Protein structure, a genetic constraint on glycosylation

Benjamin P. Kellman,[1,2,3,4,5,*] Daniel Sandoval,[6] Olga O. Zaytseva,[7] Kelly Brock,[8] Sabyasachi Baboo,[9] Daniela Nachmanson,[2,3,10] Edward B. Irvine,[5,11] Sanne Schoffelen[12], Erick Armingol,[1,2,3] Nathan Mih,[2,3] Yujie Zhang,[1] Mia Jeffris,[1,2] Philip Bartels,[6] Thi Nguyen,[6] Amy Tam,[8] Sarah Gasman,[13] Shlomi Ilan,[5] Samuel William Canner,[4] Isaac Shamie,[1,2,3] Jolene K. Diedrich,[9] Xiaoning Wang,[9] Esther van Woudenbergh,[5] Meghan Altman,[14] Anthony Aylward,[2,3] Bokan Bao,[1,2,3] Andrea Castro,[2,3] James Sorrentino,[1,2,3] Austin W.T. Chiang,[1,2,15] Matt Campbell,[16] Yannic Bartsch,[5] Patricia Aguilar-Calvo,[13,17,18] Christina Sigurdson,[13,15] Galit Alter,[5] Gordan Lauc,[7] John R. Yates III,[9] Bjørn Gunnar Rude Voldborg[12], Debora Marks,[8,19] Frederique Lisacek,[20,21] Nathan E. Lewis[1,2,3,20,*]

[1] Department of Pediatrics, University of California, San Diego, La Jolla, CA 92093, USA
[2] Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093, USA
[3] Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, CA 92093, USA
[4] Augment Biologics, San Fransico, CA
[5] Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA, USA
[6] Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA 92093, USA
[7] Genos Glycoscience Research Laboratory, Borongajska 83H, Zagreb 10000, Croatia
[8] Department of Systems Biology, Harvard Medical School, Boston, MA, USA
[9] Department of Molecular Medicine, The Scripps Research Institute, La Jolla, California 92037, United States
[10] Moores Cancer Center, University of California San Diego, 3855 Health Science Drive, San Diego, CA, 92093, USA
[11] Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[12] National Biologics Facility, Department of Biotechnology and Biomedicine, Technical University of Denmark, Lyngby, Denmark.
[13] Department of Pediatrics, Boston Medical Center, Boston, MA, USA
[14] Department of Chemistry and Biochemistry, University of California San Diego, 9500 Gilman Drive, La Jolla, CA, USA
[15] Immunology Center of Georgia & Department of Medicine, Augusta University, Augusta, GA, United States
[16] Institute for Glycomics, Griffith University, Queensland, Australia
[17] Department of Pathology, University of California San Diego, 9500 Gilman Drive, La Jolla, CA, USA
[18] Department of Neurobiology, The University of Alabama at Birmingham, Birmingham, AL 35243
[19] Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, United States
[20] Proteome Informatics Group, Swiss Institute of Bioinformatics, CH-1227 Geneva, Switzerland
[21] Computer Science Department & Section of Biology, University of Geneva, route de Drize 7, CH-1227, Geneva, Switzerland
[20] Center for Molecular Medicine, Complex Carbohydrate Research Center, and Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA, USA

Abstract

Unlike DNA, RNA, and protein biosynthesis, dogma describes glycosylation as primarily determined by intrinsic cellular limitations, such as glycosyltransferase expression and precursor availability. However, this cannot explain the commonly-observed differences between glycans on the same protein. By examining site-specific glycosylation on diverse human proteins, we detected associations between protein structure and glycan structure, broadly generalizable to human-expressed glycoproteins. Through structural analysis of site-specific glycosylation data, we found protein-sequence and structural features consistently correlated with specific glycan features. To quantify these relationships, we present a new amino acid substitution matrix describing "glycoimpact", i.e., the association of primary protein structure and glycosylation. High-glycoimpact amino acids co-evolve with glycosites, and glycoimpact is high when estimates of amino acid conservation and variant pathogenicity diverge. We report thousands of disease variants near glycosites with high-glycoimpact, including several with known links to aberrant glycosylation (e.g., Oculocutaneous Albinism, Jakob-Creutzfeldt disease, Gerstmann-Straussler-Scheinker, and Gaucher's Disease). Finally, glycoimpact quantification is validated by studying oligomannose-complex glycan ratios on HIV ENV, differential sialylation on IgG3 Fc, differential glycosylation on SARS-CoV-2 Spike, and fucose-modulated function of a tuberculosis monoclonal antibody. Finally, to test the causality of protein-glycan associations, we created 5 glycoimpact-designed novel Rituximab variants, 4 of which substantially changed glycoprofiles as predicted. In all, we report that site-specific glycan biosynthesis is influenced by underlying protein structure, enabling glycan structure prediction and genetic sequence-guided glycoengineering.

