Solid-State NMR Investigation of Viral Fusion Glycoprotein 41 (gp41)

by

Chloe A. Morgan

B.A. Biochemistry
New York University, 2017
New York City, New York

Submitted to the Department of Chemistry
in Partial Fulfilment of the Requirements for the Degree of

Master of Science
in Biological Chemistry

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2019

© 2019 Massachusetts Institute of Technology. All rights reserved

Signature of Author: ________________________________

Department of Chemistry
August 12, 2019

Signature redacted

Certified by: ________________________________

Mei Häng
Professor of Chemistry
Thesis Supervisor

Signature redacted

Accepted by: ________________________________

Róbert W. Field
Haslam and Dewey Professor of Chemistry
Chairman, Departmental Committee on Graduate Students

Signature redacted
Solid-State NMR Investigation of Viral Fusion Glycoprotein 41 (gp41)

by

Chloe A. Morgan

Submitted to the Department of Chemistry
on August 6, 2019 in Partial Fulfilment of the
Requirements for the Degree of
Master of Science in Biological Chemistry

Abstract

Protein-mediated membrane fusion is an integral part of numerous cellular processes and viral entry. The HIV envelope glycoprotein mediates viral entry into target cells by fusing the viral envelope with the cell membrane. This process requires large-scale and multi-step conformational changes of the viral fusion protein gp41. Our current understanding of the mechanisms of protein-induced membrane structural changes involved in viral entry is incomplete because the hydrophobic N-terminal fusion peptide (FP) and C-terminal transmembrane domain (TMD) of gp41 have resisted structure determination. In our study, we expressed a gp41 construct, “short NC”, containing both hydrophobic termini, including the FP, the fusion-peptide proximal region (FPPR), the membrane-proximal external region (MPER), and the TMD, together with a truncated water-soluble ectodomain linking the N and C termini.

In order to probe the membrane-bound topology and conformation of gp41, we reconstituted “short NC” gp41 into a virus-mimetic membrane for solid-state NMR experiments. $^{13}$C chemical shifts of $^{13}$C isotopically labeled residues indicated that the C-terminal MPER-TMD region is predominantly $\alpha$-helical, whereas the N-terminal FP-FPPR exhibits $\beta$-sheet character. Water and lipid $^1$H polarization transfer to the protein revealed that the TMD is well inserted into the membrane, while the FPPR and MPER are more hydrated and exposed to the membrane surface. Importantly, we observed correlation signals between the FP-FPPR and the MPER, providing direct evidence that the ectodomain is sufficiently collapsed to bring the N- and C-terminal hydrophobic domains into close proximity. These results support a hemifusion-like structural model of gp41 in which its ectodomain forms a partially folded hairpin that places the FPPR and MPER on the opposing surfaces of two lipid membranes that are in close proximity.
Contents

1. Abstract 2
2. Acknowledgements 4
3. Background 5
4. Fully hydrophobic HIV glycoprotein 41 (gp41) adopts a hemifusion-like conformation in phospholipid bilayers
   4.1. Introduction 8
   4.2. Materials and Methods 10
   4.3. Results 15
   4.4. Discussion and Future Work 26
   4.5. Supplemental Figures 31
   4.6. References 33
5. Supplemental Protocols
   5.1. gp41 short NC expression and purification 37
2 Acknowledgements

Many thanks to Professor Mei Hong for her mentorship. Her enthusiasm for solid-state NMR and high standards for research encouraged me to challenge the quality of my own work and engage in more rigorous science.

Thank you to the members of my thesis committee, Professors Matthew Shoulders and Gabriela Schlau-Cohen, for their thoughtful feedback and guidance while I prepared my thesis.

The members of the Hong Group were an enormous help during my years in the lab. Thanks are due to Dr. Myungwoon Lee for introducing me to the gp41 project, Shiva Mandala for his sound advice, Matt Elkins for knowing a fix to every equipment problem, Marty Gelenter for his sense of humor, and Dr. Byungsu Kwon for his endless positivity. Thanks to Aurelio Dregni, not only for having an astounding amount of NMR knowledge, but also for being a generous and compassionate friend. Thanks to Dr. Matthias Roos, Pyae Phyo, and Alex Shcherbakov for their positive contributions to the lab environment.

There are a number of people that have been there for me since the beginning of my academic journey. Thank you to my parents, my brother Robert, and my friends Alex and Nira. Their excitement and encouragement helped me through the toughest parts of this endeavor, and for that I can never thank them enough.

Finally, I want to thank my partner, Dr. James Banigan, and my precious dog Noesy. Their love, support, and understanding are the bedrock on which all of my work is built.
3 Background

Human immunodeficiency virus (HIV) remains a pressing global health issue. Despite recent medical advancements in preventing viral transmission and suppressing viral load, an effective vaccination has remained out of reach. The current preferred methods of preventing the spread of the virus are physical barriers such as condoms and pre-exposure prophylaxis (PrEP). PrEP, a combination of two antiretroviral medications, is specifically targeted toward those at high risk of contracting the disease, such as individuals with HIV-positive sexual partners. However, PrEP can carry a considerable treatment burden to patients, as the medication carries a risk of renal impairment, immune reconstitution syndrome, and a number of more minor adverse reactions such as diarrhea, depression, and fatigue. The medication can carry a substantial cost, and patients must meet with a physician once every three months for HIV testing\textsuperscript{1,2}. Individuals diagnosed with HIV face even greater burdens in accessing antiretroviral therapy, suppressing their viral load, and maintaining good health. Of those diagnosed with HIV, 11\% are uninsured\textsuperscript{3}, further inhibiting their access to appropriate treatments. Prevention and vaccination are the ideal strategies for combating HIV/AIDS.

Membrane fusion is a critical and ubiquitous cellular process that occurs in a number of different contexts and is carried out by proteins embedded in the cell or viral membrane. These proteins can take on a variety of structures, but all achieve the same function of merging two separate lipid bilayers. One common example is the SNARE protein complex in mammals, which is composed of more than 60 individual proteins that together enable synaptic vesicles to dock with the membranes of presynaptic neurons\textsuperscript{4}. Other fusion processes enable exocytosis of vesicles containing hormones and neurotransmitters.

Membrane fusion is also necessary for viral transfection. HIV, like many viruses, enters cells via a fusogenic glycoprotein found in its viral envelope. This protein assembly, known as gp160, is synthesized as a homotrimer, post-translationally cleaved into proteins gp120 and gp41, and incorporated into the viral envelope as it buds from the host cell’s surface\textsuperscript{5}. gp120 binds to the CD4 receptor on the surface of human immune cells and dissociates, exposing the hydrophobic N-terminal domain of gp41 known as the fusion protein (FP). The current understanding of this fusion mechanism is that the FP inserts into the host cell membrane while the C-terminal transmembrane domain (TMD) remains anchored in the viral envelope. gp41 then bends into a hairpin structure,
bringing the viral envelope and cell membrane into close proximity. The FP and the C-terminal transmembrane domain (TMD) change the membrane landscape and facilitate fusion in ways that are not yet well understood. The protein then continues through hemifusion, a step of fusion that seems to be common across all enveloped viruses. In this intermediate on the fusion pathway, the outer leaflets of the viral and cell membranes merge, a perturbation of membrane topology that results in negative Gaussian curvature. In our study, which is described in the following chapter, we explored some of these steps of HIV fusion in more detail.

In order to study gp41 in a native-like environment, we turned to solid-state nuclear magnetic resonance spectroscopy (ssNMR). In general, NMR is a useful technique for probing and determining molecular structures and dynamics. One of the great benefits of ssNMR over solution-state NMR is the ability to study large protein complexes and proteins embedded in lipid bilayers. This alleviated the need to fully solubilize the protein in detergents or isotropic bicelles, which may not appropriately mimic the native environment. Therefore, solid-state NMR is an appropriate method to use when studying membrane proteins such as gp41. Magic-angle spinning (MAS) solid-state NMR spectroscopy is one of the predominant forms of ssNMR: in MAS ssNMR, the samples are mechanically spun at the magic angle. This averages out anisotropic interactions in the sample to their isotropic values, allowing for the collection of high-resolution, solution-like NMR spectra.

