Structural basis for peptide substrate specificities of glycosyltransferase GalNAc-T2

Sai Pooja Mahajan[1], Yashes Srinivasan[2], Jason W. Labonte[1,3], Matthew P. DeLisa[4], Jeffrey J. Gray[1,5]

[1] Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, Maryland 21218, United States

[2] Department of Bioengineering, University of California, Los Angeles, Los Angeles, California 90095, United States

[3] Department of Chemistry, Franklin & Marshall College, Lancaster, Pennsylvania 17604, United States

[4] Robert Frederick Smith School of Chemical and Biomolecular Engineering, Department of Microbiology, and Nancy E. and Peter C. Meinig School of Biomedical Engineering, Biochemistry, Molecular and Cell Biology, Cornell University, Ithaca, New York 14853, United States

[5] Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, Maryland, United States

Abstract

The polypeptide *N*-acetylgalactosaminyl transferase (GalNAc-T) enzyme family initiates *O*-linked mucin-type glycosylation. The family constitutes 20 isozymes in humans—an unusually large number—unique to O-glycosylation. GalNAc-Ts exhibit both redundancy and finely tuned specificity for a wide range of peptide substrates. In this work, we deciphered the sequence and structural motifs that determine the peptide substrate preferences for the GalNAc-T2 isoform. Our approach involved sampling and characterization of peptide–enzyme conformations obtained from Rosetta Monte Carlo-minimization–based flexible docking. We computationally scanned 19 amino acid residues at positions –1 and +1 of an eight-residue peptide substrate, which comprised a dataset of 361 (19x19) peptides with previously characterized experimental

GalNAc-T2 glycosylation efficiencies. The calculations recapitulated experimental specificity data, successfully discriminating between glycosylatable and non-glycosylatable peptides with an accuracy of 89% and a ROC-AUC score of 0.965. The glycosylatable peptide substrates *viz.* peptides with proline, serine, threonine, and alanine at the –1 position of the peptide preferentially exhibited cognate sequon-like conformations. The preference for specific residues at the −1 position of the peptide was regulated by enzyme residues R362, K363, Q364, H365 and W331, which modulate the pocket size and specific enzyme-peptide interactions. For the +1 position of the peptide, enzyme residues K281 and K363 formed gating interactions with aromatics and glutamines at the +1 position of the peptide, leading to modes of peptide-binding sub-optimal for catalysis. Overall, our work revealed enzyme features that lead to the finely tuned specificity observed for a broad range of peptide substrates for the GalNAc-T2 enzyme. We anticipate that the key sequence and structural motifs can be extended to analyze specificities of other isoforms of the GalNAc-T family and can be used to guide design of variants with tailored specificity.

Introduction

In higher organisms, O-linked *N*-acetylgalactosamine (GalNAc) glycosylation (or mucin-type glycosylation) is an abundant and essential post-translational modification. This type of glycosylation is initiated by a family of glycosyltransferases (GTs) known as polypeptide *N*-acetylgalactosaminyltransferases or GalNAc-Ts (also referred to as GALNTs). These enzymes transfer a GalNAc sugar from a donor uridine di-phosphate (UDP) nucleotide to the hydroxyl group of a threonine or serine residue of an acceptor peptide. This transfer is the first committed step of mucin-type O-glycosylation, and these enzymes, therefore, define the sites of O-glycosylation. The resulting O-linked GalNAc is further extended to one of the four common core structures, which can be subsequently extended to give mature linear or branched glycans.[1,2] Aberrant O-glycosylation is a well-known marker of many cancers and has also been linked to developmental and metabolic disorders.[3,4]

In humans, the GalNAc-T family constitutes 20 isozymes. The unusually large number of isoforms for glycosylation is unique to O-glycosylation, and the multiplicity is conserved in mammalian

evolution, suggesting that cell or tissue specific isoforms have specialized functions.[5] The isoforms exhibit specific substrate preferences that vary with isoenzyme surface charge, prior neighboring long-range and short-range glycosylation patterns and the sequence of the acceptor peptide substrate. Over the last two decades, the peptide substrate preferences for a large number of isoforms have been established by *in vitro* studies.[6–8] The peptide substrate is characterized by a sequence motif (or sequon), Thr/Ser–Pro–X–Pro (T/SPXP), where T/S is the site of glycosylation (position 0). This sequon is the only conserved consensus motif modified by all isoforms except T7 and T10. The proline at the +3 position of the sequon is supported by a conserved structural motif, *viz.,* the "proline pocket" in the enzyme's peptide binding groove in all isoforms that bind the T/SPXP motif.[9–11] For the remaining positions in the sequon, most isoforms exhibit overlapping yet selective preferences for different amino acid residues. For example, at the −1 position with respect to the glycosylation site, T1 favors aromatics[12] and T12 prefers bulky non-polar residues;[13] whereas T2 exhibits very little to no activity for these amino acids and instead prefers threonine, proline, alanine, and serine. Yet both T1 and T2 glycosylate the sequon TTP[12] (with threonine at −1 and proline at +1 positions). These observations have led to the hypothesis that GalNAc-Ts exhibit both redundancy and finely tuned specificity for a wide range of peptide substrates.

While there is ample experimental data on the peptide substrate specificities of various isoforms, the molecular basis for observed peptide substrate specificities is not well understood. Computational work, so far, has been focused on understanding the mechanism of sugar transfer,[14,15] conformational changes in the flexible loop in the catalytic domain,[16] and the effect of the flexible linker connecting the catalytic and lectin domains.[11] None of the computational studies so far have examined the amino acid preferences at different positions on the peptide. Computational studies can pinpoint key positions and structural motifs on an isoform that contribute to peptide substrate specificity. These sequence and structural motifs can be studied across isoforms to reveal more general patterns, to modulate enzyme specificity, and to gain insight into the consequences of enzyme and substrate mutations implicated in aberrant glycosylation, (*e.g.*, colorectal cancer associated mutations of GalNAc-T12[17]) paving the way for rational design of specific drugs/inhibitors.

In this work, we seek to understand the sequence and structural motifs that determine the peptide substrate preferences for the GalNAc-T2 isoform. Our immediate goal is to recapitulate experimentally determined specificity in terms of glycosylation efficiency for sequon variations at positions −1 and +1 (19 amino acids tested for each position), as reported by Kightlinger *et al.,*[12] and to understand which structural motifs best explain experimentally observed trends. To recapitulate experimentally observed specificities for a large dataset, we need an efficient, high-throughput computational method that can capture the key mechanisms of enzymatic catalysis.

Enzymatic catalysis relies primarily on selective transition state stabilization, ground state (reactants) destabilization, dynamics, and active-site gating.[18,19] In practice, these effects occur at different length- and time- scales and therefore cannot be accurately captured by a single method. Hybrid quantum mechanics/molecular mechanics (QM/MM) simulations have been able to recapitulate catalytic proficiency for many enzymes such as Kemp eliminases[20] or xylase[cite Mayes] as they are well-suited to characterize the transition state. QM/MM simulations, however, are not suitable to capture binding or dynamics over longer timescales and are prohibitively expensive for a larger dataset. Other factors that determine the stability of the transition state are electrostatic- and shape- complementarity at the peptide-enzyme interface. Electrostatic complementarity can be captured by various computational techniques (*e.g.*, Monte Carlo (MC) or molecular dynamics (MD)-based methods with Poisson-Boltzmann electrostatics or other continuum electrostatics models) at different length-scales. Other effects are determined by the thermodynamics of the enzyme-peptide interactions. To achieve a lower free energy of activation,[18,21] the enzyme must stabilize the transition state selectively relative to the reactants. Additionally, if the product is too stable in the enzyme's active site, product release becomes the catalytic rate-limiting step. This thermodynamic description demands the use of methods that capture multiple states (reactants, products and transition states).[22,23] Furthermore, dynamics is important in many catalytic mechanisms, from small vibrations that lead to rate-promoting motions[24] to large conformational changes and rearrangements in the molecular structure.[25] Active site gating is another important mechanism for catalysis by which key residues outside the active site regulate access to the active site.[26,27] These thermodynamic and kinetic effects, primarily in the nanosecond to microsecond timescales, can be captured

faithfully by MD simulations though such simulations can be computationally prohibitive for comparing a large number of substrates. An alternative to MD simulations are Monte Carlo-minimization[28] (MCM) approaches, which are computationally faster and can be reliably used to determine thermodynamically stable native-like states.

Rosetta-based MCM computational protocols, notably, pepspec,[29] sequence-tolerance[30,31] and MFPred[32] have previously been used for predicting the sequence profiles of peptides recognized by various multi-specific protein recognition domains (PRDs) such as PDZ, SH2, SH3, kinases, and proteases. All protocols rely on MCM sampling and aim to approximate the stabilization of the substrate-bound state or transition state in the enzyme's peptide binding groove. The transition state is approximated by known cognate sequon conformations in the enzyme's active site (based on crystal structures and/or homology modeling) with additional constraints to preserve important structural motifs pertaining to the transition state, when available. In the absence of constraints, this approach is equivalent to evaluating the stabilization of the substrate-bound state.[33] MCM allows for faster sampling facilitating the scanning of a large number of amino acid residues at multiple positions of the peptide substrate. All three protocols achieve impressive accuracy in predicting experimentally observed profiles for many PRDs. However, since all three methods are developed with the broad goal of predicting sequence specificity profiles for a range of PRDs, the accuracy of prediction may not be sufficient to pinpoint subtle differences in specificity for a specific target of interest. For example, the sequence-tolerance protocol pre-calculates the interactions between all interacting residues ignoring changes in conformation of the peptide in the protein's binding pocket. All three methods are inferior at predicting specificity for HIV−1 protease, which has a relaxed specificity profile and a preference for small hydrophobic residues, similar to GalNAc-T2. Additionally, all three protocols employ limited backbone sampling, prohibiting the free conformational sampling of the peptide in the binding groove. For the more targeted goal of designing a peptide inhibitor to discriminate between two similar PDZ domains, Zheng *et al.*[34] employ extensive conformational sampling using the full-fledged flexpepdock protocol[35] along with the CLASSY method to achieve a solution with desired specificity and affinity goals. Similarly, Pethe *et al*.[36] were able to obtain significantly improved prediction accuracies for proteases (including HIV−1 protease) compared to previous methods

(MFPred, sequence tolerance and pepspec) by employing machine learning and a discriminatory score based on geometric features, interface score terms from Rosetta and electrostatic score terms from Amber. In another work, Pethe *et al.*[37] used supervised learning on experimentally obtained deep-sequencing data and information from structure-based models to chart the specificity landscape of 3.2 million substrate variants of the viral protease HCV.

