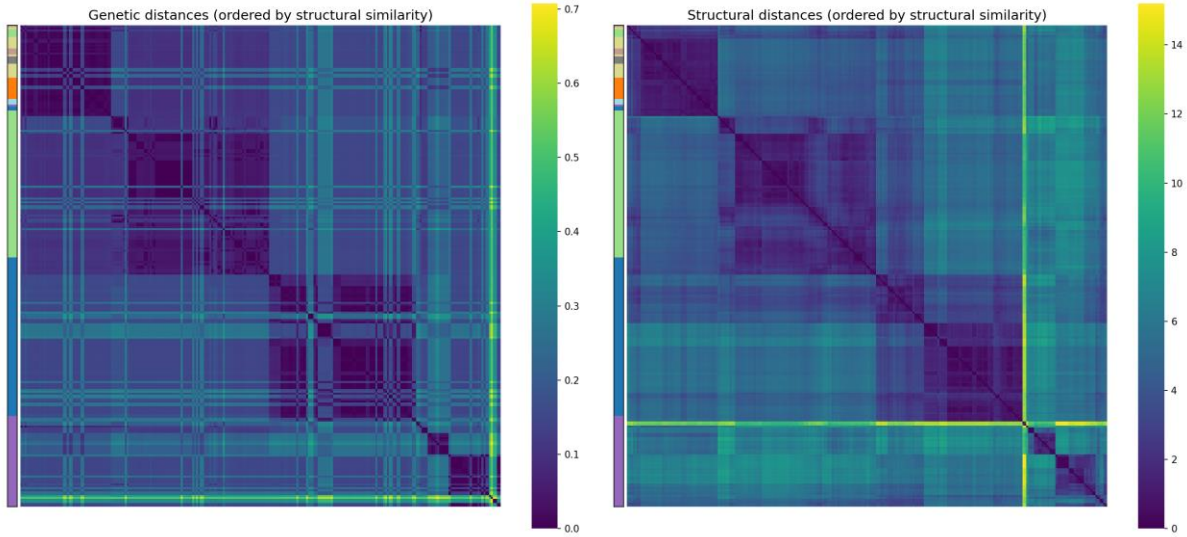

Figure 3. Heatmaps of genetic and structural distance matrices.

 doi:10.20944/preprints202604.0291.v1

3.6. Outlier Analysis

Ordering the data by structural similarity revealed potential structural anomalies. Five structures have a mean Euclidean distance to all other objects exceeding the threshold of two standard deviations (Figure 3, Table 6). These structures can be considered potential structural outliers. Interestingly, four of them belong to subtypes H1N1 (6Z2T, 6GOL) and H2N2 (8TP4, 2WRF). Since the classification accuracy for these subtypes is 100%, these structures are not misclassified; however, their spatial conformation deviates significantly from typical representatives of their subtype. This may be due to features of the specific strain (e.g., mutations in epitopes), structure quality, or crystallization conditions. Structure 5DUT (H5N1) also appears in the outlier list, making it an interesting target for further structural and functional analysis.

Table 6. Potential structural outliers.

PDB ID	Subtype	Mean distance to other structures	Threshold value
6Z2T	H1N1	12.12	6.62
6GOL	H1N1	11.71	6.62
5DUT	H5N1	7.01	6.62
8TP4	H2N2	6.75	6.62
2WRF	H2N2	6.67	6.62

Thus, the analysis confirms that the proposed embeddings are not merely an abstract feature space for classification, but a compact and informative representation reflecting both evolutionary relationships and unique structural features of hemagglutinin.

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 hemagglutinin (HA) contains all the necessary information for discriminating influenza A virus subtypes. In contrast to existing methods (ClassyFlu [1], INFINITy [2]) that work exclusively with DNA or amino acid sequences, the proposed EpitopeGNN approach is the first to use a graph representation of the protein spatial fold and graph neural network mechanisms. This allows it to account not only for the local sequence but also for the global topology of residue interactions and their physicochemical properties, which is critical for understanding antigenic differences.

