PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Deciphering gp120 sequence variation and structural dynamics in HIV neutralization phenotype by molecular dynamics simulations and graph machine learning

Journal:	*PROTEINS: Structure, Function, and Bioinformatics*
Manuscript ID	Prot-00399-2021
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	03-Nov-2021
Complete List of Authors:	Li, Yi; College of Mathematics and Computer Science Dali University Dali China Guo, Yu-Chen; College of Mathematics and Computer Science Dali University Dali China Cheng, Hong-Han; College of Mathematics and Computer Science Dali University Dali China Zeng, Xin; College of Mathematics and Computer Science Dali University Dali China Zhang, Xiao-Ling; College of Mathematics and Computer Science Dali University Dali China sang, peng; College of Agriculture and Biological Science Dali University Dali China Chen, Ben-Hui; College of Mathematics and Computer Science Dali University Dali China Yang, Li-Quan; College of Agriculture and Biological Science Dali University Dali China
Key Words:	Envelope glycoprotein gp120, HIV neutralization phenotype, molecular dynamics simulation, graph machine learning, sequence-structure-dynamics of protein

Deciphering gp120 sequence variation and structural dynamics in HIV neutralization phenotype by molecular dynamics simulations and graph machine learning

Yi Li[1], Yu-Chen Guo[1], Hong-Han Cheng[1], Xin Zeng[1], Xiao-Ling Zhang[1], Peng Sang[2], Ben-Hui Chen[1*], and Li-Quan Yang[2*]

1. College of Mathematics and Computer Science, Dali University, Dali, China.
2. College of Agriculture and Biological Science, Dali University, Dali, China.

* Corresponding authors: bhchen_dali@163.com (B.H.C.), ylqbioinfo@gmail.com (L.Q.Y.).

Abstract

HIV exploits the sequence variation and structural dynamics of the envelope glycoprotein gp120 to evade the immune attack of neutralization antibodies, contributing to various HIV neutralization phenotypes. Although the HIV neutralization phenotype has been experimentally characterized, roles of rapid sequence variability and significant structural dynamics of gp120 are not well understood. Here, 45 prefusion gp120 from different HIV strains belong to three tiers of sensitive, moderate, and resistant neutralization phenotype are structurally modeled by homology modeling and investigated by molecular dynamics simulations and graph machine learning. Our results show that the structural deviations, population distribution, and conformational flexibility of gp120 are related to the HIV neutralization phenotype. Per-residue dynamics indicate the relationship between the local regions, especially in the second structural regions with high flexibility may be responsible for HIV neutralization phenotypes. Moreover, a graph machine learning model with the attention mechanism was selected to explore inherent representation related to the classification of the HIV neutralization phenotypes, leading to further distinguish the strong related gp120 sequence variation together with structural dynamics in HIV neutralization phenotype. Our study not only deciphers gp120 sequence variation and structural dynamics in the HIV neutralization phenotype but also provides a methodological framework to explore complex relationships between sequence, structure, and dynamics of protein by combining molecular dynamics simulations and machine learning.

KEYWORDS

Envelope glycoprotein gp120, HIV neutralization phenotype, sequence, structure, and dynamics of protein, molecular dynamics simulation, graph machine learning

1. INTRODUCTION

The entry of HIV into target cells is mediated by the molecular interaction between the virus and the cell via the envelope glycoprotein gp120, which is the only virus-encoded protein exposed on the surface of the virion [1]. To avoid the attack of immune neutralization antibodies, gp120 undergoes rapid sequence variation and significant structural dynamics [2]. These features of gp120 pose major obstacles to the development of effective antivirus vaccines [3]. In terms of sequence variation, approximately half of the gp120 surface has more than 10% genetic variability to confuse the host immune system [4]. For structural dynamics, the prefusion gp120, which adopts a conformation of immune defense, has significant structural flexibility to mask the binding sites of the receptor and coreceptor [5]. The sequence variation and structural dynamics of gp120 make HIV present diverse adaptability under different physiological or experimental environments [6]. Therefore, in the environment of interest, the knowledge of sequence patterns that are associated with structural dynamics will inform our understanding of the molecular mechanism of HIV viral entry and immune escape to improve the development of antiviral drugs [7][8].

Based on the experimental assessments in a panel of genetically diverse HIV-positive plasma pools, HIV isolates are usually classified as neutralization ‘tier’ phenotype [6]. Current tiered categorization of the neutralization phenotype has suggested that HIV isolates exhibit a spectrum of neutralization phenotypes that can be divided into 1 to 3 tiers, corresponding to the sensitive, moderate, and resistant to neutralization antibodies, respectively.[9] Tier 1 is the most sensitive neutralization phenotype and often comes from laboratory-adapted strains without immune selection pressure. Although tier 1 was further classified into two subtypes, e.g. tier 1A and tier 1B, in some studies[9], it is feasible and efficient to simply use tier 1 in terms of neutralization sensitivity. Tier 2 exhibits a moderate neutralization phenotype and represents most circulating strains. Tier 3, isolated from the primary clinical sample, displays a least sensitive or most resistant neutralization phenotype.

Although the HIV neutralization phenotype has been experimentally characterized, the roles of sequence variability and structural dynamics of gp120 are not well understood. Our previous theoretical study of gp120 with the extremely sensitive and resistant neutralization preliminarily revealed the molecular mechanism of HIV neutralization phenotype by employing the molecular dynamics (MD) simulations [10]. It shows that the neutralization-sensitive gp120 from Tier 1 has higher global structural dynamics and richer conformational substates than the neutralization-resistant gp120 from Tier 3, promoting the neutralization-sensitive gp120 into a more flexible conformational state that interacts with the receptor and coreceptor. Although MD can provide an atomic exploration of protein conformational dynamics, it is still time-consuming and computationally expensive to perform a series of long-term simulations on GPU-accelerated clusters for protein with a usual size (approximately 500 amino acids). Moreover, the modern post-simulation trajectory analysis methods, such as root mean square deviation (RMSD), root mean square fluctuation (RMSF), principal component analysis, and free energy landscape (FEL), only extract statistical information during the simulation process, and cannot provide a learnable and trainable model with predictive functions.

In recent years, machine learning (ML), especially deep learning, has gained attention from academia and industry because of its surprisingly great potential in bioinformatics applications such as protein structure modeling [11]. Deep learning algorithms exploit training data to discover potential patterns and build relational mapping models to make predictions. Some well-known deep learning algorithms, such as classic convolutional neural network (CNN), recurrent neural network (RNN), and other neural-network-based models, have been widely used in biological sequence and structure analysis [12]. It should be noted that using ML to help optimize MD simulations gradually becomes a new trend in molecular science research. Wang et al. combined MD simulations with ML to predict the impacts of mutations-induced variation of protein-ligand binding affinity [13]. Bignon et al. explored the complex combinatorial chemistry of tandem DNA lesions repair and more generally local multiple damaged sites of the utmost significance in radiation chemistry by using MD simulations and ML [14]. Degiacomi mined the protein conformational space by training a neural network based on the MD simulations to supplement the existing conformations of the protein [15]. However, there are few attempts to probe the relationship between the sequence variation and structural dynamics of protein by employing ML and MD simulations. Moreover, for non-Euclidean data such as proteins, the application of graph ML in MD simulations is rarely reported.

In this study, we combine MD simulations and graph ML to probe the roles of rapid sequence variability and significant structural dynamics of gp120 in the HIV neutralization phenotypes. Randomly selected from an experimental panel [9], 45 prefusion gp120 structural models from different HIV strains belong to three tiers of sensitive, moderate, and resistant neutralization phenotype were built. Subsequently, these models were subjected to MD simulations followed by the comparative analysis of the structural deviations, conformational flexibility, and population distribution. In addition, a graph ML model with the attention mechanism was used to explore inherent representation related to the classification of the HIV neutralization phenotype, leading to further distinguish the strong related gp120 sequence variation together with structural dynamics in HIV neutralization phenotype. Our study not only deciphers gp120 sequence variation and structural dynamics in the HIV neutralization phenotype but also provides a methodological framework to explore complex relationships between sequence, structure, and dynamics of protein by combining MD simulations and graph ML.

