Supporting information for:

Deciphering the Glycosylation Code

Christopher R. Ellis and William G. Noid*

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802

E-mail: wnoid@chem.psu.edu

Abstract

This document provides supporting information for the manuscript "Deciphering the Glycosylation Code." In particular, this document details the simulation protocol, provides results for two additional simulations, describes glycosylation-induced structural correlations, discusses specific carbohydrate-peptide interactions, and further analyzes the sequons from the glycoprotein dataset.

*To whom correspondence should be addressed

Computational methods

Molecular dynamics simulations

Our simulations consider seven peptide systems with the general sequence: Ace1-Ile2-Thr3-Pro4-Asn5-Xxx6-Thr7-Yyy8-Ala9-NH_2. Table S1 presents the name and sequence of the simulated systems. The initial configurations for simulations of the Gly6-Trp8 peptide and corresponding N-linked glycopeptide were generated using the PRODRG server.[S1] The starting configurations for simulations of the Gly6-Ala8 and Ala6-Trp8 peptides were generated using Pymol.[S2] The initial configurations for simulations of the Gly6-Ala8, Ala6-Trp8, and Ala6-Ala8 glycopeptides were generated by either removing the indole sidechain of Trp8 and/or adding a methyl group to Gly6 of the Gly6-Trp8 glycopeptide, as appropriate.

All simulations were performed with the Gromacs 4.5.3 simulation suite.[S3,S4] The Gromacs stochastic dynamics algorithm with the v-rescale thermostat[S5] and Parrinello-Rahman barostat[S6] (with time constants of 0.5 ps and 20 ps, respectively,) were employed to sample the constant NPT ensemble in all production simulations. Peptide and carbohydrate interactions were modeled with the OPLS-AA force field[S7,S8] and the SPC/E water model was used.[S9] The Asn-carbohydrate linkage required an additional eighteen bonded parameters that were determined from chemically similar interactions. Tables S2, S3, and S4 list the additional parameters for the bond-stretching, angle-bending, and dihedral torsional potentials, respectively.[S10] The parameters were obtained from existing OPLS potentials for chemically similar bonds, angles, and dihedrals. Electrostatic interactions were calculated with the particle mesh Ewald method[S11] using a 0.08 nm Fourier grid spacing. Short-ranged van der Waals interactions and also the real space contribution to the electrostatic interactions were truncated at 1.2 nm. Bonds containing hydrogen atoms were constrained using the LINCS algorithm[S4] and a 2 fs integration time step was employed. Periodic boundary conditions[S12] were used in all simulations.

Each simulation system was initially prepared as follows: 1) solvation in a rhombic dodec-

ahedron with sides of length 4.75 nm; 2) steepest descent energy minimization; 3) simulated annealing from 0 to 298 K over 500 ps in the constant NVT ensemble; and 4) equilibration for 2 ns in the constant NPT ensemble at $T = 298$ K and $P = 1$ bar. The resulting configuration was replicated 59 times and each replica was annealed to a temperature between 298 and 495 K that was determined by the T-REMD server.[S13] Finally, replica-exchange molecular dynamics (REMD)[S14] simulations were performed for 110 ns with 60 replicas in the constant NPT ensemble. Exchanges were attempted every 1 ps[S15] with the corresponding acceptance exchange probability of 38-42% over the entire temperature range. The simulated results were determined from the configurations sampled every 1 ps by the low temperature replica over the final 100 ns of the simulations. The peptide stereochemistry and the glycosidic linkage were correctly preserved at all temperatures. In addition, both carbohydrates remained in stable $^4C_1$ conformations throughout the duration of all simulations.

Analysis of glycoprotein dataset

In order to corroborate our simulations of short glycopeptides, we analyzed the Structural Assesment of Glycosylation Sites (SAGS) database of glycoprotein structures.[S16–S18] This database identifies glycosylated sequons that are present within structures from the protein databank (PDB). We analyzed the structures adopted by the protein backbone in these glycosylated sequons. We eliminated from our analysis the 118 sequons for which the PDB file indicated that the proximal glycan (i.e., the glycan covalently linked to the Asn sidechain) had been modified by a fucose residue.[S19,S20] In addition, we also eliminated from our analysis all glycosylated sequons for which the PDB file lacked structural information for backbone atoms within 3 residues of the glycosylation site. We visually inspected the structures for all glycosylated sequons of the form Pro-Asn-Xxx-Thr/Ser and Xxx-Asn-Gly-Thr/Ser (where Xxx is any residue) and eliminated from our analysis one additional sequon for which the PDB file lacked the N-linked glycan. However, we did not visually inspect all of the remaining sequons. The calculations that are presented in the main text were performed with the

remaining 1524 glycosylated sequons. Further below, we provide additional analysis of the SAGS dataset to complement the results of the main text. However, as explicitly noted below, some of these supplemental figures include statistics for these fucosylated sequons.

Calculations of secondary structures[S21,S22] were performed with Gromacs utilities.[S3,S4] We identified β-turns in the SAGS database by first protonating the relevant structures and then identifying donor-acceptor pairs with a distance $r_{\mathrm{DA}} \leq 0.35$ nm and an acceptor-donor-proton angle $\theta_{\mathrm{ADH}} \leq 40^{\circ}$. Since these structures were not energy minimized and included possible inaccuracies in the atomic positions, we used slightly relaxed criterion for the donor-acceptor-proton angle, rather than the slightly more stringent criterion (i.e., $\theta_{\mathrm{ADH}} \leq 30^{\circ}$) used to analyze the simulations.[S23]

Results

Torsional preferences and structural correlations

In analyzing the conformational preferences of individual residues, we defined helical conformations for Pro4 by $-70^{\circ} \leq \psi \leq +70^{\circ}$. We defined helical conformations for other residues with the same criterion for ψ and distinguished between left- and right-handed conformations according to $\phi > 0^{\circ}$ and $\phi < 0^{\circ}$, respectively. We identified other regions of the Ramachandran map as extended.

Figure S1 analyzes the effect of glycosylation upon the torsional preferences of two additional peptide simulations. The four rows correspond to the residues Pro4, Asn5, Gly6/Ala6, and Thr7, respectively.

The left column presents Ramachandran plots sampled in the simulation of the non-glycosylated Gly6-Ala8 peptide. These Ramachandran maps of the Gly6-Ala8 peptide appear very similar to the corresponding maps for the Gly6-Trp8 peptide. (See column 1 of Figure 1 in the main text.) In the Gly6-Ala8 peptide, Pro4 and Asn5 are primarily extended, Gly6 rotates freely and samples left- ($\phi > 0^{\circ}$) and right- ($\phi < 0^{\circ}$) handed conformations with

similar probability, and Thr7 equally samples helical and extended configurations. In comparison to the Gly6-Trp8 peptide simulation, the largest difference appears to be that Thr7 samples slightly more extended conformations in the Gly6-Ala8 peptide simulation.

The right column presents Ramachandran plots sampled in the Ala6-Ala8 glycopeptide simulation. The core residues of the Ala6-Ala8 glycopeptide sample nearly identical Ramachandran maps to the corresponding residues of the Ala6-Trp8 glycopeptide. In the Ala6-Ala8 glycopeptide, Pro4 is primarily extended, Asn5 equally samples helical and extended configurations, Ala6 is primarily helical, and Thr7 adopts both helical and extended conformations.