Here, we sought to understand specificity determinants for a specific isoform of the GalNAc-T family and to pinpoint sequence and structural motifs in the enzyme that explain fine-tuning of specificity. To this end, we developed a customized Rosetta-based protocol[38,39] that allowed us to model structures of all 361 peptide sequons (19×19) with the GalNAc-T2 enzyme and computationally determine the sequon preference for the GalNAc-T2 isoform. Our protocol was similar in spirit to earlier protocols in that it docks the peptide substrate into the enzyme's active site. However, unlike pepspec,[29] sequence-tolerance[30,31] and MFPred[32] and similar to the protocol of Zheng *et al.,*[34] we allowed fully flexible peptide sampling (as opposed to limited or no backbone sampling) followed by clustering and analysis of the sampled low energy decoys. Our strategy relied on characterizing the peptide binding to the enzyme with a range of structural features at the interface as a function of the amino acid residues at the +1 and −1 positions. Using our methodology, we were able to identify features that recapitulated high-quality experimental specificity data for GalNAc-T2. Extensive peptide backbone sampling revealed that the peptide binding groove of GalNAc-T2 stabilized multiple competing conformations/states – some leading to efficient glycosylation and others hampering it. Furthermore, multiple stable states suggested that kinetics might play an important role in determining specificity and the possibility of fine-tuning specificity by modulating the relative stability of these states to discriminate between peptide substrates for an isoform and across isoforms. Overall, our work reveals key residues on the enzyme that determine peptide substrate preferences at various sequon positions.

Results

Clustering of low interaction energy decoys reveals that peptides exhibit multiple competing low-energy conformations

We studied all 361 (19x19) sequons obtained by scanning 19 amino acids (all amino acids except cysteine) at positions −1 and +1 with respect to the modified threonine (Thr_0). The experimentally determined glycosylation efficiencies for all of these sequons was determined by Kightlinger *et al*.[12] and replotted in Figure *1*A. For each sequon, we started with the co-crystal structure of the peptide and UDP-sugar bound to the enzyme (pdb ids: 4d0z and 2ffu, respectively).[16] We mutated the residues at the −1 and +1 position of the peptide to the target sequon and repacked and minimized the nearby side chains to obtain a starting enzyme-peptide configuration for the target sequon. We then subjected this starting structure to MCM sampling of rigid-body displacements and peptide torsion angles in two stages - a low-resolution centroid stage with simulated annealing followed by a high-resolution all-atom stage to generate 2,000 structures (decoys) per sequon (see Methods). We used the shorthand notation X_N for amino acid 'X' (denoted by 1-letter code) and sequon position 'N' and a sequon with the shorthand notation $X_{-1}TX_{+1}$. For example, P_{-1} denotes amino acid proline at the −1 position, M_{+1} denotes a methionine at the +1 position and $P_{-1}TM_{+1}$ denotes a sequon or peptide with P_{-1} and M_{+1}. For brevity, we also refer to peptides or sequons containing residue X at position N as "X_N peptides" or "X_N sequons" respectively.

Preliminary analysis of the decoys showed that many peptides exhibited multiple stable states with comparable energies of interaction (interaction energy) between the peptide and enzyme. In Figure *1*B, we show funnel plots for four randomly chosen sequons obtained from MCM sampling of the peptide substrate with the respective sequons in the enzyme's peptide-binding groove. For all four sequons in Figure 1B, we observed multiple clusters of low interaction energy decoys, or "funnels." For example, for sequon $T_{-1}TQ_{+1}$, we observed two distinct funnels at $RMSD_{peptide}$ (the root mean square deviation of C_α carbons of the peptide backbone with respect to the peptide in the crystal structure) values of about ~ 0.65 Å and 1.25 Å with comparable lowest interaction energies. Overall, 57% (205/361) of the sequons exhibited two significant clusters. 39/205 (19.0%) and 81/205 (39.5%) sequons exhibited lowest energy states for each cluster within 0.5 and 1.0 Rosetta energy units (or REUs) respectively, of each other, underscoring the importance of considering both states (SI Figure 1).

To characterize multiple low-energy conformations and to construct a more complete picture of the landscape of structural conformations sampled by the peptide substrate in the enzyme cavity, we developed the computational flow summarized in Figure 1C. For each sequon, we selected the top-10%-scoring decoys (by interaction energy) from MCM sampling and then clustered them using three features. The first feature, $RMSD_{peptide}$, characterized decoys on the basis of the similarity of the peptide backbone conformation and position of the peptide in the crystal structure. Next, the distance between the hydroxyl group of T_0 and the anomeric carbon (C_1) on the sugar tracked the distance for the new glycosidic linkage. Finally, the distance between the amide group of T_0 on the peptide and the oxygen of the β-phosphate group of UDP ($O_{\beta\text{-}PO4}$) was a reaction coordinate characterizing a transition-state-stabilizing hydrogen bond between the backbone amide of T_0 and UDP.[14]

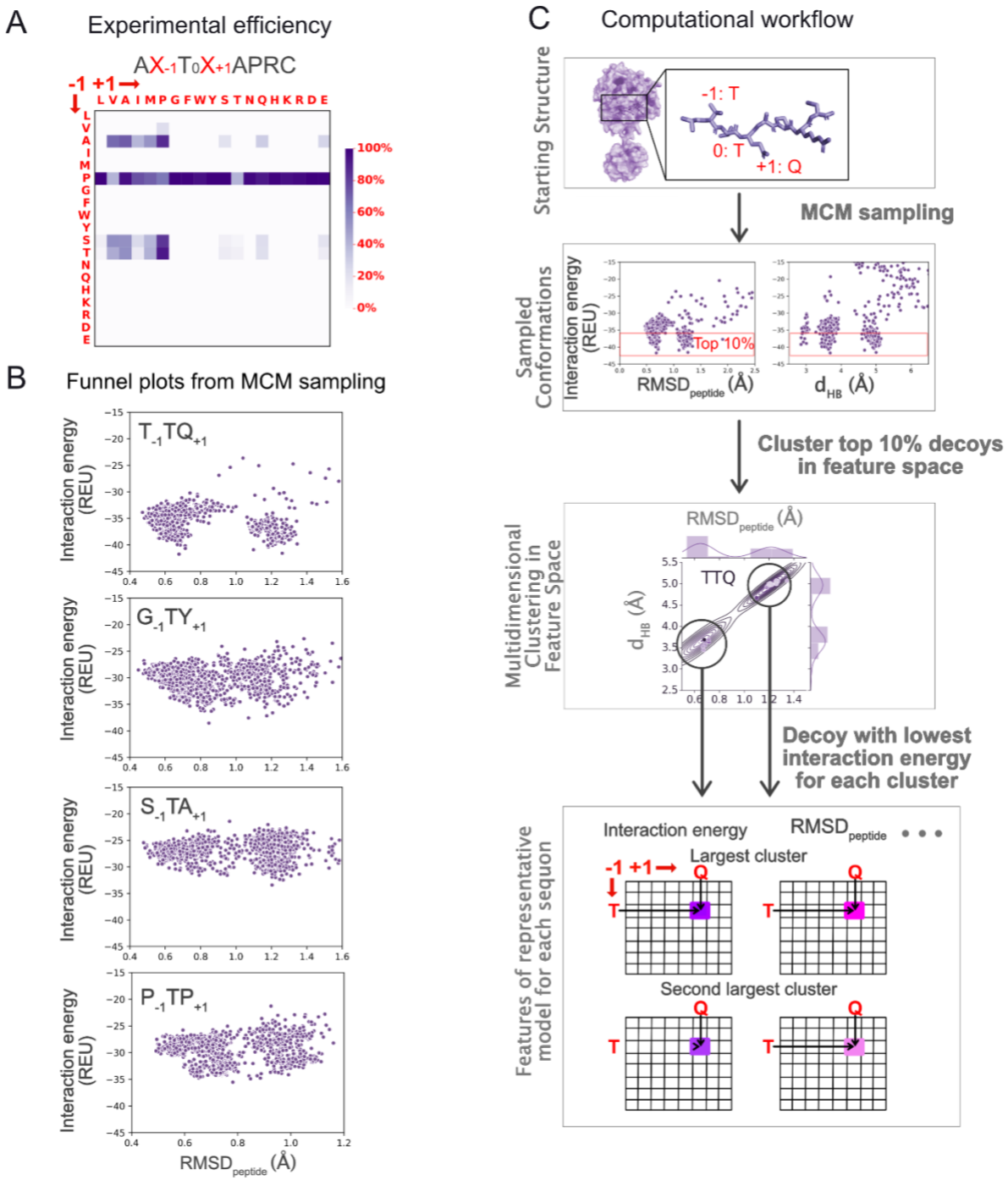

Figure 1. Computational workflow to determine glycosylation efficiency of glycosyltransferase GalNAc-T2 for peptide substrates obtained by scanning 19 amino acid residues (all except cysteine) at positions −1 and +1 of the acceptor peptide. (A) Experimentally determined glycosylation efficiencies (efficiency data replotted from Kightlinger *et al.*[12]). (B) Funnel plots from MCM sampling for four sequons. Each point represents one structural model, or "decoy," at its corresponding RMSD from the reference structure and the interaction energy calculated by

Rosetta. (C) Steps in the computational workflow to characterize enzyme–peptide interactions for a representative sequon, $T_{-1}TQ_{+1}$, with T at the −1 position and Q at the +1 position.

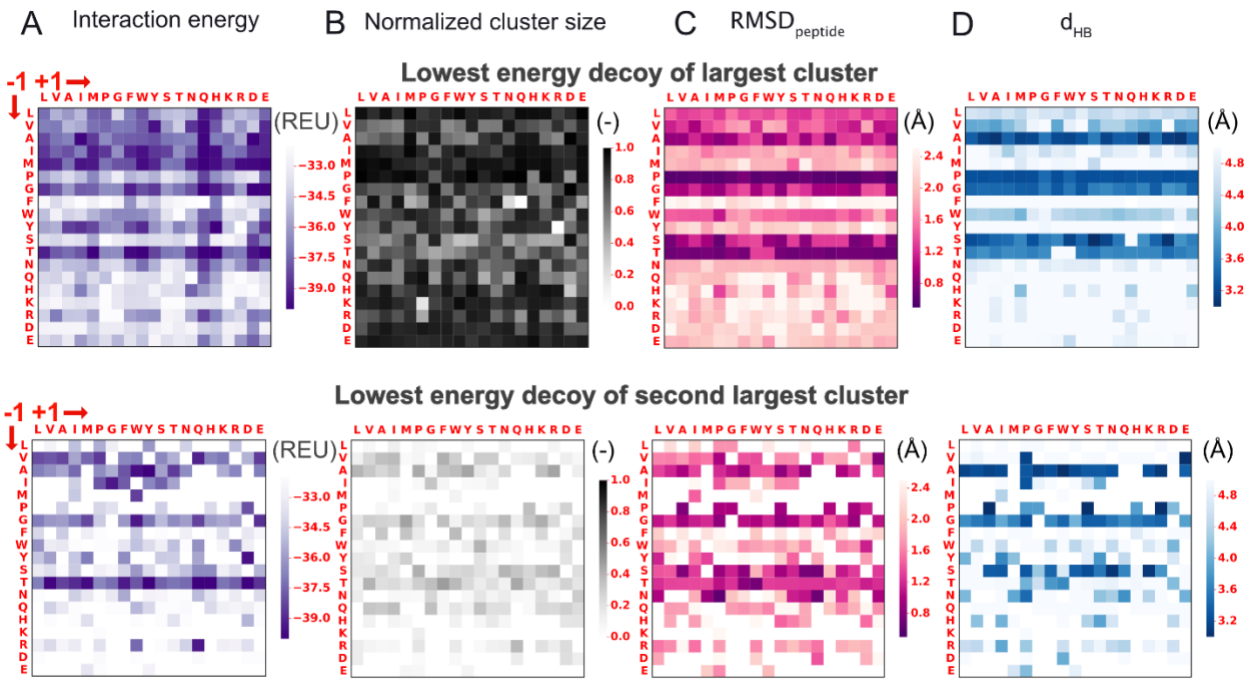

Figure 2. Characterization of the lowest-energy representative conformation for the top two clusters in Rosetta runs (top and bottom). (A) Interaction energy. (B) Normalized cluster size. (C) $RMSD_{peptide}$. (D) d_{HB} of the largest and second-largest clusters characterized by the lowest interaction energy decoy for each cluster.

