nature chemical biology

ARTICLES

https://doi.org/10.1038/s41589-018-0051-2

Design of glycosylation sites by rapid synthesis and analysis of glycosyltransferases

Weston Kightlinger[1,2,7], Liang Lin[2,3,7], Madisen Rosztoczy[3], Wenhao Li[3], Matthew P. DeLisa[4,5], Milan Mrksich[1,2,3,6*] and Michael C. Jewett[1,2*]

Glycosylation is an abundant post-translational modification that is important in disease and biotechnology. Current methods to understand and engineer glycosylation cannot sufficiently explore the vast experimental landscapes required to accurately predict and design glycosylation sites modified by glycosyltransferases. Here we describe a systematic platform for glycosylation sequence characterization and optimization by rapid expression and screening (GlycoSCORES), which combines cell-free protein synthesis and mass spectrometry of self-assembled monolayers. We produced six N- and O-linked polypeptide-modifying glycosyltransferases from bacteria and humans in vitro and rigorously determined their substrate specificities using 3,480 unique peptides and 13,903 unique reaction conditions. We then used GlycoSCORES to optimize and design small glycosylation sequence motifs that directed efficient N-linked glycosylation in vitro and in the *Escherichia coli* cytoplasm for three heterologous proteins, including the human immunoglobulin Fc domain. We find that GlycoSCORES is a broadly applicable method to facilitate fundamental understanding of glycosyltransferases and engineer synthetic glycoproteins.

Protein glycosylation is the post-translational attachment of oligosaccharides (glycans), most commonly at asparagine (N-linked) or serine and threonine (O-linked) amino acid side chains[1,2]. Glycosylation is found in all domains of life and plays critical roles in cellular function[2]. Glycosylation is also present in 70% of approved or preclinical protein therapeutics[3] and has profound effects on protein stability[4], immunogenicity[5], and potency[6], motivating the close study and intentional engineering of glycosylation sites and structures[7]. Production of glycoproteins within native hosts often results in structural heterogeneity, limits titers and genetic tractability, and constrains the diversity of glycans that can be produced[8–10]. These difficulties have led to the development of highly engineered glycosylation systems within mammalian cells[11], yeast[12], and bacteria[8,9], as well as in vitro[6,13] to produce more homogeneous human-like glycans for therapeutics[12], bacterial glycans for vaccines[9], and synthetic glycans for fundamental biology studies[6,10].

Despite these advances, major glycoengineering challenges and gaps in understanding of natural glycosylation systems still remain due in large part to a lack of high-throughput methods for synthesis and detailed biochemical characterization of glycosyltransferases (GTs), the enzymes that attach and elaborate glycans on proteins. GTs are the catalytic nodes of natural glycosylation systems and are the parts from which synthetic glycosylation systems are constructed; yet, less than 1% of putative GTs have been biochemically characterized[14], and far fewer are understood at sufficient depth to be useful in biocatalysis[15]. Typically, studies of GT specificity require expression and purification of the enzyme from cells; screening by incorporation of radioactively or chemically labeled sugars[16,17], antibody detection[17,18], or liquid chromatography (LC) separation[19]; and validation by MS (usually LC–MS/MS)[20]. Existing methods are particularly problematic for characterizing GTs that attach glycans to polypeptides. These polypeptide GTs (ppGTs) include the O-linked polypeptide *N*-acetylgalactosaminyltransferase (ppGalNAcT), O-linked *N*-acetylglucosamine transferase (OGT), and oligosaccharyltransferase (OST) enzyme families. Such enzymes are of particular interest because they determine which sites on a protein are glycosylated and constrain the possible glycoforms that can be installed. A recently discovered ppGT called *N*-glycosyltransferase from the bacterial pathogen *Actinobacillus pleuropneumoniae* (NGT), has elicited a great deal of interest for biocatalysis[9,21,22] because it is a soluble, cytoplasmic enzyme that can efficiently install an N-linked glucose on N-X-S/T amino acid sequence motifs resembling those in eukaryotic proteins using uracil-diphosphate-glucose (UDP-Glc) as a sugar donor[23]. Although pioneering efforts by several groups reported several protein and peptide substrates that can be modified by NGT[19,20,24–27], current methods for GT analysis limit investigations of NGT as well as other ppGTs to only dozens of unique peptide substrates. These methods undersample the vast amino acid sequence space available for modification, providing incomplete information of amino acid preferences at each position surrounding the glycosylation site and the interdependency of amino acids at these positions, which are required for a full understanding of GTs in natural systems and for the rational design of efficient protein glycosylation sites.

Here we report a generalizable and systematic strategy for glycosylation sequence characterization and optimization by rapid expression and screening (GlycoSCORES). GlycoSCORES couples expression by *E. coli*–based cell-free protein synthesis (CFPS) to functional characterization of GTs with self-assembled monolayers for matrix-assisted desorption/ionization (SAMDI) mass spectrometry. This workflow enables high-throughput, label-free, quantitative analysis of peptide glycosylation without time-consuming cell lysis and protein purification. We apply the GlycoSCORES workflow to the study of NGT, to two previously uncharacterized NGT

[1]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, USA. [2]Center for Synthetic Biology, Northwestern University, Evanston, IL, USA. [3]Department of Biomedical Engineering, Northwestern University, Evanston, IL, USA. [4]Robert Frederick Smith School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, NY, USA. [5]Department of Microbiology, Cornell University, Ithaca, NY, USA. [6]Department of Chemistry, Northwestern University, Evanston, IL, USA. [7]These authors contributed equally: Weston Kightlinger, Liang Lin. *e-mail: milan.mrksich@northwestern.edu; m-jewett@northwestern.edu

homologs, human ppGalNAcT1 and ppGalNAcT2, and to human OGT (hOGT) using 3,480 unique acceptor peptides and 13,903 unique reaction conditions. We demonstrate the utility of GlycoSCORES for glycoprotein engineering by rigorously optimizing NGT acceptor sequences to inform the design of improved glycosylation sites. We identify several small glycosylation tag sequence motifs termed "GlycTags" (originally described by Imperiali[28], DeLisa[29], and others) and used them to direct efficient glycosylation of several target proteins including the *E. coli* immunity protein Im7, the *H. influenzae* autotransporter protein (HMW1ct), and the constant region (Fc) of a human immunoglobulin (IgG1) antibody. We find that glycosylation efficiencies of GlycTag sequences within proteins closely mirrored trends observed from GlycoSCORES analysis of peptides. Upon synthesis and glycosylation by NGT in the cytoplasm of living *E. coli*, protein glycosylation sites that were redesigned according to a GlycoSCORES-derived GlycTag sequence were modified more efficiently than naturally occurring glycosylation sequences and a previously identified NGT glycosylation consensus sequence[20].

Results

Development of GlycoSCORES for characterization of NGT. We selected NGT as the primary GT model to demonstrate the GlycoSCORES framework (Fig. 1) because of its potential for biocatalysis and our hypothesis that a deeper analysis of the NGT acceptor substrate sequence space would enable the rational redesign of protein glycosylation sites for improved modification by NGT. Because difficulties in protein expression and purification are key challenges for GT characterization, we chose to use *E. coli*–based crude lysate CFPS for GT expression in the GlycoSCORES workflow. CFPS can rapidly produce gram per liter quantities of many complex proteins, is compatible with liquid-handling robotics for direct coupling to our SAMDI analysis pipeline, and enables enzyme quantification and functional analysis without cell lysis, affinity tags, or purification[30]. *E. coli* lysates also lack native protein glycosylation activity, providing a blank canvas for bottom-up engineering and characterization. CFPS reactions were used to express soluble NGT at 20 °C for 20 h, at which point protein synthesis was >95% complete (Supplementary Fig. 1 and Supplementary Table 1). Soluble NGT was quantified as 814 ± 97 μg/mL by [^{14}C]leucine incorporation and visualized by SDS–PAGE autoradiogram (Supplementary Fig. 2).

We next developed a SAMDI method for high-throughput analysis of NGT peptide specificity. The SAMDI method uses alkanethiolate self-assembled monolayers (SAMs) to capture enzyme reaction substrates and products, which are purified on-chip, detected, and quantified by matrix-assisted laser desorption/ionization mass spectrometry (MALDI–MS). We previously demonstrated that the SAMDI assay could be used to profile the substrate specificities of several enzyme classes including deacetylases[31], acetyltransferases[32], and DNA ligases[33]. We also reported a SAMDI screening of GT activities using monolayers presenting 24 immobilized sugars[15]. Previous works[34,35] by the Sabine group have also reported the activity of the polypeptide *N*-acetylgalactosylamine transferase 2 (ppGalNAcT2) on immobilized peptides, though with limited numbers of substrates. Here, we synthesized peptide sequence libraries to test NGT activity by solid-phase peptide synthesis (SPPS) using a C-terminal Cys residue for specific immobilization via a Michael addition reaction onto SAMs that present maleimide groups against a background of tri(ethylene glycol) groups (Fig. 2a). We then performed in vitro glycosylation (IVG) reactions by adding NGT and a UDP-Glc sugar donor to peptide libraries. Completed IVG reactions were transferred onto SAMDI plates for immobilization onto 384 maleimide SAM spots followed by rinsing and MALDI–MS of the SAMs, which provided distinct mass peaks for alkyl disulfides terminated in unmodified and glycosylated peptides. Integration of these peaks (with adjustments for ionization suppression due to glycosylation; see Methods and Supplementary Table 2) provided the glycosylation efficiency of each reaction (Fig. 2a). We validated this approach using NGT purified from BL21(DE3) *E. coli* (Supplementary Fig. 3) and truncated versions of the previously used NGT peptide substrate GSDQNATF[19]. We observed the appearance of a new MS signal at + 162 *m/z* (the mass of a glucose moiety) in peptides as short as CNAT and NATC (Supplementary Fig. 4).

Using the methods above, we then created the complete GlycoSCORES platform by reacting NGT synthesized in crude lysate CFPS with peptide substrates in solution, which were then captured on-chip and directly analyzed by SAMDI (Fig. 2a). We evaluated the activity of NGT produced in CFPS against a peptide library having the sequence $X_{-1}NX_{+1}TRC$, where X_{+1} and X_{-1} are one of 19 amino acids (Cys is excluded) (Fig. 2b and Supplementary Fig. 5). The glycosylation efficiency of these peptide sequences varied from no observed activity to nearly 100% conversion. NGT

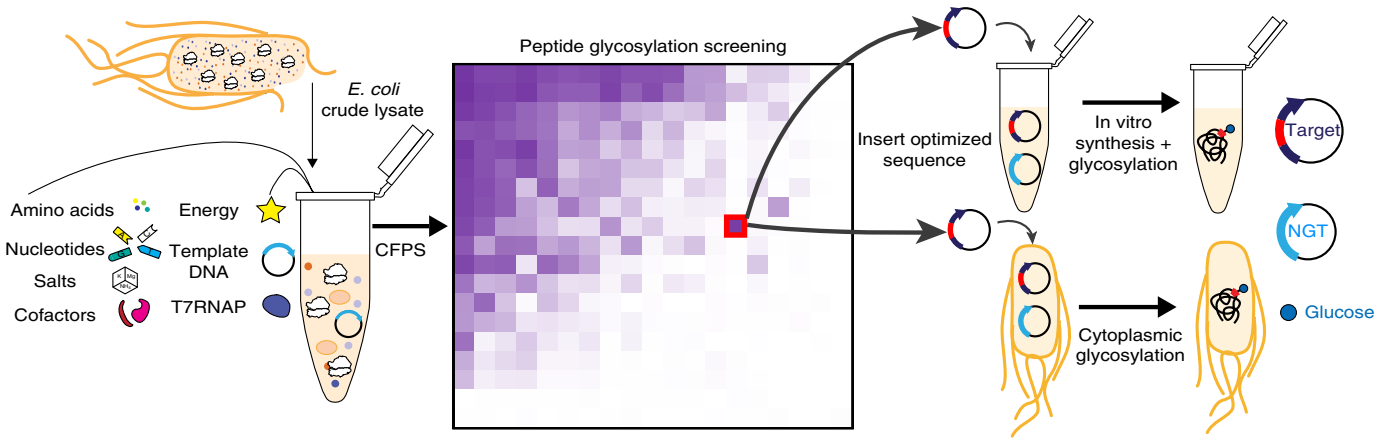

Fig. 1 | Strategy for characterizing and designing glycosylation sites. Peptide acceptor and sugar donor preferences of *N*-glycosyltransferases from *A. pleuropneumoniae* (NGT), *Mannheimia haemolytica,* and *Haemophilus ducreyi*; two human polypeptide *N*-acetylgalactosyltransferases; and the human O-GlcNAc transferase were characterized with 3,480 unique peptides and 13,903 unique reactions using SAMDI and enzymes produced by CFPS. Optimized sequences from NGT were used to design glycosylation sites on three heterologous proteins. Proteins were synthesized and glycosylated in vitro and in the cytoplasm of living *E. coli*.