Table S5 demonstrates that glycosylation of the Gly6 peptides introduces strong correlations between the conformation of Pro4 and the conformations sampled by residues 5-7. When Pro4 adopts helical conformations in simulations of the Gly6 glycopeptides, Asn5 samples helical conformations, Gly6 samples the basin at $\phi \approx 150°, \psi \approx \pm 180°$, and Thr7 is extended in more than 80% of configurations. However, prior to glycosylation, no such strong correlations are observed for the non-glycosylated Gly6-Trp8 and Gly6-Ala8 peptides. Moreover, Table S5 also demonstrates that glycosylation of the Ala6 peptide does not introduce such strong correlations. When Pro4 is helical in the Ala6 glycopeptide simulations, Asn5 and Thr7 adopt helical and extended configurations in roughly 60% and 55% of conformations, respectively.

Simulated structural motifs

The Imperiali group previously performed extensive biophysical studies that characterized the impact of glycosylation upon a short peptide system that differs from the simulated Gly6-Trp8 octapeptide only in the addition of an artificial ornithine residue before Ile2.[S24–S27] They demonstrated that glycosylation triggered a "conformational switch" from Asx-turns to β-turns.[S24,S27] Asx-turns are defined by a hydrogen bond[S23] between the sidechain carbonyl of Asn5 and the backbone amide of Thr7, while β-turns are defined by a hydrogen bond

between the backbone carbonyl of Thr3 and the backbone amide of Gly6. Table S6 analyzes the population of these structural motifs in our REMD simulations. Although none of the systems form a single stable structure, the changes in sampling are quite consistent with the prior experimental observations.

Table S6 indicates that, prior to glycosylation, Asx-turns are one of the most stable structural motifs for all of the nonglycosylated peptides. However, glycosylation destabilizes Asx-turns for each system by a factor of 2-4. Interestingly, the Gly6 $\rightarrow$ Ala6 mutation stabilizes Asx-turns by a factor of 2-4 for each simulated system. Consequently, Asx-turns remain among the most stable conformations sampled by the Ala6 glycopeptides.

Table S6 also indicates that β-turns are only rarely sampled (3-4%) in the REMD simulations of the nonglycosylated octapeptides. Glycosylation stabilizes β-turns for all peptide systems. However, while β-turns remain a minor conformer ($\approx$10%) for the Ala6 glycopeptides, β-turns become the major conformer ($\approx$30%) sampled by the Gly6 glycopeptides. Surprisingly, despite lacking an aromatic group to stack with the glycan, β-turns are even more stable for the Gly6-Ala8 glycopeptide than for the Gly6-Trp8 glycopeptide.

Simulated free energy surfaces

Figure S2 presents the free energy surfaces (FES's) sampled by the Gly6-Ala8 peptide and the Ala6-Ala8 glycopeptide as a function of θ and φ. The hinging angle, θ, characterizes the compaction of the peptide and is defined by the α carbons of Ile2, Asn5, and Trp8/Ala8. The twisting angle, φ, characterizes the peptide twist and is defined by the (negative of the) pseudodihedral angle that is formed by the α carbons of Ile2, Pro4, Gly6/Ala6, and Trp8/Ala8.

The left panel of Figure S2 demonstrates that the Gly6-Ala8 peptide FES is very similar to the Gly6-Trp8 peptide FES. Both Gly6 peptides sample a highly disordered ensemble of fairly extended conformations with a slightly negative twist on average. Additionally, the two most stable regions of the Gly6-Ala8 peptide correspond to Asx-turns.

The right panel of Figure S2 demonstrates that the Ala6-Ala8 glycopeptide FES is similar to the Ala6-Trp8 glycopeptide FES. However, removal of the indole ring stabilizes φ_- β-turns. As discussed in the following section, φ_- β-turns are not stabilized by aromatic-carbohydrate stacking interactions and removal of the indole ring actually stabilizes these conformations.

The main text identifies two classes of β-turns. These β-turns all contain a hydrogen bond[S23] between Thr3 and Res6. We identify β-turns with $\varphi > 0°$ as φ_+ β-turns, while we identify β-turns with $\varphi < 0°$ as φ_- β-turns.

Interactions stabilizing β-turns

Our previous simulation study[S10,S28] suggested that aromatic-glycan stacking interactions contributed to stabilizing β-turns for the Gly6-Trp8 glycopeptide. Strikingly, the Trp8 $\rightarrow$ Ala8 mutation that eliminates aromatic stacking interactions actually stabilizes β-turns in these glycopeptide simulations. Moreover, the effects of glycosylation appear to be significantly mitigated, even in the presence of the Trp8 indole sidechain, when Gly6 is mutated to Ala6. Thus, aromatic-glycan stacking interactions are neither necessary nor sufficient for significantly stabilizing β-turns in the glycopeptide simulations. Accordingly, we next investigated the interactions that stabilize β-turns in our REMD simulations.

Aromatic-Glycan stacking interactions

Figure S3 characterizes the direct interactions between the disaccharide and Trp8/Ala8 (position +3 in the glycosylation sequence). The left and right columns describe interactions with the proximal and distal glycan, respectively. Panels A and B present simulated configurations that demonstrate these direct interactions. In panels C-F, the black and red curves present distributions sampled by the REMD simulation of the Gly6-Trp8 and Gly6-Ala8 glycopeptides, respectively. The solid black and solid red curves correspond to the native state (NS) configurations, i.e., primarily φ_+ β-turns, in the eye of the corresponding FES, which is indicated by the red box in Figure 2B of the main text and corresponds to $30° < \theta < 70°$

and $5^\circ < \varphi < 70^\circ$. The dashed black and dashed red curves correspond to the non-native configurations, i.e., the remaining configurations, sampled by the two Gly6 glycopeptides.

Panels C and D of Figure S3 present the simulated distributions for the distance, R_{COM}, from the center of the proximal and distal glycans, respectively, to either the center of the Trp8 indole ring or the Ala8 β-carbon. The peaks at $R_{COM} \leq 0.5$ nm correspond to direct contact between the two groups. Panels E and F of Figure S3 indicate the simulated distributions of hydrophobic solvent accessible surface area (SASA) for the proximal and distal glycans, respectively.

The solid black curves in Figure S3 indicate that, when the Gly6-Trp8 glycopeptide samples φ_+ β-turns, Trp8 usually directly contacts the distal glycan and occasionally contacts the proximal glycan. In the remainder of the Gly6-Trp8 glycopeptide ensemble, Trp8 interacts only weakly with the disaccharide. In particular, the φ_- β-turns do not form stable stacking interactions. Consequently, relative to competing non-native conformations, the native φ_+ β-turns significantly reduce the hydrophobic SASA for both the proximal and distal glycans. Therefore, the direct stacking interaction and the commensurate burial of the glycan hydrophobic surface appear to contribute to stabilizing φ_+ β-turns for the Gly6-Trp8 glycopeptide.

The solid red curves in Figure S3 indicate that, when the Gly6-Ala8 glycopeptide samples native φ_+ β-turns, the Ala8 β-carbon also directly contacts the proximal or distal glycan. Somewhat surprisingly, in φ_+ β-turns, the Ala8 β-carbon appears to contact the disaccharide even more frequently than the Trp8 indole ring. (It should be noted, though, that the β-carbon of the Trp8 sidechain may form similar interactions with the disaccharide. Therefore, it is not necessarily the case that the disaccharide interacts more favorably with the Ala8 sidechain than with the Trp8 sidechain.) The dashed red curves indicate that, in the remainder of the Gly6-Ala8 glycopeptide ensemble, the Ala8 sidechain does not interact with either glycan. As expected, in comparison to the Gly6-Trp8 glycopeptide glycan, the Gly6-Ala8 glycopeptide glycan is more solvent exposed in both native and non-native states. However,

in native φ_+ β-turns, the direct interaction between the disaccharide and the Ala8 sidechain does significantly reduce the disaccharide hydrophobic SASA. Consequently, despite lacking an aromatic sidechain, a "stacking" interaction between the glycan and the Ala8 sidechain may also contribute to stabilize φ_+ β-turns via hydrophobic solvation forces.