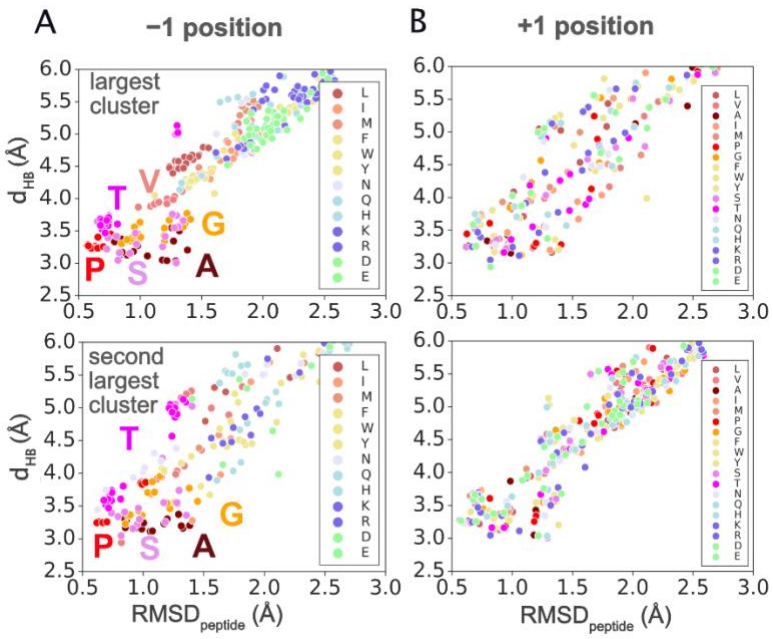

Figure 3. Lowest energy decoys belonging to the largest and second largest clusters for all sequons colored by (A) the residue at the −1 position of the sequon and (B) the residue at the +1 position of the sequon.

We characterized the lowest-energy decoy for the largest and second-largest clusters obtained for each sequon and plotted heatmaps to show the distribution of the lowest interaction energy (Figure 2A), normalized cluster size (Figure 2B) , RMSD$_{peptide}$ (Figure 2C) and d_{HB} (Figure 2D) for all 361 sequons (see also SI Figure 2A). We also characterized the clusters by the decoy representing the center of the cluster, the average over all decoys with interaction energies within 1 REU and the average over the five decoys in the cluster with the lowest interaction energies. All strategies resulted in similar heatmaps (SI Figure 3) and hence, going forward, we represented a cluster by the lowest energy decoy belonging to that cluster.

Horizontal stripes emerging across the RMSD$_{peptide}$ and d_{HB} heatmaps (Figure 2C-D) suggested that sequons with the same amino acid at the −1 position (horizontal axis) exhibited similar RMSD$_{peptide}$ and d_{HB} values. To probe whether the low-energy conformations exhibited by various peptides depends on the identity of the residue in a position-specific manner, we plotted the RMSD$_{peptide}$ and d_{HB} of the lowest energy decoy for the two largest clusters for each sequon

colored by the amino acid residue at the −1 (Figure 3A) and +1 (Figure 3B) positions. It is apparent in Figure 3A that sequons with the same amino acid residue at the −1 position, especially A, G, T, P, S and V, were grouped or clustered together. The clustering or grouping suggests that sequons with the same amino acid at the −1 position exhibited similar conformations or low-energy states. Similar grouping was not observed for sequons with the same residue at the +1 position (Figure 3B). This high-level analysis of low-energy conformations for the entire dataset suggested that the −1 position plays a dominant role in determining the low-energy conformation(s) exhibited by a sequon and that the +1 position contributed in a secondary capacity.

In the following sections, we present hypotheses to explain how each position (−1, +1) contributed in a characteristic manner to determine the low-energy conformations exhibited by a sequon and how these conformations, in turn, related to experimentally determined specificities. We characterized the low-energy conformations by a selected set of relevant features. To compare our predictions with experiments, we employed logistic regression and, unless otherwise indicated, we labeled all sequons with experimental glycosylation efficiencies greater than 10% (efficiency threshold) as glycosylatable and those with efficiencies less than 10% as un-glycosylatable, in line with previous work[32]. In SI Table 1, we have tabulated the area-under the curve (AUC) of the receiver operating curve (ROC) for a range of features and efficiency thresholds.

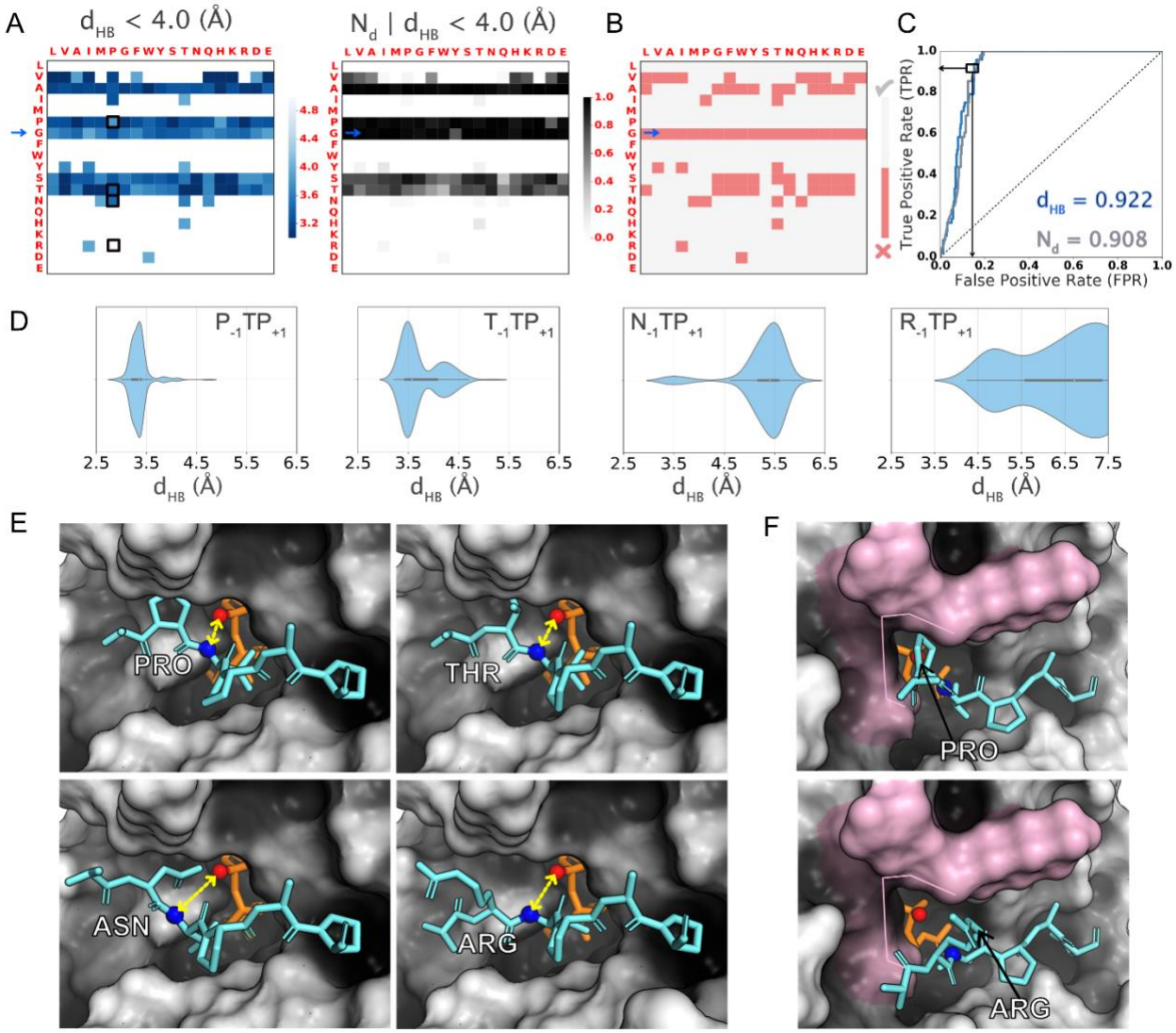

Figure 4. Substrate specificity based on TS stabilizing hydrogen bond criterion with d_{HB} < 4 Å. (A) Heatmaps of (left panel) d_{HB} distances of the lowest-interaction-energy decoy belonging to a cluster with the cluster centroid satisying the criterion and (right panel) fraction of decoys (N_d) satisfying criterion. (B) Binary glycosylatibility predicted correctly (grey) and incorrectly (coral red) by d_{HB} based on the d_{HB} < 4 Å criterion. (C) ROC curve for d_{HB} distances and N_d satisfying criterion. (D) Violinplot of distribution of d_{HB} distances sampled by the top-scoring 10% decoys for four representative sequons ($P_{-1}TP_{+1}$, $T_{-1}TP_{+1}$, $N_{-1}TP_{+1}$ and $R_{-1}TP_{+1}$). (E) Lowest interaction energy decoys for four sequons ($P_{-1}TP_{+1}$, $T_{-1}TP_{+1}$, $N_{-1}TP_{+1}$ and $R_{-1}TP_{+1}$ – black boxes in the heatmap in (A). (F) Pocket-like cavity formed by enzyme residues (pink surface) that contacts the amino acid at the −1 position on the peptide. d_{HB} is calculated between the amide nitrogen (blue sphere) of T_0 on peptide(aquamarine) and the $O_{\beta\text{-}PO4}$ (red sphere) on UDP (orange). d_{HB} is shown with double-ended yellow arrows.