Our paper fills several important gaps in the existing gp41 literature. Many structural studies have used constructs that exclude the hydrophobic transmembrane domain (TMD) and fusion peptide in order to facilitate crystallization. The omittance of these domains make it impossible to investigate the relationship and packing of the TMD and FP with the ectodomain. Additionally, NMR studies of past constructs have been performed non-native-like conditions such as detergent bicelles or micelles, neither of which faithfully replicate the complex lipid bilayer into which gp41 is inserted in vivo. These non-native-like conditions interfere with the protein’s ability to generate membrane curvature and undergo conformational changes crucial to facilitating virus-cell fusion. The construct used in our work contains full hydrophobic domains and was reconstituted into a lipid mixture that mimics the composition of the cell and viral membranes. This ensured that experiments were performed under conditions closely resembling the native environment of gp41.
We sought to investigate and confirm the secondary structural character of gp41’s individual domains using two-dimensional $^{13}$C-$^{13}$C correlation experiments. We showed that the N-terminal domains of the protein take on a β-strand conformation, while the C-terminal domains are α-helical. Using these data on secondary structure, we can treat certain residues as reporters of each domain, as some residues are mostly confined to one domain. For example, 9 of 11 glutamine residues are found in the ectodomain. This allowed us to garner more information on domain-specific hydration and interdomain contacts. We found that the N-terminal domains, while less hydrated than the ectodomain of the protein, are more exposed to water than the C-terminal domains. Long-range correlations between two sequentially disparate domains in the protein suggest that the two hydrophobic termini of the protein (the FP and TMD) are in close proximity. This may shed light on the process of hemifusion and lays the foundation for more detailed studies of both “short NC” and other constructs of gp41.
4 Fully hydrophobic HIV glycoprotein 41 (gp41) adopts a hemifusion-like conformation in phospholipid bilayers

*Authors: Myungwoon Lee*, Chloe Morgan*, and Mei Hong
*equal contribution

4.1 Introduction

The HIV virus enters sensitive cells using its envelope glycoprotein complex, gp160. The protein is biosynthesized as a homotrimer, and is transported to the cell surface after proteolytic cleavage into two subunits, gp120 and gp41 \(^8,9\). Receptor binding and pH changes trigger a cascade of conformational changes, which merge the viral envelope and the cell membrane \(^1\). The first change involves unfolding of the gp41 structure to expose and insert an N-terminal fusion peptide (FP) into the host cell membrane while the C-terminal TMD remains anchored in the viral envelope. This extended intermediate then bends into a helical hairpin, forming a trimer of hairpins, which brings the cell membrane and the viral envelope into close proximity \(^1,2\). Subsequently the membrane-interacting FP and TMD disrupt the two bilayers in ways that are not yet well understood, which progressively merge the outer and inner leaflets of the two bilayers, eventually forming a fusion pore.

The six-helix-bundle (6HB) structure formed by the trimer of hairpins is diagnostic of the end point of virus-cell fusion, and has been observed for the ectodomain of a number of viral fusion proteins \(^1,2\). For gp41, the NHR lies in the interior of the 6HB as a parallel coiled coil while the CHR drapes on the exterior, antiparallel to the NHR \(^13,14\). However, to date, no crystal structures of this post-fusion state have included the membrane-active TMD and FP segments. Instead, these domains are either removed from the protein construct to enable crystallization, replaced by a different anchor, or retained but not detected due to disorder. Therefore, it remains unknown whether the 6HB of the ectodomain continues into the lipid membrane to place the FP and TMD in close proximity. This question is further amplified by the fact that gp41, like some of the other viral fusion proteins, contains a membrane-proximal external region (MPER) N-terminal to the TMD, which has the propensity to lie on the membrane surface \(^15,16\). How this MPER packs with the FP and TMD in the post-fusion state is so far unknown.
Extensive NMR studies of the conformations of the FP and TMD of viral fusion proteins have shown that these two domains are not always \( \alpha \)-helical in the biologically relevant environment of lipid bilayers. For example, in cholesterol-containing lipid membranes, the gp41 FP assembles into antiparallel \( \beta \)-sheets and is shallowly inserted into the membrane\(^{17-19} \). This conformation is seen both in the presence and absence of the ectodomain\(^{18,20} \). In comparison, the TMD is more stably \( \alpha \)-helical in lipid bicelles\(^{21,22} \) as well as in lipid bilayers, but the oligomeric state is sensitive to the membrane environment. In cholesterol-containing lipid bilayers, the MPER-TMD forms a trimeric umbrella-like structure where the helical TMD spans the bilayer while the helical MPER lies on the membrane surface\(^{15} \). In comparison, the TMD of the parainfluenza virus 5 (PIV5) fusion protein F is conformationally plastic even in phospholipid bilayers: in negative-curvature PE membranes, the two ends of the TMD adopt \( \beta \)-sheet conformations, but in lamellar PC membranes the entire TMD is \( \alpha \)-helical\(^ {23} \). Given the ability of the FP and TMD of viral fusion proteins to adopt \( \beta \)-sheet conformations under certain conditions, it is important to determine the membrane-bound structure of these two domains in the presence of each other and in the presence of the ectodomain.

In addition to the conformation of the individual domains, the three-dimensional fold of full-length gp41 and how it evolves along the fusion pathway remains poorly understood. Due to the difficulty of producing full-length gp41 and other viral fusion proteins with sufficiently high order for crystal structure determination, or in sufficient quantity to facilitate NMR studies in phospholipid membranes, most studies of viral fusion proteins have involved truncating the protein length and using simpler membrane-mimetics such as bicelles or micelles. Since the function of viral fusion proteins centers on protein conformational changes and membrane-curvature generation, truncated protein constructs that deviate from the native protein’s ability to undergo conformational changes and membrane-mimetic solvents that do not fully reproduce the bilayer curvature can interfere with mechanistic understanding of virus-cell fusion.

Recently, several studies were reported in which both the ectodomain and the membrane-interacting domains of gp41 were included. A solution NMR study of full-length gp41 bound to DPC micelles showed \( \alpha \)-helical structures for the N-terminal FP, the fusion peptide proximal region (FPPR), the NHR, and the immunodominant loop. However, no signals were observed for the C-terminal CHR, MPER, and TMD\(^ {24} \), indicating that the C-terminal segments of the protein undergo slow or intermediate-timescale motion. Analytical ultracentrifugation data of a similar
construct that lacks the FP showed that the protein is trimeric in DPC micelles, suggesting that the NHR may be the trimerization core of the protein. In addition, an NHR-CHR only construct was found to be monomeric, suggesting that the hydrophobic domains are required for trimerization. Thus, although these studies approach full-length gp41, the minimum structural requirement for trimerization and the three-dimensional fold of the full-length protein are still inconclusive.

In this study, we investigate the structure of a gp41 construct that contains both the N- and C-terminal hydrophobic domains. Specifically, this construct spans the full FP-FPPR and MPER-TM domains and connects the two with a shortened NHR-CHR segment to simplify spectral assignment and analysis (Table 1). We incorporated this “short NC” gp41 construct into a cholesterol-containing virus-memetic lipid membrane and measured conformation-dependent $^{13}$C chemical shifts, membrane insertion depths, and three-dimensional fold of the protein. We show that when both hydrophobic termini are present, the FP-FPPR segments have a predominantly $\beta$-strand conformation whereas the C-terminal MPER-TMD is $\alpha$-helical. The $\beta$-strand FP is only partially inserted into the membrane, whereas the $\alpha$-helical TMD is well inserted into the lipid bilayer. Moreover, we observed long-range correlations between the MPER and FP-FPPR, providing evidence for close proximity of the two hydrophobic termini of the protein. Based on these data, we propose a partially bent hairpin-like structural topology that is associated with two lipid membranes, thus suggesting a hemifusion-like intermediate.