The thin blue and green lines present simulated distributions sampled by the Ala6-Trp8 and Ala6-Ala8 glycopeptides, respectively. In contrast to the Gly6 glycopeptides, the Ala6-Trp8 and Ala6-Ala8 glycopeptides did not adopt a metastable native state. Panels C and D demonstrate that the Trp8 sidechain interacts occasionally with the proximal glycan in the Ala6-Trp8 glycopeptide simulation, while the Ala8 sidechain does not interact with either glycan in the Ala6-Ala8 glycopeptide simulation. Moreover, Panels E and F demonstrate that the disaccharides experience similar solvent exposure in the Ala6 glycopeptide simulations and in the non-native ensembles of the corresponding Gly6 glycopeptides. Thus, in terms of direct interactions between Res8 and the disaccharide, the Ala6 glycopeptide ensembles appear quite similar to the non-native ensembles sampled by the corresponding Gly6 glycopeptides.

In the Ala6-Trp8 glycopeptide simulations, the Trp8 indole ring does demonstrate modest stacking interactions with the proximal glycan. Specifically, the two regions centered at $\theta \approx 150°, \varphi \approx -90°$, and $\theta \approx 115°, \varphi \approx +100°$ on the Ala6-Trp8 glycopeptide FES exhibit significant carbohydrate-aromatic stacking. As a consequence, in the Ala6-Trp8 glycopeptide simulation, the burial of the disaccharide is quite similar to that observed in non-native conformations sampled by the Gly6-Trp8 glycopeptide. In contrast, in the Ala6-Ala8 glycopeptide simulation, the Ala8 sidechain never interacts with either glycan. Consequently, the disaccharide experiences greatest solvent exposure in the Ala6-Ala8 glycopeptide simulations.

Hydrogen bonding

Table S7 demonstrates that peptide-glycan hydrogen bonds also contribute to stabilize φ_+ β-turns. In simulations of the Gly6-Trp8 glycopeptide, the disaccharide forms hydrogen bonds with peptide residues 6-8 in 65% of φ_+ β-turn conformations. The majority of these hydrogen bonds are formed between the proximal glycan and the backbone of Thr7. Table S7 indicates that these hydrogen bonds are significantly more stable for the Gly6-Ala8 glycopeptide. This increase in hydrogen bonding presumably compensates for the loss of aromatic-glycan stacking interactions after the Trp8→Ala8 mutation. However, these hydrogen bonds appear to be inaccessible to the Ala6 glycopeptides. Table S7 also demonstrates that φ_- β-turns actually form fewer glycan-peptide hydrogen bonds than other non-native conformations. Moreover, the hydrogen bonds that do form in φ_- β-turns primarily involve Thr3, which is on the other side of the glycosylation site.

Steric interactions

The preceding calculations demonstrated that aromatic-glycan stacking interactions are neither necessary nor sufficient to significantly stabilize β-turns for short glycopeptides. In contrast, the mutation of Gly6 to Ala6 much more profoundly impacts the glycopeptide ensemble. Figure S4 characterizes the steric effects exerted by the Ala6 sidechain upon the conformations sampled by the simulated glycopeptides.

Panel A of Figure S4 presents a representative φ_+ β-turn configuration sampled in the Gly6-Trp8 glycopeptide simulation. The red and gray spheres indicate the van der Waals radii for the Asn5 carbonyl oxygen and the Gly6 α proton, respectively. The highlighted α proton corresponds to the location of the Ala6 β-carbon in the Ala6 peptides. Clearly, the two atoms are eclipsed such that the Ala6 β-carbon would significantly destabilize this configuration.

Panels B and C of Figure S4 present the simulated distributions for the distance, r, between the Asn5 carbonyl oxygen and either the Gly6 α proton or the Ala6 β carbon for the

Gly6-Trp8 (black), Gly6-Ala8 (cyan), and Ala6-Trp8 (red) glycopeptide. Panels B and C of Figure S4 present the distributions sampled in the entire simulation and in β-turn configurations, respectively. Both Gly6 glycopeptide simulations sample bimodal distributions with subpopulations that correspond to r $<$ 0.3 nm and r $>$ 0.3 nm. Panels D and E of Figure S4 present a scatter plot on the (θ,φ) FES and on the Gly6 Ramachandran map, respectively, for these two subpopulations of β-turns from the Gly6-Trp8 glycopeptide simulation. In these scatter plots, black and red points indicate subpopulations of β-turn conformations with r $<$ 0.3 nm and r $>$ 0.3 nm, respectively. Clearly, the r $<$ 0.3 nm subpopulation corresponds to φ_+ β-turns with Gly6 $\phi > 0°$, while the r $>$ 0.3 nm subpopulation corresponds to φ_- β-turns with Gly6 $\phi < 0°$. In contrast, the Ala6-Trp8 glycopeptide simulation only samples configurations with r $>$ 0.3 nm, which explains why the Ala6 glycopeptides do not sample φ_+ β-turns.

Impact of fucosylation

As noted above, the proximal glycan is fucosylated in 118 of the 1642 glycosylated sequons that we analyzed from the SAGS dataset. Figure S5 characterizes the conformation of the protein backbone in these 118 glycosylated sequons. The top figure presents a scatter plot of these conformations as a function of the backbone bend θ and twist φ, in analogy to Figure 3 of the main text. Figure S5 suggests that the attachment of a single fucose residue does not dramatically alter the conformational distribution sampled by glycosylated sequons. In particular, the structure ensemble of Figure S5 appears quite similar to that observed in Figure 3 of the main text for glycosylated sequons that are not fucosylated. However, the three cases in which two fucose residues are linked to the proximal glycan do appear quite distinct. In these three cases, the protein backbone samples highly extended conformations.

Comparison of Thr and Ser in position +2

The main text presents simulations for a series of glyco/peptides with the central sequon Pro-Asn-Gly/Ala-Thr-Trp/Ala. These simulations are compared with the conformations adopted by glycosylated sequons from the SAGS database. Our bioinformatic analysis primarily considered sequons of the form Pro/Zzz-Asn-Gly/Yyy-Thr/Ser-Xxx, where Zzz indicates any residue in position -1, Yyy indicates any residue other than Gly in position +1, and Xxx indicates any residue. However, the bioinformatic analysis of the main text did not distinguish between sequons with Thr or Ser in position +2. Figure S6 compares the conformations adopted by glycosylated sequons with either Thr or Ser in this position. The top 4 panels present scatter plots of the conformations adopted by different classes of glycosylated sequons in the SAGS dataset as a function of the bending θ and twisting φ angles, in analogy to Figure 3 of the main text. In each panel, red X's and green X's indicate conformations of sequons with Thr or Ser, respectively, at position +2 within a given glycosylated context, while the legends indicate the frequency of each sequon within the SAGS database. The black dots in panels A and C indicate conformations sampled from the Gly6-Trp8 glycopeptide simulation, while the black dots in panels B and D indicate conformations sampled from the Ala6-Trp8 glycopeptide simulation.