Recapitulation of amino acid specificity trends for the −1 position

Sampling of TS-critical hydrogen bond recapitulates specificity trends with a 92% discriminative capacity

In QM/MM simulations of the glycosylation of the EA2 peptide by GalNAc-T2, Gomez *et al.* characterized a hydrogen bond between the backbone amide of Thr_0 and the β-phosphate group on UDP [14]. They proposed that the hydrogen bond stabilizes the transition state (TS) in "a general catalytic strategy used in peptide O-glycosylation by retaining glycosyltransferases". Hence, our first hypothesis was that a criterion for successful glycosylation is the ability of a peptide to exhibit a low-energy conformation with d_{HB} distances compatible with the proposed hydrogen bond. Thus, in Figure 4A, we show the heatmap of the of the lowest interaction energy decoys. Applying a 4.0 Å threshold to the 19×19 grid of sequons (Figure 4B) split the sequons into those that do not meet this condition (*i.e.,* > 4.0 Å,) and those that exhibited a representative low-energy conformation compatible with hydrogen bonding between the peptide and UDP.

When compared with experimental results (Figure 1A), this criterion discriminated well between substrate peptides and non-substrates of the enzyme (Figures 4B). A ROC analysis (Figure 4C) showed that the d_{HB} based-criterion had an AUC value of 0.922, meaning that this metric had a 92.2% chance of correctly distinguishing the glycosylatable sequons from the non-glycosylatable sequons. If instead of d_{HB} threshold we used the fraction of the of decoys that satisfied the d_{HB} criterion, the AUC was 0.908.

Amino acid residues with larger side chains are excluded from "−1 pocket" of the enzyme

To understand the structural basis for the observed d_{HB} trends, we considered specific sequons and the lowest-energy conformations sampled by them. First, we note that in Figure 4A, peptides preferentially exhibited low-energy, highly populated states (higher fraction of decoys) with d_{HB} distances compatible with hydrogen-bonding when amino acid residues with smaller side chains such as proline, alanine, glycine, serine, or threonine were present at the −1 position. In Figure 4D, we show the d_{HB} distances sampled by top 10% low-energy decoys for four representative sequons – $P_{-1}TP_{+1}$, $T_{-1}TP_{+1}$, $N_{-1}TP_{+1}$ and $R_{-1}TP_{+1}$. Sequons with P_{-1} and T_{-1} preferentially sampled conformations with d_{HB} < 4.0 Å, whereas those with N_{-1} or R_{-1} preferentially sampled higher d_{HB} distances. In Figure 3D, we show the structures of the lowest energy conformation (largest

cluster) for sequons with P_{-}, T_{-1}, N_{-1} and R_{-1}. We observed that the sequons P_{-1} or T_{-1} fit in the pocket-like cavity in the enzyme's peptide binding groove (Figure 4E, top panel) whereas, sequons with N_{-1} or R_{-1} were excluded from this cavity due to steric hinderance thereby resulting in larger distances (Figure 4E, bottom panel). See SI Figure 4 for lowest-energy conformations and d_{HB} distances for D_{-1} peptides as aspartate is a similar size range as amino acid residues threonine and proline. Hence, the structural basis for peptides to preferentially sample low-energy conformations compatible with sampling of the proposed TS stabilizing hydrogen bond was the *relative size of the side chain of the amino acid at the −1 position and the "−1 pocket" on the enzyme* (highlighted in Figure 4F; discussed in more detail in subsequent sections).

Notably, sequon $N_{-1}TP_{+1}$ (Also $V_{-1}TP_{+1}$; SI Figure 4) exhibited two significant clusters (Figure 2A, Figure 4A), the smaller one (normalized cluster size ~ 10%) exhibiting distances compatible with hydrogen bonding (Figure 4A [right panel]) and the larger one (normalized cluster size ~ 90%) with comparable interaction energy exhibiting larger distances. Experimentally this sequon was non-glycosylatable. This points to the importance of the TS-stabilizing hydrogen bond for efficient peptide glycosylation.

Characterization of the d_{HB} distance correlated with undetectable glycosylation in experimental assays for sequons with larger amino acid residues, as peptides/sequons with larger amino acids did not meet the hydrogen-bonding criteria and assumed conformations at distances farther from the UDP-GalNAc donor, making the reaction less likely. However, using d_{HB} as the sole criteria for specificity incorrectly classified $G_{-1.}$ (Figure 4C; blue arrow). Since d_{HB} generated some false positives, especially G_{-1} peptides, the ability of the peptide to assume conformations amenable to the formation of the TS-stabilizing hydrogen bonding may be a necessary but not sufficient condition to determine specificity.

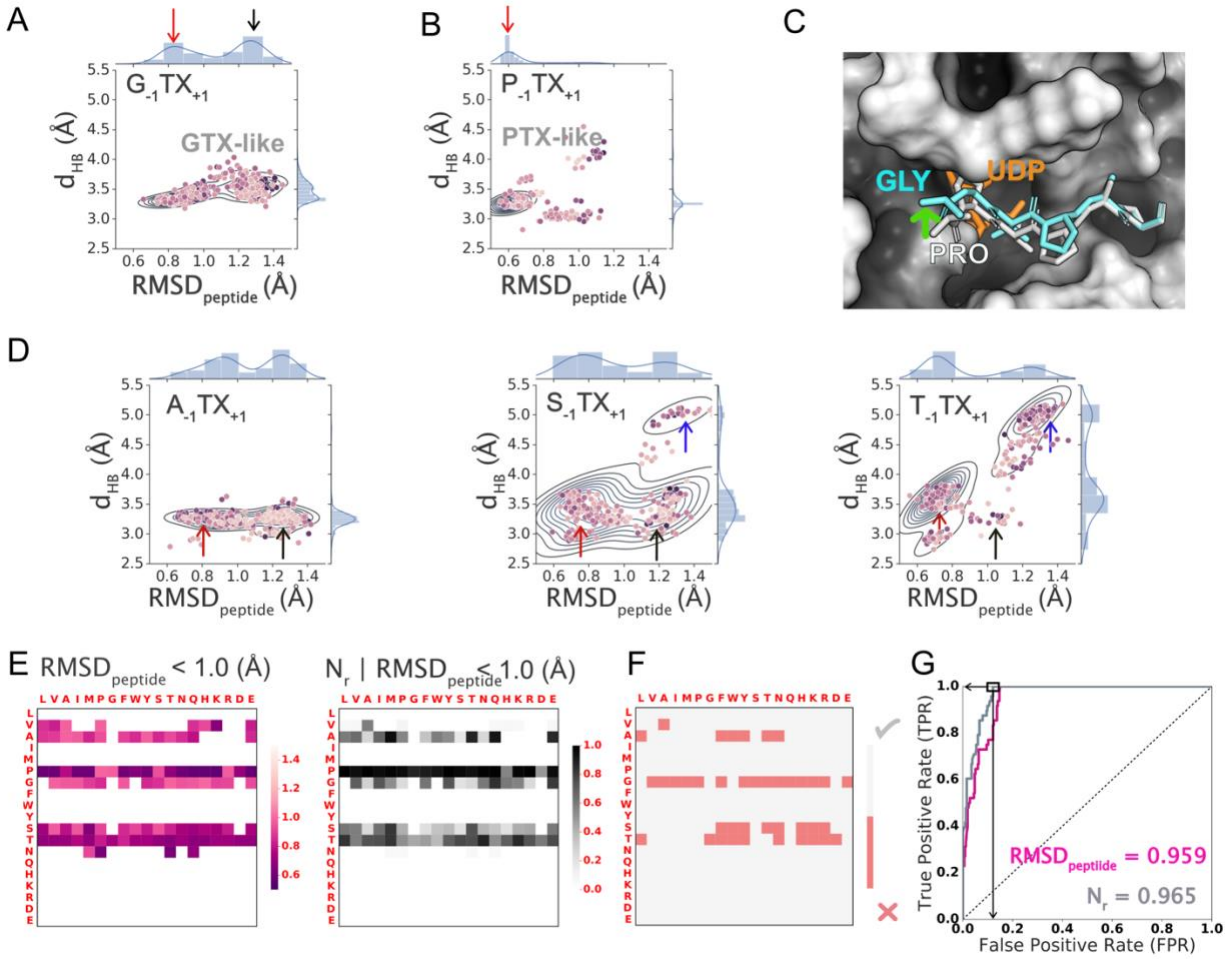

Figure 5. Substrate specificity based on $RMSD_{peptide}$ < 1.0 Å criterion. Joint and marginal probability densities of Top 10%(200/2000) sequons for all peptides with fixed amino acids. (A) G_{-1} and B) P_{-1} and all amino acid residues at X_{+1}; Top 1% (20/2000) decoys per sequon shown as points where darker color indicates lower interaction energy. (C) Lowest interaction energy decoy for representative sequons $P_{-1}TP_{+1}$ (white; $RMSD_{peptide}$ < 1.0 Å) superposed with that for $G_{-1}TP_{+1}$ (aquamarine; $RMSD_{peptide}$ > 1.0 Å) in the enzyme's peptide binding groove. (D) Joint and marginal probability densities of Top 10%(200/2000) sequons for all peptides with fixed amino acids A_{-1}, S_{-1}, T_{-1} and all amino acid residues at X_{+1}; Top 1% (20/2000) decoys per sequon shown as points where darker color indicates lower interaction energy. (E) Heatmap of $RMSD_{peptide}$ (left panel) of the lowest energy decoy per sequon, fraction of decoys (N_r) satisfying RMSD criterion (right panel) for $RMSD_{peptide}$ < 1.0 Å. (F) Binary glycosylatibility predicted correctly (grey) and incorrectly (coral red) by N_r based on the RMSDpeptide < 1.0 Å criterion at a True Positive Rate of 1.0. (G) ROC curve for $RMSD_{peptide}$ (magenta) and N_r(grey) satisfying $RMSD_{peptide}$ < 1.0 Å. "GTX-like" state is marked with a black arrrow and "PTX-state" is marked with a red arrow; the blue-arrow indicates a third state distinct from PTX- and GTX-like states.

G_{-1} results in distinct low-energy states characterized by higher $RMSD_{peptide}$ values

To probe why G_{-1} peptides may be unglycosylatable even though they satisfy the metric, we examined the joint distribution of the $RMSD_{peptide}$ and d_{HB} sampled by top 10% of decoys for all (*i.e.*, averaged over all 19 amino acid at the +1 position) G_{-1} and P_{-1} peptides (see SI Figures 5-9 for plots of all 19 sequons for G_{-1} and P_{-1}.). While G_{-1} peptides exhibited two low-energy states (Figure 5A), P_{-1} peptides primarily exhibited a single, low $RMSD_{peptide}$ state (Figure 5B). We refer to the low $RMSD_{peptide}$ state ($RMSD_{peptide}$ < 1.0 Å and < 4.0 Å) as the PTX-like state and to the higher $RMSD_{peptide}$ ($RMSD_{peptide}$ ≥ 1.0 Å and < 4.0 Å) state as the GTX-like state.