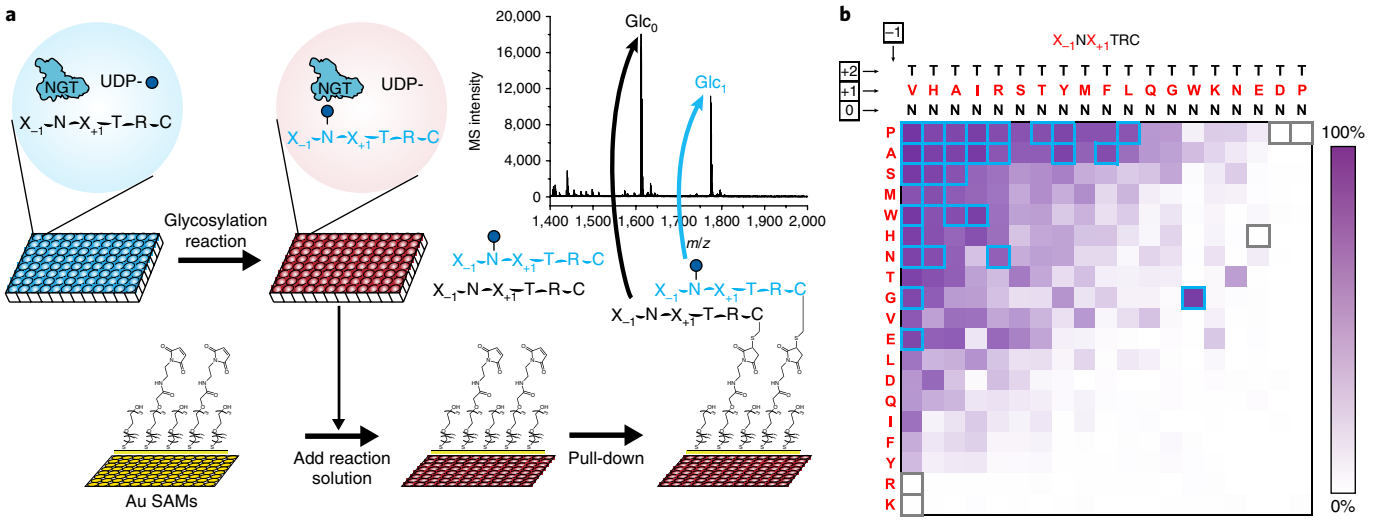

Fig. 2 | GlycoSCORES workflow and application to X_{-1} and X_{+1} position screening of NGT substrates. a, Scheme of GlycoSCORES. NGT produced in CFPS was reacted with a peptide acceptor and a sugar donor in 384-well plates. Reaction products were applied directly to self-assembled maleimide-functionalized alkanethiolate monolayers on gold (Au) islands. Peptides were captured to the monolayer by a terminal Cys residue via Michael addition. A MALDI–MS spectrum of this monolayer shows the addition of a single glucose residue (+162 Da) on a representative peptide. **b**, NGT peptide acceptor preferences were determined at the X_{-1} and X_{+1} positions relative to the modified Asn residue using the scheme in **a**. All NGT heat maps are shaded by percent conversion of peptide substrates and arranged by the mean glycosylation efficiency of each row and column. NGT sequences showing high modification (blue boxes, > 66%) and low modification (gray boxes, <3%) were chosen for subsequent screens toward an optimized GlycTag. Glycosylation reactions were performed with 50 μM peptide, 2.5 mM UDP-glucose, and 0.2 μM CFPS NGT incubated at 30 °C for 3 h. Conversion efficiencies were determined by integration of substrate and product mass spectra peaks with adjustment for measured relative ionization factor (RIFs) in Supplementary Table 2. Numerical values of modification efficiencies from SAMDI-MS spectra acquired from $n = 2$ peptide immobilizations are shown in Supplementary Fig. 5.

preferred Pro and Ala at the X_{-1} position and Val, His, Ala, and Ile at the X_{+1} position. Low modification was observed for Lys and Arg (basic residues) at the X_{-1} position, for Pro (a conformationally constrained residue) at the X_{+1} position, and for Asp and Glu (acidic residues) at the X_{+1} position (Fig. 2b). Our results extend recent findings that charged amino acids adjacent to the NGT modification site are generally disfavored[24] by showing that positively charged residues are most disfavored in the X_{+1} position whereas negatively charged residues are most disfavored in the X_{-1} position.

Interestingly, the preference of NGT for a certain amino acid at a given sequence position is dependent on nearby amino acids. For example, Trp in the X_{+1} position is generally not well tolerated; however, the motif GNWTRC was among the most efficiently glycosylated sequences (Fig. 2b). These unexpected dependencies necessitate the sampling of large, combinatorial libraries to fully understand glycosylation preferences, which requires a high-throughput assay such as GlycoSCORES. Importantly, NGT produced in CFPS showed similar peptide selectivity and specific activity to purified NGT across the $X_{-1}NX_{+1}TRC$ library (Supplementary Fig. 6). Control IVG reactions performed with CFPS reactions synthesizing superfolder green fluorescent protein (sfGFP) rather than NGT showed no peptide modification (Supplementary Fig. 7).

Though we focused on peptide specificity, we also showed the breadth of the platform by screening NGT activity with six different nucleotide-activated sugar donors: UDP-Glc, UDP-galactose (UDP-Gal), UDP-*N*-acetylglucosamine (UDP-GlcNAc), UDP-*N*-acetylgalactosamine (UDP-GalNAc), guanosine-diphosphate mannose (GDP-Man), and UDP-xylose (UDP-Xyl). Consistent with previous results[19], we found that NGT transferred UDP-Glc with the highest efficiency and UDP-Gal and UDP-Xyl with much lower efficiencies (Supplementary Fig. 8). We also tested the interdependency of the sugar donor on the $X_{-1}NX_{+1}T$ peptide library selectivity. Although the peptide selectivity remained comparable in each of the UDP-Glc, UDP-Gal, and UDP-Xyl donors, sugar donor identity did influence relative X_{-1} and X_{+1} residue preferences in some cases. For example, NGT is less tolerant of Trp at the X_{-1} position when transferring xylose compared to glucose or galactose.

Study of NGT homolog and human O-linked GT peptide specificities. To demonstrate the utility of GlycoSCORES for analysis of uncharacterized glycosyltransferases, we synthesized NGT homologs from the bacterial pathogens *Mannheimia haemolytica* and *Haemophilus ducreyi* (MhNGT and HdNGT) and found that they are in fact NGTs; we then determined their specificity on all possible acceptor sequences within the canonical NGT target sequence of X_{-1}-N-X_{+1}-S/T (Supplementary Figs. 9–11). Given sequence identities of 68%–76% (Supplementary Fig. 10), these enzymes show striking similarities to NGT acceptor sequence preferences. This may indicate two-fold evolutionary pressure for these enzymes to modify designated target proteins but not modify and interfere with essential cytoplasmic proteins. The discovery of NGTs with conserved specificities in these organisms motivates further studies to understand the roles these enzymes in their pathogenesis.

We also applied the GlycoSCORES workflow to determine the peptide specificities of two important O-linked human glycosylation enzymes, ppGalNAcT1 and ppGalNAcT2. These enzymes install the first sugar of mucin-like glycans, which effect the development of several cancers[36] and aberrant lipid metabolism[37]. We produced ppGalNAcT1 and ppGalNAcT2 in CFPS (Supplementary Fig. 12) and characterized them with a saturated X_{-1}-T-X_{+1}-P peptide library (Fig. 3 and Supplementary Fig. 13). In addition to corroborating previous investigations of the specificity of these enzymes on peptides in vitro[37–40] and on proteins in vivo[37,41], the throughput of GlycoSCORES allowed us to simultaneously vary both the X_{-1} and X_{+1} positions, which was difficult using conventional strategies, and obtain quantitative readouts for each combination. Our data

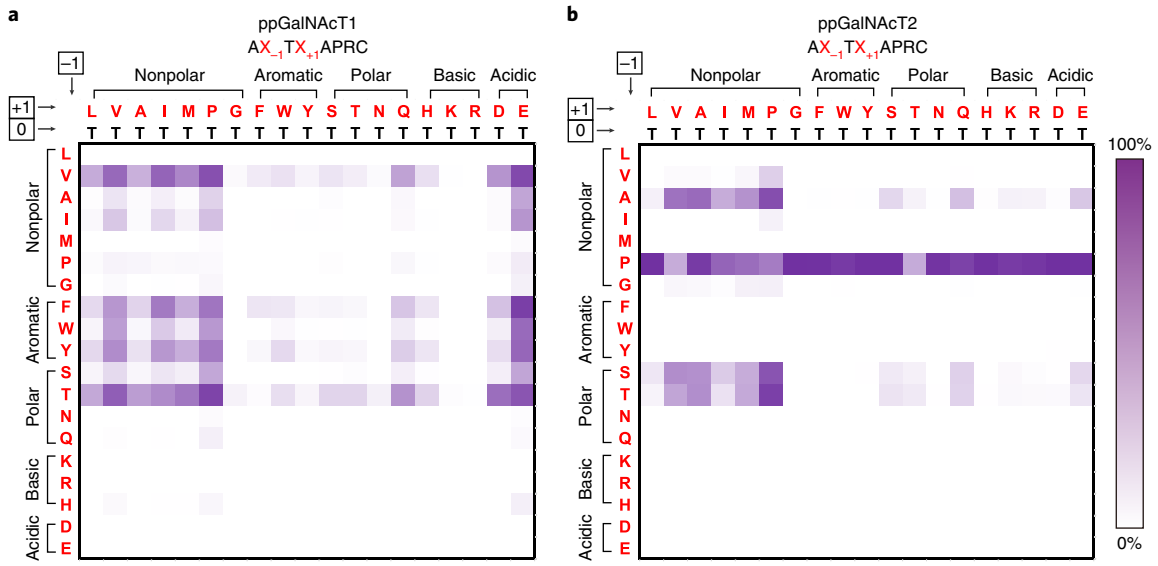

Fig. 3 | Using GlycoSCORES to determine peptide specificity of human ppGalNAcTs. a Specificity of ppGalNAcT1 produced in CFPS on a peptide array of $AX_{-1}TX_{+1}APRC$ designed from peptide AATPAP[38]. **b**, The same peptide array with ppGalNAcT2 produced in CFPS. Reaction conditions: 100 μM peptide, 1 mM UDP-GalNAc, and 0.024 μM CFPS ppGalNAcT1 in **a** or 0.04 μM CFPS ppGalNAcT2 in **b** incubated at 37 °C for 1 h. Isoforms ppGalNAcT1 and ppGalNAcT2 share 40% sequence identity (Uniprot Clustal Omega Alignment[50]) and showed major differences in peptide preferences. Peptide heatmaps in **a** and **b** are shaded by relative intensities of peptide substrates and glycosylated products averaged from $n = 2$ SAMDI-MS spectra from separate peptide immobilizations. Peptide maps annotated with numerical values and a negative control library generated using CFPS without ppGalNAcTs are shown in Supplementary Fig. 13.

led to unexpected results in the specificities of ppGalNAcT1 and ppGalNAc2. For example, we discovered that intolerance of aromatic residues adjacent to the glycosylation site of ppGalNAcT2 can be overcome by the presence of a Pro in the X_{-1} position. However, the preference of ppGalNAcT2 for Pro is reduced in cases of Val and Thr in the X_{+1} position, which themselves would predict relatively good substrates (Fig. 3). The ability to investigate ppGalNAcT specificity at this depth could enhance glycosylation site prediction and advance efforts to design isoform-specific substrates and substrate mimics.