The table below the figure provides statistical analysis for each scatter plot. In particular, this table indicates the number of instances of each sequon within the SAGS dataset and characterizes the fraction of theses sequons that are compact ($\theta < 65°$), extended ($\theta > 135°$), have positive twist ($\varphi > 0°$), and form β-turns (irrespective of whether they are φ_+ or φ_- β-turns). We note that this figure includes statistics from fucosylated sequons. According to Figure S5, we expect that this should have minimal impact upon the observed trends.

In each case there are roughly twice as many sequons with Thr in position +2, as previously observed in prior analyses of the SAGS database.[S16,S18] Nevertheless, the figure suggests that, for each class of sequon, the difference between Thr or Ser at residue +2 has relatively modest impact upon the conformation adopted by the glycosylated sequon. Most

importantly, irrespective of the residue at position +2, the presence of Gly at position +1 significantly increases the sampling of φ_+ β-turns. Moreover, the combination of Pro at position -1 and Gly at position +1 further increases this tendency.

Otherwise, we observe relatively modest differences between glycosylated sequons with Thr or Ser at position +2. The figure suggests that glycosylated sequons with Thr at position +2 sample a slightly broader distribution of conformations than glycosylated sequons with Ser at this position. Panel B indicates that, in the case that Pro is in position -1 and Gly is not in position +1, 2 of the 38 sequons with Thr at position +2 sample conformations near the Gly6-Trp8 native state, while 0 of the 13 sequons with Ser at this position do so. However, this may simply reflect the very limited statistics that are available for the Pro-Asn-Yyy-Ser sequon rather than any intrinsic biophysical difference. The table indicates that glycosylated sequons with Ser at position +2 demonstrate a slightly greater tendency for sampling β-turn structures in the case that Pro is in position -1, but not otherwise.

Nevertheless, these differences appear to be of secondary significance relative to the trends discussed in the main text. Consequently, in order to analyze a larger and more diverse dataset, we did not distinguish between Thr or Ser in position +2 for our bioinformatic analysis in the main text.

Analysis of Pro-Asn-Xxx sequons

Our simulations demonstrated that the Gly6 peptides (corresponding to Pro-Asn-Gly-Thr sequons with Gly at position +1) readily adopted φ_+ β-turns upon glycosylation, while Ala6 peptides (corresponding to sequons with Ala at position +1) did not sample this conformation. Figure S4 explicitly demonstrates that steric clashes between the Asn backbone carbonyl oxygen and the Ala sidechain β carbon preclude the Ala6 peptides from sampling φ_+ β-turns. Since any amino acid other than Gly has a sidechain β carbon, we hypothesized that steric clashes with the β carbon at position +1 would generally destabilize φ_+ β-turns for Pro-Asn-Yyy-Thr/Ser sequons whenever Yyy is not Gly. The green X's in Figure 3B of

the main text support this hypothesis with structural analysis of the SAGS database.

Figure S7 more thoroughly analyzes the conformations adopted by glycosylated Pro-Asn-Xxx-Thr/Ser sequons as a function of the identity of the Xxx residue at position +1. The top two panels indicate, as a scatter plot of the backbone bend θ and twist φ, the conformations sampled for the eight most common residues at position +1 and also for Cys. The table below the figure provides a statistical analysis of the structures adopted by all instances of residues that appear at least four times in this sequon. This analysis includes fucosylated structures, since Figure S5 suggests that these structures should not alter the basic trends. This analysis also does not discriminate between Thr or Ser at position +2, since Figure S6 suggests that the difference between these residues is not critical for our analysis.

Gly is the most common residue at position +1 in Pro-Asn-Xxx-Thr/Ser sequons. Gly accounts for more than 20% of all cases and is almost twice as common as Lys, which is the next most common residue in this position. Due to the very limited statistics, it is difficult to make strong statements regarding tendencies adopted by the remaining amino acids in this sequon. Aromatic and negatively charged residues are almost entirely excluded from this position. Interestingly, Lys is eight times more common than Arg, while Asn is present four times but Gln is excluded. In both of these cases, the more compact sidechain is significantly more common. In contrast, the larger Thr residue is three times more common in this position than the smaller Ser residue. However, these trends may simply be due to differences in the efficiency of glycosylation rather than any biophysical differences associated with glycan-protein interactions.

Clearly, Gly has the greatest tendency for adopting compact conformations with positive twist. While glycosylated Pro-Asn-Gly-Thr/Ser sequons readily adopt φ_+ β-turns, the SAGS database contains only three glycosylated Pro-Asn-Yyy-Thr/Ser sequons that adopt φ_+ β-turn conformations with $\varphi > 0^\circ$ with Yyy being different from Gly. In one of these three sequons, Lys is in the +1 position and φ is only very slightly positive. Very interestingly, the other two sequons that form φ_+ β-turns both have Cys in the +1 position and both

adopt conformations that map to the Gly6 glycopeptide native state. Moreover, these are the only two instances of glycosylated Pro-Asn-Cys-Thr/Ser sequons and, in both cases, the Cys residue forms a disulfide bond. In contrast, Leu, Lys, and Ile demonstrate the greatest tendency for forming φ_- β turns. Leu and Ile only sample conformations with negative twist. Ala and Ile demonstrate an interesting tendency for sampling both compact and extended conformations. Thr and Val also show a tendency for sampling extended conformations. Nevertheless, we emphasize that these observations are all based upon a very small dataset. Accordingly, we adopted a simpler treatment of grouping amino acids based upon the presence or absence of a sidechain in the main text.

References

(S1) Schuttelkopf, A.; van Aalten, D. *Acta Crystallogr.* **2004**, *D60*, 1355–1363.

(S2) Schrödinger, LLC, *The PyMOL Molecular Graphics System, Version 1.5*; 2010.

(S3) Van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A.; Berendsen, H. *J. Comp. Chem.* **2005**, *26*, 1701–1718.

(S4) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theor. Comp.* **2008**, *4*, 435–447.

(S5) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*.

(S6) Parrinello, M.; Rahman, A. *J. Chem. Phys.* **1982**, *76*, 2662–2666.

(S7) Jorgensen, W.; Maxwell, D.; TiradoRives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(S8) Kony, D.; Damm, W.; Stoll, S.; van Gunsteren, W. *J. Comp. Chem.* **2002**, *23*, 1416–1429.

(S9) Berendsen, H.; Grigera, J.; Straatsma, T. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

(S10) Ellis, C. R.; Maiti, B.; Noid, W. G. *J. Am. Chem. Soc.* **2012**, *134*, 8184–8193.

(S11) York, D.; Darden, T.; Pedersen, L. *J. Chem. Phys.* **1993**, *99*, 8345–8348.

(S12) Allen, M. P.; Tildesley, D. P. *Computer Simulation of Liquids*; Oxford Press: New York, NY USA, 1987.

(S13) Patriksson, A.; van der Spoel, D. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2073–2077.

(S14) Sugita, Y.; Okamoto, Y. *Chem. Phys. Let.* **1999**, *314*, 141–151.

(S15) Sindhikara, D. J.; Emerson, D. J.; Roitberg, A. E. *J. Chem. Theor. Comp.* **2010**, *6*, 2804–2808.

(S16) Petrescu, A.; Milac, A.; Petrescu, S.; Dwek, R.; Wormald, M. *Glycobiology* **2004**, *14*, 103–114.

(S17) Petrescu, A.-J.; Wormald, M. R.; Dwek, R. A. *Curr. Opin. Struc. Biol.* **2006**, *16*, 600–607.

(S18) Petrescu, S., Ed. *Glycosylation*; InTech: Rijeka, Croatia, 2012.

(S19) Becker, D. J.; Lowe, J. B. *Glycobiology* **2003**, *13*, 41R–53R.