Introduction

DNA, RNA, and protein synthesis follow DNA and RNA templates. In contrast, glycosylation is understood as primarily regulated by the cellular environment.[3–6] While glycosylation varies across species, cell types and cell lines, specific glycosites on proteins reproducibly host different glycans, termed glycosite-specific microheterogeneity.[7] Microheterogeneity suggests local protein structure may also constrain glycosylation.

Many studies have identified specific examples of how primary protein structure or sequence can influence glycosylation patterns. First, the N-glycosylation sequon[8] defines where glycans covalently attach, i.e., to asparagines (N) with a downstream (N+2) serine (S) or threonine (T) separated by any amino acid (X) except proline (NX[S/T]). Variation at N+1 impacts glycan complexity[9] and accessibility impacts glycosite occupancy.[10] Glycosylation of sequons ending in threonine is ~40 times more efficient than those with serine.[11] Studies have identified specific examples wherein primary protein structure or sequence can influence glycosylation patterns. Upstream of the glycosite, a phenylalanine (F) to alanine (A) substitution in human IgG3 can increase bi-galactose structures with a core-fucose.[12] Additionally, influenza evolves hemagglutinin glycosylation sites to facilitate immune evasion.[13] Glycoprotein-determined glycan evolution, used by HIV for immune evasion, has been leveraged to engineer better vaccine epitopes.[14] Tools like GlycoSiteAlign[15] and mutagenesis studies have offered expansions of the founding sequon structure (NX[S/T]) including the enhanced aromatic sequon—an aromatic residue upstream of the glycosite (N-2) that can influence glycan complexity[16] with a variable impact given the N+1 variation.[17] Several studies since 2002 have shown variable success in predicting glycosites from sequons and hidden features using machine learning,[18–21] though none have predicted beyond glycosite presence or the root monosaccharide.

Beyond a protein primary structure (i.e., protein sequence), secondary structures (e.g., β-sheets and α-helixes) can influence glycosylation[22] while tertiary structures (concavity, accessibility and hydrophobicity) also impact glycosite occupancy.[10] Crucial bottlenecks in glycan processing such as, ERManI (MAN1B1) and Golgi Mannosidase IA (MAN1A1), have been co-crystalized with a Man9 glycan to study specific glycoconjugate features favoring or inhibiting this interaction.[23–26] A study of >150 glycoproteins revealed differential glycosylation as a function of protein structure, wherein low accessibility sites impacted core fucosylation and branching;[27] similar accessibility constraints predicted oligomannose on the SARS-CoV-2 spike protein.[28] Furthermore, a structural analysis found FUT8 activity is impacted by glycosite depth.[29] Site-specific kinetics for glycosylation of PDI (Protein Disulfide

Isomerase)[30–32] and five additional glycoproteins[33] also showed that protein structure influences occupancy. Recent studies have aimed to graft glycans to proteins[20,34–36], providing valuable tools for exploring mechanistic and steric associations between glycans and folded proteins. Yet, they do not explore generalizable associations between glycan and protein structure features that can be mapped back to the genome.

Here, we identified associations between glycosylation and local protein structures (1D to 3D). We did this by analyzing trends in site-specific glycosylation[37] and comparing biosynthetic precursors of observed glycans[38] with local glycoprotein structure features.[39] We call associations between protein structures and glycan structures "***Intra-molecular Relations***" (IMR); IMR are quantified using metrics associating protein structures and glycans structures (e.g., regression models, correlation, or Fisher Exact test) . Further, we call the expected difference in glycan structure following protein sequence or structure changes, "***Glycoimpact***;" the difference in IMR associated with protein sequence or structure defining the change. Glycoimpact is detected and validated here by comparison to evolutionary substitution matrices, variant pathogenicity scores, and glycosite co-evolution. Further, glycosite-proximal pathogenic variants correspond to higher glycoimpact substitutions. Finally, glycoimpact accurately predicts changes in glycan complexity, galactosylation, sialylation, and functional glycosylation. These results show IMR represent generalizable associations between protein structure and glycosylation (**Figure** 1). Consequently, IMR suggest protein structure can constrain glycosylation, providing increased clarity of factors impacting microheterogeneity.