We further demonstrated the broad applicability of GlycoSCORES by characterizing hOGT, which has been implicated in neurodegeneration and insulin resistance disease states[42,43]. We expressed hOGT in CFPS (Supplementary Fig. 14) and analyzed 19 amino acid substitutions for each position of a proposed OGT targeting sequence of PPVSR[16] (Supplementary Fig. 15). Specificity results were largely consistent with previous analyses of human glycosylation sites and peptide screens[16,44–46], such as the preference for V in the X_{-3} position and for A in the X_{+2} position. When we extended the sequence to the X_{-4} position, we discovered an overall increase in modification and a strong preference for aromatic residues. This screen provides proof of concept for future studies of hOGT variants, which can be rapidly produced in CFPS and probed for peptide specificity at high depth and throughput.

Optimization of NGT peptide acceptor sequences. To show that GlycoSCORES can be used for rigorous acceptor-peptide sequence optimization for increased modification of whole proteins, we sought to develop optimized GlycTag sequences for NGT comprised of six amino acids that could be efficiently modified in the context of whole proteins. Because the total number of 6-mer sequences is prohibitively large, we iteratively tested preferences in the X_{+2}, X_{-2}, and X_{+3} positions across a set of sequences informed by previous libraries (Fig. 4 and Supplementary Figs. 5 and 16). To determine the preference of NGT for amino acids at the X_{+2} position, we prepared an array of 380 peptides based on the motif $(X_{-1}NX_{+1})X_{+2}RC$ by selecting 20 efficiently glycosylated $X_{-1}NX_{+1}T$ sequences from Fig. 2b and resynthesizing each with 19 amino acids in the X_{+2} position (Fig. 4a). As was previously reported[19,24], we found that Thr or Ser is required in the X_{+2} position for efficient modification by NGT, with a preference for Thr over Ser. We also evaluated a complete $X_{-1}NX_{+1}S$ peptide library and did not find significant differences in X_{-1} and X_{+1} residue preferences compared to the $X_{-1}NX_{+1}T$ peptide library (Supplementary Fig. 17).

We next evaluated NGT activity for residues in the X_{-2} position using 40 sequences selected from the $X_{-1}NX_{+1}X_{+2}RC$ screens shown in Fig. 2b and Fig. 4a. To allow synergistic and compensatory mutations that would lead to overall greater activity in later screens, we balanced preservation of sequence diversity while narrowing our search toward an optimized sequence. This narrowing approach is justified by the observation that sequence modification trends generally matched those predicted by earlier screens (see blue, gray, and black boxes and text in Figs. 2b and 4). We resynthesized the 40 sequences from Fig. 2b and Fig. 4a with 19 amino acids in the X_{-2} position and performed GlycoSCORES characterization of a 760-peptide library of the form $X_{-2}(X_{-1}NX_{+1}X_{+2})RC$ (Fig. 4b). We found that Gly, Asn, and Tyr were preferred at the X_{-2} position, whereas Lys gave low activity. This library showed robust modification of the GNWT motif even when a nonpreferred residue was present in the X_{-2} position. Although Pro was a preferred residue in the X_{-1} position of the $X_{-1}NX_{+1}X_{+2}RC$ libraries (Figs. 2b and 4a and Supplementary Fig. 5), we found that most of these sequences were poorly modified with the addition of an X_{-2} residue, especially when the X_{-2} residue was Asn (Fig. 4b and Supplementary Figs. 5 and 16). This effect is likely related to the conformational constraints of Pro.

Finally, we synthesized and evaluated a library of 1,140 peptides having the sequence $(X_{-2}X_{-1}NX_{+1}X_{+2})X_{+3}RC$ with 19 amino acids in the X_{+3} position (Fig. 4c) based on 60 sequences selected from Fig. 4b. We discovered efficient glycosylation of peptides containing Thr, Met, and Phe in the X_{+3} position, but not peptides with Pro, Asp, and Glu at this position. We found that 59 acceptor peptides had greater than 70% modification efficiency after exposure to just

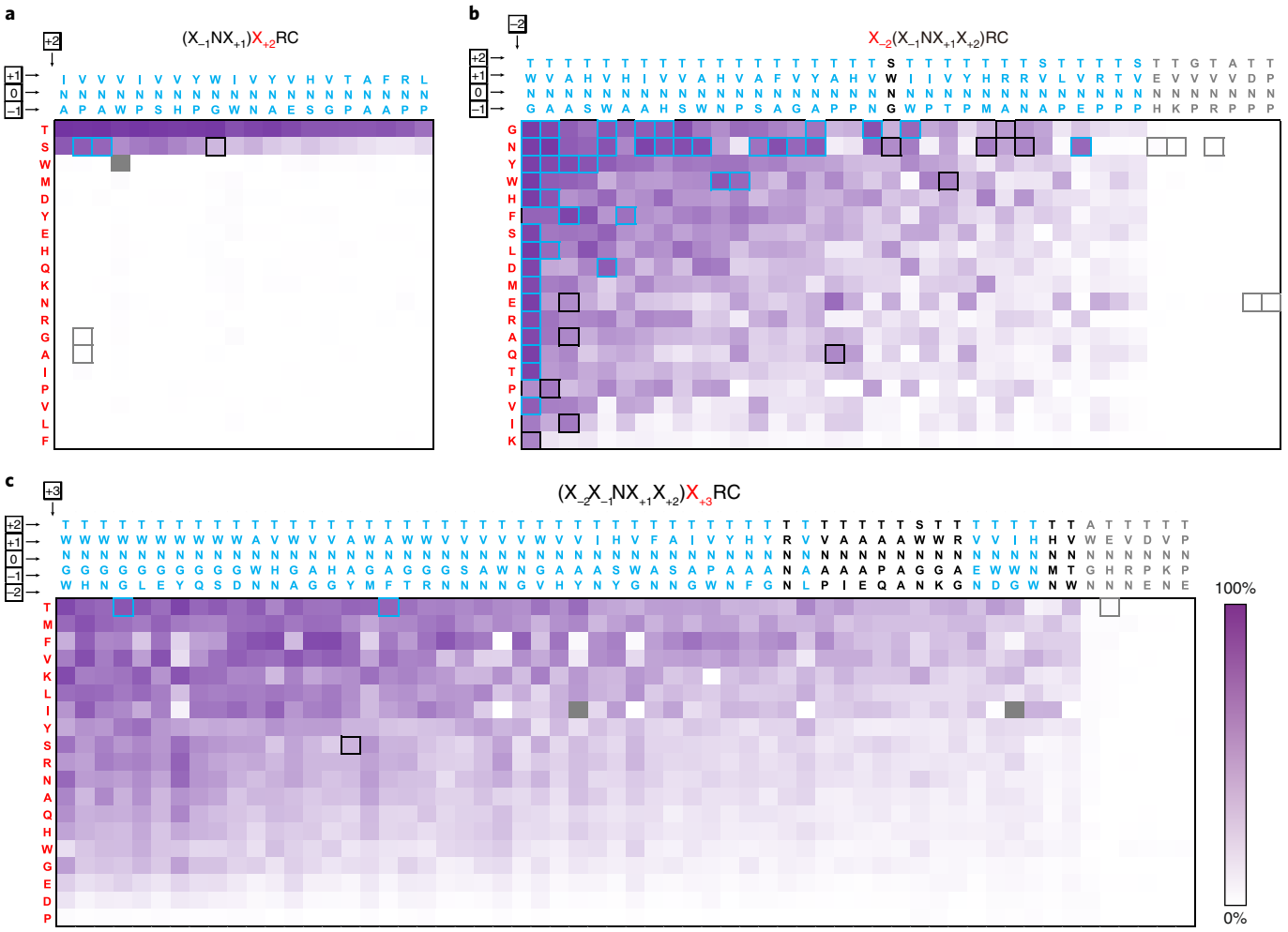

Fig. 4 | GlycoSCORES X_{+2}, X_{-2}, and X_{+3} position peptide specificity screening of NGT. a, Highly modified peptide sequences from Fig. 2b were tested with 19 amino acids in position X_{+2} relative to modified Asn. Only Thr and Ser showed efficient modification. **b**, Sequences from Fig. 2b and **a** were tested with 19 amino acids in the X_{-2} position. Only the GNWT sequence showed efficient modification with all amino acids at the X_{-2} position. **c**, Sequences from **b** were tested with 19 amino acids in the X_{+3} position. Sequences showing high (blue boxes, >66%), medium (black boxes, 33–66%), and low (gray boxes, <3%) modification efficiencies were chosen for subsequent screens toward an optimized GlycTag. Four sequences from **c** with varying levels of modification were selected for installation into proteins in Figs. 5 and 6. Peptide sequences for which modification efficiencies could not be determined due to poor peptide solubility are shown as filled gray squares in the heat map. Reactions in **a**,**b** were completed with 0.2 μM NGT for 3 h. Reaction conditions were altered to 0.025 μM NGT for 1 h in **c** to avoid saturation. Supplementary Fig. 16 shows all peptide maps completed with 0.1 μM NGT for 1 h. Numerical values of modification efficiencies from SAMDI-MS spectra acquired from $n = 2$ separate peptide immobilizations are shown in Supplementary Fig. 5. Control peptide arrays reacted with CFPS reactions without NGT are shown in Supplementary Fig. 7.

0.025 μM NGT for 1 h (Supplementary Fig. 5). We chose one efficiently modified 6-mer sequence (YANATT) to test NGT preference for the residue that undergoes glycosylation. We found that only peptides with Asn at the X_0 position showed detectable modification (Supplementary Fig. 18).

GlycTags enable efficient protein glycosylation in vitro. From the peptide screens described above, we hypothesized that preferred peptide substrates found by GlycoSCORES could be glycosylated by NGT when engineered into whole proteins. We chose the sequence GGNWTT as a model because it was found to be more efficiently glycosylated than any previously studied[19,20,23,47] 6-mer sequence that we tested (Supplementary Fig. 19). The preference of NGT for the GGNWTT sequence was unexpected because of the presence of Trp (the largest amino acid) in the X_{+1} position at the center of the required N-X-S/T glycosylation motif. We grafted the GGNWTT sequence as a GlycTag into the internal loop of Im7 at Ala28 (ref. [28]) (Im7-6) and developed an IVG method using enzymes and target proteins from CFPS (Methods and Supplementary Fig. 20). We first validated our method by recapitulating the two-enzyme, native glycosylation system of *A. pleuropneumoniae* on the HMW1ct-WT[47] target protein (Supplementary Fig. 20). We then assembled IVG reactions containing Im7-6 and NGT synthesized in separate CFPS reactions and a UDP-Glc sugar donor (Fig. 5a and Supplementary Fig. 21). We purified and analyzed the reaction product with LC–time-of-flight (LC–TOF) mass spectrometry. We found that Im7-6 (containing the GGNWTT GlycTag) was efficiently modified with a single hexose residue (Fig. 5a). Modification was not observed when NGT was absent (Fig. 5a) or when the acceptor residue was mutated from Asn to Gln (Supplementary Fig. 21).