4.2 Materials and Methods

Expression and purification of gp41

The amino acid sequence and residue distribution of this “short NC” gp41 (heretofore designated as gp41) is shown in Table 1. We expressed and purified uniformly (U) $^{13}$C, $^{15}$N-labeled protein and 1,3-$^{13}$C labeled protein for this study. The U-$^{13}$C, $^{15}$N-labeled gp41 was expressed in E. coli Rosetta pLysS (DE3) cells (Novagen) while the 1,3-$^{13}$C-labeled protein was expressed in E. coli Lemo21 (DE3) cells (New England BioLabs). Bacteria were grown in 2 L of LB medium at 37°C until A$_{600}$ reached 0.5. The cells were then centrifuged at 7,000 rpm for 10 min. For the 1,3-$^{13}$C-labeled protein, the cell pellet was resuspended in 1 L of M9 minimal media containing 3 g of 1,3-$^{13}$C-labeled glycerol and 1 g of $^{15}$N-labeled ammonium chloride. 100 mg of
unlabeled His, Glu, Gln were added to the media to prevent these residues from being isotopically labeled. For the U-\(^13\)C-labeled protein, the cell pellet was resuspended in 1 L of M9 minimal media containing 3 g of U-\(^13\)C-labeled glucose and 1 g of \(^15\)N-labeled ammonium chloride. 100 mg of unlabeled His, Lys, Met, and Leu were added to reverse-label these residues. Each cell pellet was resuspended in the M9 media and equilibrated at 37°C for 30 min. Once the \(A_{600}\) increased by \(~10\)%, the temperature was decreased to 25°C and 0.5 mM of IPTG was added to induce protein expression for 18-20 hrs. After protein expression, the cells were centrifuged at 7,000 rpm for 10 min at 4°C.

**Table 1.** Amino acid sequence and distribution of gp41 short NC.

<table>
<thead>
<tr>
<th>Residues</th>
<th>FP</th>
<th>FPPR</th>
<th>HR</th>
<th>MPER</th>
<th>TM</th>
<th>CT</th>
<th>Tag</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Asn</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Asp</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Arg</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Phe</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Gly</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>Glu</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Gln</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>His</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Ile</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Leu</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>19</td>
</tr>
<tr>
<td>Lys</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Met</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Pro</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Ser</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Thr</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Trp</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Tyr</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Val</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>16</td>
<td>12</td>
<td>37</td>
<td>19</td>
<td>21</td>
<td>11</td>
<td>26</td>
<td>142</td>
</tr>
</tbody>
</table>

The cell pellet was resuspended in 40 mL of lysis buffer at pH 8.0 (50 mM Tris-HCl, 200 mM NaCl) and sonicated for 30 min in an ice bath. The lysed cells were centrifuged at 11,500
rpm and 4°C for 45 min. The protein was primarily found in the inclusion body. The pellets were resuspended in lysis buffer and sonicated one more time for 30 min in an ice bath. The lysed cells were centrifuged at 11,500 rpm and 4°C for 45 min. The inclusion body was washed with lysis buffer to remove any remaining soluble fractions.

The crude protein was purified using nickel affinity column chromatography. The inclusion body was dissolved in 8 M urea and 1% SDS and stirred overnight at room temperature. After 1 hr of centrifugation at 11,500 rpm and 25°C, the supernatant was loaded onto the Ni²⁺ column (Biorad). The column was placed in a rotator at room temperature for 3 hrs to allow protein binding. The column was then washed stepwise by 1) 8 M urea and 1% SDS; 2) 4 M urea, 0.5% SDS, and 20 mM imidazole; 3) 2 M urea, 0.2% SDS, and 20 mM imidazole; and 4) 0.2% SDS. The protein was eluted using 250 mM imidazole and 0.2% SDS. All washing and elution solutions were prepared in pH 8.0 lysis buffer. The yield after the column purification was ~2 mg protein per liter of M9 medium. To remove SDS, we dialyzed the eluent in a dialysis bag with a molecular weight cutoff of 10.0–10.5 kDa against 1 L of water for 5 days with water change every 12 hrs. The precipitated protein was collected and lyophilized to obtain a dry powder.

**Membrane protein sample preparation**

The purified gp41 was reconstituted into three different lipid membranes in this work: POPE, DOPE, and a composite membrane designated as VMS(-), which consists of POPC : POPE : POPS : cholesterol at molar ratios of 35 : 15 : 20 : 30. The protein : lipid molar ratio was 1:60 for all samples. For ³¹P NMR experiments to examine the impact of the protein on membrane curvature and hydration, we reconstituted 3 mg of unlabeled gp41 into all three membranes with the 1:60 of protein : lipid molar ratio. For 2D correlation experiments to investigate the protein structure, we reconstituted the 7 mg of 1, 3-¹³C-labeled and 4 mg of U-¹³C-labeled protein into the VMS(-) membrane.

Phospholipids were dissolved in chloroform while the protein was dissolved in hexafluoroisopropanol (HFIP). The two solutions were mixed and the solvents were removed under nitrogen gas, then the sample was lyophilized overnight. The dried protein-lipid film was resuspended in pH 7.5 HEPES buffer (10 mM HEPES-NaOH, 1 mM EDTA, 0.1 mM NaN₃) and subjected to seven freeze-thaw cycles between liquid nitrogen and 35°C water bath to create homogeneous vesicles. The vesicle solutions were spun at 40,000 rpm using a Beckman SW60Ti
rotor at 4°C for 6 hrs to obtain wet membrane pellets, then allowed to equilibrate in a desiccator to 40% water by mass. The equilibrated membrane pellets were spun into a MAS rotor through a pipette tip.

**Lipid mixing assays**

VMS(-) and POPE membranes were used to measure peptide-induced lipid mixing. Lipids were dissolved in chloroform and dried under nitrogen gas. The dried lipid film was resuspended in 10 mM HEPES buffer (pH 7.5), freeze-thawed 15 times between liquid nitrogen and a 35°C water bath, then extruded 15-20 times through 100 nm membranes to produce homogeneous large unilamellar vesicles (LUVs). Fluorescently labeled vesicles containing 2 mol% of the fluorescent lipid NBD-PE (1,2-dipalmitoyl-sn-glycero-3-phosphoethanolamine-N-(7-nitro-2,1,3-benzoazadazol-4-yl)) and 2 mol% of the quenching lipid Rh-PE (1,2-dipalmitoyl-sn-glycero-3-phosphoethanolamine-N-(Lissamine rhodamine B sulfonyl)) were prepared using the same method. Unlabeled and labeled vesicles were mixed at a 9:1 mole ratio with a total lipid concentration of 150 μM. Purified gp41 was dissolved in TFE or formic acid, then added to the lipid vesicle solution to reach a protein:lipid molar ratio of 1:20. A HORIBA Fluoromax-P fluorimeter was used to measure fluorescence at an excitation wavelength of 465 nm and an emission wavelength of 530 nm. Each measurement was carried out in 2 mL of LUV solution under continuous stirring with a time increment of 1 s. 20 μL of 10% Triton X-100 was added to the 2 mL solution to measure the maximum fluorescence, $F_{max}$. We measured the initial fluorescence before ($F_0$) and after the addition of the peptide ($F_t$). The percent lipid mixing was obtained as $\%$ mixing = $[(F_t - F_0) / (F_{max} - F_0)] \times 100$, where $F_t$ is the fluorescence intensity at time $t$.

**Circular dichroism experiments**

We investigated the global conformation of gp41 in DPC and lipid bilayers using circular dichroism (CD) experiments. For DPC samples, the protein was dissolved in 0.5 ml of 0.5% DPC solution to a final concentration of 4 μM. The POPC/POPG membrane samples were prepared in the same way as the SSNMR samples with a P/L molar ratio of 1:60 and the final protein concentration was 25 μM. The protein-containing membrane and protein-free-membrane samples were subjected to bath sonication for 30 min after 7 cycles of freeze-thaw to obtain a sufficiently homogeneous vesicle solution. CD spectra were measured at room temperature on an AVIV 202
spectrometer using a 1 mm path-length quartz cuvette. The spectra of the protein-free membrane samples were subtracted from the spectra of the protein-containing samples to obtain the pure-protein spectrum.

**Solid-state NMR experiments**

Solid-state NMR spectra were measured on Bruker spectrometers at $^1$H Larmor frequencies of 400, 600 and 800 MHz. $^{13}$C chemical shifts were externally referenced on the TMS scale to the adamantane CH$_2$ signal at 38.48 ppm or the 14.0 ppm peak Met Cε peak of the model peptide, N-formyl-Met-Leu-Phe-OH (MLF). $^{31}$P chemical shifts were referenced to the hydroxyapatite $^{31}$P signal at 2.73 ppm on the phosphoric acid scale.