Amino acid residues with smaller side chains are sub-optimal for −1 pocket of the enzyme

In Figure 5C, we have superposed the PTX-like and the GTX-like states. In the GTX-like state, the backbone was "shifted up" with respect to that of the PTX-like state (Figure 5C; green arrow). For G_{-1} peptides, the small size of the glycine residue allowed multiple configurations in the −1 pocket of the enzyme, all of which still made the TS stabilizing hydrogen bond (i.e., < 4.0 Å). Consequently, we also observed the GTX-like state for A_{-1} and S_{-1} (shorter side chains) peptides (Figure 5D; black arrow) but not for T_{-1} peptides. Instead, T_{-1} peptides exhibited a third state (Figure 5D; blue arrow) which we discuss later. While the d_{HB} metric explained why sequons with larger side chains at the −1 position were non-glycosylatable, the $RMSD_{peptide}$ metric may explain why certain sequons with smaller side chains at the −1 position may not be suitable for glycosylation.

$RMSD_{peptide}$ metric improves sequon specificity predictions for G_{-1} peptides

P_{-1} peptides, irrespective of the amino acid at the +1 position, experimentally exhibited high glycosylation efficiencies and also primarily exhibited the low $RMSD_{peptide}$, PTX-like state. This leads us to hypothesize that besides the TS stabilizing hydrogen bond (characterized by d_{HB}), the second factor that determined the glycosylatability of a sequon was the precise positioning of the peptide in the enzyme's peptide binding groove, *i.e.*, how close the peptide backbone was, spatially and conformationally, to the cognate sequon peptide conformation in the crystal structure, as characterized by $RMSD_{peptide}$. We postulated that the PTX-like state with $RMSD_{peptide}$ < 1.0 Å lead to successful glycosylation (reactive state) whereas all other conformations or states with $RMSD_{peptide}$ ≥ 1.0 Å e.g. the GTX-like state did not lead to glycosylation (non-reactive).

Hence, we used the sampling of the PTX-like state by the top-scoring decoys, quantified by the $RMSD_{peptide}$ of the lowest energy decoy of the largest cluster and the normalized size of the largest cluster, as the second criterion for successful glycosylation. This criterion improved prediction for sequons that exhibited low-energy conformations for the GTX-like state, ($A_{-1}TH_{+1}$, $A_{-1}TG_{+1}$, $S_{-1}TG_{+1}$, $G_{-1}TG_{+1}$, etc.), including G_{-1} peptides and was able to correctly classify many such peptides ($G_{-1}TL_{+1}$, $G_{-1}TG_{+1}$, $G_{-1}TA_{+1}$ etc.) as non-glycosylatable (Figure 5 E, F). When compared with experimental results, this criterion, based on the fraction of decoys that satisfy the criterion, gave an ROC AUC value of ~ 0.965 (Figure 5F).

However, since both sequons that were glycosylatable (*e.g.* ATA, STA, TTQ) and non-glycosylatable (*e.g.* GTA, ATH, TTY, STY) exhibited non-reactive states, the criterion based on $RMSD_{peptide}$ was not sufficient to correctly classify all peptides, especially for sequons that exhibited both reactive and non-reactive states with similar interaction energies and/or similar fraction of decoys.

Amino acid residue at the −1 position dictates the low-energy conformations and glycosylatibility for the majority of the sequons

The analysis of the low-energy conformations characterized by d_{HB} and $RMSD_{peptide}$ lead to the following observations. For the majority of sequons, those with K_{-1}, R_{-1}, F_{-1}, Y_{-1}, W_{-1}, D_{-1}, E_{-1}, Q_{-1}, N_{-1}, H_{-1}, I_{-1}, M_{-1}, L_{-1}, or V_{-1}, the peptide primarily sampled non-reactive low-energy conformations with d_{HB} > 4.0 Å and $RMSD_{peptide}$ > 1.0 Å. For a small fraction of sequons (P_{-1} peptides), the peptide primarily sampled a reactive, cognate-sequon like state (or PTX-like state) with d_{HB} < 4.0 Å and RMSDpeptide <1.0 Å. For both of these categories that primarily sample one state—either the non-reactive state or the reactive-state—the computational predictions based on either hypothesis ($RMSD_{peptide}$ or d_{HB}), agreed quite well with experimental data. These observations underscore the importance of the residue at the −1 position in determining the low-energy conformations and, consequently, the glycosylatability for the majority of the sequons (~ 15 x 19 = 285 out of 361 peptides).

Recapitulation of amino acid specificity trends for the +1 position

For G_{-1}, A_{-1}, S_{-1}, T_{-1} sequons (4 x 19 = 76 out of 361), the peptide sampled both reactive and non-reactive states with comparable interaction energies. For many of these sequons, the computational predictions based on the effect of the −1 position did not accurately recapitulate experimental observations. Hence, for G_{-1}, A_{-1}, S_{-1}, T_{-1}, to recapitulate experimental glycosylation trends, we must consider the effect of the +1 position.

Amino acid at the +1 position confers secondary effects that modulate effects of the −1 position

To investigate the effect of the +1 amino acid residue for G_{-1}, A_{-1}, S_{-1} and T_{-1} peptides, we considered the variation in sampling and interaction energy of the GTX-like state. Figure 6A shows these fractions for a subset of sequons for which the +1 position was critical, i.e. G_{-1}, A_{-1}, S_{-1}, and T_{-1}. For T_{-1} peptides, no sequon exhibited the GTX-like state for a significant fraction of the decoys. For A_{-1}, S_{-1} and G_{-1} peptides, G_{+1} and D_{+1} significantly increased the propensity to sample (indicated by a large fraction of decoys and low interaction energies) the GTX-like state. Furthermore, for A_{-1} peptides, H_{+1}, K_{+1}, R_{+1}, and S_{+1} also resulted in a large fraction of decoys exhibiting the GTX-like state. We summarize the changes in classification accuracy in SI Table 2 and SI Figure 10B upon inclusion of a GTX-like state-based criterion. Hence, the +1 position, in these specific cases, enhanced the sampling of the non-reactive, GTX-like state, modulating the glycosylatability of a peptide in a capacity secondary to the −1 position.

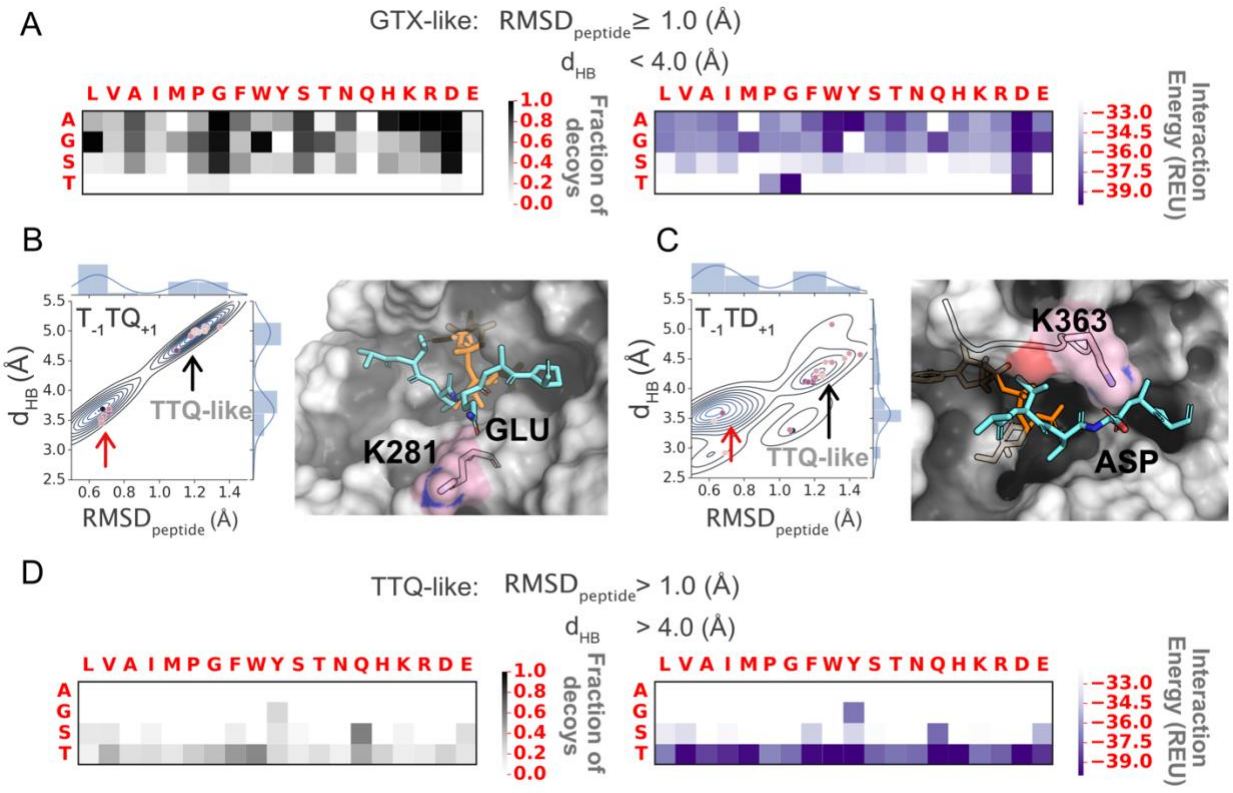

Figure 6. Secondary effects of the amino acid at the +1 position on conformations sampled by the peptide (A) GTX-like state. (B) Joint and marginal probability densities of top 10%(200/2000) sequons for $T_{-1}TQ_{+1}$ (left panel) and lowest energy decoy for TTQ-like state for sequon $T_{-1}TQ_{+1}$ state, where Q_{+1} position interacts with K281. (C) Joint and marginal probability densities of top 10%(200/2000) sequons for $T_{-1}TD_{+1}$ (left panel) and lowest energy decoy for TTQ-like state for sequon $T_{-1}TD_{+1}$ state, where D_{+1} position interacts with K363. (D) TTQ-like state. Top 1% (20/2000) decoys per sequon shown as points in (B) and (C) where darker color indicates lower interaction energy. "TTX-like" state is marked with a black arrrow and "PTX-state" is marked with a red arrow.

Residues glutamine, glutamate, aspartate and the aromatics at the +1 position interact with residues K363/K281 on the enzyme to form competing states

To understand the variation in glycosylatability for T_{-1} peptides with the +1 position, we considered the sequons $T_{-1}TQ_{+1}$, $T_{-1}TF_{+1}$, $T_{-1}TY_{+1}$ and $T_{-1}TW$. Experimentally, $T_{-1}TQ_{+1}$ was glycosylatable with ~20% activity, whereas $T_{-1}TF_{+1}$, $T_{-1}TY_{+1}$ and $T_{-1}TW$ were non-glycosylatable. All four sequons exhibited the PTX-like state (red arrow in Figure 6B and SI Figure 11). These

sequons additionally exhibited a second, low-energy state with $RMSD_{peptide}$ > 1.0 and > 4.0 Å (black arrow in Figure 6B and SI Figure 11). We designated this state the TTQ-like state since it was highly pronounced for the $T_{-1}TQ_{+1}$ sequon. In the TTQ-like state, the residue T_{-1} occupied the −1 pocket similar to the PTX-like and GTX-like states, while the residue Q_{+1} interacted with residue K281 on the enzyme, which lies at the rim of the peptide-binding groove (Figure 6B). The interaction between the residues Q_{+1} and K281 pulled the peptide backbone away from the catalysis site (Figure 6B), resulting in a non-reactive state that competed with the reactive PTX-like state.