We next investigated how the modification efficiencies of peptides in the GlycoSCORES screening correlated with the modification efficiencies of these sequences in the context of whole proteins. In addition to GGNWTT, we investigated three sequences from GlycoSCORES peptide screens, including FANATT, which showed a high glycosylation efficiency (~75%); YANATS, which showed

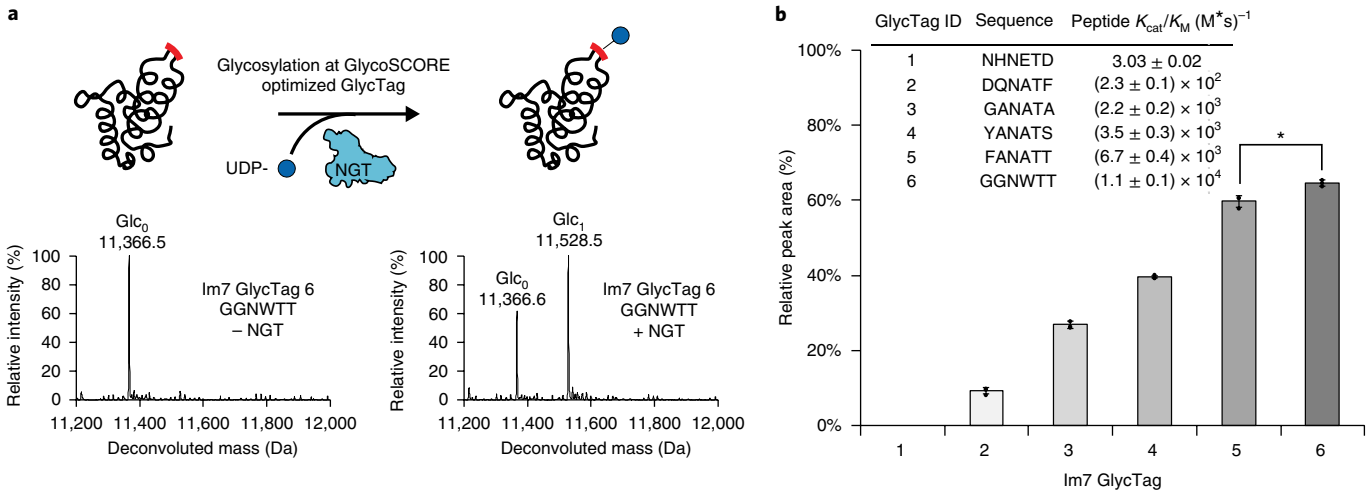

Fig. 5 | In vitro synthesis and glycosylation of Im7 with GlycoSCORES-identified sequences. Im7 GlycTag variants and NGT were synthesized in CFPS and then combined with UDP-glucose in an IVG reaction (workflow shown in Supplementary Fig. 21). **a**, Representative LC–TOF spectra from analysis of $n=3$ IVG reactions generated by maximum entropy deconvolution of the Im7-6 variant containing the GlycoSCORES-optimized GlycTag GGNWTT after Ni-NTA purification from IVG reactions with and without NGT. Representative deconvoluted spectra from all sequences and controls are shown in Supplementary Fig. 21. Deconvolution mass errors as well as chromatogram peak retention times and quantification of relative peak area for all samples are shown in Supplementary Tables 3 and 4. **b**, Relative peak areas of $Glc_1/(Glc_0+Glc_1)$ for Im7 variants containing different GlycTags. The results correlate with kinetics data measured by SAMDI for corresponding peptide sequences (inset table). Relative peak areas were calculated from extracted ion chromatograms of the three most abundant charge states based on theoretical average masses (see Methods). Mean and s.d. of $n=3$ IVG reactions are shown. $^*P=0.016$ as determined by two-tailed t-test. Details of kinetic constants are shown in Supplementary Fig. 22. Sequences positioned at the Im7 internal loop are flanked by spacer sequences of the form $ATT(X_{-2}X_{-1}NX_{+1}X_{+2}X_{+3})AGG$; see Supplementary Note 1 for sequence details. Shading of column graph areas indicates increasing relative peak areas and therefore greater glycosylation efficiencies.

a medium modification efficiency (~36%) and was used in a previous peptide study of NGT[23]; and NHNETD, which showed no detectable modification (Supplementary Fig. 5). For comparison to previous studies, we also analyzed a biological consensus sequence for NGT glycosylation (GANATA) derived from an LC–MS/MS study by Naegeli and co-workers in which NGT was expressed in the cytoplasm of *E. coli*[20], as well as the optimized PglB GlycTag sequence (DQNATF)[28], which has been used for study of NGT glycosylation at the peptide level[19]. We determined Michaelis–Menten constants for these sequences along with the GGNWTT sequence used in Fig. 5a using SAMDI (Fig. 5b and Supplementary Fig. 22). The relative values of k_{cat}/K_M correlate with GlycoSCORES conversion efficiencies observed in Fig. 4c and Supplementary Fig. 19. For each of the sequences in Fig. 5b, we synthesized a corresponding Im7 variant containing these sequences at the Ala28 internal loop. To minimize the effects of surrounding amino acids and secondary structure on the glycosylation site, we added flexible flanking sequences around all GlycTags inserted into Im7 based on a biological consensus sequence in the form of $ATT\text{-}X_{-2}X_{-1}NX_{+1}X_{+2}X_{+3}\text{-}AGG$[20] (Supplementary Note 1). The average protein expression yields of all Im7 variants was 624 ± 67 μg/mL (Supplementary Fig. 23). We performed IVG reactions and analyzed glycosylation using LC–TOF after purification. Modification was quantified by relative peak areas as in previous literature[48], using dominant charge states of the glycosylated and aglycosylated forms of the protein (Fig. 5b and Supplementary Tables 3 and 4). Relative peak areas and deconvoluted mass spectra (Supplementary Fig. 21) of Im7 variants correlated well with the k_{cat}/K_M values of peptide sequences with identical ranking (Fig. 5b). Of these sequences, GGNWTT showed the highest k_{cat}/K_M value for the peptide substrate and the most efficient modification within the Im7 protein. We also found that Im7-6 can be homogeneously glycosylated by increasing the concentration of CFPS-derived NGT in the IVG reaction to 4 μM (Supplementary Fig. 21). Therefore, we chose GGNWTT as our optimized GlycTag for site-directed protein glycosylation in vivo.

Efficient protein glycosylation in the *E. coli* cytoplasm. Next, we investigated the use of the GlycoSCORES-derived GGNWTT GlycTag to direct efficient modification of heterologous proteins in the cytoplasm of living *E. coli* by redesigning the internal protein glycosylation site at Asn297 in human Fc (Fig. 6a). NGT was co-expressed with Fc variants containing the naturally occurring sequence (QYNSTY), the biological consensus sequence (GANATA)[20], and our GlycoSCORES optimized GlycTag (GGNWTT) in vivo. As with Im7, flexible sequences flanked the site. Using a two-plasmid system in BL21(DE3) *E. coli*, we induced expression of the Fc target protein with IPTG and then did so for NGT with arabinose. We found that our GlycoSCORES-derived GlycTag (GGNWTT) enabled the most efficient glycosylation followed by the biological consensus sequence (GANATA)[20] and the naturally occurring sequence (QYNSTY) in Fc (Fig. 6b,c and Supplementary Fig. 24). We used this system to produce homogeneously glycosylated Fc in *E. coli* by extending the co-expression time of the engineered GGNWTT variant with NGT for 4 h (Supplementary Fig. 24). This same engineered variant of Fc could also be synthesized in CFPS and be efficiently glycosylated in vitro (Supplementary Fig. 25). We completed a similar analysis using a variant of HMW1ct and found that engineered HMW1ct targets also showed trends predicted by GlycoSCORES peptide characterization (Supplementary Fig. 19), with the optimized GlycTag GGNWTT showing the highest modification, followed by GANATA[20] and the naturally occurring NINATS sequence (Supplementary Fig. 26). We observed similar expression levels of NGT and Fc or HMW1ct variants across all strains, indicating that differences in glycosylation efficiency are due to NGT sequence specificity rather than differences in expression (Supplementary Fig. 27).

Discussion

This paper describes the GlycoSCORES platform, a cell-free approach for rapid determination of GT peptide specificity to improve fundamental understanding of glycosylation systems and

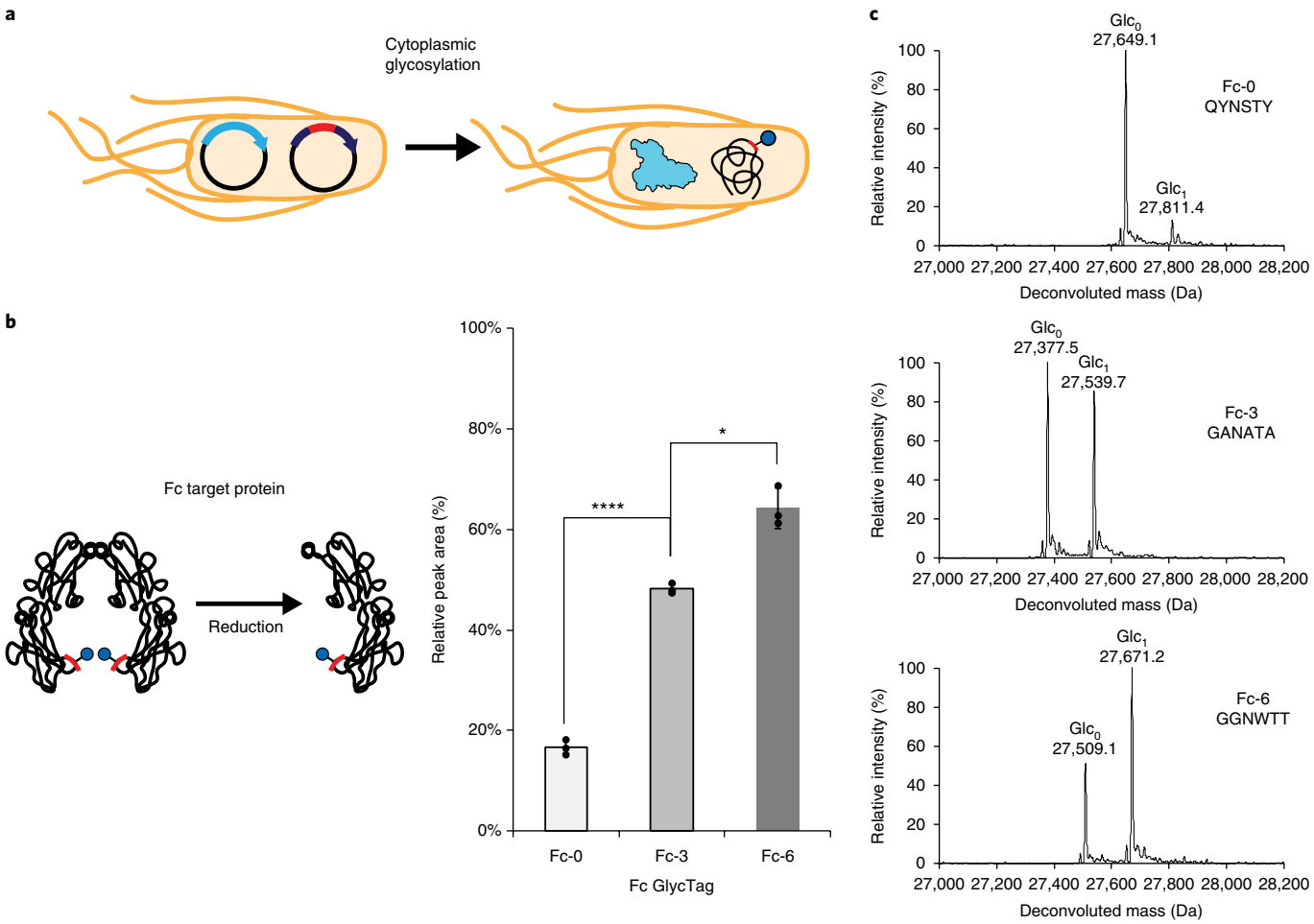

Fig. 6 | Site-directed cytoplasmic glycosylation of human Fc using GlycoSCORES optimized sequences. a, Workflow for cytoplasmic glycosylation in living *E. coli* by co-expression of NGT and target protein GlycTag variants. **b**, Relative peak areas of $Glc_1/(Glc_0 + Glc_1)$ for Fc containing naturally occurring (0), biological consensus (3), and GlycoSCORES optimized sequences (6) at the Asn297 internal loop position with flanking sequences as in $ATT(X_{-2}X_{-1}NX_{+1}X_{+2}X_{+3})AGG$. The GlycoSCORES-identified GlycTag (GGNWTT) showed higher relative peak area, indicating greater glycosylation efficiency. Fc was treated with DTT for disulfide reduction before LC–TOF analysis. Relative peak areas were calculated from extracted ion chromatograms of the eight most abundant peaks based on theoretical average masses for Fc (see Methods). Mean and s.d. of $n = 3$ cell cultures are shown. $^{*}P = 0.017$ and $^{****}P = 0.000027$ as determined by two-tailed *t*-tests. Shading of bar areas indicates increasing relative peak areas and therefore greater glycosylation efficiencies. **c**, Representative LC–TOF spectra from analysis of $n = 3$ cell cultures generated by maximum entropy deconvolution are shown on the right (see Methods). Representative deconvolution spectra including N/Q substitution and no-NGT controls are shown in Supplementary Fig. 24. Deconvolution mass errors as well as chromatogram peak retention times and quantification of relative peak area for all samples are shown in Supplementary Tables 3 and 4. A similar analysis of engineered HMW1ct variants showed similar results, with the GlycoSCORES optimized GlycTag providing more efficient glycosylation (Supplementary Fig. 26).

guide the efficient glycosylation of diverse proteins in vitro and in vivo. By using more than 3,480 unique peptide substrates and 13,903 unique reaction conditions, GlycoSCORES enabled, to our knowledge, the most complete substrate characterization of any ppGT thus far (Supplementary Table 5). This data set further facilitated the selection of efficiently modified NGT peptide substrates (Supplementary Figs. 5, 16, and 19), for example, GGNWTT, which was found to direct efficient glycosylation of Im7, HMW1ct, and Fc proteins. Looking forward, our data set could also be used to make informed, single mutations to improve modification. Future work could also explore how glycosylation of these optimized sequences is affected by protein structure. We further generalized the GlycoSCORES approach to discover two NGT homologs in pathogenic bacteria with conserved peptide specificities, to show complex specificity differences between human ppGalNAcTs, and to demonstrate a proof of principle for high-throughput analysis of hOGT specificity. These innovations result from the ability to screen substrate residues more completely and determine synergistic residue combinations that are conventionally not tested.