2. RESULTS

2.1. Structural architecture of the prefusion gp120

As shown in Figure 1A, the structural model of the prefusion gp120 is represented by the atomic coordinate of gp120 in the full-length crystal structure of HIV-1 prefusion envelope glycoprotein (PDB ID: 5FYK) [5]. The structural architecture of gp120 is composed of a relatively conserved functional core, a hydrophobic subdomain centered on the bridging sheet (β2, β3, β20, and β21), and five variable regions (V1–V5) located in the peripheral areas [16]. The hydrophobic subdomain from the stem of the V1/V2 region to the V3 loop plays a regulatory role in mediating the conformational dynamics of gp120 [17]. There are three layered structural regions above the hydrophobic subdomain [18]. The first layer is located between β$\overline{3}$ and β0 and mainly contains α0; the second layer is located between β1 and β5 and mainly contains α1; the third layer is located between β7 and β25 and mainly contains α5.

In the peripheral region, the V1/V2 region presents different structural characteristics. There is a "Greek key" structural folding composed of four β-strands (βA-βD) in the V1/V2 region [19]. The variable loop V1 is located between βA and βB, V2 is located between βC and βD, and the connecting loop L1 is located between βB and βC. Extending from the V1/V2 region and the V3 loop, a 4-stranded β sheet, called the bridging sheet, has been considered as a regulatory switch for conformational transitions of gp120 [20].

Based on the experimental assessment of HIV neutralization phenotype with a broad range of genetic and geographic diversity [9], 45 gp120 structural models from different HIV strains with sensitive, moderate, and resistant neutralization phenotype were randomly selected and constructed by homology modeling, respectively. As shown in Figure 1B, all gp120 models from three neutralization phenotypes exhibit similar structural architecture without some loops have different spatial orientations. These constructed structural models have been validated and used for MD simulations and structure-based ML.

2.2. Structural Deviations of gp120

To quantitatively evaluate the structural deviation of gp120 with different neutralization phenotypes during the simulations, the time evolution of backbone RMSD values (Figure 2) relative to the template structure (PDB ID: 5FYK) were calculated. For the neutralization-sensitive, moderate, and resistant gp120, different relaxation time (about 10, 5, and 3 ns, respectively) is required to reach the equilibrium stage of simulations. A longer relaxation time means that the neutralization-sensitive gp120 has higher structural dynamics, which makes it easier to be exposed and attacked by neutralization immune antibodies [21]. In contrast, the structure of neutralization-resistant gp120 is easier to achieve stability which can be generally considered as a neutralization-resistant conformation after long-term immune selection [22]. For neutralization-moderate gp120, the relaxation time in the middle matches its transition position in the neutralization phenotype of HIV.

After reaching a relatively stable RMSD equilibrium, the neutralization-sensitive, moderate, and resistant gp120 fluctuate in a similar structural region around the average RMSD values of 0.47, 0.44, and 0.42 nm, respectively. Average RMSD is positively correlated with the degree of neutralization phenotype, indicating that gp120 with a more neutralization sensitive phenotype has a higher structural deviation. Except for the subtle differences in the RMSD timing changes, the RMSD values of gp120 with different neutralization phenotypes are distributed in the range of about 0.3–0.7 nm, indicating that gp120 can cover sufficient structural deviations to maintains the basic viral infection function while showing different neutralization phenotypes.

2.3. Dynamics of the Hydrophobic Subdomain

A recent study has pointed out that the hydrophobic subdomain composed of the V1/V2 region, the V3 loop, and the bridging sheet plays a regulatory role in mediating the conformational dynamics of gp120, contributing to the neutralization phenotype of HIV [17]. Therefore, the Rg of this hydrophobic subdomain was calculated to quantitatively analyze the pressure of the surrounding environment on it.

The time evolution of the Rg of the hydrophobic subdomain was plotted in Figure 3. For all phenotypic gp120, the Rg of the hydrophobic subdomain shows a downward trend, which is due to the Brown attack of the surrounding molecules, especially solvent water. However, the hydrophobic subdomains of different phenotypic gp120 shrink to different degrees. For neutralization-sensitive, moderate, and resistant gp120, the minimum value of the Rg is about 2.58, 2.6, and 2.62 Å, respectively. It demonstrates that gp120 with a more neutralization-sensitive phenotype has a looser hydrophobic subdomain and is therefore easy to shrink by the surrounding environment. More neutralization-resistant gp120 has a more stable hydrophobic subdomain to maintain its stable conformation.

2.4. Conformational Population Distribution of gp120

Characterizing by the RMSD relative to the template structure and the Rg of the hydrophobic subdomain, the conformational population distribution of gp120 with different neutralization phenotypes can be evaluated from the perspective of energetics. Three FELs (Figure 4) were constructed by projecting MD trajectories into subspace spanned by the RMSD and the Rg. All FELs present a funnel-like free energy surface with irregular and divergent edges. The gp120 which has a more neutralization-sensitive phenotype occupies a larger FEL range and has more complex edges, indicating that it exhibits more conformational flexibility. In contrast, neutralization-moderate or resistant gp120 has a relatively concentrated FEL population distribution.

For the neutralization-sensitive gp120, there is a dominant free energy basin and two other metastable conformational distribution regions (in purple, free energy below about -12 kJ/mol). Although the neutralization-moderate gp120 has one main and two attached metastable regions, their areas are smaller than that of the neutralization-sensitive gp120, especially for the metastable regions. The neutralization-resistant gp120 has only one dominant free energy basin. The shrinking of the global minimum free energy region and the reduction of metastable regions indicate that the conformational population distribution of gp120 is positively correlated with the HIV neutralization phenotype.

2.5. Conformational Flexibility of gp120

The RMSF value (Figure 5) of the per-residue C_α atom at structurally equivalent residue positions was calculated to characterize the conformational flexibility of each gp120. All RMSF profiles among the three phenotypic gp120 have similar features. Regular secondary structure regions have lower RMSF values, while the N-/C-termini and surface-exposed loops show higher RMSF values. Especially in the V1/V2 region, there is a typical conformational flexibility distribution despite undergoing significant residue mutations, deletions, or insertions. Separated by the βA-D, relatively large RMSF values can be observed in the V1 loop (between βA and βB), the L1 loop (between βB and βC), and the V2 loop (between βC and βD).

However, the conformational flexibility of subregions seems to have different roles among the three phenotypic gp120. As the neutralization phenotype becomes more resistant, both α1 and α2 show a trend toward lower RMSF values. This indicates that increasing conformational rigidity of the peripheral regular secondary structure may be one of the strategies to enhance the neutralization resistance. The V3 loop exhibits higher conformational flexibility in some neutralization-sensitive gp120, but it is significantly inhibited in neutralization-moderate and resistant gp120, indicating that the V3 loop perturbs the hydrophobic core to make gp120 more likely to be exposed to immune recognition. As a regulatory switch for conformational transitions of gp120[20], the β20-β21 hairpin has a slightly higher RMSF value in the sensitive gp120, implying that forcing gp120 to transfer to various conformations may provide an opportunity to eliminate the virus.

2.6. Building datasets of three phenotypic gp120

According to the characteristics of the protein data, different data preprocessing methods were used to process the sequence, structure, and MD simulation data of the three phenotypic gp120 resulting in five datasets (Table 1). Using natural language processing methods as a reference, the amino acid sequence of three phenotypic gp120 is regarded as a one-hot coding sequence (Table 1 Datasets Seq) for classification training [23]. For the structural datasets (Table 1 Datasets Stru and Snap), a graph-based model is introduced by transferring the structure of gp120 into the graph data. In the Stru dataset, only the gp120 models constructed by homology modeling were used. For MD simulations, 15, 30, and 60 dynamic structural snapshots were extracted from the 30-ns MD simulation trajectory of each gp120 at different time intervals (2, 1, and 0.5 ns), respectively, to construct the dynamic graph (Table 1 Datasets Snap_2, _1, and _0.5).

Table 1：Datasets of three phenotypic gp120

Datasets	Amounts	Total	Data Model
Seq	3*15*1	45	Sequence
Stru	3*15*1	45	Static graph
Snap_2	3*15*15	675	Dynamic graph
Snap_1	3*15*30	1350	Dynamic graph
Snap_0.5	3*15*60	2700	Dynamic graph

2.7. Selecting machine learning models for the classification of three phenotypic gp120

To choose a suitable model to predict the classification of three HIV neutralization phenotypes, we tested several ML models on different datasets (Figure 6). For the Datasets Seq, the Gated Recurrent Unit (GRU) [24] is trained to learn the contextual representation of the gp120 sequences by maximizing the conditional probability. After transforming the gp120 structures into graphs, two graph ML models, i.e. Graph Convolution Network (GCN) [25] and Graph Isomorphism Network (GIN) [26], were employed to handle the structure-based graph data model. Taking into account the three performance indicators of Accuracy, AUC, and F1 (Table 2), the performance of the GCN model is significantly better than the GRU but slightly worse than the GIN. It reveals that graph ML models (GCN and GIN) can effectively extract complex spatial features from protein structures to obtain better prediction performance in the classification of three HIV neutralization phenotypes. It should be noted that the lower model performance indicators may be because the number of training data (45 sequences or structures) is too small to learn enough classification features for the ML model. However, the comparison of ML models based on sequence or structure provides direction guidance for us to choose the graph ML models as the following training ML model.