Static $^{31}$P spectra of lipid membranes in the absence and presence of gp41 were measured to investigate the effects of the protein on membrane curvature. 2D $^1$H-$^{31}$P correlation (HETCOR) spectra under MAS were measured to investigate the membrane surface hydration in the presence of the protein$^{23,30-32}$. These HETCOR spectra were measured at 298 K under 5 kHz MAS using a 100 ms $^1$H mixing time.

Conformation-dependent $^{13}$C chemical shifts and inter-residue correlations were measured using 2D $^{13}$C-$^{13}$C correlation experiments with CORD spin diffusion$^{33}$. Mixing times were 55 ms and 300 ms for the U-$^{13}$C labeled protein and 100 ms for the 1,3-$^{13}$C labeled protein. The spectra were measured at 263 K under 13.5 kHz MAS.

To quantify protein dynamics, we conducted 2D $^{13}$C-$^1$H DIPSHIFT experiments at 303 K under 7 kHz MAS$^{34,35}$. A $^1$H FSLG homonuclear decoupling sequence$^{36}$ was used during the C-H dipolar evolution period. The time-dependent $^{13}$C-$^1$H dipolar evolution intensities were fitted to obtain the dipolar couplings. The observed couplings were then divided by the FSLG scaling factor of 0.577 to obtain the true couplings. The order parameter $S_{CH}$ was calculated as the ratio of the true coupling to the rigid-limit one-bond dipolar coupling strength of 22.7 kHz.

To investigate the depth of insertion of the protein in the lipid membrane, we measured 2D $^1$H-$^{13}$C correlation spectra that correlate the water and lipid $^1$H chemical shifts with the protein $^{13}$C chemical shifts$^{37}$. These spectra were measured at 303 K in the liquid-crystalline phase of the membrane under 14.5 kHz MAS. A $^1$H T$_2$ filter of 0.45 ms x 2 was used to suppress the $^1$H magnetization of the relatively rigid protein$^{37,38}$. A $^1$H spin diffusion mixing period of 25 ms to 225 ms was used to transfer the water and lipid $^1$H magnetization to the protein. To investigate the
residue-specific water accessibility of gp41, we also measured water-edited 2D $^{13}$C-$^{13}$C correlation spectra $^{39,40}$. The experiment uses a soft Gaussian 90° pulse centered at the water $^1$H chemical shift of 5.1 ppm to selectively excite the water magnetization. A 9 ms $^1$H spin diffusion mixing period was then used to transfer the polarization from water to the protein. The spectrum was measured at 263 K under 10.5 kHz MAS. Due to frictional heating, the estimated sample temperature was about 273 K, as verified by the water $^1$H chemical shift of 5.1 ppm.

4.3 Results

gp41 short NC has fusion activity and generates membrane curvature

To determine whether this ectodomain-shortened gp41 has fusion activity, we conducted lipid mixing assays. Fluorescence spectra (Fig. 1a) of dye-labeled vesicles mixed with unlabeled vesicles show an increase of the fluorescence intensities within 10 minutes, indicating that the two vesicle populations are well mixed by the protein. This mixing is observed for both POPE vesicles and VMS(-) vesicles at low pH, but at neutral pH minimal lipid mixing is detected for the POPE sample. This result is consistent with the known higher fusion activity of gp41 at acidic pH, which has been attributed to stronger affinity of the protein for the lipids at low pH $^{41-43}$. Between the two lipid membranes, the POPE sample shows higher fusion than the VMS(-) membrane, suggesting that gp41 has stronger fusion activity towards membranes that possess spontaneous negative curvature.

To assess the global conformation of the protein, we measured the CD spectra of gp41 bound to DPC micelles and lipid vesicles. The DPC-bound protein exhibits a strong $\alpha$-helical signature at both low and high pH (Fig. 1b). In POPC/POPG membranes, the protein has somewhat reduced helicity, and the helical content is higher in the presence of higher concentrations of negatively charged POPG lipids.

To investigate whether this gp41 construct causes membrane curvature and dehydration, we measured static $^{31}$P spectra and 2D $^1$H-$^{31}$P correlation spectra under MAS (Fig. 2). The VMS(-) and POPE membranes show a uniaxial powder pattern that is characteristic of lamellar membranes, and are unperturbed by the protein (Fig. 2a). In comparison, the protein caused a strong isotropic peak to the inverse-hexagonal-phase DOPE membrane, indicating that gp41 strongly alters the curvature of the membrane that possesses significant negative spontaneous
curvature. The protein-induced isotropic peak is indicative of negative Gaussian curvature \(^{44,45}\), which is the membrane topology necessary for hemifusion and fusion-pore formation. The generation of this isotropic peak in the DOPE membrane is similar to the action of the PIV5 TMD, which has been shown to cause lipid cubic phases to DOPE membranes \(^{45}\). 2D \(^{2}H-^{31}P\) correlation spectra of all three membranes in the presence of gp41 show clear water cross peaks with the lipid phosphate headgroups (Fig. 2b), indicating that all membranes are well-hydrated in the presence of the protein.

![Graphs showing lipid mixing and CD spectra](image)

**Figure 1.** Characterization of the lipid-mixing activity and global conformation of the gp41 construct. (a) Percent lipid mixing of POPE membranes (left) and VMS(-) membranes (right) in the presence of gp41. The 7.5 uM protein was used with 1:20 of protein : lipid molar ratio. The protein causes lipid mixing at low pH but not high pH. (b) CD spectra of gp41 bound to DPC micelles (left) and POPC/POPG membranes (right). 4 uM protein was dissolved in 0.5% DPC solution or 25 uM protein was mixed with POPC/POPG lipid with 1:60 of protein : lipid molar ratio. The protein is predominantly \(\alpha\)-helical in both environments.
Figure 2. Effects of gp41 on membrane curvature and hydration. (a) $^{31}$P static spectra of VMS(-), POPE, and DOPE membranes in the absence (black) and presence (red) of the protein. 3 mg of unlabeled gp41 protein was bound to each membrane with 1:60 of protein : lipid molar ratio. The protein did not perturb the VMS(-) membrane and only exerted a small impact on the POPE membrane curvature. In comparison, the protein caused a strong isotropic peak in the DOPE membrane, indicating the generation of negative Gaussian curvature. (b) 2D $^1$H-$^{31}$P correlation spectra of the three protein-bound membranes under MAS. All three membranes show water – $^{31}$P cross peaks, indicating that the membrane surface is well hydrated.

**Domain-specific backbone conformations of gp41**

To investigate the backbone conformation of membrane-bound gp41, we measured 2D $^{13}$C-$^{13}$C correlation spectra of both U-$^{13}$C labeled protein and the more sparsely 1,3-$^{13}$C labeled protein. The amino acid sequence of the protein (Table 1) indicates that certain residue types are enriched in one of the domains. For example, the FP contains six out of the nine Ala residues in the protein, whereas the FPPR contains three out of four Thr residues in the protein. Conversely, the MPER harbors five out of six Trp residues, whereas the TMD contains half of the Ile and Val residues in the protein. Therefore, the chemical shifts of these residue types serve as useful reporters of the backbone conformation of these domains when single-site resolution is difficult to obtain due to the broad linewidths of this protein in the cholesterol-rich membrane.