We observed a similar interaction for residue D_{+1} (sequons $T_{-1}TD_{+1}$ and $P_{-1}TD_{+1}$), however, due to a shorter side chain compared to Q_{+1}, it was in a better position to interact with K363 residue (Figure 6C, SI Figure 11).

In Figure 6D, we show the sampling of the TTQ-like state for A_{-1}, G_{-1}, S_{-1}, and T_{-1} peptides. The TTQ-like state was exhibited primarily by S_{-1} and T_{-1} peptides. F_{+1}, Y_{+1}, W_{+1}, E_{+1}, Q_{+1}, H_{+1}, and D_{+1} exhibited highly stabilized TTQ-like states. We also observed the TTQ-like state for non-polar residues such as methionine and isoleucine at the +1 position, a result of non-polar interactions of the +1 side chain with the K281 side chain.

To compare the stability of the PTX-like and TTQ-like states for sequons that exhibited both states, we computed the difference between the lowest-energy decoys for the two state (SI Figure 12C). For sequons $T_{-1}TD_{+1}$, $T_{-1}TW_{+1}$ and $T_{-1}TY_{+1}$, the lowest interaction energy of the TTQ-like state was about 2 REU lower than that of the PTX-like state. For $T_{-1}TQ_{+1}$, the difference was small (−0.2 REU), and for $T_{-1}TE_{+1}$, the PTX-like state was more stable by 2.4 REU. The relative stabilization of the PTX-like state over the TTQ-like state as measured by interaction energy correlated with higher experimental glycosylation efficiencies for sequons $T_{-1}TQ_{+1}$ (~ 22%) and $T_{-1}TE_{+1}$ (~13%) compared to $T_{-1}TD_{+1}$ (3%) and $T_{-1}TX_{+1}$ (0%), where X was an aromatic residue.

To quantify the interaction energy of different amino acid residues at the +1 position to specific residues on the enzyme, we computed the pairwise energies of interaction between the residue at the +1 position and the enzyme (SI Figure 13) and, as expected, found that the residues that

exhibit the TTQ-like state interact favorably with residues K281 or K363 on the enzyme. On the other hand, residues P_{+1} and A_{+1} did not interact with K281 or K363 residues on the enzyme. The lack of interaction with K281 or K363 residues on the enzyme suggested that sequons $T_{-1}TP_{+1}$, $T_{-1}TA_{+1}$, $S_{-1}TP_{+1}$ and $S_{-1}TA_{+1}$ had no propensity for the TTQ-like state and may explain the high glycosylation efficiencies observed for these sequons. We summarize the changes in classification accuracy in SI Table 2 and SI Figure 10C upon inclusion of a TTQ-like state-based criterion.

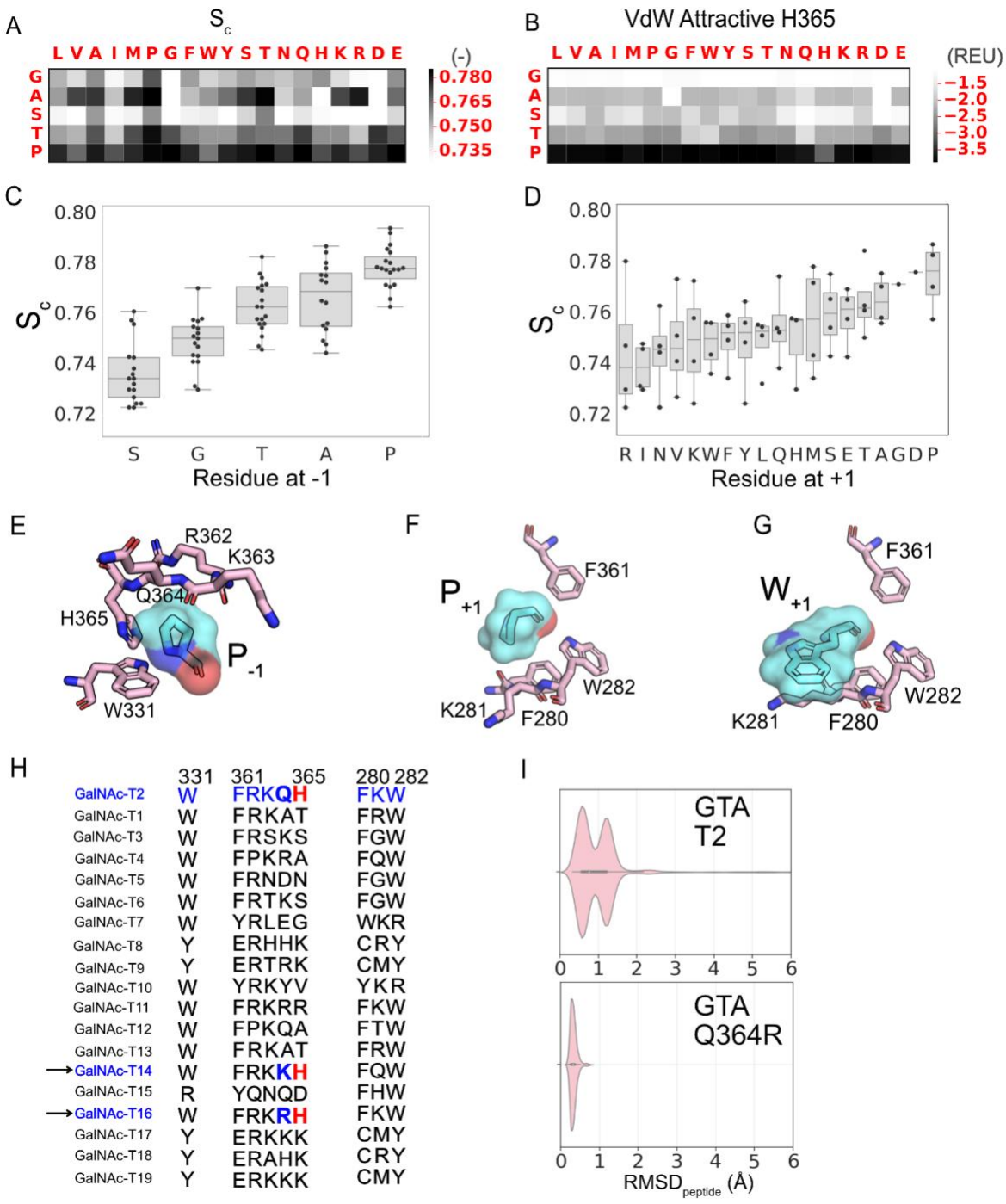

Figure 7. Characterization of enzyme–peptide interactions for top 10 decoys A_{-1}, P_{-1}, G_{-1}, S_{-1}, T_{-1} peptides. (A) Median S_c. (B) Attractive component of the van der Waals (VdW) potential in Rosetta score function between the residue at the −1 position and H365 on the enzyme. (C) Distribution of median S_c values as a function of the residue at −1 position. (D) Distribution of median S_c values as a function of the residue at +1 position. (E) +1 pocket of at the enzyme peptide interface with H365 on the enzyme (pink) interacting with the proline at the −1 position on the peptide (aquamarine). (F) Residues 280, 281, 282 and 361 on the enzyme (pink) interacting with proline

at the +1 position on the peptide (aquamarine). (G) Residues 280, 281, 282 and 361 on the enzyme (pink) interacting with tryptophan at the +1 position on the peptide (aquamarine). (H) Multiple sequence alignment of isoform T2 with other isoforms for the residues at the enzyme-peptide interface for +1 and −1 positions on the peptide. (I) Violinplots for $RMSD_{peptide}$ distributions sampled for sequon GTA for isoform T2 (top) and a variant T2 (Q364R) (bottom). Residue numbering based on GalNAcT2 Uniprot entry Q10471.

Characterization of the peptide-enzyme interface

Shape complementarity and hydrogen bonding contribute to the finely tuned specificities at the −1 and +1 positions

Our analysis so far focused on analyzing the landscape of low-energy conformations exhibited by the peptides and on recapitulating the experimentally observed specificity trends as a function of the amino acid at the −1 and +1 position of the sequon. In this process, we discovered the dominant modes of interaction between the peptide and the enzyme that possibly lead to various reactive (PTX-like state) and non-reactive (GTX-like and TTQ-like state) conformations. Comparison between experimental data and computational predictions also revealed that a majority of sequons that were glycosylatable exhibited a PTX-like conformation. In this section, we characterized the PTX-like state to decipher the structural basis for the variation of specificity within the subset of peptides that exhibited this state.

First, we calculated the shape complementarity statistic,[40] S_c, for the enzyme-peptide interface for all sequons for top 10 decoys (lowest interaction energies) that satisfied the $RMSD_{peptide}$ < 1.0 Å criterion (Figure 7A). We found that P_{-1} peptides exhibited the highest shape complementarity at the peptide-enzyme interface (Figure 7C). The P_{-1} residue packed against the planar interface formed by a histidine residue at position 365 on the enzyme (Figure 6E). We further characterized the residue-wise and pairwise interaction energies at the interface. The P_{-1} residue exhibited significantly higher attractive van der Waals energy (fa_atr in Rosetta) with H365 of the enzyme than any other residues at the −1 position (Figure 7B). The P_{-1} residue exhibited generally higher

attractive van der Waals energies with all enzyme residues at the interface (SI Figure 14). T_{-1}, S_{-1}, and A_{-1} residues exhibited energies (SI Figure 14, Figure 7B) and shape complementarities (Figure 7A, C) that varied to a significant extent with the residue at the +1 position. This suggested that the +1 position may additionally contribute to anchoring the peptide in the binding cavity for these sequons.

For the +1 position, proline exhibited the highest shape complementarity (Figure 7D) in the "+1 pocket" formed by three aromatics F280, W282 and F361 stabilized by favorable interactions between the partially positively charged proline ring and the partially negatively charged π faces of aromatic side chains (Figure 7F, SI Figure 15). Also, similar to the TTQ-like state, sequons with aromatics, glutamine, glutamate and non-polar residues other than alanine, proline and glycine at the +1 position interacted with K281 on the enzyme (Figure 7G).

For T_{-1} and S_{-1} residues, the PTX-like state was additionally stabilized by a hydrogen bond between the hydroxyl side chain and the backbone carboxyl of the arginine residue (R362) on the enzyme (SI Figure 16).