Table 2: Performance of the sequence and graph machine learning models training on sequence and structure databases for the classification of three phenotypic gp120.

Models	Datasets	Accuracy	AUC	F1
GRU	Seq	0.480±0.194	0.587±0.124	0.304±0.175
GCN	Stru	0.575±0.117	0.595±0.105	0.304±0.170
GIN	Stru	0.635±0.130	0.577±0.185	0.352±0.271

2.8. Comparison of different gp120 dynamics information in GIN

After testing and comparing different ML models for the classification of three phenotypic gp120, we chose GIN as the basic model. Unlike previous learning on static structures of gp120 (dataset Stru), dynamic structural snapshots (datasets Snap_2, Snap_1, and Snap_0.5) extracted from MD simulations at different time intervals (2.0, 1.0, and 0.5 ns) are used in the GIN model to explore the contribution of dynamic information to the classification of three phenotypic gp120. It can be seen from Table 3 that the performance of the GIN model has been greatly improved after introducing dynamics information through

MD simulations. The increase in training data directly leads to a significant improvement in the performance of the ML model. This shows that using MD simulations to expand the difficult-to-obtain protein structure data is a strategy that can be applied to ML training. As the sampling interval decreases, the accuracy, AUC, and F1 of the GIN model are getting higher and higher, indicating that it is beneficial to expand training data with a smaller sampling interval through MD simulations to exploit more dynamic conformational information of gp120. However, the ML model performance does not continue to increase as the time interval decreases. It is conceivable that a smaller time interval will not further improve the performance of the ML model.

Table 3: Performance of the graph machine learning model (GIN) training on static structures of gp120 and dynamic structural snapshots at different time intervals for the classification of three phenotypic gp120.

Datasets	Model	Time intervals	Accuracy	AUC	F1
Stru	GIN	-	0.635±0.130	0.577±0.185	0.352±0.271
Snap_2	GIN	2.0 ns	0.732±0.041	0.744±0.085	0.503±0.128
Snap_1	GIN	1.0 ns	0.785±0.032	0.798±0.056	0.587±0.123
Snap_0.5	GIN	0.5 ns	0.822±0.018	0.840±0.097	0.715±0.114

2.9. Comparison of different protein features in GIN

In graph ML, nodes use one-hot encoding and node embedding to introduce their features into the graph model. In the case of the protein graph model, one-hot encoding uses only an N-bit status register to encode 20 amino acids, while node embedding employs a high-dimensional vector to store physical, chemical, and other properties for each amino acid. In this paper, one-hot encoding maps 20 natural amino acid names into integer values and then converts the integer values into binary vectors. For the node embedding, 14 physicochemical properties (e.g. charge, hydropathy, polarity, volume, etc. in Table S1) of different amino acids and the 3D coordinates of the C_α atom for each residue were transformed into the feature vectors of nodes to enrich the characteristics of nodes in the graph model.

As shown in Table 4, two protein features were used to train the graph ML model on dataset Snap_0.5 by using the GIN model. Compared with the method of one-hot encoding, the method of node embedded achieved better performance. The prediction performance of the GIN model on dataset Snap_0.5 is improved by more than 10%. The results show that appropriate protein features can significantly improve the performance of the ML model.

Table 4: Performance of the graph machine learning model (GIN) training on dynamic structural snapshots at time interval 0.5 ns for the classification of three phenotypic gp120 with different protein features.

Datasets	Model	Features	Accuracy	AUC	F1
Snap_0.5	GIN	One-hot encoding	0.822±0.018	0.840±0.097	0.715±0.114
Snap_0.5	GIN	Node embedded	0.971±0.003	0.986±0.010	0.934±0.036

2.10. Classification-related attention score from GIN with the attention mechanism

To analyze the classification-related information of the HIV neutralization phenotype, the graph ML model with an attention layer (GIN+ATT, Figure 7) was constructed by introducing the attention mechanism

[27] into the basic GIN model. After aggregating node features into graph features in the GIN model, the attention layer was employed to distribute attention scores into different nodes along the gradient descent direction of the classification for the HIV neutralization phenotype. During the training process, the attention score of each node was automatically adjusted, leading to more attention to the important amino acid residues for the classification. Although the classification performance of the GIN+ATT model is weakened after adding the attention mechanism (Table 5), the attention score can indicate the residue or node that plays an important role in the classification. A small difference in classification performance may be a potential signal of the stronger generalization ability of the model. In addition, the attention score of each residue can be used as an indicator to determine the crucial mutations in various sequences of gp120 and to measure the biological contribution of these mutations to the HIV neutralization phenotype.

Table 5: Predictive performance of our graphic-ML model.

Dataset	Model	Accuracy	AUC	F1
Snap_0.5	GIN	0.971±0.003	0.986±0.010	0.934±0.036
Snap_0.5	GIN+ATT	0.941±0.012	0.983±0.007	0.930±0.026

Similar to other attention-based models, attention scores can capture local contributions related to classification in the graph ML model. The attention scores (Figure 8) of the per-residue C_{α} atom at structurally equivalent residue positions were calculated to evaluate the contribution of every residue in gp120 to the classification of the HIV neutralization phenotype. For all attention scores among the three phenotypic gp120, the N-termini, V1/V2 region, and V3 loop have larger scores, suggesting that these local regions play a vital role in identifying the HIV neutralization phenotype. The large attention score of the V1/V2 region and the V3 loop, together with the higher RMSF of these regions, indicate that high sequence variability and conformational flexibility are the molecular basis of the HIV neutralization phenotype. It should be noted that different attention score distributions and amplitude changes can further reveal more subtle residue contributions. For example in the V1/V2 region, larger attention scores locate at the V1 loop, βD, and β3 for the neutralization-sensitive gp120 (Figure 8, red lines), βB and the L1 loop for the neutralization-moderate gp120 (Figure 8, green lines), and βC for the neutralization-resistant gp120 (Figure 8, blue lines). Compared to the neutralization-sensitive gp120, C-termini seems to contribute more to the classification of the HIV neutralization phenotype in the neutralization-moderate and resistant gp120 and thus shows a larger attention score. Similarly, the β20-β21 hairpin also has a more important classification contribution in the neutralization-resistant gp120. These results all indicate that the attention scores can provide a structured explanation of gp120 for the classification of the HIV neutralization phenotype.

3. DISCUSSION

Establishing the relationship between sequence, structure, and function is the core issue in protein research. The sequence variation and structural dynamics of the envelope glycoprotein gp120 determine its function in viral infection, resulting in various HIV neutralization phenotypes [28]. Although MD simulations of gp120 with the extremely sensitive and resistant neutralization phenotypes preliminarily revealed the molecular mechanism of HIV neutralization phenotype [29], broad-spectrum theoretical studies of gp120 with different neutralization phenotypes are still missing. In this paper, 45 pre-fusion gp120 structural models from different HIV strains belong to three tiers of sensitive, moderate, and resistant neutralization phenotype were built and subjected to MD simulations. The comparative analysis of the structural deviations, conformational flexibility, and population distribution together shows that the molecular dynamics properties of gp120 are positively correlated with the HIV neutralization phenotype.

Together with the kinetic of gp120 in the CD4-complex and CD4-free state [29], CD4-binding effects on the conformational dynamics of gp120 [30][31], and conformational transitions of gp120 [32], these results in this paper provide a systematic description of the molecular dynamics of gp120. Although gp120 is a trimer under physiological conditions, it is still reasonable to use the gp120 subunit structure for research because small-angle X-ray scattering data revealed that the single gp120 domain in solution closely resembles that of the gp120 subunit in the context of the oligomeric viral spike/trimer [33]. In addition, the glycosylation effect was ignored in the MD simulations. Considering a single variable parameter (sequence variation) in the same experimental environment (prefusion gp120) can better focus on the purpose of research. Moreover, previous MD simulation studies on gp120 with glycosylated and non-glycosylated variable loops showed no significant differences in molecular fluctuations [34].