We measured 2D $^{13}$C-$^{13}$C correlation spectra using $^{13}$C spin diffusion mixing times of 55 ms and 300 ms for the U-$^{13}$C-labeled protein and 100 ms for the sparsely 1,3-$^{13}$C-labeled protein (Fig. 3, Fig. S1). In all three spectra, characteristic chemical shifts of many amino acid residues allowed the evaluation of the conformations of the different domains of the protein. We use the
cross peak intensities in the 55 ms and 100 ms spectra to estimate the relative number of residues in different secondary structures, while the 300 ms spectrum is useful for identifying inter-residue cross peaks. In the 1,3-\textsuperscript{13}C labeled spectrum (Fig. 3a), a β-sheet Thr Cα-Cγ cross peak at (59.0, 19.4) ppm and a Cβ-Cγ cross peak at (69.0, 19.4) ppm are detected, and are well resolved from a weaker α-helical Thr Cα/Cβ-Cγ cross peak at (65.5, 19.9) ppm. The sheet : helix intensity ratio is about 2.7 : 1. Since three Thr residues exist in the FPPR domain while one Thr resides in the MPER, this intensity distribution suggests that the FPPR predominantly adopts the β-sheet conformation. Consistently, a β-sheet Thr Cβ-Cα cross peak is stronger than the α-helical cross peak in the 55 ms spectrum of the U-\textsuperscript{13}C labeled protein (Fig. S1), supporting the assignment of the β-strand conformation for the FPPR. The 55 ms spectrum of the U-\textsuperscript{13}C labeled protein also shows resolved β-sheet and α-helical Ala Cα-Cβ cross peaks at (48.7, 21.3) ppm and (53.0, 26.0) ppm, respectively (Fig. S1), and the β-sheet intensity is about 2-fold stronger than the α-helical peak. Since six out of the nine Ala residues in the protein are located in the FP-FPPR regions, this intensity distribution again suggests that the FP-FPPR has a β-strand conformation.

The conformation of the MPER-TMD region can be assessed through Ile cross peaks. Ile Cα-Cγ1 and Cα-Cγ2 cross peaks are observed at α-helical chemical shifts such as (63.5, 15.3) ppm (Fig. 3a, Fig. S1), and these peaks are 2-3 fold stronger than the corresponding peaks at β-sheet chemical shifts such as (57.6, 15.8) ppm. Since six out of eight Ile residues lie in the MPER and TMD, this intensity distribution indicates that the MPER-TMD region is predominantly α-helical. In the carbonyl region of the 55 ms 2D spectrum (Fig. S1), an α-helical Ala Cβ-C’ cross peak at (16.0, 178.0) ppm and an Ile Cγ2-C’ cross peak at (15.3, 175.1) ppm have an intensity ratio of about 1 : 3. The MPER-TMD region contains 2 Ala and 6 Ile residues. Thus, this intensity ratio lends further support to the assignment of a predominantly α-helical conformation to the MPER-TMD segment.
Figure 3. 2D $^{13}$C-$^{13}$C correlation spectra of VMS(-) membrane-bound gp41 for conformational analysis. (a) 100 ms spin diffusion spectrum of $^{1,3}$C-labeled protein. Sparse $^{13}$C labeling increases the spectral resolution and thus facilitates resonance assignment. Residue-type assignment is indicated together with the secondary structure motif. H: helix; C: coil; S: sheet. Green dashed rectangles denote cross peaks whose intensities are used to deduce the conformational propensity of the various domains. (b) 300 ms spectrum of $U^{13}$C-labeled protein. Cross peaks are assigned in pink, black, and blue for $\alpha$-helical, random coil, and $\beta$-sheet chemical shifts, respectively. Both spectra were measured at 263 K. 7 mg and 4 mg of protein were used with 1:60 of protein: lipid molar ratio for $^{1,3}$C-labeled and $U^{13}$C-labeled gp41, respectively.
With 300 ms $^{13}$C spin diffusion, we detected a number of inter-residue cross peaks (Fig. 3b). Although these inter-residue $^{13}$C-$^{13}$C cross peaks cannot be sequentially assigned without 3D $^{15}$N-$^{13}$C correlation data, most of these cross peaks correspond to residue types that exist as sequential pairs in the protein. Thus, we attribute them to sequential contacts. The majority of these inter-residue cross peaks are observed at $\alpha$-helical chemical shifts, whereas the $\beta$-sheet inter-residue cross peaks are fewer and weaker. This is consistent with the fact that sequential distances in $\beta$-sheets are much longer than those in $\alpha$-helices. We observed inter-residue Ala-Val cross peaks at both $\alpha$-helical and $\beta$-sheet chemical shifts, consistent with the fact that sequential Ala-Val residue pairs exist in both FP and TMD segments. However, an additional Ala C$\alpha$-Gly C$\alpha$ cross peak is clearly observed at the $\beta$-sheet chemical shift of (48.7, 42.6) ppm. The only region of the protein that contains sequential Gly-Ala pairs is the FP-FPPR, which harbors three Gly-Ala and Ala-Gly pairs (Table 1). Therefore, the presence of this $\beta$-sheet cross peak provides unambiguous evidence that the $\beta$-sheet Ala residues occur in the N-terminal FP-FPPR region, while the $\alpha$-helical Ala cross peaks are only located in the MPER-TMD domain.

Taken together, the relative intensities of $\alpha$-helical and $\beta$-sheet cross peak intensities as well as inter-residue correlations found in these 2D spectra support the conclusion that the MPER and TMD of gp41 are predominantly $\alpha$-helical whereas the FP and FPPR mainly adopt a $\beta$-strand conformation when bound to the cholesterol-rich membrane.

**Membrane-bound topology of gp41 from lipid-protein and water-protein polarization transfer**

To investigate the depth of insertion of gp41 into the lipid bilayer, we measured 2D $^1$H-$^{13}$C correlation spectra of the 1,3-$^{13}$C-labeled protein in VMS(-) membranes. $^1$H spin diffusion mixing times of 25 to 225 ms were used to transfer the lipid- and water-$^1$H magnetization to the protein. Fig. 4 shows a representative 2D spectrum measured with 25 ms $^1$H spin diffusion. Comparing the lipid acyl chain CH$_2$ cross section and the water cross section, we find that the 53 ppm C$\alpha$ peak for various $\beta$-strand residues is preferentially enhanced in the water cross section compared to the lipid cross section. In contrast, the $\alpha$-helical C$\alpha$ signal at $\sim$56 ppm is stronger in the lipid cross section than the water cross section. This difference indicates that the $\alpha$-helical domains of the protein are more deeply inserted into the membrane than the $\beta$-strand conformation. The buildup curves of water and lipid to the protein (Fig. 4c) quantify this observation. In particular, the lipid-
to-protein buildup curves differ between different residue types. For example, the α-helical Ile Cα shows the fastest lipid buildup intensities, whereas the α-helical Ser Cβ signal at 61 ppm shows the slowest buildup from the lipid chains among all residues. Helical Ile residues are mainly located in the TMD whereas Ser residues are mostly found in the NHR-CHR domain. Therefore, these lipid ¹H magnetization transfer rates indicate that the TMD is well inserted into the lipid bilayer while the ectodomain is exposed to water. Interestingly, the β-strand Thr peaks at 59 ppm and 69 ppm, as well as the β-strand signals at 53 ppm, show slow lipid buildup, indicating that the N-terminal β-strand domains are not deeply inserted into the membrane. Compared to the lipid-to-protein ¹H polarization transfer rates, the water buildup curves show less distinction between different residues. Despite the lower spatial resolution, the trend is approximately opposite that of the lipid to protein buildup rates: the α-helical peaks at 56 and 64 ppm show slower water polarization transfer than the β-sheet peaks at 53 and 59 ppm, consistently indicating that the α-helical C-terminus of the protein is more inserted into the membrane while the β-sheet rich N-terminus is more exposed to water.

To obtain more residue-specific information about the water accessibility of the different domains of the protein, we measured a water-edited 2D ¹³C-¹³C correlation spectrum of the VMS(-) bound protein (Fig. 5a). The intensity ratios between the water-edited spectrum and the control spectrum are indicated. We found that β-strand cross peaks such as those of Thr and Ser exhibit higher water-transferred intensities than the α-helical cross peaks such as those of Ile. For Ala residues, the β-sheet cross peak intensity is also higher than the α-helical intensity. Thus, the water-edited 2D spectrum is consistent with the ¹H-¹³C 2D correlation spectra in indicating that the β-strand FP-FPPR domains are more exposed to water than the α-helical MPER-TMD. Fig. 5b compares the water-edited spectral intensities for residues enriched in different domains of the protein. The highest water accessibility is found for coil residues, which are enriched in the NHR and CHR domains. At the other extreme are α-helical Val, Ile and Gly residues which are enriched in the TMD and which show the lowest water-transferred intensities, indicating that the TMD is the most inserted into the membrane. The helical Asn residues, which are only found in the MPER, show moderately higher hydration than the TMD residues, consistent with the notion that the MPER is not as well inserted into the bilayer as the TMD. In comparison, β-strand Ser, which are enriched in the FPPR, show a relatively high water-transferred intensity of 25%, while the β-strand Ala, Val, Gly, and Ile residues in the FP have lower water-transferred intensities of 13%, 16%, 21%.
11%, an 24%, respectively. Therefore, the FPPR is more exposed to water than the FP, suggesting that this region may lie on the membrane surface.