Results

UniCarbKB site-specific glycosylation is representative of human glycosites in PDB.

We first tested if glycan structure correlates with protein structure. To do this, we curated a Protein-Glycan Dataset (PGD), a compendium of experimentally-measured and expert-curated site-specific glycosylation on human glycoproteins collected from the UniCarbKB[41] and GlyConnect[37] databases (description and URLs, **see Methods**). Glycan structures were decomposed into glycan substructures (i.e., biosynthetic intermediates using GlyCompare[38]). Glycosite-proximal protein structure features were extracted from protein sequence and 3D structure models. "Sequence proximal" was defined as five residues C- or N-terminus to the glycosite and spatial proximity was defined by minimum distance—the minimum 3D Euclidean distance between the nearest atoms in each residue. The resulting dataset

includes 111 human glycoproteins (98 glycoproteins with N-glycosylation sites and 38 glycoproteins with O-glycosylation sites) 306 glycosylation sites and 4,263 glycosylation events (3,563 N- and 700 O-linked glycosylation events) (Supplementary **Figure** 1). In this work, we focused training on human glycoproteins and validated predictions to human-expressed glycoproteins. Human and human expressed glycoproteins were used to assert consistency in the underlying biosynthetic machinery.

We verified that annotated glycosites in PGD are representative of all glycosite structures in the human secretome;[42] all secretome glycosites fall within a dimensionality reduction trained only on PGD glycosite structures (**Figure** 2a, see **Supplementary Results**, Supplementary **Figure** 2). Thus, PGD can support generalizable conclusions about protein-glycan associations.

Site-specific glycosylation data contain significant associations between protein sequence, 3D protein structures, and glycosylation

We next examined the associations between glycan substructures (e.g., bisecting GlcNAc) and glycosite-proximal protein features (e.g., proximal tyrosine) within PGD. Substantial and significant (Supplementary **Figure** 3a) associations between glycan substructures and protein features were termed "intramolecular relations" (**IMR**) (Figure 1c). Fisher exact test selected 10,111 IMR with 9,296 positive and 815 negative associations (**Figure** 2b).

Approximately 20% of IMR predict glycosylation from protein structure with high confidence

The PGD contains several highly deterministic amino acids (AA) specific IMR for which the presence or absence of a glycan substructure is highly determined by the presence or absence of a proximal AA (i.e., Pr(glycan substructure | proximal amino acid) is 0.999-1 or 0-0.001). Overall, confidence in glycan presence increases when a specific protein structure feature is present (Supplementary **Figure** 4). For IMR with a spatially proximal AA (within 6Å), 20.2% are highly deterministic of glycan substructures. Additionally, 32.4% and 17.5% of down- and upstream AA IMR (+/-5 AA) are highly deterministic. The certainty with which an AA predicts a glycan structure decreases substantially when proximal AA are absent (Supplementary **Figure** 4**a**). The high-confidence IMR count is proportional to the number of unique substructures revealing that these IMR are not dominated by small numbers of glycan motifs, notable substructures (Supplementary **Figure** 4**b**).

Among the highly deterministic protein-glycan relations (1,725 N-glycosylation events), 1,553 glycans contain a N-acetylglucosamine (GlcNAc) on the α-1,6-mannose branch (Glc2NAc(β1-6)Man(α1-6)Man(β1-4)Glc2NAc), indicative of complex N-glycosylation. All 75 high-confidence IMR with a

downstream tryptophan (W) include glycans with the hybrid/complex substructure. These observations suggest that W would be sufficient to escape oligomannose structures. Similarly, in 454 O-glycosylation high-confidence IMR, 237 contain 6-sialyl-N-acetyllactosamine (Neu5Ac(α2-6)Gal(β1-3)GalNAc). Of six events containing a sequence-proximal W, each also contained a sialyl-T antigen. Thus, several IMR link protein features, such as proximal W, with specific glycan features.

Glycosite-proximal, sequence, structure, and chemistry correlate with glycan structure

To generate testable hypothetical protein structural constraints on glycan biosynthesis, we modeled IMR using hierarchical regression to minimize relevant bias. We quantified specific IMR using univariate logistic generalized estimation equations (GEE) to probe the glycan-protein co-occurrences in PGD and to control for protein identity effects (see **Methods**). A list of GEE odds ratios (OR) (each describing the association between a protein structure and one of several glycan substructures) describes typical glycosylation near a given protein structure; the OR list indicates the "expected substructure abundance/absence" like an "expected glycoprofile" given a proximal protein structure.