Although MD simulations can help explore detailed changes in the molecular structure of viral proteins, it is still impossible and unnecessary to perform long-term MD simulations for large-scale HIV strains. Moreover, modern post-simulation trajectory analysis methods (e.g. RMSD, RMSF, and FEL) only extract statistical information during the MD simulations to summarize the rules of molecular motion. In issues such as the sequence variation and structural dynamics of the envelope glycoprotein gp120 in the HIV neutralization phenotype, MD simulations are difficult to provide global statistical information to help distinguish the potential relationship between gp120 and the HIV neutralization phenotype. Therefore, using ML will be a good attempt to explore complex relationships between the sequence, structure, and dynamics of the protein. However, for the application of ML to protein research, there is a major obstacle that ML models require a lot of data (especially protein structure data that is difficult to obtain) for training. Many attempts solely on protein sequence data for ML training have been reported, but using protein structure data to construct ML models will capture more direct protein function information because the function of a protein is closely related to its structure. In this paper, MD simulations help expand the difficult-to-obtain experimental data into protein datasets for ML training.

After testing different types of data sets (the sequence, structure, and dynamic structure) and ML models (GRU, GCN, and GIN), the dynamic graph representation of protein from the MD simulations were used to train the graph ML model (GIN and GIN+ATT) in this paper. Compared to protein sequences, the use of the structure graph of the protein can improve the performance of ML models. A large number of biological data including the protein is often non-Euclidean form, which has complex topological structures and contains multiple features. GCN and GIN have an excellent performance in processing these data by following the neighborhood aggregation scheme and calculating the representation vector of nodes by recursively aggregating and transforming the representation vector of adjacent nodes. Thanks to the increase of training data and the supplement of dynamic information, the introduction of dynamic graphs significantly improves the performance of ML. In this paper, the advantages of MD simulations and graph ML have been cleverly utilized to decipher the role of envelope glycoprotein gp120 in HIV neutralization phenotype from two aspects of the sequence variation and structural dynamics. It is conceivable that this method that utilizes ML models, especially deep learning, to help analyze calculated data, such as MD simulations holds great research potential.

4. MATERIALS AND METHODS

4.1. Sequence selection and alignment

All HIV sequences were obtained from the UniProtKB database (http://www.uniprot.org) with the accession IDs listed in Table 6. By removing the segments corresponding to the signal peptide and gp41, the primary sequence of gp120 was extracted and used as the target to construct structural models of gp120 with different neutralization phenotypes.

4.2. Structural model construction and validation

The atomic coordinate of gp120 in the full-length crystal structure of HIV prefusion envelope glycoprotein (PDB ID: 5FYK) at 3.11 Å resolution [5] in the RCSB PDB database (https://www.pdbus.org) was used as the template to construct the structural models. The high sequence identity (higher than 66%, Table S1) between each selected gp120 sequence and the template ensures the reliability of homology modeling, which is implemented in the MODELLER [35] (version 9.17) package. For each structural model of gp120, 20 candidates were generated and the one with the lowest molecular probability density function score was selected. All models were validated using the SAVES server (https://services.mbi.ucla.edu/SAVES), in which algorithms of ERRAT [36], PROVE [37], and VERIFY3D [38] suggest that the constructed model and the template have a similar structural quality score (Table S1) which demonstrated that the gp120 structural models have good stereochemical quality. To obtain structurally equivalent positions, all gp120 models were aligned by using the multiple sequence alignment tool MUSCLE [39] within the MEGA [40]. Regular secondary structural elements and variable regions were labeled according to the standard HIV envelope glycoprotein crystal structures (PDB IDs: 1G9M [41] and 3JWD [18]).

4.3. Molecular dynamics simulations

Each structural model of gp120 was individually solvated using TIP3P water molecules [42] in a dodecahedron box with a solute-wall minimum distance of 0.8 nm and a periodic boundary condition. Sodium and chloride ions were added to obtain a 150-mM salt concentration. All simulations were performed by employing GROMACS (version 5.1.4) software [43] with the AMBER99SB-ILDN49 force field [44] at the GPU-accelerated clusters. After energy minimization with steepest descent algorithm and position-restrained simulations with decreasing harmonic positional restraint force constants on the heavy atoms of the protein, 30-ns production MD run for each gp120 was performed with the following protocol: integration time was set as 2 fs due to LINCS [45] algorithm was used to constrain bond lengths involving hydrogen atoms; the partial-mesh Ewald (PME) [46] method and a twin-range cut-off strategy were used to calculate the long-range electrostatic interactions and van der Waals interactions respectively; protein and non-protein components were independently coupled to a 300 K and 1 atm with an external bath; the system coordinates were saved every 2 ps.

The package MDTraj [47] was employed to calculate RMSD, Rg, and RMSF. FELs were constructed by $F(s) = -k_B T \ln(N_i/N_{max})$, where k_B is Boltzmann's constant, T is the simulation temperature, N_i is the population of bin i and N_{max} is the population of the most populated bin.

4.4. Datasets construction

The gp120 sequence data adopted one-hot encoding and filled with the longest sequence length of 500. For the gp120 structure, a graph-based model was introduced by selecting the C_α atom of each residue as the node and connecting the edge if the distance of the pairwise C_α atom is less than 8 Å. For each residue, 14 amino acid properties (Table S2) and the 3D coordinates of the C_α atom for each residue were used as the feature vector of each node.

4.5. Machine learning

The GRU model is implemented by the Pytorch framework. The Deep Graph Library (DGL, https://www.dgl.ai) was employed to train and validate graph ML models, i.e. GCN and GIN. When testing the above model, the scikit learn package (https://scikit-learn.org) is used to split data into train/test sets and k-fold (k = 10) cross-validation training strategy. All models were trained for 100 epochs, and the learning rate was set to 0.001.

In the GIN+ATT model, the hidden representation H_v (H_v=[h_1, h_2, …, h_n], where n is the number of nodes and w is the learnable weight) of all nodes obtained from the last layers of the GIN model were used to construct the attention mechanism by using the following calculation:

$$M = \mathrm{Relu}\,(H_v), \tag{1.1}$$

$$\alpha = softmax(w^T M), \tag{1.2}$$

$$H'_v = H_v \alpha^T. \tag{1.3}$$

The α of formula 1.2 is the attention vector. Through the calculation of formula 1.3, the graph node representation H'_v adjusted by attention weight is obtained. As the first part of the model, the GIN layer is used to extract the potential spatial features of each node. The feature vectors of nodes from the last layers of the GIN model were passed through the attention layer to aggregate as graph representation vectors. The attention layer is implemented by the trainable parameter module of Pytorch.

To measure the performance of the model, three different evaluation indicators, i.e. the accuracy, the area under the ROC curve (AUC), and the F1 score, were used. The calculation method can be obtained from the confusion matrix (Table 7) and formula 2.1.

Table 7: The confusion matrix

Factual / Forecast	0	1
0	TN	FN
1	FP	TP

$$\text{Accuracy } = \frac{TP + TN}{TP + TN + FN + FP} \tag{2.1}$$

ROC curves were plotted with the true positive rate (TPR) and the false positive rate (FPR) as abscissa and ordinate respectively. The definitions are as formula 2.2-2.3.

$$\text{TPR} = \frac{TP}{TP + FN} \tag{2.2}$$

$$\text{FPR} = \frac{FP}{TN + FP} \tag{2.3}$$

The area under the ROC curve (AUC) can help to comprehensively measure the performance of the model. F1 score is the harmonic average of recall (TPR) and precision (P), and its calculation method is shown in formula 2.4-2.5.

$$\text{P} = \frac{TP}{TP + FP} \tag{2.4}$$

$$\frac{2}{F} = \frac{1}{P} + \frac{1}{TPR} \tag{2.5}$$

5. CONCLUSION

In this study, 45 prefusion gp120 structural models from different HIV strains belong to three tiers of sensitive, moderate, and resistant neutralization phenotypes were built and investigated by MD simulations and graph ML to probe the roles of rapid sequence variability and significant structural dynamics of gp120 in the HIV neutralizing phenotype. By integrating homology modeling, MD simulations, and graph ML, we constructed a reliable model for HIV neutralization phenotype. Statistical analysis of gp120 dynamics (e.g. RMSF) and classification-related attention scores together revealed potential structural factors that may affect HIV neutralization phenotype. Our study not only deciphers gp120 sequence variation and structural dynamics in the HIV neutralization phenotype but also provides a methodological framework to explore complex relationships between the sequence, structure, and dynamics of the protein by combining MD simulations and ML.