**Figure 4.** Depth of insertion of 1,3-13C-labeled gp41 in VMS(-) membranes from lipid- and water-1H polarization transfer to the protein. The sample contained 7 mg of protein with the 1:60 of protein : lipid molar ratio. (a) Representative 2D 1H-13C correlation spectrum, measured with 25 ms 1H spin diffusion at 303 K. (b) 13C cross sections at the lipid CH2 (orange) and water (blue) 1H chemical shifts. The β-strand Cα peak at ~52 ppm is much higher in the water cross section than in the lipid cross section, indicating that the β-strand segment is preferentially hydrated and exposed to the membrane surface. (c) Lipid-to-protein and water-to-protein 1H polarization transfer intensities as a function of mixing time. The α-helical peaks from the TMD (red and magenta) show fast lipid polarization transfer while the α-helical Ser that is enriched in the ectodomain shows slow lipid buildup intensities.
Figure 5. Water accessibility of gp41 from 2D water-edited $^{13}$C-$^{13}$C correlation spectra. (a) Full and water-edited 2D spectra of U-$^{13}$C labeled gp41, measured at 263 K. The sample contained 4 mg of protein and the protein : lipid molar ratio was 1:60. The water-edited spectrum was measured using a $^1$H spin diffusion mixing time of 9 ms. Assignments are shown in pink, grey and blue for $\alpha$-helical, random coil, and $\beta$-sheet chemical shifts, respectively. (b) Hydration values for residues in the various domains of gp41. The $\alpha$-helical MPER-TMD segments show lower hydration than the FP and FPPR segments.

Long-range correlations between the FPPR and the MPER domains

To determine whether this gp41 short NC construct adopts a hairpin fold, we measured a 500 ms 2D $^{13}$C-$^{13}$C correlation spectrum. We focused on the aromatic-aliphatic region (Fig. 6), which contains characteristic Trp sidechain $^{13}$C chemical shifts that are well resolved from other
residues’ signals. These resolved signals include Trp Cε3 (117 ppm), Cη2 (122 ppm), and Cδ1 (124 ppm) peaks. In addition, characteristic Phe Cγ and Cδ1 chemical shifts are also resolved. At this long mixing time, many inter-residue correlations are observed, most of which can be attributed to sequential residues in the protein. However, two clear and uniquely assigned cross peaks are observed between β-sheet Ala Cα at 49 ppm and Trp Cδ1 (124 ppm) and Cη2 (122 ppm). Because β-sheet Ala residues are found only in the FP and FPPR domains while the α-helical Trp residues are only found in the MPER, these cross peaks indicate that the N-terminal FP-FPPR must lie in close proximity to the C-terminal MPER. Thus, the protein adopts a hairpin-like fold that places the N- and C-termini within $^{13}$C spin diffusion reach of about 1 nm.

Membrane-bound gp41 undergoes small-amplitude local motions

To investigate the dynamics of membrane-bound gp41, we measured $^{13}$C-$^1$H dipolar couplings at 303 K using the 2D $^{13}$C-$^1$H DIPSHIFT experiment. Fig. 7 shows representative $^{13}$C-$^1$H dipolar dephasing curves, which correspond to relatively large C-H order parameters of 0.69 – 0.89 for all labeled sites. These large order parameters indicate that protein is overall immobilized in the VMS(-) membrane and likely oligomerized, and only exhibits local segmental motion. The 61.3-ppm peak of α-helical Ser Cβ exhibits a relatively weak dipolar coupling with an order parameter of 0.60. We attribute this to higher mobility of the NHR-CHR ectodomain.
Figure 6. Aromatic-aliphatic region of the 500 ms 2D $^{13}$C-$^{13}$C correlation spectrum of membrane-bound gp41. Two cross peaks between β-sheet Ala and α-helical Trp (shown in blue) are observed, indicating that the protein adopts a hairpin-like fold that places the N-terminal FP-FPPR in close proximity with the C-terminal MPER. Other inter-residue cross peaks (shown in pink) can be attributed to sequential residue pairs. The spectrum was measured at 263 K. The sample contained 4 mg of protein and the protein: lipid molar ratio was 1:60.
Figure 7. Mobility of membrane-bound gp41 detected from $^{13}$C-$^1$H dipolar couplings measured using $^{1,3-^{13}}$C-labeled gp41 at 303 K. The sample contained 7 mg of protein with the 1:60 of protein : lipid molar ratio. (a) The first slice of the $^{13}$C dimension of the 2D DIPSHIFT spectrum, showing chemical shift assignment. (b) Representative C-H dipolar dephasing curves of Cα sites. Residues of different conformations show similar $S_{	ext{CH}}$ order parameters of 0.8 to 0.9, indicating relatively uniform and small-amplitude motion of the various domains.

4.4 Discussion and Future Work

The gp41 protein used in this study is designed to probe the post-fusion or hemifusion state of the full-length protein by retaining the two essential hydrophobic termini in full length while shortening the water-soluble ectodomain. The cholesterol-containing complex phospholipid membrane VMS(-) was chosen to mimic the composition of eukaryotic cell membranes as well as the virus envelope. This choice minimizes potential non-native effects of detergent micelles and lipid bicelles such as high membrane curvature and low membrane viscosity, which may perturb
the protein structure. The above results provide new insights into the conformations and three-dimensional fold of gp41, and both validate and revise previous structural conclusions of gp41 obtained using shorter peptides.

The measured $^{13}$C chemical shifts indicate that the N-terminal FP-FPPR domains have significant β-sheet character while the C-terminal MPER-TMD domains are predominantly α-helical. Residues that are enriched in the MPER-TMD such as Ile are predominantly α-helical, whereas residues enriched in the FP-FPPR such as Thr and Ala exhibit β-sheet chemical shifts. This result is in good agreement with previous studies of isolated FP and MPER-TMD peptides in lipid bilayers. In particular, it indicates that the β-sheet conformation of the FP that has been reported for isolated FP in cholesterol-containing membranes $^{17}$ persists in the full-length protein.

In a previously investigated gp41 construct that contains the FP and NHR-CHR but lacks the MPER-TMD, SSNMR data found the FP to adopt a β-sheet conformation $^{18}$. However, in DPC micelles, both the isolated FP $^{46}$ and FP in full-length gp41 $^{24}$ are α-helical. The isolated FP is also α-helical in non-cholesterol-containing membranes $^{47}$. Therefore, the FP conformation is sensitive to the membrane environment but is largely independent of the presence or absence of the MPER-TMD: the helical conformation is promoted by micelles and by the avoidance of cholesterol, suggesting that high membrane curvature and low membrane viscosity favor helix formation.

Compared to the FP, the C-terminal MPER-TMD domains are stably α-helical in a variety of membrane environments $^{16,22,48}$ both as an isolated peptide and in the near full-length protein. For example, a recent SSNMR study found that an MPER-TMD peptide forms a trimeric helix-turn-helix structure, with the MPER lying on the membrane surfaces while the three TMD helices span the membrane $^{15}$. These results indicate that a helical structure at the C-terminus is important for gp41 to maintain its trimeric structure and for its anchoring in the virus envelope through multiple conformational changes during virus-cell fusion.