The shape complementarity and pairwise-energies point to a −1 pocket that was highly specific for the P_{-1} residue. This observation aligns well with the high experimental glycosylation efficiencies measured for P_{-1} peptides.

Sequence motifs at the −1 pocket hint at modes of specificity modulation across isoforms T2, T14 and T16

The −1 pocket on the enzyme plays an important role in screening for optimally-sized side chains at the −1 position of the sequon. This pocket is primarily formed by residues R362, K363, Q364, H365 and W331. These residues determine the size and chemical composition of the −1 pocket. The residues R362, K363, Q364, and H365 reside on the flexible, semi-conserved catalytic loop[41] of the enzyme. The flap-like loop can additionally contribute to the variability of the −1 pocket size across the GalNAc-T isoforms. The H365 residue is conserved in all three isoforms (T2, T14, T16) of the GalNAc-T family that show a strong preference for P_{-1} (Figure 7H). Residues K363 and Q364 reside at the point of entry for the −1 residue on the peptide. Variation of amino acid

residue at these positions could possibly allow for variation in the size of the amino acid preferred by an isoform at the −1 position of the sequon. For example, isoforms T14 and T16 which are evolutionary most proximal to the T2 isoform have residues lysine or arginine at position 364. Both isoforms, unlike the T2 isoform, preferred G_{-1}; indicative of a −1 pocket suitable for smaller sidechains. In fact, when we repeated MCM sampling of the T2 isoform with the Q364R mutation, for the G_{-1} position, we observed a complete shift towards conformations with $RMSD_{peptide}$ < 1.0 Å (PTX-like state) and the elimination of the GTX-like state, suggesting a possible strategy for varying the peptide substrate preference of various isoforms (Figure 7I).

Energy-based predictors incorrectly classify P_{-1} peptides as non-glycosylatable

Energy is a commonly used metric in determining specificity of peptide substrates (eg. pepspec, sequence_tolerance and MFPred). In this work, we characterized each cluster by interaction energy and used it as the metric for choosing the "top N" decoys for further analysis. However, we find, for the purpose of prediction of specificity trends for the T2 enzyme, interaction energy, by itself, was a weaker predictor than other metrics (Table *1*). The significantly lower AUC values based on interaction energy, compared to $RMSD_{peptide}$ (0.959) and fraction of decoys (0.969), were due to the fact that P_{-1} and S_{-1} peptides bind the enzyme with significantly lower interaction energies than A_{-1}, G_{-1} or T_{-1} peptides (Figure 2A).

For comparison with other energy based approaches, we applied the MFPred method by Rubenstein *et al.*[32] to obtain the specificity profile for GalNAc-T2 (SI Figure 17). We obtained an AUC score of 0.76 with this approach, which was significantly lower than the AUC scores obtained in this work (Table *1*, SI Table 3). Notably, MFPred did not correctly classify P_{-} or G_{-1} peptides though it accurately predicted the preference for T_{-1} and S_{-1} residues. The low classification accuracy of energy-based predictors suggested that the stability of the peptide-enzyme complex or the interaction-energy at the interface, by itself, was a weak indicator of efficient catalysis by GalNAcT-2. In fact, since selective stabilization of the transition state over the reactants is important for catalysis, the over-stabilization of the reactant state (indicated by higher interaction energies) may increase the free energy of activation (difference between the energy

of reactant and the transition state) thereby preventing the reaction from proceeding forward or slowing down the kinetics of reaction.

Table 1. Summary of AUC scores for predictions based on features with a range of glycosylation efficiency thresholds below which a peptide is classified as non-glycosylatable

Feature	**Experimental glycosylation efficiency threshold (%)**	**AUC Score**
$RMSD_{peptide}$ (largest cluster)	0.10	0.959
d_{HB} (largest cluster)	0.10	0.922
$RMSD_{sequon}$ (largest cluster)	0.10	0.955
Interaction Energy (largest cluster)	0.10	0.566
$RMSD_{peptide}$ (largest cluster)	0.55	0.967
d_{HB} (largest cluster)	0.55	0.954
$RMSD_{sequon}$ (largest cluster)	0.55	0.977
Fraction of decoys (d_{HB} < 4.0 Å)	0.10	0.908
Fraction of decoys ($RMSD_{peptide}$< 1.0 Å)	0.10	0.965
Interaction Energy (d_{HB} < 4.0 Å)	0.10	0.875
Interaction Energy ($RMSD_{peptide}$< 1.0 Å)	0.10	0.908
Shape Complementarity	0.10	0.853
Shape Complementarity ($RMSD_{peptide}$< 1.0 Å)	0.10	0.924
MFPred	0.10	0.760

Discussion

In this work, we attempted to understand the structural basis for the peptide substrate preferences of the T2 isoform of the GalNAc-T family. We expect this work to be useful in

understanding how the preference for different peptide substrates is modulated across the 20 isozymes of this family.

We used a flexible backbone protocol with MCM sampling which resulted in more than one low-energy peptide conformation/state in the vicinity of the starting peptide conformation obtained from the crystal structure. Most existing protocols for determining peptide specificity of PBDs employed limited backbone sampling, generating ensembles close to the starting structures (pepspec, MFPred) and usually employing additional constraints to sample TS-like conformations. While these studies have been quite successful at predicting specificity trends, a wealth of information can be garnered from sampling the peptide landscape without imposed constraints. Our work benefitted from the availability of crystal structures for the peptide-enzyme complex but may be less accurate in the absence of crystal structures. Our approach also suffered from inaccuracies in the Rosetta energy function, the limitations of MCM sampling, and the use of implicit solvation models to name a few. We further note that an MD-based simulation, though computationally prohibitive for a large dataset, may be better suited for generating thermodynamically accurate ensembles and for characterizing the density of multiple stable states.

We find that for the T2 isoform, the −1 position on the peptide strongly determined the glycosylation efficiency. Residues R362, K363, Q364 and H365 on the catalytic loop and residue W331 on the enzyme formed the −1 pocket and selected for amino acids threonine, proline, serine, alanine or glycine at the −1 position. For sequons with residues that did not fit this pocket, the peptide was not able to form a hydrogen bond with UDP that has been proposed to stabilize the TS. We further found that this pocket was especially favorable for recognizing peptides with proline at the −1 position as demonstrated by high degree of shape complementarity irrespective of the amino acid at the +1 position and highly favorable interactions between H365 and proline. Hence, by modulating the size and other biophysical aspects of this pocket, the specificity for the −1 position can be potentially modulated. These structural and sequence features are especially relevant for specificity modulation across isoforms as the GalNAc-T family can glycosylate a wide range of amino acids at the −1 position.

We additionally found that residues K281 and K363 acted as gating residues and interacted with certain amino acid residues on the peptide, such as Q_{+1} and D_{+1}, leading to low energy states that may compete with the reactive state. Hence, the specificity for the +1 position may be modulated by altering the lysine residues at positions 281 and 363 on the enzyme. Similar to the −1 position, such variation in specificity for the +1 position was already observed in the GalNAc-T family as certain isoforms were capable of efficiently glycosylating D_{+1}.

We have identified key structural motifs in this work that may be important for designing more promiscuous forms of the enzyme or tailored forms with specificities different from those seen in the 20 naturally occurring isoforms. Furthermore, since many members of the GalNAc-T family have been associated with various cancers, the sequence and structural motifs identified in this work may be used to decipher mutations that may cause aberrant glycosylation.

Methods

Starting structure for enzyme–peptide complex

The starting structure of the enzyme–peptide complex was obtained from two crystal structures—the active conformation of the enzyme from the crystal structure of the complex (pdb id: 2ffu) and bound peptide (mEA2), manganese and UDP-GalNAc-5S from the crystal structure of the complex with the modified GalNAc (pdb id: 4d0z). The sugar is absent from the first structure, so we used the second structure for the non-hydrolyzed sugar still covalently bound to UDP. While the sugar bound to UDP in 4d0z has a modification (sulfur instead of oxygen in the ring), it aligns exactly (SI 12) with 2ffu with the additional sugar. To generate the starting structure for each sequon, we started from the crystal structure of the complex with the peptide $A_{-2}X_{-1}T_0X_{+1}A_{+2}P_{+3}R_{+4}C_{+5}$ from the work by Kightlinger *et al.* [12] (aligned to the mEA2 peptide) instead of the mEA2 peptide, were X is one of 19 amino acid residues (all amino acid residues except cysteine). Residues at positions −1 and +1 (denoted by Xs) were mutated to the target sequon for all 361 sequons studied in the work by Kightlinger *et al.*[12] using MutateResidue mover followed by side chain repacking and minimization using the PackRotamersMover. No backbone motion is allowed at this stage.

Rosetta protocol for generating decoys

The glycosylation protocol is based on the flexpepdock protocol[42] with a few modifications. The foldtree includes a jump across the peptide–enzyme interface similar to a peptide or protein–protein docking protocol. The protocol gives the option for an alternative foldtree that is centered on a user-specified residue on the peptide. This alternate foldtree allows the anchoring of glycosylated peptides at glycosylated positions for peptides with multiple glycosylated sites. We do not report any results for successive glycosylation of glycosylated peptides in this work. Additionally, the protocol supports the addition of constraints to preserve catalytic motifs in the active site. There are two main stages in the glycosylation protocol – 1) Low-resolution sampling with the centroid score 2) High-resolution refinement with the all-atom ref2015 score function. In the low-resolution phase, we use simulated annealing for enhanced sampling of the peptide. We vary the temperature from 2.0 to 0.6 in Rosetta temperature units (kT) over 30 Monte Carlo (MC) cycles. For each temperature cycle of simulated annealing, we use 50 inner MC cycles are used for perturbation followed by minimization in rigid body (across enzyme–peptide interface) and torsional(peptide) space. “Small” and “shear” movers from Rosetta are used for torsional sampling of the peptide [43]. We implemented rigid body perturbation with the RigidBodyPerturbMover. The final pose from the low-resolution stage is passed to the high-resolution stage. In the high-resolution stage, the attractive and repulsive potential weights are ramped down and up respectively over 10 outer cycles. Similar to the low-resolution stage, we apply rigid body sampling across the enzyme–peptide interface and torsional sampling of the peptide backbone followed by minimization and Metropolis criterion. Additionally, both rigid body moves (30 cycles) and torsional moves (30 cycles) are accompanied by side chain repacking of the peptide side chains every cycle and the interface side chains every 3rd cycle by the packer[43]. We used the default distance of 8 Å to define the interface. The protocol allows user-specified interface distance. We found the interface distance of 8 Å to be computationally most efficient as the runs slow down with larger distances (*e.g.* 12 Å). Additionally, the run was terminated if the peptide moved more than a user-specified distance away from the enzyme–peptide interface (default 8 Å). The backbone of the enzyme is fixed throughout sampling. We generated 2000

decoys per sequon. Larger number of decoys (8000) were not found to be particularly advantageous.

This protocol will be made available in the Rosetta suite as an application. See Supplementary Information for the steps to run the protocol.