When combined with recent advancements in the elaboration of the single glucose residue installed by NGT to human-like glycans using chemoenzymatic transglycosylation techniques[13,27] and polysialic acids using a fully biosynthetic approach[21], the deep specificity data and demonstration of highly efficient GlycTags shown in this work may open the door to diverse applications of NGT-based synthetic glycosylation systems just as the design and implementation of OST GlycTags[28] enabled the improvement of biopharmaceuticals and an array of studies using the bacterial OST, PglB, to produce vaccines and therapeutics in *E. coli*[8,9]. NGT systems may complement OST-based methods, as they do not require export out of the cytoplasm or lipid-associated substrates[9].

Given the versatility of CFPS for rapid, parallelized expression of diverse enzymes and target proteins and the throughput of SAMDI for rapid detection of glycosylation without radioactively

or chemically modified sugars or antibodies, we anticipate that GlycoSCORES will be applied to a broad range of ppGT investigations of interest in the glycoengineering community, including the further characterization of the ppGalNAcTs, OGTs, and OSTs (which have been recently shown to be produced in CFPS using protein nanodiscs[49]). GlycoSCORES is also uniquely suited to the engineering of glycosylation enzymes for alternative specificities in vitro, obviating the need for in vivo selection schemes, which have been challenging to develop for glycan modification. Specifically, CFPS reactions can be performed in 96/384-well plates with linear templates, substrate concentrations can be rigorously controlled, and any peptide mass addition can be detected and quantified by SAMDI. An example application is the synthesis and screening of diverse NGT homologs and engineered variants (building off recent work on NGT mutants[13,24]) to install GlcNAc onto proteins using a single enzyme or identifying ppGTs that can modify a specific amino acid sequence of interest.

In summary, the GlycoSCORES workflow provides a versatile platform for characterizing and engineering GTs. We expect this platform to enable a deep, quantitative understanding of glycosylation systems and advance compelling biotechnology applications.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41589-018-0051-2.

Received: 24 October 2017; Accepted: 7 March 2018;
Published online: 7 May 2018

References

1. Khoury, G. A., Baliban, R. C. & Floudas, C. A. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci. Rep.* **1**, 90 (2011).
2. Helenius, A. & Aebi, M. Intracellular functions of N-linked glycans. *Science* **291**, 2364–2369 (2001).
3. Sethuraman, N. & Stadheim, T. A. Challenges in therapeutic glycoprotein production. *Curr. Opin. Biotechnol.* **17**, 341–346 (2006).
4. Elliott, S. et al. Enhancement of therapeutic protein in vivo activities through glycoengineering. *Nat. Biotechnol.* **21**, 414–421 (2003).
5. Chung, C. H. et al. Cetuximab-induced anaphylaxis and IgE specific for galactose-α-1,3-galactose. *N. Engl. J. Med.* **358**, 1109–1117 (2008).
6. Lin, C.-W. et al. A common glycan structure on immunoglobulin G for enhancement of effector functions. *Proc. Natl Acad. Sci. USA* **112**, 10611–10616 (2015).
7. Clausen, H., Wandall, H.H., Steentoft, C., Stanley, P. & Schnaar, R.L. in Essentials of Glycobiology. (eds. A. Varki et al.) 713–728 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2015).
8. Valderrama-Rincon, J. D. et al. An engineered eukaryotic protein glycosylation pathway in *Escherichia coli*. *Nat. Chem. Biol.* **8**, 434–436 (2012).
9. Keys, T. G. & Aebi, M. Engineering protein glycosylation in prokaryotes. *Curr. Opin. Syst. Biol.* **5**, 23–31 (2017).
10. Wang, L.-X. & Davis, B. G. Realizing the promise of chemical glycobiology. *Chem. Sci.* **4**, 3381–3394 (2013).
11. Yang, Z. et al. Engineered CHO cells for production of diverse, homogeneous glycoproteins. *Nat. Biotechnol.* **33**, 842–844 (2015).
12. Li, H. et al. Optimization of humanized IgGs in glycoengineered *Pichia pastoris*. *Nat. Biotechnol.* **24**, 210–215 (2006).
13. Xu, Y. et al. A novel enzymatic method for synthesis of glycopeptides carrying natural eukaryotic N-glycans. *Chem. Commun. (Camb.)* **53**, 9075–9077 (2017).
14. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495 (2014).
15. Ban, L. et al. Discovery of glycosyltransferases using carbohydrate arrays and mass spectrometry. *Nat. Chem. Biol.* **8**, 769–773 (2012).
16. Pathak, S. et al. The active site of O-GlcNAc transferase imposes constraints on substrate sequence. *Nat. Struct. Mol. Biol.* **22**, 744–750 (2015).
17. Ortiz-Meoz, R. F., Merbl, Y., Kirschner, M. W. & Walker, S. Microarray discovery of new OGT substrates: the medulloblastoma oncogene OTX2 is O-GlcNAcylated. *J. Am. Chem. Soc.* **136**, 4845–4848 (2014).
18. Robinson, P. V., Tsai, C. T., de Groot, A. E., McKechnie, J. L. & Bertozzi, C. R. Glyco-seek: ultrasensitive detection of protein-specific glycosylation by proximity ligation polymerase chain reaction. *J. Am. Chem. Soc.* **138**, 10722–10725 (2016).
19. Naegeli, A. et al. Substrate specificity of cytoplasmic N-glycosyltransferase. *J. Biol. Chem.* **289**, 24521–24532 (2014).
20. Naegeli, A. et al. Molecular analysis of an alternative N-glycosylation machinery by functional transfer from *Actinobacillus pleuropneumoniae* to *Escherichia coli*. *J. Biol. Chem.* **289**, 2170–2179 (2014).
21. Keys, T. G. et al. A biosynthetic route for polysialylating proteins in *Escherichia coli*. *Metab. Eng.* **44**, 293–301 (2017).
22. Cuccui, J. et al. The N-linking glycosylation system from *Actinobacillus pleuropneumoniae* is required for adhesion and has potential use in glycoengineering. *Open Biol.* **7**, 160212 (2017).
23. Schwarz, F., Fan, Y. Y., Schubert, M. & Aebi, M. Cytoplasmic N-glycosyltransferase of *Actinobacillus pleuropneumoniae* is an inverting enzyme and recognizes the NX(S/T) consensus sequence. *J. Biol. Chem.* **286**, 35267–35274 (2011).
24. Song, Q. et al. Production of homogeneous glycoprotein with multi-site modifications by an engineered N-glycosyltransferase mutant. *J. Biol. Chem.* **292**, 8856–8863 (2017).
25. Gross, J. et al. The Haemophilus influenzae HMW1 adhesin is a glycoprotein with an unusual N-linked carbohydrate modification. *J. Biol. Chem.* **283**, 26010–26015 (2008).
26. Kawai, F. et al. Structural insights into the glycosyltransferase activity of the *Actinobacillus pleuropneumoniae* HMW1C-like protein. *J. Biol. Chem.* **286**, 38546–38557 (2011).
27. Lomino, J. V. et al. A two-step enzymatic glycosylation of polypeptides with complex N-glycans. *Bioorg. Med. Chem.* **21**, 2262–2270 (2013).
28. Chen, M. M., Glover, K. J. & Imperiali, B. From peptide to protein: comparative analysis of the substrate specificity of N-linked glycosylation in *C. jejuni*. *Biochemistry* **46**, 5579–5585 (2007).
29. Fisher, A. C. et al. Production of secretory and extracellular N-linked glycoproteins in *Escherichia coli*. *Appl. Environ. Microbiol.* **77**, 871–881 (2011).
30. Carlson, E. D., Gan, R., Hodgman, C. E. & Jewett, M. C. Cell-free protein synthesis: applications come of age. *Biotechnol. Adv.* **30**, 1185–1194 (2012).
31. Kuo, H. Y., DeLuca, T. A., Miller, W. M. & Mrksich, M. Profiling deacetylase activities in cell lysates with peptide arrays and SAMDI mass spectrometry. *Anal. Chem.* **85**, 10635–10642 (2013).
32. Kornacki, J. R., Stuparu, A. D. & Mrksich, M. Acetyltransferase p300/CBP associated Factor (PCAF) regulates crosstalk-dependent acetylation of histone H3 by distal site recognition. *ACS Chem. Biol.* **10**, 157–164 (2015).
33. Kim, J. & Mrksich, M. Profiling the selectivity of DNA ligases in an array format with mass spectrometry. *Nucleic Acids Res.* **38**, e2 (2010).
34. Laurent, N. et al. Enzymatic glycosylation of peptide arrays on gold surfaces. *ChemBioChem* **9**, 883–887 (2008).
35. Laurent, N. et al. SPOT synthesis of peptide arrays on self-assembled monolayers and their evaluation as enzyme substrates. *ChemBioChem* **9**, 2592–2596 (2008).
36. Hussain, M. R., Hoessli, D. C. & Fang, M. N-acetylgalactosaminyltransferases in cancer. *Oncotarget* **7**, 54067–54081 (2016).
37. Schjoldager, K. T. et al. Probing isoform-specific functions of polypeptide GalNAc-transferases using zinc finger nuclease glycoengineered SimpleCells. *Proc. Natl Acad. Sci. USA* **109**, 9893–9898 (2012).
38. Yoshida, A., Suzuki, M., Ikenaga, H. & Takeuchi, M. Discovery of the shortest sequence motif for high level mucin-type O-glycosylation. *J. Biol. Chem.* **272**, 16884–16888 (1997).
39. 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).
40. Kong, Y. et al. Probing polypeptide GalNAc-transferase isoform substrate specificities by in vitro analysis. *Glycobiology* **25**, 55–65 (2015).
41. Steentoft, C. et al. Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J.* **32**, 1478–1488 (2013).
42. Wang, A. C., Jensen, E. H., Rexach, J. E., Vinters, H. V. & Hsieh-Wilson, L. C. Loss of *O*-GlcNAc glycosylation in forebrain excitatory neurons induces neurodegeneration. *Proc. Natl Acad. Sci. USA* **113**, 15120–15125 (2016).
43. Yang, X. et al. Phosphoinositide signalling links O-GlcNAc transferase to insulin resistance. *Nature* **451**, 964–969 (2008).
44. Liu, X. et al. A peptide panel investigation reveals the acceptor specificity of O-GlcNAc transferase. *FASEB J.* **28**, 3362–3372 (2014).
45. Chalkley, R. J., Thalhammer, A., Schoepfer, R. & Burlingame, A. L. Identification of protein O-GlcNAcylation sites using electron transfer dissociation mass spectrometry on native peptides. *Proc. Natl Acad. Sci. USA* **106**, 8894–8899 (2009).
46. Lazarus, M. B., Nam, Y., Jiang, J., Sliz, P. & Walker, S. Structure of human O-GlcNAc transferase and its complex with a peptide substrate. *Nature* **469**, 564–567 (2011).
47. Choi, K. J., Grass, S., Paek, S., St Geme, J. W. III & Yeo, H. J. The *Actinobacillus pleuropneumoniae* HMW1C-like glycosyltransferase mediates

N-linked glycosylation of the *Haemophilus influenzae* HMW1 adhesin. *PLoS One* **5**, e15888 (2010).
48. Haselberg, R., de Jong, G. J. & Somsen, G. W. Low-flow sheathless capillary electrophoresis-mass spectrometry for sensitive glycoform profiling of intact pharmaceutical proteins. *Anal. Chem.* **85**, 2289–2296 (2013).
49. Schoborg, J. A. et al. A cell-free platform for rapid synthesis and testing of active oligosaccharyltransferases. *Biotechnol. Bioeng.* **115**, 739–750 (2018).
50. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).