CONFLICT OF INTERESTS

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The source data (sequence files, structural models, C_α trajectories, RMSD, and RMSF data) and scripts are available at https://github.com/liyigerry/gp120_role.

Funding information

National Natural Sciences Foundation of China, Grant/Award Number: 62003068 and 31960198
Yunnan Fundamental Research Projects, Grant/Award Number: 2019FD014 and 2019FB021
Scientific Research Fund Project of Yunnan Education Department, Grant/Award Number: 2020Y0487

REFERENCES

[1] Chen B. Molecular Mechanism of HIV-1 Entry. Trends Microbiol 2019;27:878–91. https://doi.org/10.1016/j.tim.2019.06.002.

[2] Pancera M, Zhou T, Druz A, Georgiev IS, Soto C, Gorman J, et al. Structure and immune recognition of trimeric pre-fusion HIV-1 Env. Nature 2014;514:455–61. https://doi.org/10.1038/nature13808.

[3] Wyatt R, Sodroski J. The HIV-1 envelope glycoproteins: Fusogens, antigens, and immunogens. Science (80-) 1998;280:1884–8. https://doi.org/10.1126/science.280.5371.1884.

[4] Munro JB, Mothes W. Structure and Dynamics of the Native HIV-1 Env Trimer. J Virol 2015;89:5752–5. https://doi.org/10.1128/jvi.03187-14.

[5] Stewart-Jones GBE, Soto C, Lemmin T, Chuang GY, Druz A, Kong R, et al. Trimeric HIV-1-Env Structures Define Glycan Shields from Clades A, B, and G. Cell 2016;165:813–26. https://doi.org/10.1016/j.cell.2016.04.010.

[6] Montefiori DC, Roederer M, Morris L, Seaman MS. Neutralization tiers of HIV-1. Curr Opin HIV AIDS 2018;13:128–36. https://doi.org/10.1097/COH.0000000000000442.

[7] Rouse BT, Sehrawat S. Immunity and immunopathology to viruses: what decides the outcome? Nat Rev Immunol 2010;10:514–26. https://doi.org/10.1038/nri2802.

[8] Gnanakaran S, Daniels MG, Bhattacharya T, Lapedes AS, Sethi A, Li M, et al. Genetic signatures in the envelope glycoproteins of HIV-1 that associate with broadly neutralizing antibodies. PLoS Comput Biol 2010;6. https://doi.org/10.1371/journal.pcbi.1000955.

[9] Seaman MS, Janes H, Hawkins N, Grandpre LE, Devoy C, Giri A, et al. Tiered Categorization of a Diverse Panel of HIV-1 Env Pseudoviruses for Assessment of Neutralizing Antibodies. J Virol 2010;84:1439–52. https://doi.org/10.1128/JVI.02108-09.

[10] Li Y, Deng L, Ai S-M, Sang P, Yang J, Xia Y-L, et al. Insights into the molecular mechanism underlying CD4-dependency and neutralization sensitivity of HIV-1: a comparative molecular dynamics study on gp120s from isolates with different phenotypes. RSC Adv 2018;8:14355–68. https://doi.org/10.1039/C8RA00425K.

[11] Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature 2020;577:706–10. https://doi.org/10.1038/s41586-019-1923-7.

[12] Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform 2017;18:851–69. https://doi.org/10.1093/bib/bbw068.

[13] Wang DD, Ou-Yang L, Xie H, Zhu M, Yan H. Predicting the impacts of mutations on protein-ligand binding affinity based on molecular dynamics simulations and machine learning methods. Comput Struct Biotechnol J 2020;18:439–54. https://doi.org/10.1016/j.csbj.2020.02.007.

[14] Bignon E, Gillet N, Chan C-H, Jiang T, Monari A, Dumont E. Recognition of a tandem lesion by DNA bacterial formamidopyrimidine glycosylases explored combining molecular dynamics and machine learning. Comput Struct Biotechnol J 2021;19:2861–9. https://doi.org/10.1016/j.csbj.2021.04.055.

[15] Degiacomi MT. Coupling Molecular Dynamics and Deep Learning to Mine Protein Conformational Space. Structure 2019;27:1034-1040.e3. https://doi.org/10.1016/j.str.2019.03.018.

[16] Liu J, Bartesaghi A, Borgnia MJ, Sapiro G, Subramaniam S. Molecular architecture of native HIV-1 gp120 trimers. Nature 2008. https://doi.org/10.1038/nature07159.

[17] Da LT, Lin M. Opening dynamics of HIV-1 gp120 upon receptor binding is dictated by a key hydrophobic core. Phys Chem Chem Phys 2019;21:26003–16. https://doi.org/10.1039/c9cp04613e.

[18] Pancera M, Majeed S, Ban YEA, Chen L, Huang CC, Kong L, et al. Structure of HIV-1 gp120 with gp41-interactive region reveals layered envelope architecture and basis of conformational mobility. Proc Natl Acad Sci U S A 2010;107:1166–71. https://doi.org/10.1073/pnas.0911004107.

[19] McLellan JS, Pancera M, Carrico C, Gorman J, Julien JP, Khayat R, et al. Structure of HIV-1 gp120 V1/V2

domain with broadly neutralizing antibody PG9. Nature 2011;480:336–43. https://doi.org/10.1038/nature10696.

[20] Herschhorn A, Gu C, Moraca F, Ma X, Farrell M, Smith AB, et al. The β20–β21 of gp120 is a regulatory switch for HIV-1 Env conformational transitions. Nat Commun 2017;8:1049. https://doi.org/10.1038/s41467-017-01119-w.

[21] Guttman M, Cupo A, Julien J-P, Sanders RW, Wilson IA, Moore JP, et al. Antibody potency relates to the ability to recognize the closed, pre-fusion form of HIV Env. Nat Commun 2015;6:6144. https://doi.org/10.1038/ncomms7144.

[22] Wei X, Decker JM, Wang S, Hui H, Kappes JC, Wu X, et al. Antibody neutralization and escape by HIV-1. Nature 2003;422:307–12. https://doi.org/10.1038/nature01470.

[23] Ofer D, Brandes N, Linial M. The language of proteins: NLP, machine learning &amp; protein sequences. Comput Struct Biotechnol J 2021;19:1750–8. https://doi.org/10.1016/j.csbj.2021.03.022.

[24] Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. Proc. 2014 Conf. Empir. Methods Nat. Lang. Process., Stroudsburg, PA, USA: Association for Computational Linguistics; 2014, p. 1724–34. https://doi.org/10.3115/v1/D14-1179.

[25] Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks 2016.

[26] Xu K, Hu W, Leskovec J, Jegelka S. How Powerful are Graph Neural Networks? 2018.

[27] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv. Neural Inf. Process. Syst., vol. 2017- Decem, 2017, p. 5999–6009.

[28] Shaik MM, Peng H, Lu J, Rits-Volloch S, Xu C, Liao M, et al. Structural basis of coreceptor recognition by HIV-1 envelope spike. Nature 2019;565:318–23. https://doi.org/10.1038/s41586-018-0804-9.

[29] Li Y, Deng L, Liang J, Dong G-H, Xia Y-L, Fu Y-X, et al. Molecular dynamics simulations reveal distinct differences in conformational dynamics and thermodynamics between the unliganded and CD4-bound states of HIV-1 gp120. Phys Chem Chem Phys 2020;22:5548–60. https://doi.org/10.1039/C9CP06706J.

[30] Li Y, Guo Y-C, Zhang X-L, Deng L, Sang P, Yang L-Q, et al. CD4-binding obstacles in conformational transitions and allosteric communications of HIV gp120. Biochim Biophys Acta - Biomembr 2020;1862:183217. https://doi.org/10.1016/j.bbamem.2020.183217.

[31] Li Y, Deng L, Yang L-Q, Sang P, Liu S-Q. Effects of CD4 Binding on Conformational Dynamics, Molecular Motions, and Thermodynamics of HIV-1 gp120. Int J Mol Sci 2019;20:260. https://doi.org/10.3390/ijms20020260.

[32] Li Y, Zhang X-L, Yuan X, Hou J-C, Sang P, Yang L-Q. Probing intrinsic dynamics and conformational transition of HIV gp120 by molecular dynamics simulation. RSC Adv 2020;10:30499–507. https://doi.org/10.1039/D0RA06416E.