Quantitative analysis of embedding clustering confirmed high separation quality: the overall silhouette coefficient was 0.495, and for most subtypes (H1N1, H2N2, H3N2, H5N1, H9N2) the average silhouette scores exceeded 0.46, reaching 0.96 for H9N2. The proximity of the centroids of H7N3, H7N7 and H7N9 (distances 0.7–1.5) and the partial overlap of these clusters in t-SNE are explained by the small number of samples (4 and 15) and inevitable distortions upon dimensionality reduction. At the same time, the high classification accuracy and the absence of errors on the test set indicate that in the original 128-dimensional embedding space these subtypes remain completely 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.

 doi:10.20944/preprints202604.0291.v1

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.

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 (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 α2,6 vs. avian α2,3).

5. Conclusions

1. An original graph neural network (**EpitopeGNN**) was developed for classifying hemagglutinin subtypes of influenza A virus, utilizing the three-dimensional protein structure and physicochemical features of amino acid residues.

2. On a test set of 249 structures (8 subtypes), the model achieved 100% accuracy; five-fold cross-validation confirmed the stability of the result (97.6 ± 2.3%).

3. The proposed method complements traditional sequence-based classifiers and opens up opportunities for structural monitoring of virus evolution, searching for structural analogs, and predicting functional properties of hemagglutinin (e.g., receptor specificity).

4. Further development of the model involves incorporating evolutionary information (PSSM), expanding the dataset with predicted structures (AlphaFold, ESMFold, etc.), and applying the embeddings to related tasks.

5. Moreover, the structural embeddings generated by EpitopeGNN may serve as a foundation for predicting functional properties such as 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., 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., O.E.; project administration, A.A., O.E., 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/. 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

 doi:10.20944/preprints202604.0291.v1

deposited in a public GitHub repository: https://github.com/andytimoffilim/epitope-gnn-ha-subtype-classification (accessed on 25 March 2026).

Conflicts of Interest: The authors declare no conflicts 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
PDB	Protein Data Bank
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, doi:10.1371/journal.pone.0084558.
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, doi:10.1111/irv.13096.
3. Ceccarelli, F.; Giusti, L.; Holden, S.; Liò, P. Integrating Structure and Sequence: Protein Graph Embeddings via GNNs and LLMs. *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods - ICPRAM* **2024**, doi:10.5220/0012453600003654.
4. Chatzianastasis, M.; Dasoulas, G.; Vazirgiannis, M. Geometric Self-Supervised Pretraining on 3D Protein Structures using Subgraphs. . *arXiv preprint arXiv:2406.14142.* **2024**, doi:10.48550/arXiv.2406.14142.
5. Begue, S.C.; Leonardi, E.; Tosatto, S.C.E. Decoding protein structures with residue interaction networks. *Trends in biochemical sciences* **2025**, *50*, 1072-1085, doi:10.1016/j.tibs.2025.08.006.
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, doi:10.1016/j.compbiomed.2024.108083.
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 research* **2000**, *28*, 235-242, doi:10.1093/nar/28.1.235.
8. Shu, Y.; McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* **2017**, *22*, doi:10.2807/1560-7917.ES.2017.22.13.30494.
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. *Journal of molecular biology* **1996**, *256*, 623-644, doi:10.1006/jmbi.1996.0114.
10. Kyte, J.; Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology* **1982**, *157*, 105-132, doi:10.1016/0022-2836(82)90515-0.
11. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **1974**, *185*, 862-864, doi:10.1126/science.185.4154.862.
12. Brandes, U. A Faster Algorithm for Betweenness Centrality. . *Journal of Mathematical Sociology* **2001**, *25*, 163-177, doi:10.1080/0022250X.2001.9990249.
13. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph attention networks. . *6th International Conference on Learning Representations (ICLR).* **2018**.

 doi:10.20944/preprints202604.0291.v1

14. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **2004**, *32*, 1792-1797, doi:10.1093/nar/gkh340.
15. Rost, B. Twilight zone of protein sequence alignments. *Protein engineering* **1999**, *12*, 85-94, doi:10.1093/protein/12.2.85.