(S20) Ma, B.; Simala-Grant, J. L.; Taylor, D. E. *Glycobiology* **2006**, *16*, 158R–184R.

(S21) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–2637.

(S22) Joosten, R. P.; Beek, T. A. H. T.; Krieger, E.; Hekkelman, M. L.; Hooft, R. W. W.; Schneider, R.; Sander, C.; Vriend, G. *Nuc. Acids Res.* **2011**, *39*, D411–D419.

(S23) Luzar, A.; Chandler, D. *J. Chem. Phys.* **1993**, *98*, 8160–8173.

(S24) O'Connor, S.; Imperiali, B. *J. Am. Chem. Soc.* **1997**, *119*, 2295–2296.

(S25) Imperiali, B.; Rickert, K. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 97–101.

(S26) O'Connor, S.; Imperiali, B. *Chem. Biol.* **1998**, *5*, 427–437.

(S27) Bosques, C.; Tschampel, S.; Woods, R.; Imperiali, B. *J. Am. Chem. Soc.* **2004**, *126*, 8421–8425.

(S28) Ellis, C. R.; Maiti, B.; Noid, W. G. *J. Am. Chem. Soc.* **2014**, *136*, 8485–8485.

Tables

Table S1: Simulated systems.

Name	Sequence
Gly6-Trp8-peptide	Ace1-Ile2-Thr3-Pro4-Asn5-Gly6-Thr7-Trp8-Ala9-NH_2
Gly6-Ala8-peptide	Ace1-Ile2-Thr3-Pro4-Asn5-Gly6-Thr7-**Ala8**-Ala9-NH_2
Gly6-Trp8-glyco.	Ace1-Ile2-Thr3-Pro4-**Asn5(GlcNAc)**$_2$-Gly6-Thr7-Trp8-Ala9-NH_2
Gly6-Ala8-glyco.	Ace1-Ile2-Thr3-Pro4-**Asn5(GlcNAc)**$_2$-Gly6-Thr7-**Ala8**-Ala9-NH_2
Ala6-Trp8-peptide	Ace1-Ile2-Thr3-Pro4-Asn5-**Ala6**-Thr7-Trp8-Ala9-NH_2
Ala6-Trp8-glyco.	Ace1-Ile2-Thr3-Pro4-**Asn5(GlcNAc)**$_2$-**Ala6**-Thr7-Trp8-Ala9-NH_2
Ala6-Ala8-glyco.	Ace1-Ile2-Thr3-Pro4-**Asn5(GlcNAc)**$_2$-**Ala6**-Thr7-**Ala8**-Ala9-NH_2

Table S2: Bond Stretching Constants.

New Type	OPLS Type	r_{eq}(nm)	k_b(kJ/mol)
N-CO	N-CT2	0.14490	282001.6

Table S3: Angle Bending Constants.

New Type	OPLS Type	θ_{eq}(degrees)	k_θ(kJ/mol)
C-N-CO	C-N-CT2	121.9	418.400
H-N-CO	H-N-CT2	118.4	317.984
N-CO-HC	N-CT2-HC	109.5	292.880
N-CO-CT	N-CT2-CT	109.7	669.440
N-CT-CO	N-CT-CT	109.7	669.440
N-CO-OS	CT-CT-OS	109.5	418.400

Table S4: Dihedral Angle Constants: The dihedral bonded interactions use OPLS Fourier dihedral types that have been translated to Gromacs Ryckaert-Bellemans form.

New Type	OPLS Type	C1	C2	C3	C4
CT-C-N-CO	CT-C-N-CT	30.28798	-4.81160	-25.47638	0.0
CO-N-C-O	CT2-N-C-O	25.47638	0.0	-25.47638	0.0
C-N-CO-CT	C-N-CT2-CT	15.70255	31.75656	-3.66936	-43.78975
C-N-CO-OS	C-N-CT-OS	-3.13800	-3.13800	6.27600	0.0
H-N-CO-OS	HC-CT-C-OS	0.27615	0.82844	0.0	-1.10458
N-CO-CT-CT	N-CT2-CT-CT	-0.76567	2.70705	4.02501	-5.96639
HC-CO-CT-HC	HC-CT-CT-OH	0.62760	1.88280	0.0	-2.51040
HC-CO-CT-N	HC-CT-CT-N	0.97069	2.91206	0.0	-3.88275
N-CT-CT-OH	N-CT2-CT-OH	9.89307	-4.71746	3.67774	-8.85335
CO-CT-N-C	CT2-CT-C-N	-9.49768	-6.36386	8.89936	6.96218
N-CO-OS-CT	CO-OS-CT-CT	1.71544	2.84512	1.04600	5.60656

Table S5: Conditional probabilities of configurations with α-helical configurations of Pro4. This table quantifies the conditional probability that Asn5, Res6, and Thr7 are located in the α-helical region, the forbidden region ($\phi \approx 150°, \psi \approx \pm 180°$), and β-sheet region of Ramachandran space, respectively, given that Pro4 is α-helical.

	P(Pro4$_\alpha$)	P(Asn5$_\alpha$|Pro4$_\alpha$)	P(Res6$_{\phi\approx150°,\psi\approx\pm180°}$|Pro4$_\alpha$)	P(Thr7$_\beta$|Pro4$_\alpha$)
Gly6-Trp8 peptide	10.3	45.6	38.7	31.0
Gly6-Ala8 peptide	10.2	46.6	33.7	53.4
Gly6-Trp8 glyco.	47.5	87.4	82.1	80.9
Gly6-Ala8 glyco.	55.5	90.3	82.6	83.5
Ala6-Trp8 peptide	9.8	53.7	0.0	29.6
Ala6-Trp8 glyco.	21.5	54.9	0.0	52.4
Ala6-Ala8 glyco.	25.1	67.1	0.0	56.9

Table S6: Secondary structure populations. This table quantifies the percent occupancy of Asx-turns, β-turns, and various regions of the FES.

	Asx-Turns	β-turns (% $\varphi+$)	$\theta < 65$	$\theta > 135$	$\varphi > 0$
Gly6-Trp8 peptide	12.0	2.9 (46.6)	14.2	26.7	44.1
Gly6-Ala8 peptide	9.3	4.2 (38.7)	15.7	24.3	46.3
Gly6-Trp8 glyco.	3.7	27.1 (88.7)	38.9	14.3	64.7
Gly6-Ala8 glyco.	4.5	30.8 (86.7)	46.2	12.6	66.9
Ala6-Trp8 peptide	29.9	3.5 (2.9)	7.3	25.1	30.2
Ala6-Trp8 glyco.	12.2	8.6 (2.1)	7.5	26.7	40.3
Ala6-Ala8 glyco.	15.6	13.1 (6.7)	10.5	24.1	37.6

Table S7: Carbohydrate-peptide hydrogen bonding. This table quantifies the probability for forming residue-carbohydrate hydrogen bonds within $\varphi+$ β-turns, $\varphi-$ β-turns, and other remaining configurations.