[33] Guttman M, Garcia NK, Cupo A, Matsui T, Julien JP, Sanders RW, et al. CD4-induced activation in a soluble HIV-1 Env trimer. Structure 2014;22:974–84. https://doi.org/10.1016/j.str.2014.05.001.

[34] Yokoyama M, Naganawa S, Yoshimura K, Matsushita S, Sato H. Structural dynamics of HIV-1 envelope GP120 outer domain with V3 loop. PLoS One 2012;7. https://doi.org/10.1371/journal.pone.0037530.

[35] Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinforma 2016;54:5.6.1-5.6.37. https://doi.org/10.1002/cpbi.3.

[36] Colovos C, Yeates TO. Verification of protein structures: Patterns of nonbonded atomic interactions. Protein Sci 1993;2:1511–9. https://doi.org/10.1002/pro.5560020916.

[37] Pontius J, Richelle J, Wodak SJ. Deviations from standard atomic volumes as a quality measure for protein crystal structures. J Mol Biol 1996;264:121–36. https://doi.org/10.1006/jmbi.1996.0628.

[38] Bowie JU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional stucture. Science (80-) 1991;253:164–70. https://doi.org/10.1126/science.1853201.

[39] Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004. https://doi.org/10.1093/nar/gkh340.

[40] Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 2018. https://doi.org/10.1093/molbev/msy096.

[41] Kwong PD, Wyatt R, Majeed S, Robinson J, Sweet RW, Sodroski J, et al. Structures of HIV-1 gp120 envelope glycoproteins from laboratory-adapted and primary isolates. Structure 2000. https://doi.org/10.1016/S0969-2126(00)00547-5.

[42] Price DJ, Brooks CL. A modified TIP3P water potential for simulation with Ewald summation. J Chem Phys 2004;121:10096–103. https://doi.org/10.1063/1.1808117.

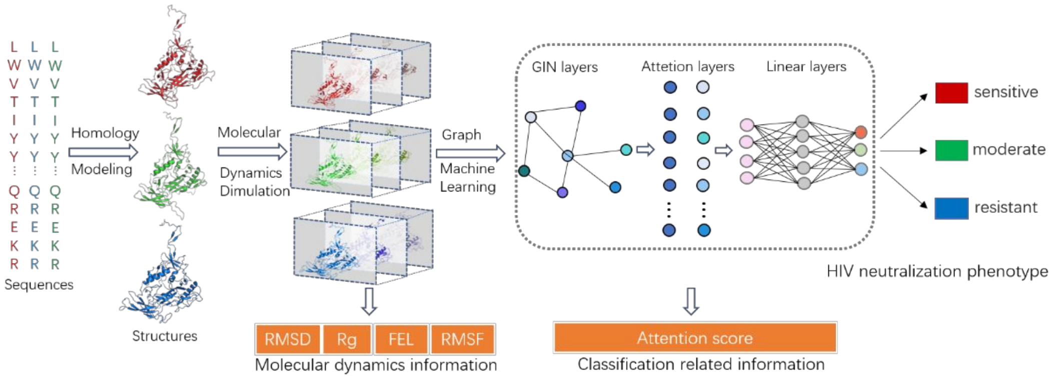
[43] Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 2015;1–2:19–25. https://doi.org/10.1016/j.softx.2015.06.001.

[44] Aliev AE, Kulke M, Khaneja HS, Chudasama V, Sheppard TD, Lanigan RM. Motional timescale predictions by molecular dynamics simulations: Case study using proline and hydroxyproline sidechain dynamics. Proteins Struct Funct Bioinforma 2014;82:195–215. https://doi.org/10.1002/prot.24350.

[45] Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: A Linear Constraint Solver for molecular simulations. J Comput Chem 1997;18:1463–72. https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

[46] Darden T, York D, Pedersen L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. J Chem Phys 1993;98:10089–92. https://doi.org/10.1063/1.464397.

[47] McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX, et al. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. Biophys J 2015;109:1528–32. https://doi.org/10.1016/j.bpj.2015.08.015.



Graphic abstract

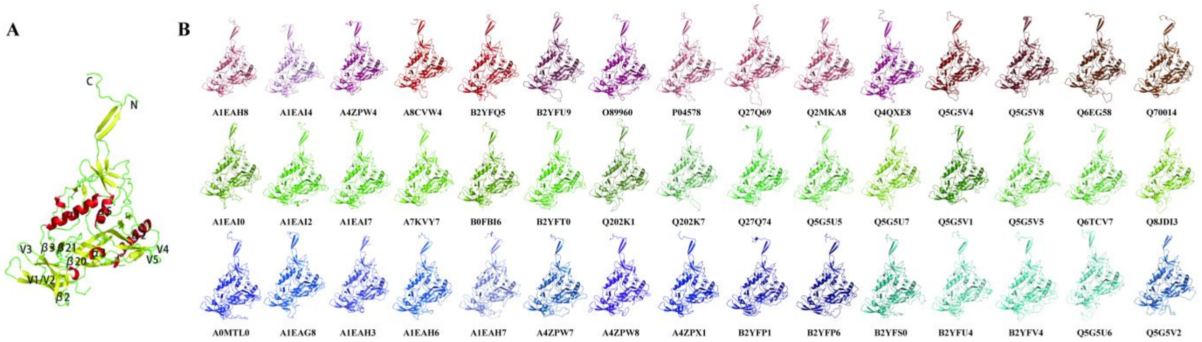


Fig. 1. The structural architecture of gp120. (A) Structure of prefusion gp120 (PDB ID: 5FYK). (B) Structural models of the neutralization-sensitive (reds), moderate (greens), and resistant (blues) gp120 from different HIV isolates with the accession ID of UniProtKB database.

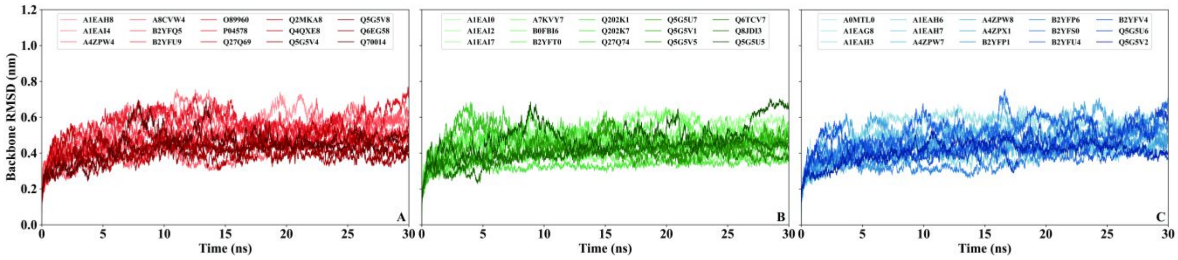


Fig. 2. Time evolution of backbone root-mean-square deviation (RMSD) values of the neutralization-sensitive (A), moderate (B), and resistant (C) gp120 relative to the template structure (PDB ID: 5FYK) during the molecular dynamics simulations.

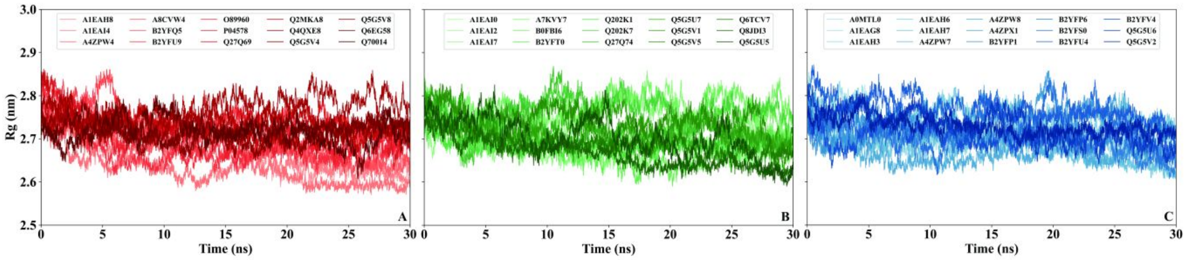


Fig. 3. Time evolution of radius gyration (Rg) of the hydrophobic core in the neutralization-sensitive (A), moderate (B), and resistant (C) gp120 during the molecular dynamics simulations.