Clustering and analysis of decoys

The top 10% decoys (200/2000) were clustered using the dbscan clustering algorithm[44,45] in sklearn with parameters set to eps=0.3 (maximum distance between samples for one to be considered in the neighborhood of the other) and min_samples=10 (number of samples in the neighborhood of a point to be considered a core point). We also tested kmeans clustering but found dbscan clustering to be more robust at assigning clusters.

Calculation of features

We report two RMSD metrics in this work – $RMSD_{peptide}$ and $RMSD_{sequon}$. Both metrics are calculated over backbone Cα atoms only with respect to the backbone of the peptide in the starting structure. For $RMSD_{peptide}$, RMSD is calculated for all peptide positions. For $RMSD_{sequon}$, RMSD is calculated for positions −1 to +3 (XTXAP). Shape Complementarity is calculated using the shape complementarity calculator in PyRosetta[46] as described in Supplementary Information. The interaction energy at the enzyme–peptide interface is calculated as the difference between the ref2015 score for the bound complex and the ref2015 score for the enzyme (includes the UDP-sugar molecule) and the peptide, separated from the complex without repacking side chains.

Specificity prediction with MFPred

We follow the protocol outlined in the MFPred study.[32] 1) The starting structure is relaxed. 2) The lowest energy decoy from relax step is used as a starting structure for the FastRelax protocol for each sequon. 3) The lowest energy decoy for each sequon from the FastRelax protocol is processed by the GenMeanFieldMover.

Acknowledgements

We thank Nadine L. Samara at National Institutes of Health for helpful discussions and for critically reading and editing the manuscript. We also thank Weston Kightlinger, Liang Lin and Michael C. Jewett at Northwestern University for helpful discussions.

Funding

This work was supported by National Institutes of Health grants R01-GM127578 and R01-GM078221.

References

1. Steen, P. Van den, Rudd, P. M., Dwek, R. A. & Opdenakker, G. Concepts and Principles of O-Linked Glycosylation. *Crit. Rev. Biochem. Mol. Biol.* **33**, 151–208 (1998).

2. Brockhausen, I. & Stanley, P. Chapter 10 O-GalNAc Glycans. *Essentials Glycobiol.* **1**, 1–9 (2017).

3. Sletmoen, M., Gerken, T. A., Stokke, B. T., Burchell, J. & Brewer, C. F. Tn and STn are members of a family of carbohydrate tumor antigens that possess carbohydrate-carbohydrate interactions. *Glycobiology* **28**, 437–442 (2018).

4. Kudelka, M. R., Ju, T., Heimburg-Molinaro, J. & Cummings, R. D. Simple sugars to complex disease-mucin-type O-glycans in cancer. in *Advances in Cancer Research* **126**, 53–135 (Academic Press Inc., 2015).

5. Ten Hagen, K. G., Fritz, T. A. & Tabak, L. A. All in the family: the UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferases. *Glycobiology* **13**, 1R – 16 (2003).

6. Gerken, T. A., Raman, J., Fritz, T. A. & Jamison, O. Identification of common and unique peptide substrate preferences for the UDP-GalNAc:polypeptide α-N-acetylgalactosaminyltransferases T1 and T2 derived from oriented random peptide substrates. *J. Biol. Chem.* **281**, 32403–32416 (2006).

7. Gerken, T. A., Ten Hagen, K. G. & Jamison, O. Conservation of peptide acceptor

preferences between Drosophila and mammalian polypeptide-GalNAc transferase ortholog pairs. *Glycobiology* **18**, 861–70 (2008).

8. Perrine, C. L. *et al.* Glycopeptide-preferring polypeptide GalNAc transferase 10 (ppGalNAc T10), involved in mucin-type O-glycosylation, has a unique GalNAc-O-Ser/Thr-binding site in its catalytic domain not found in ppGalNAc T1 or T2. *J. Biol. Chem.* **284**, 20387–20397 (2009).

9. Bennett, E. P. *et al.* Control of mucin-type O-glycosylation: a classification of the polypeptide GalNAc-transferase gene family. *Glycobiology* **22**, 736–56 (2012).

10. Fritz, T. A., Raman, J. & Tabak, L. A. Dynamic Association between the Catalytic and Lectin Domains of Human UDP-GalNAc:Polypeptide α-N-Acetylgalactosaminyltransferase-2. *J. Biol. Chem.* **281**, 8613–8619 (2006).

11. Lira-Navarrete, E. *et al.* Dynamic interplay between catalytic and lectin domains of GalNAc-transferases modulates protein O-glycosylation. *Nat. Commun.* **6**, (2015).

12. Kightlinger, W. *et al.* Design of glycosylation sites by rapid synthesis and analysis of glycosyltransferases article. *Nat. Chem. Biol.* **14**, 627–635 (2018).

13. Gerken, T. A. *et al.* Emerging paradigms for the initiation of mucin-type protein O-glycosylation by the polypeptide GalNAc transferase family of glycosyltransferases. *J. Biol. Chem.* **286**, 14493–14507 (2011).

14. Gómez, H. *et al.* A computational and experimental study of O-glycosylation. Catalysis by human UDP-GalNAc polypeptide:GalNAc transferase-T2. *Org. Biomol. Chem.* **12**, 2645–2655 (2014).

15. Trnka, T., Kozmon, S., Tvaroška, I. & Koča, J. Stepwise Catalytic Mechanism via Short-Lived Intermediate Inferred from Combined QM/MM MERP and PES Calculations on Retaining Glycosyltransferase ppGalNAcT2. *PLoS Comput. Biol.* **11**, (2015).

16. Lira-Navarrete, E. *et al.* Substrate-Guided Front-Face Reaction Revealed by Combined Structural Snapshots and Metadynamics for the Polypeptide *N* -

Acetylgalactosaminyltransferase 2. *Angew. Chemie Int. Ed.* **53**, 8206–8210 (2014).

17. Fernandez, A. J. *et al.* The structure of the colorectal cancer-associated enzyme GalNAc-T12 reveals how nonconserved residues dictate its function. *Proc. Natl. Acad. Sci.* **116**, 20404–20410 (2019).

18. Mak, W. S. & Siegel, J. B. Computational enzyme design: Transitioning from catalytic proteins to enzymes. *Current Opinion in Structural Biology* **27**, 87–94 (2014).

19. Kundert, K. & Kortemme, T. Computational design of structured loops for new protein functions. *Biological Chemistry* **400**, 275–288 (2019).

20. Frushicheva, M. P., Cao, J. & Warshel, A. Challenges and advances in validating enzyme design proposals: The case of kemp eliminase catalysis. *Biochemistry* **50**, 3849–3858 (2011).

21. Pauling, L. Molecular architecture and biological reactions. *Chem. Eng. News* (1946). doi:10.1021/cen-v024n010.p1375

22. Leaver-Fay, A., Jacak, R., Stranges, P. B. & Kuhlman, B. A generic program for multistate protein design. *PLoS One* **6**, e20937 (2011).

23. St-Jacques, A. D., Eyahpaise, ve C. & Chica, R. A. Computational Design of Multisubstrate Enzyme Specificity. **14**, 15 (2019).

24. Antoniou, D. & Schwartz, S. D. Protein dynamics and enzymatic chemical barrier passage. *J. Phys. Chem. B* **115**, 15147–15158 (2011).

25. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chemical Biology* **5**, 789–796 (2009).

26. Brouk, M. *et al.* The influence of key residues in the tunnel entrance and the active site on activity and selectivity of toluene-4-monooxygenase. *J. Mol. Catal. B Enzym.* **66**, 72–80 (2010).

27. Lee, H.-L., Chang, C.-K., Jeng, W.-Y., Wang, A. H.-J. & Liang, P.-H. Mutations in the

substrate entrance region of β-glucosidase from Trichoderma reesei improve enzyme activity and thermostability. *Protein Eng. Des. Sel.* **25**, 733–740 (2012).

28. Li, Z. & Scheraga, H. A. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 6611–6615 (1987).

29. King, C. A. & Bradley, P. Structure-based prediction of protein-peptide specificity in rosetta. *Proteins Struct. Funct. Bioinforma.* **78**, 3437–3449 (2010).

30. Smith, C. A. & Kortemme, T. Predicting the tolerated sequences for proteins and protein interfaces using rosettabackrub flexible backbone design. *PLoS One* **6**, (2011).

31. Smith, C. A. & Kortemme, T. Structure-Based Prediction of the Peptide Sequence Space Recognized by Natural and Synthetic PDZ Domains. *J. Mol. Biol.* **402**, 460–474 (2010).

32. Rubenstein, A. B., Pethe, M. A. & Khare, S. D. MFPred: Rapid and accurate prediction of protein-peptide recognition multispecificity using self-consistent mean field theory. *PLoS Comput. Biol.* **13**, e1005614 (2017).

33. Chaudhury, S. & Gray, J. J. Identification of structural mechanisms of HIV-1 protease specificity using computational peptide docking: implications for drug resistance. *Structure* **17**, 1636–1648 (2009).

34. Zheng, F. *et al.* Computational design of selective peptides to discriminate between similar PDZ domains in an oncogenic pathway. *J. Mol. Biol.* **427**, 491–510 (2015).

35. Raveh, B., London, N. & Schueler-Furman, O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins Struct. Funct. Bioinforma.* **78**, NA-NA (2010).

36. Pethe, M. A., Rubenstein, A. B. & Khare, S. D. Large-Scale Structure-Based Prediction and Identification of Novel Protease Substrates Using Computational Protein Design. *J. Mol. Biol.* **429**, 220–236 (2017).

37. Pethe, M. A., Rubenstein, A. B. & Khare, S. D. Data-driven supervised learning of a viral protease specificity landscape from deep sequencing and molecular simulations. *Proc.*

Natl. Acad. Sci. U. S. A. **116**, 168–176 (2019).

38. Leaver-Fay, A. *et al.* Rosetta3. in *Methods in enzymology* **487**, 545–574 (2011).

39. Leman, K. *Macromolecular modeling and design in Rosetta: new methods and frameworks*. (2019).

40. Lawrence, M. C. & Colman, P. M. Shape complementarity at protein/protein interfaces. *Journal of Molecular Biology* **234**, 946–950 (1993).

41. De Las Rivas, M. *et al.* Structural and Mechanistic Insights into the Catalytic-Domain-Mediated Short-Range Glycosylation Preferences of GalNAc-T4. *ACS Cent. Sci.* **4**, 1274–1290 (2018).

42. Raveh, B., London, N. & Schueler-Furman, O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins Struct. Funct. Bioinforma.* **78**, NA-NA (2010).

43. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein Structure Prediction Using Rosetta. *Methods Enzymol.* **383**, 66–93 (2004).

44. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. (1996).

45. Schubert, E., Sander, J., Ester, M., Kriegel, H. P. & Xu, X. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.* **42**, 1–21 (2017).

46. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–691 (2010).