	Gly6-Trp8 glyco.			Gly6-Ala8 glyco.			Ala6-Trp8 glyco.		
	$\varphi+$	$\varphi-$	Other	$\varphi+$	$\varphi-$	Other	$\varphi+$	$\varphi-$	Other
Thr3-Proximal	4.1	12.4	10.1	3.0	12.0	9.3	8.3	12.1	12.9
Gly6-Proximal	14.6	3.1	7.9	8.2	2.0	4.4	-	-	1.8
Thr7-Proximal	47.1	-	16.5	78.0	-	30.0	-	-	6.9
Trp8-Proximal	4.9	1.6	5.9	4.2	-	5.6	6.7	2.3	7.5
Ala9-Proximal	2.5	-	3.8	5.7	1.5	4.8	-	-	2.6
Thr3-Distal	-	-	-	-	-	-	-	2.7	-
Trp8-Distal	7.7	-	4.9	1.7	-	2.4	-	-	2.1
Ala9-Distal	9.0	-	5.4	16.8	1.4	8.1	-	-	2.1

Figure Captions

Figure S1 Ramachandran plots for the core residues of the Gly6-Ala8 peptide (left) and Ala6-Ala8 glycopeptide (right) REMD simulations.

Figure S2 Global free energy surfaces. The left and right panels present the simulated FES for the Gly6-Ala8 peptide and Ala6-Ala8 glycopeptide, respectively, as a function of θ and φ.

Figure S3 Carbohydrate stacking interactions. The left and right columns characterize the stacking interactions of the Res8 sidechain with the proximal and distal glycans, respectively. The solid black and solid red curves present distributions sampled by the native state of the Gly6-Trp8 and Gly6-Ala8 glycopeptides, while dashed black and dashed red curves present distributions sampled from the nonnative configurations. The thin blue and green curves present distributions from the entire Ala6-Trp8 and Ala6-Ala8 glycopeptide simulations. Panels C and D present the simulated distributions for the distance, R_{COM}, from the center of the proximal and distal glycans, respectively, to either the center of the Trp8 indole ring or the β-carbon of Ala8. Panels E and F present the simulated distributions of hydrophobic solvent accessible surface area for the proximal and distal glycans, respectively.

Figure S4 Sidechain steric interactions and their role in destabilizing φ_+ β-turns. Panel A presents a representative φ_+ β-turn configuration. The red and gray spheres indicate the van der Waal radii for the Asn5 carbonyl oxygen and the Gly6 alpha proton, respectively. Panels B and C present the simulated distribution for the distance, r, between the Asn5 carbonyl oxygen and either the Gly6 α proton or the Ala6 β-carbon sampled in the entire simulation and in β-turn configurations, respectively, for the Gly6-Trp8 (black), Gly6-Ala8 (cyan), and Ala6-Trp8 (red) glycopeptide simulations. Panels D and E present a scatter plot of the Gly6-Trp8 β-turn configurations on the FES and Gly6 Ramachandran map, respectively. The configurations are separated into two

subpopulations with r $<$ 0.3 nm (black) and r $>$ 0.3 nm (red).

Figure S5 Analysis of conformations adopted by glycosylated sequons in which the proximal glycan has been fucosylated. The black X's and red X's indicate cases that the fucose residue is linked via the O6 or the O3 oxygen, respectively, of the proximal glycan. The green X's indicate cases that fucose residues were linked via both the O6 and O3 oxygens. The X's identified by blue circles correspond to cases in which the protein backbone forms a β turn at the glycosylation site. The bottom table provides statistical analysis of these structures.

Figure S6 Comparison of structures adopted by glycosylated sequons with Thr or Ser at position +2. Panels A-D present scatter plots that describe the sequon conformation as a function of the backbone bend θ and twist φ in analogy to Figure 3 of the main text. Panel A considers Pro-Asn-Gly sequons; Panel B considers Pro-Asn-Yyy sequons; Panel C considers Zzz-Asn-Gly sequons; and Panel D considers Zzz-Asn-Xxx sequons. Zzz indicates any residue at position -1, Xxx indicates any residue at position +1, and Yyy indicates any residue other than Gly at position +1. In each panel, the red and green X's indicate sequons with Thr or Ser, respectively, at position +2. The black X's indicate simulated conformations for the Gly6-Trp8 glycopeptide (panels A and C) or the Ala6-Trp8 glycopeptide (panels B and D). In each case, the legend indicates the frequency of each sequon in the SAGS database. The table statistically analyzes the conformation distribution in each panel.

Figure S7 Frequency and conformations of glycosylated Pro-Asn-Xxx-Thr/Ser sequons.

Figures

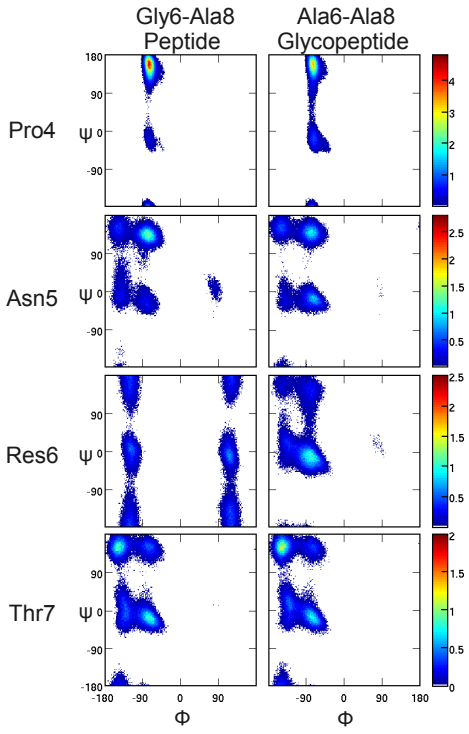


Figure S1: Ramachandran plots for the core residues of the Gly6-Ala8 peptide (left) and Ala6-Ala8 glycopeptide (right) REMD simulations.

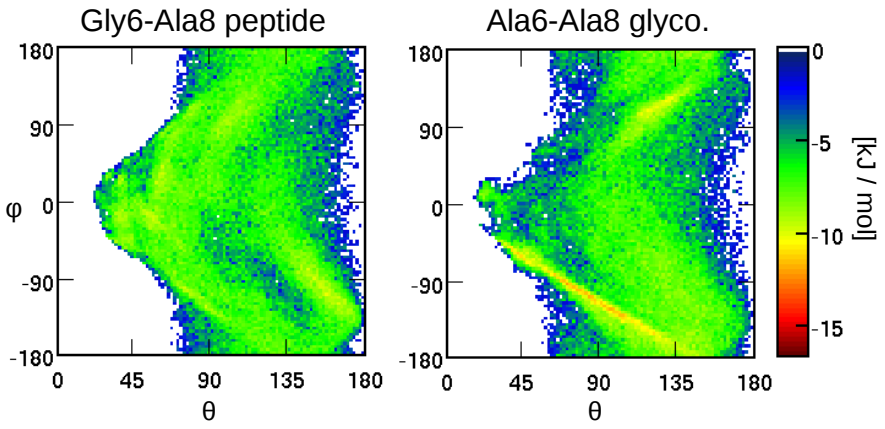


Figure S2: Global free energy surfaces. The left and right panels present the simulated FES for the Gly6-Ala8 peptide and Ala6-Ala8 glycopeptide, respectively, as a function of θ and φ.