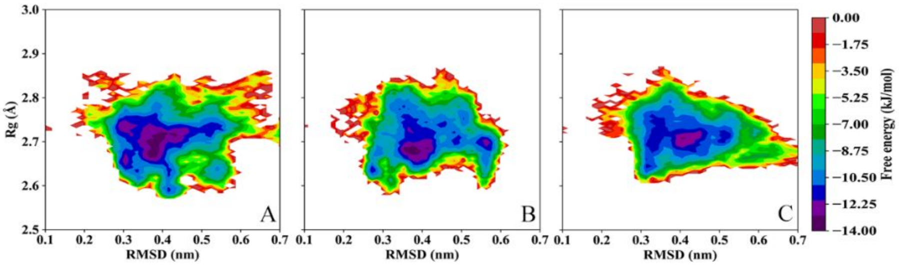


Fig. 4. Free energy landscapes of the neutralization-sensitive (A), moderate (B), and resistant (C) gp120 by projecting trajectories into the subspace of root-mean-square deviation (RMSD) relative to the template structure and the radius gyration (Rg) of the hydrophobic subdomain.

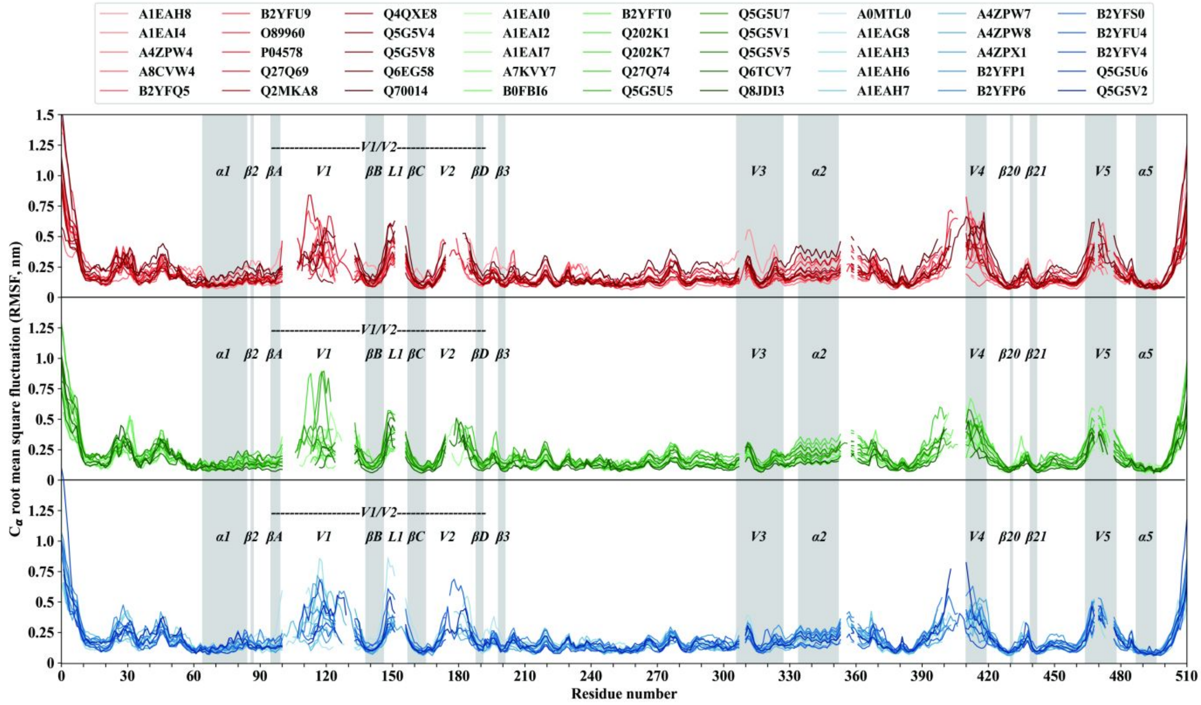


Fig. 5. C_α root-mean-square fluctuation (RMSF) values of the neutralization-sensitive (red lines), moderate (green lines), and resistant (blue lines) gp120 during the molecular dynamics simulations.

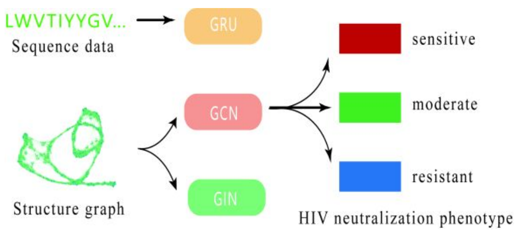


Fig. 6. The Gated Recurrent Unit (GRU) was used to handle the sequence data. The Graph Convolution Network (GCN) and Graph Isomorphism Network (GIN) were trained by the structure graph. All ML models were used for the classification of HIV neutralization phenotypes.

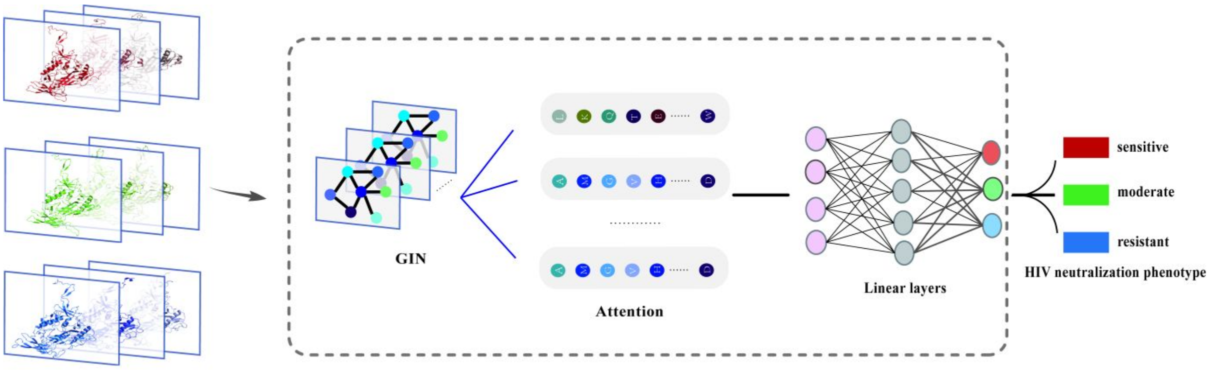


Fig. 7. Diagram of the Graph Isomorphism Network (GIN) with the attention mechanism (GIN+ATT).

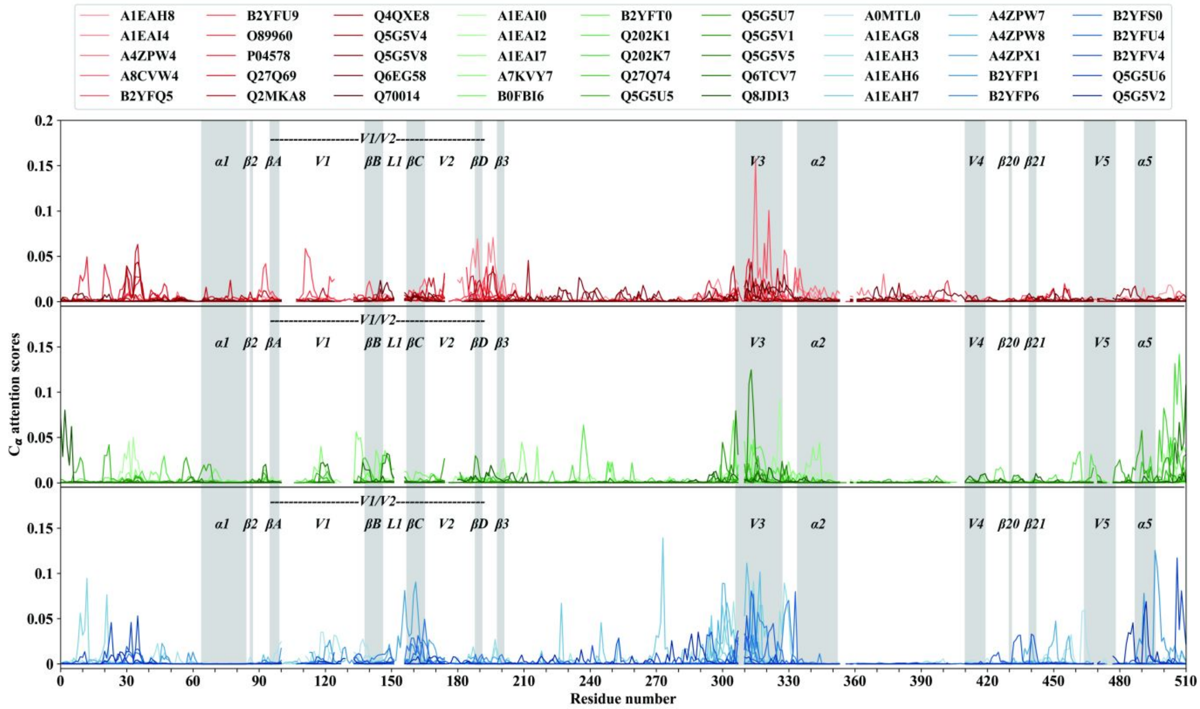


Fig. 8. C_α attention scores of the neutralization-sensitive (reds), moderate (greens), and resistant (blues) gp120.