What is the three-dimensional fold of the protein and the corresponding membrane morphology? Based on three lines of experimental evidence, we propose a partially formed hairpin structure that puts the FPPR and MPER in close proximity while keeping their neighboring FP and TMD apart; moreover, we propose that this protein fold is associated with two different lipid membranes, thus corresponding to a hemifusion-like intermediate (Fig. 8). First, the lipid and water buildup curves of the α-helical, β-sheet and random coil residues indicate that the N-terminal FP-FPPR domains are shallowly inserted into the membrane, the C-terminal TMD spans the
membrane, and the random coil residues are the most accessible to water (Fig. 4, Fig. 5). The shallow insertion of the FP-FPPR is consistent with many peptide structural studies, while the full insertion of the TMD in the membrane is consistent with studies of MPER-TMD peptides by solution and solid-state NMR. Second, the observed Trp long-range correlations with β-sheet Ala (Fig. 6) put strong constraint on the proximity of the membrane-surface MPER and FPPR. The distance separation measurable by $^{13}$C spin diffusion is expected to be within ~1 nm. Third, the lack of long-range correlations between the FP and TMD, as well as the different backbone conformations of these two domains suggest that the FP and TMD more likely interact with two different lipid bilayers rather than the same membrane. The alternative model of putting the FPPR and MPER on the same membrane surface would reduce the likelihood of spatial contact because of the trimeric nature of the protein. In comparison, when the FPPR and MPER lie on two different membrane surfaces that are brought into proximity by the bent ectodomain, the probability for Trp-Ala spatial contact should increase. Therefore, although the present data cannot exclude a reentry model, it favors a hemifusion model in which the protein bridges two lipid membranes. This model implies the presence of local membrane curvature. The $^{31}$P static spectrum of the VMS(-) membrane (Fig. 2) does not show an isotropic peak, while the DOPE membrane does. We suggest that this discrepancy may be explained by the fact that only moderate curvature is required by hemifused membranes, and that the curvature may occur to only a small fraction of lipids such as POPE in the composite membrane.

The structural model of Fig. 8 depicts two trimers at the locally curved hemifusion region. This hypothesis is made based on previous SSNMR data that the β-sheet FP assembles in an antiparallel fashion, which requires two parallel trimers to interdigitate with each other. It has been proposed that to induce sufficient membrane curvature for fusion to occur, multiple trimeric subunits must aggregate.
Figure 8. Schematic model of the membrane-bound conformation and topology of gp41. The protein is depicted as bridging two lipid bilayers in a hemifusion-like state. These two membranes correspond to the virus envelope and the cell membrane in vivo. The trimeric \( \alpha \)-helical TMD is bound to one bilayer whereas the FP is shallowly inserted into the other bilayer in a \( \beta \)-sheet rich conformation. The \( \alpha \)-helical MPER lies on the surface of one bilayer, within \( \sim 1 \) nm of \( \beta \)-sheet FPPR, which lies on the surface of the other bilayer. The water-soluble ectodomain is dynamic and may be significantly disordered. Two trimeric assemblies are depicted, to be consistent with the previous data that indicate that the FP associates as antiparallel \( \beta \)-sheets 49.

Our partial hairpin and hemifusion structural model places the FP and TMD in two different membranes. Evidence both for and against the association and interaction of FP and TMD have been reported in the literature. For example, FRET studies of lipid membranes containing both FP and TMD found heteromeric association 51. On the other hand, evidence that FP and TMD do not associate with each other has been provided by solution NMR studies of DPC-bound gp41, which found that the FP has much faster motional rates than the TMD, thus suggesting a lack of FP-TMD association 24,25. Since isolated peptides cannot recapitulate the conformational constraints imposed by the intervening segments of the FPPR, ectodomain, and MPER, while studies in DPC micelles may be perturbed by the micelle 52, the presence or absence of heteromeric association in these earlier studies should be considered with caution.

Although we prepared samples using 1,-3-\(^{13}\)C-labeling and reverse labeling to reduce dipolar couplings and peak overlaps for better spectral resolution, some of the spectral region in the \(^{13}\)C-\(^{13}\)C spectra still show overlapped peaks. Thus, we mainly utilized well resolved peaks such
as Thr, Ala, Ile, and Trp to quantitatively obtain the domain specific conformational information such as secondary structure, inter-domain interactions and water accessibility. However, overlapped peaks, especially near (50, 30) ppm in the $^{13}$C-$^{13}$C correlation spectra that we assigned as disordered coil from Glu, Gln, Leu, Lys in HR domains might contain b-sheet peaks that might be exist in the NHR domain near FPPR region because of chemical shift similarity between b-sheet and coil. In future experiments, partially deuterated or amino acid specific labeled samples could improve the quality of spectra and provide better information to clarify the HR domain conformation that we propose as a disordered domain. To further test the current structural model, longer-range distance constraints will also be highly desirable. These can be measured, for example, by incorporating paramagnetic tags or $^{19}$F labels into the FP and/or TMD regions and measuring paramagnetic relaxation enhancement or $^{19}$F-$^{19}$F distances. Sequential resonance assignment will be valuable for determining whether the ectodomain indeed bends into a hairpin, and to extract additional constraints that define the relative locations of the FPPR and MPER. Finally, mixed labeled protein samples will be important to confirm and characterize the trimeric nature of this protein.
### Table S1. $^{13}$C chemical shifts of membrane-bound gp41 assigned from 2D correlation spectra. C, H, and S in the brackets denote coil, helix and sheet conformations based on the well-known characteristic chemical shifts of amino acids.

<table>
<thead>
<tr>
<th>Residue</th>
<th>C'</th>
<th>Cα</th>
<th>Cβ</th>
<th>Cδ1</th>
<th>Cδ2</th>
<th>Cε1</th>
<th>Cε2</th>
<th>Cε3</th>
<th>Cy1</th>
<th>Cy2</th>
<th>Cη2</th>
<th>Cζ1</th>
<th>Cζ2</th>
<th>Cζ3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A(C)</td>
<td>174.3</td>
<td>49.7</td>
<td>17.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A(H)</td>
<td>178.0</td>
<td>53.0</td>
<td>16.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A(S)</td>
<td>172.7</td>
<td>48.7</td>
<td>21.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D(C)</td>
<td>175.0</td>
<td>53.8</td>
<td>40.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E(C)</td>
<td>174.3</td>
<td>56.8</td>
<td>181.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>32.8</td>
</tr>
<tr>
<td>F(H)</td>
<td>175.0</td>
<td>59.8</td>
<td>37.1</td>
<td>128.4</td>
<td>128.8</td>
<td>128.9</td>
<td>128.7</td>
<td>136.6</td>
<td>126.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F(S)</td>
<td>53.7</td>
<td>128.6</td>
<td>128.8</td>
<td>136.4</td>
<td>127.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G(H)</td>
<td>172.5</td>
<td>45.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G(S)</td>
<td>169.0</td>
<td>42.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I(C)</td>
<td>12.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>23.4</td>
</tr>
<tr>
<td>I(H)</td>
<td>175.1</td>
<td>63.5</td>
<td>35.5</td>
<td>12.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>27.3</td>
<td>15.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I(S)</td>
<td>172.5</td>
<td>57.6</td>
<td>40.4</td>
<td>12.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>26.2</td>
<td>15.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N(H)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>35.8</td>
</tr>
<tr>
<td>P(C)</td>
<td>175.2</td>
<td>62.9</td>
<td>29.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25.1</td>
</tr>
<tr>
<td>Q(C)</td>
<td>175.5</td>
<td>52.8</td>
<td>27.0</td>
<td>177.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>31.5</td>
</tr>
<tr>
<td>R(C)</td>
<td>54.3</td>
<td>24.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>26.9</td>
</tr>
<tr>
<td>R(H)</td>
<td>56.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25.7</td>
</tr>
<tr>
<td>S(C)</td>
<td>172.7</td>
<td>56.2</td>
<td>61.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S(H)</td>
<td>173.8</td>
<td>59.2</td>
<td>61.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S(S)</td>
<td>172.5</td>
<td>54.1</td>
<td>64.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T(H)</td>
<td>173.7</td>
<td>63.4</td>
<td>65.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19.9</td>
</tr>
<tr>
<td>T(S)</td>
<td>170.7</td>
<td>59.0</td>
<td>69.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19.4</td>
</tr>
<tr>
<td>V(C)</td>
<td>173.5</td>
<td>50.1</td>
<td>29.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19.3</td>
</tr>
<tr>
<td>V(H)</td>
<td>175.5</td>
<td>64.8</td>
<td>28.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19.6</td>
</tr>
<tr>
<td>V(S)</td>
<td>171.6</td>
<td>58.3</td>
<td>33.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19.1</td>
</tr>
<tr>
<td>W(H)</td>
<td>56.6</td>
<td>26.7</td>
<td>124.3</td>
<td>127.1</td>
<td>136.5</td>
<td>117.1</td>
<td>110.9</td>
<td>121.6</td>
<td>109.2</td>
<td>119.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y(C)</td>
<td>174.1</td>
<td>54.5</td>
<td>41.0</td>
<td>129.5</td>
<td>130.8</td>
<td>116.1</td>
<td>154.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y10 II</td>
<td>130.2</td>
<td>117.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure S1. 2D $^{13}$C-$^{13}$C correlation spectrum of U-$^{13}$C labeled gp41 in VMS(-) membranes. A mixing time of 55 ms was used. Amino-acid type assignment is given together with the secondary structure motif. H: helix; C: coil; S: sheet. Green dashed rectangles denote cross peaks whose intensities are used to deduce the conformational propensity of the various domains.
4.6 References