Acknowledgements

The authors acknowledge J.C. Stark and J. Hershewe for assistance with western blotting, helpful discussions, and sharing of reagents and ideas; S. Habibi for assistance with LC-TOF instrumentation; and A. Karim for helpful conversations. The authors also thank J. Kath for supply of plasmids, advice on protein expression, and critical reading of the manuscript. We also thank A. Natarajan of the Department of Microbiology at Cornell University, T. Jaroentomeechai of the Robert Frederick Smith School of Chemical and Biomolecular Engineering at Cornell University, and J. Janetzko of the Department of Chemistry and Chemical Biology at Harvard University for sharing the ppGalNAcT, Im7, and hOGT source plasmids, respectively. This work made use of the Integrated Molecular Structure Education and Research Center at Northwestern University, which has received support from the state of Illinois, the Northwestern University Office of Research and the Chemistry Department for LC-TOF instrumentation. This material is based upon work supported by the Defense Threat Reduction Agency (HDTRA1-15-10052/P00001), the David and Lucile Packard Foundation, the Dreyfus Teacher-Scholar program, and the National Science Foundation (Graduate Research Fellowship under Grant No. DGE-1324585 and MCB-1413563).

Author contributions

W.K. and L.L. designed, performed, and analyzed experiments. M.R. designed and optimized experimental protocols. W.L. helped to synthesize peptide libraries. M.M. and M.C.J. directed the studies and interpreted the data. W.K., L.L., M.P.D., M.M., and M.C.J. conceived of the study and wrote the manuscript with assistance from M.R. and W.L.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41589-018-0051-2.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.M. or M.C.J.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Solid-phase synthesis of peptide arrays. All peptide arrays were synthesized manually using 96-well filter plates (cat. no. AWFP-F20000, Arctic White LLC) as described previously[51] with some modification. All Fmoc-amino acids and Fmoc-Rink amide MBHA resins were purchased from AnaSpec, Inc. All solvents, *N,N*-dimethylformamide (DMF), dichloromethane (DCM), trifluoroacetic acid (TFA) and piperidine were purchased from Thermo Fisher Scientific. Other chemical reagents used in peptide synthesis were purchased from Sigma-Aldrich unless otherwise noted. Briefly, 10 mg of Fmoc-Rink amide MBHA resins were placed in each well of 96-well filter plates. Before adding each amino acid, N-terminal fluorenylmethyloxycarbonyl (Fmoc) was deprotected with 300 µL 20% piperidine in DMF, with 600 r.p.m. shaking for 30 min. After five washes with DMF, 300 µL 0.1 M Fmoc-amino acid, 0.125 M hydroxybenzotriazole (HOBt) and 0.1 M diisopropylcarbodiimide (DIC) were used to add the amino acid onto the resin, with 600 r.p.m. shaking for 2 h. After all amino acids were added onto the resin, Fmoc was deprotected and acetic anhydride (10% in DMF) was used to add an acetyl group on the N-terminal of peptides with 600 r.p.m. shaking for 0.5 h. The resin was washed with DCM five times and dried for 1 h, before being cleaved by 500 µL 95% TFA, 2.5% H_2O, and 2.5% triethylsilane (TES) with 400 r.p.m. shaking for 2 h. After the solvents were evaporated by flowing nitrogen overnight, remaining solids were dissolved with 600 µL H_2O and transferred to 96-well plates. After lyophilization, the peptides were redissolved in 200 µL 50 mM Tris (pH 8), transferred to 384-well plates (Ref. No. 784201, Greiner Bio One), and stored at −80 °C. All peptides had one cysteine to determine the concentration of the peptides and bind to SAMs on gold islands. In most cases, an Arg residue was included N-terminal to the Cys to provide efficient ionization in the mass spectrometry experiments.

Preparation of SAMDI plates. 384-well SAMDI plates were prepared as previously described[51] with minor modifications. Briefly, 384 islands with 50 nm Ti and 300 nm Au were prepared by evaporation and rinsed in 0.25 mM ethanolic solution containing 60% of a symmetric disulfide presenting tri(ethylene glycol) (EG3 disulfide from ProChimia Surfaces, Poland) and 40% of an asymmetric disulfide presenting one tri(ethylene glycol) and one maleimide (EG3-Maleimide disulfide from ProChimia Surfaces, Poland) at 4 °C for 2 d. The SAMDI plate was ready for use after washing with ethanol, H_2O, and ethanol and then drying with flowing nitrogen.

Profiling NGT activity with peptide arrays. A subset of 16–24 peptides were used to measure the average concentration of each peptide library (361 or 380 peptides) using the Ellman test based on the manufacturer's protocols (Gold Biotechnology). After reduction with tris(2-carboxyethyl)phosphine (TCEP) reducing gel (Thermo Fisher Scientific), 50 µM peptides were reacted with the indicated concentrations of NGT purified by Ni-NTA from living *E. coli* or by unpurified NGT produced in *E. coli* CFPS, 2.5 mM UDP-Glc in 100 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (in HEPES, pH 8) and 500 mM NaCl at 30 °C for indicated times. As a control, the same volume of CFPS after 20 h of sfGFP synthesis was used instead of NGT from CFPS. The reaction was not quenched unless otherwise stated. 2 µL TCEP reducing gel were added to each 10 µL of reaction solution and incubated at 37 °C for 1 h. 2 µL of the reduced solutions were then transferred to a 384-well SAMDI plate using Tecan 96-channel arm and incubated at room temperature (~25°C) for 0.5 h. SAMDI plates were washed with two rounds of H_2O and ethanol, and dried with flowing nitrogen. After application of 10 mg/mL of 2′,4′,6′-trihydroxyacetophenone monohydrate (THAP) matrix (Sigma-Aldrich) in acetone onto the entire SAMDI plate, an Applied Biosystems SciEx MALDI-TOF/TOF 5800 instrument was used to perform mass spectrometry on each spot. Applied Biosystems SciEx Time of Flight Series Explorer Software version 4.1.0 was used to analyze MS spectra. Generally, each 384-well IVG plate was immobilized onto two separate 384-well SAMDI plates and analyzed by MALDI separately. The modification efficiency of peptides was calculated using the following equation:

$$\%\ \text{modification} = \frac{\frac{I(P)}{\text{RIF}(P\ \text{to}\ S)}}{\frac{I(P)}{\text{RIF}(P\ \text{to}\ S)} + I(S)}$$

where $I(P)$ is the intensity of product Glc-peptide in mass spectrometry, $I(S)$ is the intensity of substrate peptide in mass spectrometry, and RIF(P *to* S) is the relative ionization factor of product to substrate. RIFs equal $IF(P)/IF(S)$ and were determined as described below. Glucose modification efficiencies for peptides are shown as heat maps in Figs. 2 and 4 and annotated with numerical values in Supplementary Figs. 4–8 and 16–19. Heat maps for comparison or characterization are arranged by amino acid properties, whereas heat maps for NGT optimization are ranked in descending order by average modification of all substrates in each row or column and ranked from left to right and top to bottom. As determined by negative controls peptide arrays using CFPS synthesizing sfGFP rather than NGT (Supplementary Fig. 7), glucose modification efficiencies of less than 0.03 were regarded as background. NGT homologs MhNGT and HdNGT were produced in *E. coli* CFPS and profiled with peptide arrays the same way as NGT (Supplementary Figs. 9–11).

Measuring relative ionization factors. Relative ionization factors (RIFs) were determined by measurements of mass spectra intensity/concentration for glycosylated and aglycosylated peptide samples. After reduction with TCEP reducing gel, 50 µM peptides were reacted with 2.5 mM UDP-Glc and 10 µM purified NGT or 0.575 µM CFPS NGT in 100 mM HEPES (pH 8) and 500 mM NaCl at 30 °C for 4 h to achieve more than 70% glucose modification (glycosylated samples). Identical reactions without UDP-Glc were used as control reactions to provide the same total concentration of peptides (aglycosylated samples). The reactions were quenched by placing the reaction plates at 60 °C for 20 min. Glycosylated and aglycosylated samples were mixed at a 1:1 ratio and reduced with TCEP reducing gel, and mass spectra for glycosylated, aglycosylated, and mixed samples were collected by SAMDI. The aglycosylated samples always showed no detectable glucose modification. Relative ionization factors were calculated using the equation below.

$$\text{RIF} = \frac{\%I\,(\text{Re})^{*}\%I\,(\text{Mix})}{\%I\,(\text{Re})^{*}\%I\,(\text{Mix}) + \%I\,(\text{Re}) - 2^{*}\%I\,(\text{Mix})}$$

Where %I(Re) is the intensity of the glycosylated product peptide ($I(P_g)$) divided by the sum of the intensities of the glycosylated product peptide ($I(P_g)$) and aglycosylated substrate peptide ($I(S_g)$) in the glycosylated samples, or $I(P_g)/(I(P_g) + I(S_g))$. %I(Mix) is the intensity of the glycosylated product peptide ($I(P_m)$) divided by the sum of the intensities of the glycosylated product peptide ($I(P_m)$) and aglycosylated substrate peptide($I(S_m)$) in a 1:1 ratio mixture of glycosylated and aglycosylated samples, or $I(P_m)/(I(P_m) + I(S_m))$. A subset of 20–24 peptides were used to measure the RIF of each length of peptide library. The RIFs of peptides for which reaction kinetics data was collected were also determined (Supplementary Table 2).

Determining sugar donor specificity of NGT. Six peptides were used to profile the monosaccharide selectivity of NGT. 50 µM peptides were reacted with 0.1–0.2 µM purified NGT, 1 mM UDP-Glc, UDP-Gal, GDP-Man, UDP-GlcNAc, UDP-GalNAc, or UDP-Xyl in 100 mM HEPES (pH 8) and 500 mM NaCl at 30 °C for 1 h, 4 h, or 21 h. After reduction with TCEP reducing gel, the percentage intensity of Glc-peptide was recorded by SAMDI. For testing of sugar donor selectivity of glucose, galactose, or xylose modification with the $X_{-1}NX_{+1}TRC$ peptide library, 1 mM UDP-Glc, UDP-Gal, or UDP-Xyl and indicated concentration of purified NGT were used. UDP-Xyl was purchased from Carbosource Services. Other sugar donors were purchased from Sigma-Aldrich.

Measuring reaction kinetics parameters of selected GlycTag peptides. Various (6–8) concentrations of selected HPLC-purified peptides were reacted with the indicated concentrations of NGT as well as 10 mM UDP-Glc in 100 mM HEPES (pH 8) and 500 mM NaCl at 30 °C for a series of reaction times (15 min to 2 h). Reactions were quenched using 2 µL of 2 M HCl per 10 µL of reaction solution. After neutralization with 2 µL of 2 M K_2CO_3 and reduction with TCEP reducing gel, the modification efficiency was determined by SAMDI. Initial reaction velocities were calculated using the slopes in the linear time-frame of each initial peptide concentration. K_M and k_{cat} and associated errors were then determined by nonlinear fitting to the Michaelis–Menten formula using OriginPro 9 software.

Using GlycoSCORES to screen peptide selectivity of human O-linked GTs produced in CFPS. To demonstrate the applicability of GlycoSCORES to the study of mammalian O-linked GTs, ppGalNAcT1, ppGalNAcT2, and hOGT were produced in *E. coli* CFPS. While hOGT was synthesized in CFPS the same way as NGT, the ppGalNAcTs were synthesized in CFPS under oxidizing conditions to allow the formation of disulfide bonds. Oxidizing conditions were achieved using standard CFPS reactions modified as described previously[52], supplemented with 14.3 µM iodoacetamide, 1 mM glutathione, 4 mM glutathione disulfide, and 3.16 µM *E. coli* disulfide bond isomerase (DsbC). For GlycoSCORES screening of ppGalNAcTs, 100 µM of each peptide from peptide array $AX_{-1}TX_{+1}APRC$ was reacted with 0.024 µM CFPS ppGalNAcT1 or 0.04 µM ppGalNAcT2, 1 mM UDP-GalNAc in 100 mM HEPES (pH 7.5), and 3 mM Mn^{2+} at 37 °C for 1 h followed by quenching with 5 mM EDTA. As a control, the same volume of CFPS after 20 h of sfGFP synthesis was used instead of ppGalNAcTs from CFPS. GlycoSCORES screening of hOGT was completed similarly, with 50 µM of each peptide variant of the sequence PPVSRC reacted with 0.062 µM hOGT made in CFPS and 2.5 mM UDP-GlcNAc in 20 mM Tris (pH 7.4), 125 mM NaCl, and 1 mM EDTA for 21 h at 37 °C. After reduction with TCEP reducing gel and maleimide capture, the relative percentage intensities of the GalNAc or GlcNAc-modified and unmodified peptides were recorded by SAMDI as described for NGT.