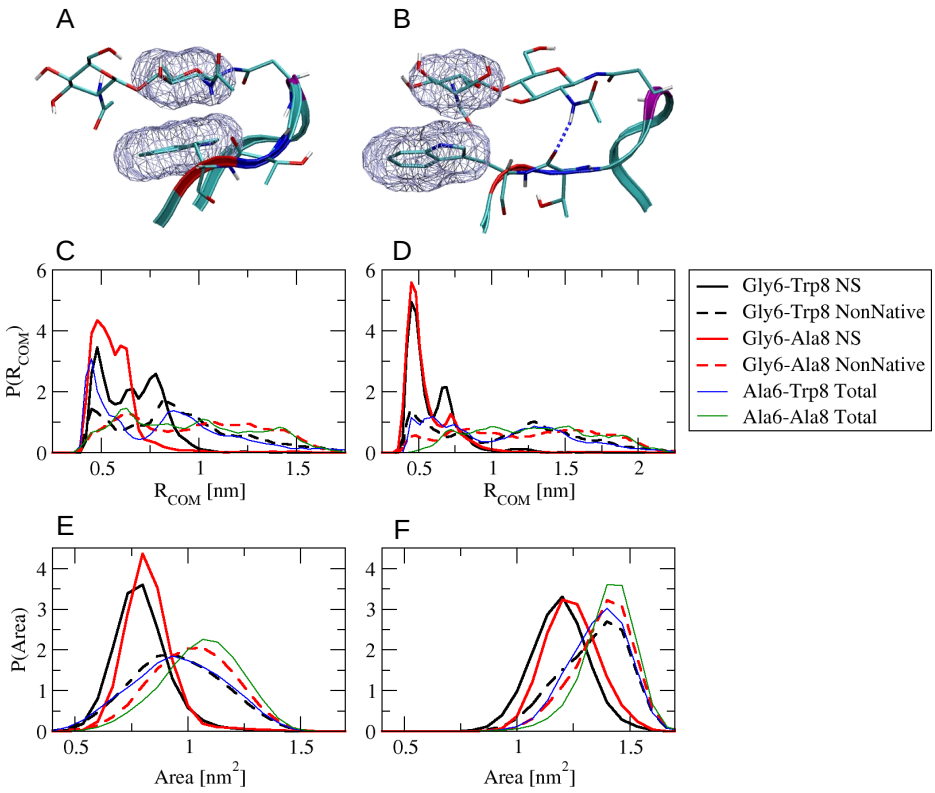


Figure S3: Carbohydrate stacking interactions. The left and right columns characterize the stacking interactions of the Res8 sidechain with the proximal and distal glycans, respectively. The solid black and solid red curves present distributions sampled by the native state of the Gly6-Trp8 and Gly6-Ala8 glycopeptides, while dashed black and dashed red curves present distributions sampled from the nonnative configurations. The thin blue and green curves present distributions from the entire Ala6-Trp8 and Ala6-Ala8 glycopeptide simulations. Panels C and D present the simulated distributions for the distance, R_{COM}, from the center of the proximal and distal glycans, respectively, to either the center of the Trp8 indole ring or the β-carbon of Ala8. Panels E and F present the simulated distributions of hydrophobic solvent accessible surface area for the proximal and distal glycans, respectively.

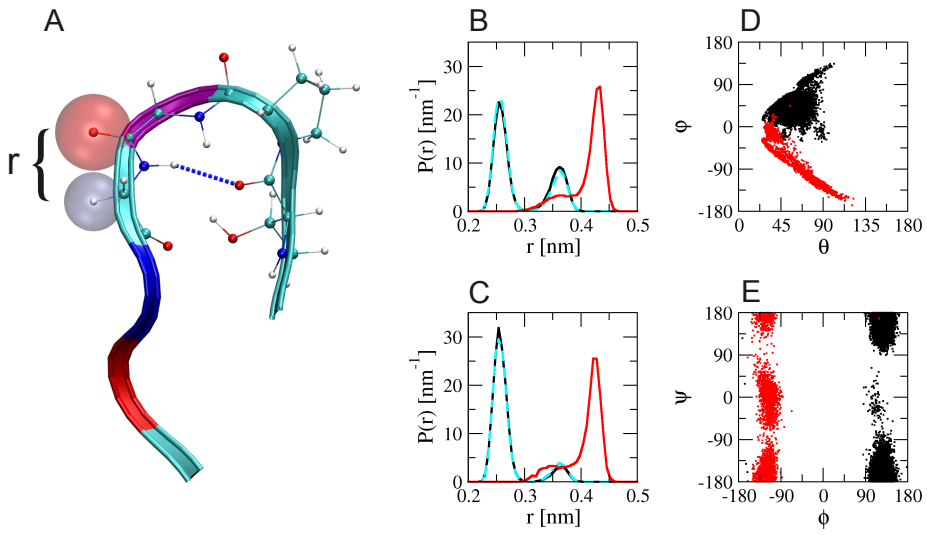
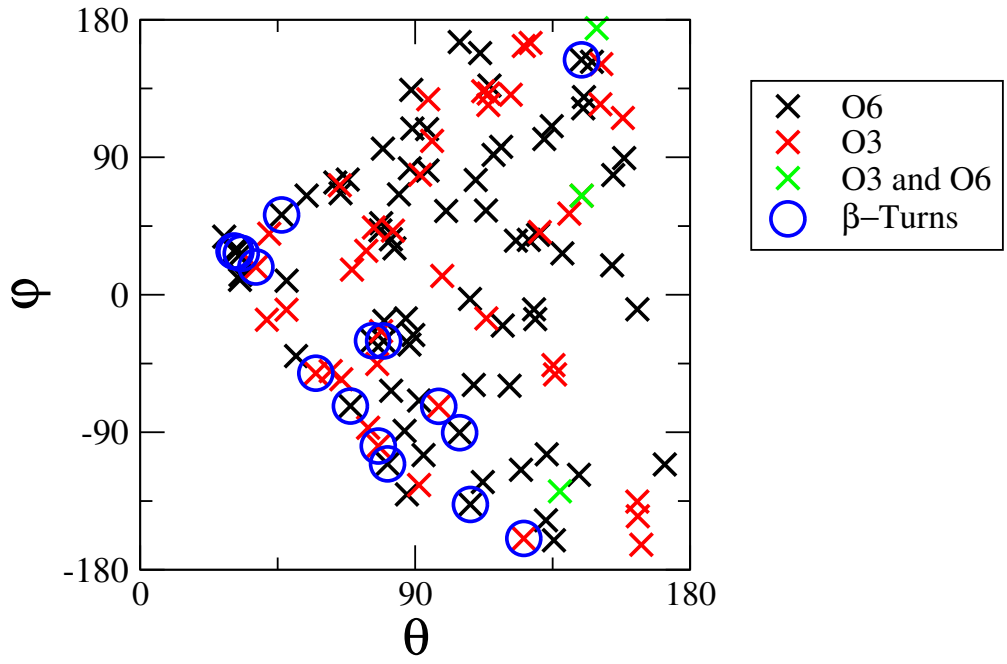
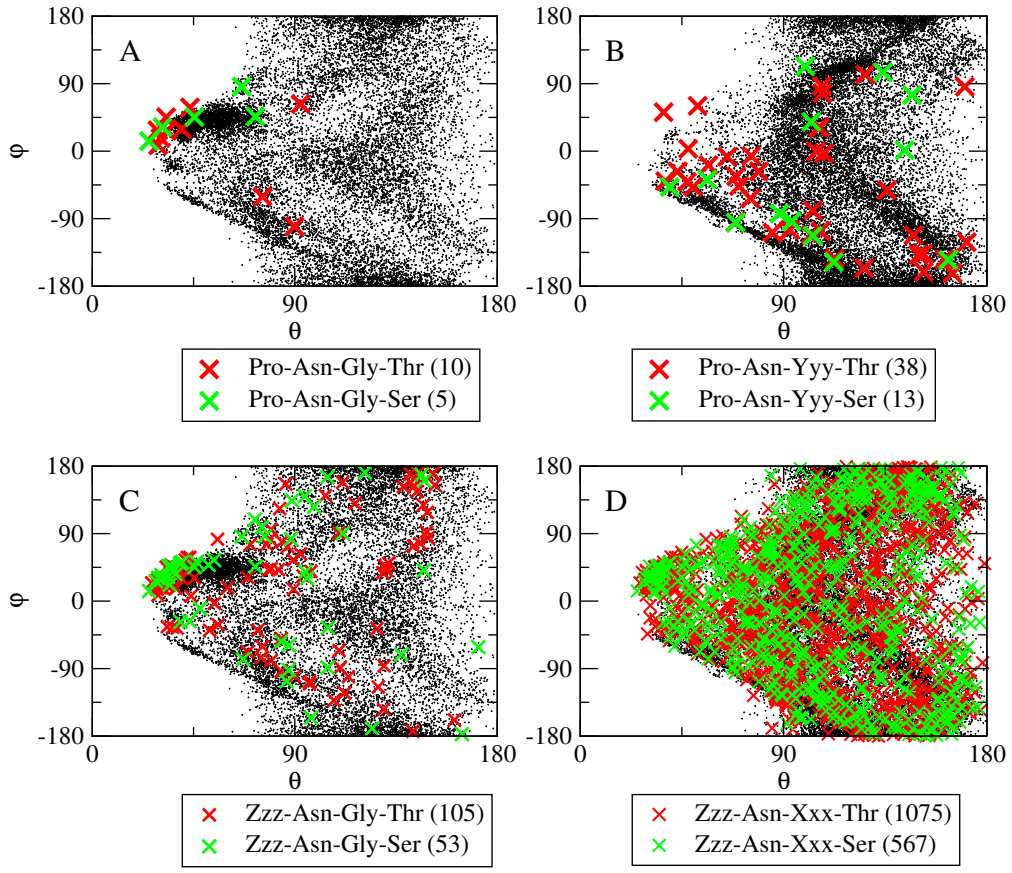