5 Supplemental Protocols

5.1 gp41 short NC expression and purification

Expression

Day 1
1. Make fresh LB and M9 media; autoclave sterilize.

**LB medium (1L)**
- 25 g of LB

**M9 medium (1L)**
- 12.9 g of Na₂HPO₄·7H₂O (MW: 268.38).
- 3 g of KH₂PO₄ (MW: 136.22).
- 0.5 g of NaCl.

2. Culture a carbenicillin and chloramphenicol plate of gp41 in the early morning using Rosetta pLysS competent cells. It will take 8~12 hrs to grow under 37 °C.

3. Add 10 μL of 100mg/mL of carbenicillin and 31mg/ml of chloramphenicol to 10 mL LB medium.

4. Take one colony and put it in 10 mL LB medium.

5. Shake at 250 rpm and 37 °C overnight.

Day 2
1. In the afternoon, add 1 mL of 100mg/mL of carbenicillin and 1mL of 31 mg/mL of chloramphenicol per 1 L of sterile LB.

2. Transfer 10 ml of overnight seed culture to 1 L of LB (0.5 L of LB in each shaker flask).

3. Incubate the flasks at 37 °C and 250 rpm to reach OD₆₀₀ ~ 0.5 (it will take 3 - 4 hours).

4. **While waiting for LB to grow**, prepare M9 medium (1L).
   - To label proteins, add 1.1 g of ^{15}NH₄Cl, 2 g of ^{13}C-glucose or 4 g of ^{13}C-glycerol.
   - 2 ml of 1M MgSO₄, 0.1 ml of 1M CaCl₂ (These should be autoclaved).
   - To block a specific amino acid from being labeled, add 100 mg of unlabeled amino acid per 1L.

5. After OD₆₀₀ reaches ~ 0.5, spin the cells at 7000rpm, 20 °C for 10 min.

6. Resuspend the cell pellets in 20-30 mL of prepared M9. Rinse the centrifuge tubes with M9 twice before pouring cells into the M9 flask. (Transfer ratio(v/v) LB : M9 = 2:1)

7. Place the flasks in the shaker and incubate ~30 min at 37 °C to stabilize cells in M9 medium.
   - Wait for OD₆₀₀ to reach 0.6. **Take a 1 mL sample here for SDS-PAGE analysis.**

8. Decrease the temperature to 25 °C and add IPTG to a final concentration of 0.5 mM.

9. Incubate the flask for ~18 hr at 25 °C.

Purification

Day 3
1. Take a 1 mL sample for SDS-PAGE analysis. Spin the cells at 7000 rpm at 20 °C for 10 min.

2. Prepare lysis buffer (50 mM Tris HCl, 200 mM NaCl, pH 8.0).

3. Re-suspend cells with 30-35 mL lysis buffer and transfer to beaker.
4. Keeping cells on ice, sonicate 5 seconds and wait 5 seconds. Repeat this for 30 minutes at ~4 amplitude.
5. Centrifuge at 12500 rpm at 4 °C for 30 min to separate inclusion body from supernatant.
6. Save the supernatant.
7. Repeat steps 3-6.
8. Wash the inclusion body with 30-35 ml of lysis buffer to wash out any soluble proteins and centrifuge at 12500 rpm at 4°C for 30 min to separate inclusion body from supernatant.
9. Save the supernatant and take a sample for SDS-PAGE analysis
10. Prepare 50 ml urea buffer (8M urea, 0.2% SDS, 50 mM Tris HCl, 200 mM NaCl, pH 8.0)
11. Re-suspend inclusion bodies (pellet) in urea buffer (~35 ml of buffer per 1L culture).
12. Stir the re-suspended inclusion bodies for ~3 hours.
13. Prepare Ni-NTA column. Wash the Ni-NTA column with ~30 ml of water and equilibrate with urea buffer (same buffer to dissolve inclusion body).
14. Centrifuge the dissolved inclusion body solution at 12500 rpm at 8 °C for 30 min to separate the cell debris from dissolved inclusion body.
15. Take a sample of the supernatant for SDS-PAGE analysis. Add supernatant to the prepared Ni-NTA column.
16. Mix well and put the column in rotator overnight at 4 °C to ensure complete binding.

**Day4.**
Prepare all the purification buffers for Ni-NTA purification. The amounts of buffer are for 1L M9 culture.

<table>
<thead>
<tr>
<th>Wash Buffer 1</th>
<th>Wash Buffer 2</th>
<th>Wash Buffer 3</th>
<th>Wash Buffer 4</th>
<th>Elution Buffer</th>
</tr>
</thead>
<tbody>
<tr>
<td>8M Urea 1 % SDS</td>
<td>4M Urea 0.5 % SDS</td>
<td>2M Urea 0.2 % SDS</td>
<td>0.2 % SDS</td>
<td>0.2 % SDS</td>
</tr>
<tr>
<td>-</td>
<td>20 mM Imidazole</td>
<td>20 mM Imidazole</td>
<td>-</td>
<td>250 mM Imidazole</td>
</tr>
<tr>
<td>50 mM Tris-HCl, 200mM NaCl, pH 8.0</td>
<td>50 mM Tris-HCl, 200mM NaCl, pH 8.0</td>
<td>50 mM Tris-HCl, 200mM NaCl, pH 8.0</td>
<td>50 mM Tris-HCl, 200mM NaCl, pH 8.0</td>
<td>50 mM Tris-HCl, 200mM NaCl, pH 8.0</td>
</tr>
<tr>
<td>40 ml x 2</td>
<td>40 ml</td>
<td>40 ml</td>
<td>40 ml</td>
<td>40 ml</td>
</tr>
</tbody>
</table>

1. Allow the column to settle and save the flow-through in a 50 ml tube.
2. Rinse out the Ni-NTA column with Wash Buffer 1 and collect fractions. Check the UV absorbance at 280 nm every 30-40 drops until absorbance is below < 0.1.
3. Wash the Ni-NTA column with Wash Buffers 2, 3, and 4 and save all fractions. Take samples of all fractions for SDS-PAGE analysis.
4. Elute the protein with the elution buffer.
   a. Elution Tube #1 (15 mL) – Check the UV absorbance at 280 nm every 10 drops. Once the absorbance reaches > 0.2 at 280 nm, switch to Elution Tube #2. Use Elution buffer as a blank.
   b. Elution Tube #2 (15 mL) – Check the UV absorbance every 15-20 drops. Once the absorbance drops back below ~0.2, switch to tube #3.
   c. Elution tube #3 (15 mL) – Check the UV absorbance every 30-40 drops. Elute an addition 2-4 mL of “dead volume” to ensure all protein is off of the column.
5. Concentrate the elution using Amicon (MWCO: 10 kDa) at 4000 rpm at 4 °C for 30 min.
6. Add the concentrated protein solution to a dialysis bag (MWCO: 10 kDa) and dialyze the protein solution against 1 L of water for 5 days by changing the water two times per day.
7. You can see the white precipitated protein after 3 days of dialysis.
8. Take out the precipitated protein to 20 mL scintillation vial and lyophilize.