Supporting Information

Table S1： 14 physicochemical properties of different amino acids.

Amino acid	alpha	beta	charge	core	hydropathy	pH	polarity	rim	surface	turn	volume	strength	disorder	high_contact
A	1.29	0.9	0	0.049	1.8	0	0	0.047	0.065	0.78	67	0	0	1
C	1.11	0.74	0	0.02	2.5	-2	0	0.015	0.015	0.8	86	1	-1	0
D	1.04	0.72	-1	0.051	-3.5	-2	1	0.071	0.074	1.41	91	0	1	1
E	1.44	0.75	-1	0.051	-3.5	-2	1	0.094	0.089	1	109	0	1	0
F	1.07	1.32	0	0.051	2.8	0	0	0.021	0.029	0.58	135	1	-1	0
G	0.56	0.92	0	0.06	-0.4	0	0	0.071	0.07	1.64	48	0	1	1
H	1.22	1.08	0	0.034	-3.2	1	1	0.022	0.025	0.69	118	0	-1	0
I	0.97	1.45	0	0.047	4.5	0	0	0.032	0.035	0.51	124	1	-1	0
K	1.23	0.77	1	0.05	-3.9	2	1	0.105	0.08	0.96	135	0	1	0
L	1.3	1.02	0	0.078	3.8	0	0	0.052	0.063	0.59	124	1	-1	1
M	1.47	0.97	0	0.027	1.9	0	0	0.017	0.016	0.39	124	1	1	0
N	0.9	0.76	0	0.058	-3.5	0	1	0.062	0.053	1.28	96	0	1	1
P	0.52	0.64	0	0.051	-1.6	0	0	0.052	0.054	1.91	90	0	1	0
Q	1.27	0.8	0	0.051	-3.5	1	1	0.053	0.051	0.97	114	0	1	0
R	0.96	0.99	1	0.066	-4.5	2	1	0.068	0.059	0.88	148	0	1	1
S	0.82	0.95	0	0.057	-0.8	-1	1	0.072	0.071	1.33	73	0	1	1

Amino acid	alpha	beta	charge	core	hydropathy	pH	polarity	rim	surface	turn	volume	strength	disorder	high_contact
T	0.82	1.21	0	0.064	-0.7	-1	0	0.064	0.065	1.03	93	0	0	1
V	0.91	1.49	0	0.049	4.2	0	0	0.048	0.048	0.47	105	1	-1	0
W	0.99	1.14	0	0.022	-0.9	1	1	0.007	0.012	0.75	163	1	-1	0
Y	0.72	1.25	0	0.07	-1.3	-1	1	0.032	0.033	1.05	141	1	-1	1

This table is taken from https://github.com/mikessh/vdjtools/blob/master/src/main/resources/profile/aa_property_table.txt

Table S2: Structural models of gp120.

	Isolate	Residues[1]	Identity[2]	RMSD[3] (nm)	VERIFY3D[4]	ERRAT[5]	PROVE[6]	MD&ML[7]
Template	5FYK_G	480	-	-	85.26%	84.7727	5.00%	No
Tier 1	A1EAH8	468	71.55%	0.684	83.33%	63.2690	8.20%	Yes
	A1EAI4	485	72.58%	0.67	84.71%	55.7235	7.00%	Yes
	A4ZPW4	470	69.10%	0.72	55.74%	58.658	8.90%	Yes
	A8CVW4	470	71.07%	0.697	77.87%	64.6421	6.30%	Yes
	B2YFQ5	477	72.26%	0.684	84.70%	59.1075	9.30%	Yes
	B2YFU9	481	68.32%	0.667	73.80%	57.3561	7.10%	Yes
	O89960	486	70.61%	0.706	81.89%	63.0252	5.80%	Yes
	P04578	478	67.08%	0.706	82.01%	59.743	6.20%	Yes
	Q27Q69	494	67.94%	0.653	84.41%	56.1728	7.20%	Yes
	Q2MKA8	476	70.00%	0.686	85.50%	68.4665	6.50%	Yes
	Q4QXE8	466	69.87%	0.691	83.48%	68.1223	6.40%	Yes
	Q5G5V4	473	69.83%	0.725	83.51%	67.0978	6.20%	Yes
	Q5G5V8	477	68.46%	0.677	87.42%	61.1940	7.80%	Yes
	Q6EG58	488	68.16%	0.742	83.47%	59.1667	6.60%	Yes
	Q70014	469	71.34%	0.76	77.40%	57.7007	6.20%	Yes
Tier 2	A1EAI0	483	69.53%	0.706	84.68%	63.2911	6.90%	Yes
	A1EAI2	489	71.84%	0.722	79.75%	58.6134	7.10%	Yes
	A1EAI7	463	71.76%	0.696	82.94%	58.6813	5.70%	Yes
	A7KVY7	469	68.27%	0.677	84.43%	59.9129	8.40%	Yes
	B0FBI6	475	69.48%	0.777	85.26%	61.9355	6.60%	Yes
	B2YFT0	486	69.96%	0.748	84.98%	58.9958	8.00%	Yes
	Q202K1	472	69.52%	0.679	77.97%	70.6897	7.20%	Yes
	Q202K7	470	71.07%	0.692	78.72%	61.1354	5.70%	Yes
	Q27Q74	471	71.04%	0.739	84.50%	61.1231	9.00%	Yes

	Q5G5U5	472	70.35%	0.67	76.06%	64.5022	7.00%	Yes
	Q5G5U7	478	68.05%	0.685	86.82%	55.8190	7.00%	Yes
	Q5G5V1	484	70.72%	0.723	85.33%	59.8739	6.10%	Yes
	Q5G5V5	476	69.48%	0.703	85.50%	63.4615	6.40%	Yes
	Q6TCV7	470	69.63%	0.674	86.81%	61.2554	6.40%	Yes
	Q8JDI3	468	72.41%	0.703	83.12%	63.0435	6.50%	Yes
Tier 3	A0MTL0	475	71.99%	0.699	77.47%	48.8223	8.90%	Yes
	A1EAG8	487	71.84%	0.656	83.37%	58.7866	6.50%	Yes
	A1EAH3	475	72.41%	0.667	81.47%	58.9247	6.40%	Yes
	A1EAH6	486	71.19%	0.772	82.72%	56.9328	7.40%	Yes
	A1EAH7	473	71.43%	0.695	90.27%	60.9914	7.40%	Yes
	A4ZPW7	490	67.61%	0.701	83.47%	59.5833	6.60%	Yes
	A4ZPW8	475	70.19%	0.716	78.32%	60.8137	5.50%	Yes
	A4ZPX1	478	67.63%	0.703	85.36%	64.8936	6.80%	Yes
	B2YFP1	487	68.65%	0.843	81.00%	55.9499	7.20%	Yes
	B2YFP6	472	68.68%	0.687	79.24%	63.6364	6.50%	Yes
	B2YFS0	478	69.55%	0.713	87.87%	66.3793	7.70%	Yes
	B2YFU4	482	67.89%	0.708	80.91%	57.5949	7.90%	Yes
	B2YFV4	493	68.61%	0.706	81.95%	55.3719	6.40%	Yes
	Q5G5U6	477	66.53%	0.721	80.71%	57.7827	5.80%	Yes
	Q5G5V2	487	68.64%	0.695	81.52%	66.9456	5.80%	Yes

1. The number of residues in gp120.

2. The identity between the target sequence and the sequence of the template (PDB ID: 5FYJ, chain G).

3. The root means square deviation (RMSD) to the template (PDB ID: 5FYJ, chain G).

4. The percentage of the residues have averaged 3D-1D score >= 0.2 calculated from the algorithm VERIFY3D.

5. Overall quality factor of the algorithm ERRAT.

6. The percentage of buried outlier protein atoms in the algorithm PROVE.

7. Whether the structural model is used for molecular dynamics (MD) simulations and machine learning (ML) training.