Figure S4: Sidechain steric interactions and their role in destabilizing φ_+ β-turns. Panel A presents a representative φ_+ β-turn configuration. The red and gray spheres indicate the van der Waal radii for the Asn5 carbonyl oxygen and the Gly6 alpha proton, respectively. Panels B and C present the simulated distribution for the distance, r, between the Asn5 carbonyl oxygen and either the Gly6 α proton or the Ala6 β-carbon sampled in the entire simulation and in β-turn configurations, respectively, for the Gly6-Trp8 (black), Gly6-Ala8 (cyan), and Ala6-Trp8 (red) glycopeptide simulations. Panels D and E present a scatter plot of the Gly6-Trp8 β-turn configurations on the FES and Gly6 Ramachandran map, respectively. The configurations are separated into two subpopulations with r $<$ 0.3 nm (black) and r $>$ 0.3 nm (red).



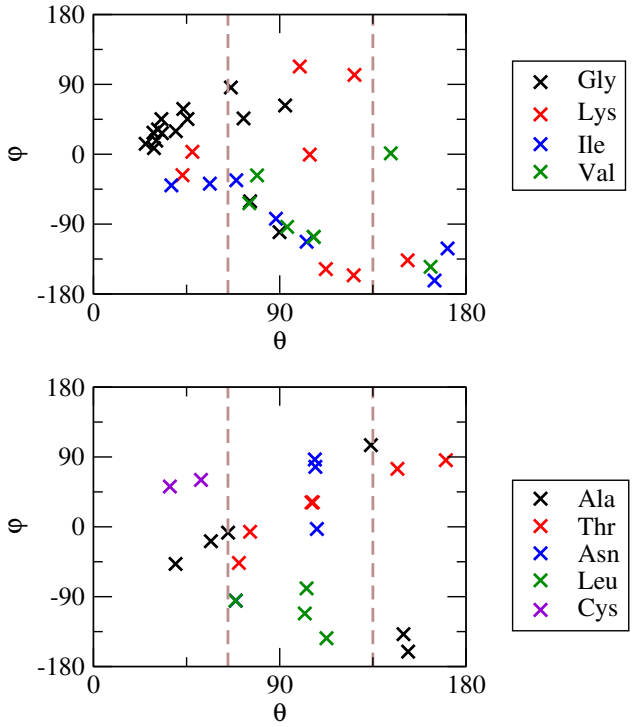
	Instances	$\theta < 65°$	$\theta > 135°$	$\varphi > 0°$	% β-turns
O6	74	14.9	16.2	59.5	14.9
O3	41	14.6	24.4	56.1	12.2
O6 and O3	3	0.0	100.0	66.7	0.0

Figure S5: Analysis of conformations adopted by glycosylated sequons in which the proximal glycan has been fucosylated. The black X's and red X's indicate cases that the fucose residue is linked via the O6 or the O3 oxygen, respectively, of the proximal glycan. The green X's indicate cases that fucose residues were linked via both the O6 and O3 oxygens. The X's identified by blue circles correspond to cases in which the protein backbone forms a β turn at the glycosylation site. The bottom table provides statistical analysis of these structures.



	Instances	$\theta < 65°$	$\theta > 135°$	$\varphi > 0°$	% β-turns
Pro-Asn-Gly-Thr	10	70.0	0.0	80.0	60.0
Pro-Asn-Gly-Ser	5	60.0	0.0	100.0	60.0
Pro-Asn-Yyy-Thr	38	23.7	23.7	23.7	26.3
Pro-Asn-Yyy-Ser	13	15.4	23.1	38.5	38.5
Zzz-Asn-Gly-Thr	105	41.9	16.2	74.3	21.9
Zzz-Asn-Gly-Ser	53	45.3	11.3	71.2	18.9
Zzz-Asn-Xxx-Thr	1075	14.6	29.4	50.8	12.9
Zzz-Asn-Xxx-Ser	567	15.9	28.0	54.8	12.3

Figure S6: Comparison of structures adopted by glycosylated sequons with Thr or Ser at position +2. Panels A-D present scatter plots that describe the sequon conformation as a function of the backbone bend θ and twist φ in analogy to Figure 3 of the main text. Panel A considers Pro-Asn-Gly sequons; Panel B considers Pro-Asn-Yyy sequons; Panel C considers Zzz-Asn-Gly sequons; and Panel D considers Zzz-Asn-Xxx sequons. Zzz indicates any residue at position -1, Xxx indicates any residue at position +1, and Yyy indicates any residue other than Gly at position +1. In each panel, the red and green X's indicate sequons with Thr or Ser, respectively, at position +2. The black X's indicate simulated conformations for the Gly6-Trp8 glycopeptide (panels A and C) or the Ala6-Trp8 glycopeptide (panels B and D). In each case, the legend indicates the frequency of each sequon in the SAGS database. The table statistically analyzes the conformation distribution in each panel.



+1 Residue	Instances	$\theta < 65°$	$\theta > 135°$	$\varphi > 0°$	% β-turns
Gly	15	66.7	0.0	86.7	60.0 (9)
Lys	8	25.0	12.5	37.5	37.5 (3)
Ile	7	28.6	28.6	0.0	57.1 (4)
Val	6	0.0	33.3	16.7	16.7 (1)
Ala	6	33.3	33.3	16.7	16.7 (1)
Thr	6	0.0	33.3	66.7	0.0
Asn	4	0.0	0.0	50.0	25.0 (1)
Leu	4	0.0	0.0	0.0	75.0 (3)
Met	3				
Cys	2				100.0 (2)
Ser	2				
Arg	1				
Glu	1				
Tyr	1				
Asp	0				
Gln	0				
His	0				
Phe	0				
Pro	0				
Trp	0				

Figure S7: Frequency and conformations of glycosylated Pro-Asn-Xxx-Thr/Ser sequons.