Plasmid construction and molecular cloning. Plasmids used in this study with sources and details are shown in Supplementary Table 1, and coding sequences with sequence context are shown in Supplementary Note 1. The wild-type Im7 coding sequence (Uniprot: IMM7_ECOLX) was PCR amplified from pBR322.Im7 and assembled into the pJL1 CFPS vector between the NdeI and SalI sites using Gibson Assembly to produce pJL1.Im7-0s. Wild-type IgG1 constant Fc region (Δ1-98 Uniprot: IGHG1_HUMAN) was synthesized by Twist Bioscience and

assembled into a variant of pET22b with redesigned restriction sites (pETBCS.NS) using restriction ligation to form pETBCS.NS.Fc-0s. The wild-type sequence for HMW1ct (Δ1-1203 GenBank: ADO96128.1) was synthesized by Life Technologies and assembled into pJL1 between NdeI and SalI sites to form pJL1.HMW1ct-WT. A variant sequence of HMW1ct with N/Q substitutions at all naturally occurring N-X-S/T sites except at N1366 was synthesized by Life Technologies and assembled into pET.BCS.NS to form pET.BCS.NS.HMW1ct-0 using restriction and ligation at NdeI and SalI sites. Variants of the N26_T31 NVAAT loop in Im7-0s, the Q178_Y183 QYNSTY naturally occurring glycosylation sequence in Fc-0s, and the N1364_S1370 naturally occurring glycosylation sequence in HMW1ct-0 were constructed by inverse PCR with 18 bp of overlapping 5′ homology and recircularized by one-piece Gibson Assembly. All variants of Im7, Fc, and HMW1ct contained C-terminal 6×His-tags. The wild-type NGT sequence (Uniprot: NGT_ACTP2) was synthesized by Twist Bioscience and assembled into the pJL1 vector between NdeI and SalI sites, into the pET.21b vector between NcoI and XhoI sites with a C-terminal 6×His-tag, and into the pMAF10 vector between NcoI and HindIII sites with a C-terminal 1×FLAG tag by Gibson Assembly. The α-1,6 glucose polymerase from *A. pleuropneumoniae* (AGT, Uniprot: GTF_ACTP7) was ordered from Twist Bioscience in pJL1 with a customized ribosome binding site designed for maximum translation initiation rate using the RBS Calculator v2.0 (ref. [53]). Codon optimized sequences for MhNGT (Uniprot: A0A0B5BRN9_MANHA) and HdNGT (Uniprot: Q7VKK3_HAEDU) were ordered from Integrated DNA Technologies with C-terminal Strep tags and placed into PJL1 using Gibson assembly. Human ppGalNAcT1 (CGAT1_HUMAN) and ppGalNAcT2 (CGAT2_HUMAN) truncated without the N-terminal 40 amino acids (Δ40) were also cloned into PJL1 using Gibson Assembly either with or without N-terminal CAT-Strep-Linker fusions. The coding sequence for hOGT (OGT1_HUMANΔ1-313) was PCR amplified from pET42a.hOGT[46] and cloned into PJL1 using Gibson Assembly.

Preparation of cell extracts for CFPS. Crude extracts for CFPS were generated from a genomically recoded release factor 1 (RF1)-deficient *E. coli* strain (*E. coli* C321.ΔA.759)[54], based on *E. coli* C321.ΔA[55]. Cell growth, harvest, and lysis were performed as described in Kwon and Jewett[56]. Briefly, *E. coli* cells were grown in 1 L of 2xYTPG (yeast extract 10 g/L, tryptone 16 g/L, NaCl 5 g/L, K_2HPO_4 7 g/L, KH_2PO_4 3 g/L, and glucose 18 g/L, pH 7.2) in a 2.5 L Tunair flask at 34 °C and 250 r.p.m. with initial inoculation to OD_{600} = 0.08. At OD_{600} = 3.0, cells were pelleted by centrifugation at 5,000 × *g* at 4 °C for 15 min. The pellets were washed three times with cold S30 buffer (10 mM Tris–acetate pH 8.2, 14 mM magnesium acetate, 60 mM potassium acetate, and 2 mM dithiothreitol (DTT)) and flash frozen on liquid nitrogen and stored at −80 °C. Cells were thawed, resuspended in 0.8 mL of S30 buffer per gram wet weight, and lysed in 1.4 mL aliquots on ice using a Q125 Sonicator (Qsonica) for three pulses (50% amplitude, 45 s on and 59 s off). After sonication, 4 µL of DTT (1 M) was added, which was followed by centrifugation at 12,000 × *g* and 4 °C for 10 min. The supernatant was incubated at 37 °C at 250 r.p.m. for 1 h for a run-off reaction and centrifuged again at 10,000 × *g* at 4 °C for 10 min. The supernatant was flash-frozen on liquid nitrogen and stored at −80 °C until use.

Cell-free protein synthesis. CFPS reactions were conducted using a PANOx-SP crude lysate system[57]. A standard reaction contained 1.2 mM ATP; 0.85 mM each of GTP, UTP, and CTP; 34 µg/mL folinic acid; 170 µg/mL of *E. coli* tRNA mixture; 16 µg/mL purified T7 RNA polymerase; 2 mM for each of the 20 standard amino acids; 0.33 mM nicotinamide adenine dinucleotide (NAD); 0.27 mM coenzyme-A (CoA); 1.5 mM spermidine; 1 mM putrescine; 4 mM sodium oxalate; 130 mM potassium glutamate; 10 mM ammonium glutamate; 12 mM magnesium glutamate; 57 mM HEPES, pH 7.2; 33 mM phosphoenolpyruvate (PEP); 13.3 µg/mL plasmid template of interest; and 27% v/v of *E. coli* cell extract. *E. coli* total tRNA mixture (from strain MRE600) and phosphoenolpyruvate were purchased from Roche Applied Science[58]. ATP, GTP, CTP, UTP, the 20 amino acids, and other materials were purchased from Sigma-Aldrich. Plasmid DNA for cell-free protein synthesis was purified from DH5-α *E. coli* strain (NEB) using ZymoPURE Midi Kit (Zymo Research). NGT and AGT were synthesized in 50 µL batch reactions in 2.0 mL microtubes and Im7, Fc, and HMW1ct-WT target proteins were synthesized in 15 µL batch reactions in 1.5 mL microtubes. The CFPS reactions were carried out at 20 °C for 20 h.

Quantification of CFPS yields. Total and soluble CFPS yields were quantified using CFPS reactions identical to those used for NGT, AGT, Im7, HdNGT, MhNGT, ppGalNAcT1, ppGalNAcT2, and hOGT synthesis supplemented with 10 µM [^{14}C]leucine (PerkinElmer). Protein quantification for triplicate CFPS reactions was completed using trichloroacetic acid (TCA) protein precipitation followed by radioactivity quantification using a Microbeta2 liquid scintillation counter (PerkinElmer) according to established protocols[59]. Soluble fractions were taken after centrifugation at 12,000 × *g* for 15 min at 4 °C. CFPS yields of sfGFP were quantified as described previously[56] using a multi-well fluorimeter (Synergy2, BioTek) and converted to microgram per milliliter yields using a previously determined standard curve based on [^{14}C]leucine incorporation assays[60].

Autoradiograms of CFPS proteins. After synthesis in [^{14}C]leucine supplemented CFPS reactions, 2 µL of each sample was loaded onto a 4–12% Bolt Bis–Tris Plus SDS–PAGE gels (Invitrogen) and run in MOPS buffer at 150 V for 70 min. The gels were stained using InstantBlue (Expedeon) and destained in water. The gels were incubated in gel drying solution (Bio-Rad), dried overnight between cellophane films in a GelAir Dryer (Bio-Rad) without heating, and exposed for 48 h on a Storage Phosphor Screen (GE Healthcare). Autoradiogram images were acquired using Typhoon FLA7000 imager (GE Healthcare). The same coomassie stained gels were imaged using a GelDoc XR + Imager for molecular weight standard references (Bio-Rad).

Production and purification of NGT from *E. coli*. NGT was purified as described previously[23] with minor modifications. Briefly, BL21(DE3) cells were transformed with pET21b.NGT plasmid by electroporation. An overnight culture was inoculated in carbenicillin (CARB) LB media. Fresh CARB LB was inoculated at initial OD_{600} = 0.08, and the cells were grown at 37 °C at 250 r.p.m. to 0.6–0.8 OD and induced with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) for 6 h at 30 °C. The cells were pelleted by centrifugation at 8,000 × *g* for 10 min at 4 °C, resuspended in Buffer 3 (20 mM Tris–HCl and 250 mM NaCl, pH 8.0), pelleted again by centrifugation at 8,000 × *g* for 10 min at 4 °C, and frozen at −80 °C. The pellets were then thawed and resuspended in 5 mL Buffer 3 with 20 mM imidazole per gram wet pellet weight and supplemented with 70 µL of 10 mg/mL lysozyme (Sigma), 1 µL benzonase (Millipore), and 1× Halt protease inhibitor (Thermo Fisher Scientific). Cells were then lysed by single pass homogenization at 21,000 psig (Avestin) and centrifuged at 15,000 × *g* for 20 min at 4 °C. The supernatant was applied to an Ni-NTA agarose column (Qiagen) equilibrated with Buffer 3 with 20 mM imidazole, washed with 10 column volumes of Buffer 3 with 40 mM imidazole, and eluted with 4 column volumes of Buffer 3 with 500 mM Imidazole. The elution was dialyzed against 50 mM HEPES 200 mM NaCl, pH 7.0, supplemented with 5% glycerol, and flash frozen at −80 °C. NGT concentration was quantified using Image Lab software version 6.0.0 densitometry with BSA standard curve after separation by SDS–PAGE, staining with InstantBlue coomassie stain, and destaining in water.

In vitro glycosylation of protein substrates. IVG reactions were assembled in standard 0.2 mL tubes from completed CFPS reactions containing targets and enzymes at concentrations determined by [^{14}C]leucine incorporation. Im7 glycosylation reactions contained 5 µM of one Im7 variant, 0.1 µM NGT, and 2.5 mM UDP-Glc in the final reaction. Each reaction contained a total of 5 µL UDP-Glc and 25 µL CFPS reaction (the remaining CFPS reaction volume up to 25 µL was filled by a completed CFPS reaction that synthesized sfGFP). Similarly, HMW1ct-WT IVG reactions contained 5 µM HMW1ct-WT, 0.1 µM NGT, and 2 µM AGT and 2.5 mM UDP-Glc in the final reaction. Each IVG reaction contained 10 µL completed CFPS reaction and 2 µL UDP-Glc. IVG reactions for Im7 and HMW1ct-WT were performed at 30 °C for 2.5 h and 16 h, respectively.

Western blotting of HMW1ct-WT. Completed HMW1ct-WT IVG reactions (1 µL) were loaded onto a 4–12% Bolt Bis–Tris SDS–PAGE gel in MOPS buffer and run at 130 V for 100 min. The gel was then transferred onto a 0.2 µM PVDF membrane (Bio-Rad) using the Trans-Blot SD semi-dry blotting system (Bio-Rad) using 80% MOPS and 20% Methanol buffer. The target protein was detected by blocking the membrane in 5% milk in PBS with 0.1% Tween 20 and then incubating with a polyclonal His antibody (Abcam, ab1187) diluted 1:7,500 in PBS with 1% milk for 45 min. The poly-α-glucose moiety installed by NGT and AGT was detected using a Concanavalin A (ConA) lectin blot using an identical gel with identical membrane and transfer conditions. The ConA blot was blocked with Carbo-free solution (Vector Laboratories) for 1 h and probed with 5 µg/mL ConA-HRP (Sigma, L6397-1MG) diluted in Carbo-free solution supplemented with 1 mM $MgCl_2$, 1 mM $MnCl_2$, 0.1% Tween, and 1 mM $CaCl_2$ for 1 h. Blots were imaged using Western-Sure Chemiluminescent substrate on an Odessey Fc (Li-Cor) imager.

Purification from in vitro glycosylation reactions. Purification of Im7 from IVG reactions was completed using Dyna-His tag beads (Thermo Fisher Scientific). The 30 µL IVGs were diluted to 120 µL in Buffer 1 (50 mM NaH_2PO_4 and 300 mM NaCl, pH 8.0) with a final concentration of 10 mM imidazole and incubated at room temperature for 5 min on a roller with 20 µL of beads. The beads were then washed with 120 µL of 20 mM imidazole in Buffer 1 four times using a 96-well plate magnetic tube rack (Life Technologies) for separations. The samples were then eluted using 30 µL of 500 mM imidazole in Buffer 1. The samples were dialyzed against Buffer 2 (20 mM NaH_2PO_4 and 150 mM NaCl, pH 7.5) in 3.5 kDa MWCO 96-well plate dialysis cassettes (Thermo Fisher Scientific). After dialysis, 10 µL was injected into LC–TOF for analysis.

Production of glycosylated proteins in cells. *E. coli* BL21(DE3) cells were transformed first with pMAF10.NGT by electroporation and selected on trimethoprim (TMP) LB agar plates. A colony was picked, prepared for calcium-chloride transformation, transformed with pETBCS.NS vectors containing Fc or HMW1ct target proteins, and selected on TMP + CARB LB agar plates. Colonies

were grown to mid-exponential phase and glycerol stocked. The glycerol stocks were used to inoculate overnight cultures in TMP + CARB LB media. Fresh cultures in TMP + CARB were inoculated at initial OD_{600} = 0.08 and grown at 37 °C at 250 r.p.m. For HMW1ct sequence variants, the target protein was induced at 0.6–0.8 OD for 1 h with 400 µM IPTG at 30 °C followed by NGT induction with 0.2% arabinose for 2 h at 30 °C. For Fc sequence variants, the target protein was induced for 2 h, which was followed by NGT induction for 30 min (unless otherwise noted) at identical inducer concentrations. The cells were then pelleted by centrifugation at 4 °C for 2 min at 10,000 × *g*, resuspended in Buffer 1, centrifuged at 4 °C for 2 min at 10,000 × *g*, frozen on liquid nitrogen, and stored at −80 °C. The pellets were thawed and resuspended in 630 µL of Buffer 1 with 10 mM imidazole and supplemented with 70 µL of 10 mg/mL lysozyme (Sigma), 1 µL benzonase (Millipore), and 1× Halt protease inhibitor (Thermo Fisher Scientific). After 15 min of thawing and resuspension, the cells were incubated for 15 min on ice, sonicated for 45 s at 50% amplitude, and then spun at 12,000 × *g* for 15 min. The supernatant was then loaded onto Ni-NTA His-tag spin columns (Qiagen) pre-equilibrated with 10 mM imidazole in Buffer 1. The columns were washed three times with 30 mM imidazole and eluted with 2 × 100 µL 500 mM imidazole. Samples were then dialyzed with 10 kDa MWCO MINI slide-a-lyzers (Thermo Fisher Scientific) overnight. Protein concentrations were quantified using Image Lab software densitometry with BSA ladder standard after separation by SDS–PAGE, 1 h stain with InstantBlue, and 1 h destain in water. Prior to injection into LC–TOF, purified Fc was incubated with 50 mM DTT for 1 h at room temperature to reduce disulfide linkages.

LC–TOF analysis of glycoprotein modification. Purified proteins from CFPS of Im7 or in vivo expression of Fc and HMW1ct were injected onto an Agilent 1200 HPLC equipped with an XBridge BEH300Å C4 3.5 µm 2.1 mm × 50 mm reverse-phase column (186004498 Waters Corporation) with a 10 mm guard column of identical packing (186007230 Waters Corporation) coupled to an Agilent 6210 A ESI–TOF mass spectrometer. The chromatographic separation method was based on the manufacturer's instructions for XBridge column with minor modifications. Solvent A was 95% H_2O and 5% acetonitrile (ACN) with 0.1% formic acid, and solvent B was 100% ACN with 0.1% formic acid. The separation was completed at a flow rate of 0.4 mL/min with a column temperature of 50 °C. Solvent conditions were held at 15.8% B for 1 min; then, the target protein of interest was eluted during a 12 min gradient from 15.8% to 65.8% B. The column was then washed and re-equilibrated using a 2 min gradient from 65.8–69.9% B, a 2 min hold at 100% B, and a 6 min hold at 15.8% B. Purified Fc after in vitro synthesis and glycosylation was injected into a Bruker Elute UPLC system, separated using the same chromatography methods as listed above, and analyzed by an Impact-II UHR TOF–MS system (Bruker Daltonics, Inc.). External calibration was completed before analysis of all proteins.

LC–TOF data analysis. Data from Agilent 6210 A was processed using Agilent Mass Hunter software version B.04.00. Methods for quantification of relative peak areas for glycosylated and aglycosylated glycoforms were adapted from previous works[48,61]. Extracted ion chromatograms (EICs) were created using theoretical values for the most dominant charge states from the glycosylated and aglycosylated samples ± 0.5 Da. Protonated charge states +12 to +14, +29 to +36, and +34 to +43 were used to quantify the relative peak areas for Im7, Fc, and HMW1ct, respectively. EIC peaks corresponding with retention times of each protein (Supplementary Table 3) were then integrated and used for quantification of relative peak areas, defined as $Glc_1/(Glc_0 + Glc_1)$. Deconvoluted spectra were produced using Agilent Mass Hunter maximum entropy deconvolution using MS peaks within *m/z* range 700–2,000 into mass ranges of 10,000–15,000 u; 25,000–30,000 u; and 32,500–37,500 u for Im7, Fc, and HMW1ct, respectively. Isotope widths were calculated by Mass Hunter for deconvolution mass ranges at 7.1, 10.5, and 11.6 u for Im7, Fc, and HMW1ct, respectively. Data from Impact-II UHR TOF-MS was performed using Bruker Compass Hystar software version 4.1. Deconvolution was performed using maximum entropy deconvolution using MS peaks within *m/z* range of 700–2,000 into a mass range of 20,000–30,000 u. Raw data was then plotted and annotated using R Studio. Deconvolutions used full mass spectra averaged across the entire peak width of the proteins of interest (encompassing the full elution of the glycosylated and aglycosylated glycoforms). Deconvoluted masses and errors compared to calculated values are shown in Supplementary Table 4.

Statistical analysis. Two-tailed Student's *t*-tests and resulting *P* values were calculated in Microsoft Excel 2016 assuming unequal variances and two-tailed distributions to assign significance to observed differences in relative peak areas for GlycTag variants of Im7, HMW1, and Fc. In these cases, $n = 3$ independent IVG reactions were performed for analysis of Im7, whereas $n = 3$ independent *E. coli* expression cultures were completed for analysis of HMW1 and Fc.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability. All data generated or analyzed during this study are included in this published article (and its supplementary files) or are available from the corresponding authors on reasonable request.

References

51. Gurard-Levin, Z. A., Scholle, M. D., Eisenberg, A. H. & Mrksich, M. High-throughput screening of small molecule libraries using SAMDI mass spectrometry. *ACS Comb. Sci.* **13**, 347–350 (2011).
52. Goerke, A. R. & Swartz, J. R. Development of cell-free protein synthesis platforms for disulfide bonded proteins. *Biotechnol. Bioeng.* **99**, 351–367 (2008).
53. Espah Borujeni, A., Channarasappa, A. S. & Salis, H. M. Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res.* **42**, 2646–2659 (2014).
54. Martin, R. W. *et al.* Cell-free protein synthesis from genomically recoded bacteria enables multisite incorporation of noncanonical amino acids. *Nat. Commun.* **9**, 1203 (2018).
55. Lajoie, M. J. *et al.* Genomically recoded organisms expand biological functions. *Science* **342**, 357–360 (2013).
56. Kwon, Y.-C. & Jewett, M. C. High-throughput preparation methods of crude extract for robust cell-free protein synthesis. *Sci. Rep.* **5**, 8663 (2015).
57. Jewett, M. C. & Swartz, J. R. Mimicking the *Escherichia coli* cytoplasmic environment activates long-lived and efficient cell-free protein synthesis. *Biotechnol. Bioeng.* **86**, 19–26 (2004).
58. Jewett, M. C., Calhoun, K. A., Voloshin, A., Wuu, J. J. & Swartz, J. R. An integrated cell-free metabolic platform for protein production and synthetic biology. *Mol. Syst. Biol.* **4**, 220 (2008).
59. Jewett, M. C. & Swartz, J. R. Rapid expression and purification of 100 nmol quantities of active protein using cell-free protein synthesis. *Biotechnol. Prog.* **20**, 102–109 (2004).
60 Hong, S. H. *et al.* Cell-free protein synthesis from a release factor 1 deficient *Escherichia coli* activates efficient and multiple site-specific nonstandard amino acid incorporation. *ACS Synth. Biol.* **3**, 398–409 (2014).
61 Jian, W., Edom, R. W., Wang, D., Weng, N. & Zhang, S. W. Relative quantitation of glycoisoforms of intact apolipoprotein C3 in human plasma by liquid chromatography-high-resolution mass spectrometry. *Anal. Chem.* **85**, 2867–2874 (2013).

nature research

Corresponding author(s): Michael Jewett

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

   Describe how sample size was determined.

   Quantification of protein production from n=3 CFPS reactions were used to enable calculation of mean and standard deviation as is regular practice in cell-free literature. SAMDI-MS spectra from n=2 separate peptide immobilizations were analyzed and quantified for each measurement displayed as heat maps. This is standard practice for high-throughput SAMDI studies and permitted simple liquid handling and high-throughput screening while allowing us to confirm that differences in peptide immobilization and MS acquisition/ ionization were less than 10 percent, indicating that sample size did not seriously affect quantification. NGT peptide substrate kinetics parameters were determined by non-linear regression of velocities at n=8 or n=6 peptide concentrations, each from n=9 peptide in vitro glycosylation reactions at various time-points. This sample scheme provided high correlation coefficients and calculation of uncertainties which are listed. Quantitative measurements of relative peak areas (protein modification) by LC-MS were determined by n=3 IVG or cell cultures allowing for calculation of standard deviation and student's t-tests for significance. Deconvolutions of MS spectra representative of samples used for quantification and n=2 spectra for samples used as controls or for qualitative assessment of protein modification (as specifically listed in figure and supplementary figure legends) are also provided. The use of n=3 samples for quantification and n=1 for identification is regular practice in LC-MS studies.

2. Data exclusions

   Describe any data exclusions.

   No data was excluded

3. Replication

   Describe the measures taken to verify the reproducibility of the experimental findings.

   All attempts at replication were successful.

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   Samples were organized by experimental variables then characterized fully and reported in completion. Therefore, samples were not randomized.

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   Blinding was not relevant because we used peptide libraries to fully characterize glycosyltransferase activity. Animal or human participants were not used in this study.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	Commercial software used includes: Image Lab version 6.0.0 build 25 (Bio-rad Laboratories, Inc.), Agilent Mass Hunter B.04.00 (Agilent Technologies, Inc.), Bruker Compass HyStar 4.1 (Bruker Daltonics, Inc.), OriginPro 9 (OriginLab Corporation), and Time of Flight Series Explorer Software version 4.1.0 (Applied Biosystems SciEx). No custom software was used.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.	All unique materials will be available through a materials transfer agreement (MTA)

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	We used a commercially available poly-His antibody Abcam ab1187 conjugated to horseradish peroxidase (HRP) for chemiluminescent analysis of western blots. This product was was validated by the vendor. No secondary antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used.
b. Describe the method of cell line authentication used.	No eukaryotic cell lines were used.
c. Report whether the cell lines were tested for mycoplasma contamination.	No eukaryotic cell lines were used.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No eukaryotic cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.	There were no animals used in this study.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	There were no human participants in this study.