We discovered 1,715 significant N-glycan IMR (FDR<0.1, $|\log(OR)|>0.1$), many of which were associated with spatially proximal (N+6Å) and sequence-proximal AA (N+/-5 residues). Stratifying sequence-proximal effects, we found almost twice as many IMR involving upstream (N-5, N-terminal) than downstream (N+5, C-terminal) AA. Among the downstream AA effects, tryptophan (W), alanine (A), serine (S), and F are most impactful (n = 99, 55, 55, and 48 IMR respectively). W also has many IMR when downstream or spatially proximal (n = 26). Spatially proximal arginine (R) (n = 70) and downstream glutamine (Q) (n = 35) are the largest effectors. Finally, glycosylation sites on turns have many IMR (n = 61) (**Figure** 2c).

Turn-associated IMR include >3-fold increases in di- and tri-sialylated tetra-antennary and >2-fold increases in mono- and di-galactosylated structures with core fucose; all positively correlated structures have at least one galactose (Gal) while not all are core fucosylated. Increased complexity at a high-exposure glycosite (i.e., turn) is consistent with prior results demonstrating an inverse association between complexity and depth[27,43]. Structurally proximal Q is associated with a >20-fold increase in monosialylated triantennary structures and a 10-fold decrease in tetraantennary structures (**Figure** 3**a**). Histidine (H), threonine (T), and valine (V) show correlation with increasing GalNAc[4S] (**Figure** 3**b**). Expanding on **Figure** 3a-b, we further interrogated IMR for specific protein structure features (i.e., proximal amino acid or local secondary structure) by comparing the IMR odds ratio to monosaccharide count for several monosaccharides (**Figure** 3**c**). A biclustering highlights at least two major groups of

glycan features differentially impacted by different protein structure influences, mirroring the N-glycan/O-glycan dichotomy in glycosylation. Providing clues to the elusive O-glycosylation site, proximal A is negatively correlated with Gal and GlcNAc but positively correlated with N-acetylgalactosamine (GalNAc). Conversely, T and H are positively associated with GlcNAc and Gal but negatively correlated with β-GalNAc. As expected, GlcNAc and complex-glycans follow similar trends to Neu5Ac. However, Neu5Ac, GlcNAc and Gal trends diverge near proline (P), cysteine (C), and V; these amino acids may act as limiters of high-complexity (**Figure** 3**c**).

Amino acid changes can predict glycan structural changes, "Glycoimpact"

The difference in expected glycosylation associated with two distinct amino acids can indicate the expected change in glycosylation following substitution between these amino acids (AA). As mentioned previously, the expected impact of substitution on glycosylation, is termed "glycoimpact." Such variations are estimated by comparing expected glycoprofiles (see above) for two AA-pairs (e.g. proximal valine and isoleucine). Specifically, glycoimpact is measured as the difference (z-score normalized Euclidean distance) in expected glycoprofiles between each AA-pair observed near glycosites of two protein structures (**Figure** 3, see **Methods**).

Many AA-substitutions are glycoimpactful on glycan biosynthesis, as exemplified by the F to W substitution (**Figure** 3**d-e, Supplementary Results**). Upstream F is associated with bi- and tri-antennary (>3-fold) while upstream W is associated with tetraantennary terminal Gal (>2-fold) implicating this substitution in branching (**Figure** 3**d**). Additionally, spatially proximal F is strongly associated with increased sialylation (>10-fold) implicating this amino acid in glycan maturation (**Figure** 3**e**). These predictions suggest that W to F substitution has a large impact on glycoprofiles. Relevant substitution events are highlighted as "glycoimpactful" (significantly high glycoimpact) and structurally ambivalent (BLOSUM>0) (**Figure** 3**f**); those high-glycoimpact (red) and low structural impact (thick line) are substitutions of particular note as potential avenues of non-structurally impactful glycan modulation.

Glycoimpact explains divergence of BLOSUM and PAM substitution matrices

We call the glycoimpact AA-substitution matrix the BLOSUM-PAM Orthology matrix, BLAMO x:y, where x and y refer to the log odds ratio and FDR thresholds respectively. Sub-threshold odds ratios, those insignificant or unsubstantial by the x:y threshold, are excluded from the glycoimpact calculation. **Figure** 3f displays a subset of BLAMO 0.5:0.1 relations. Comparisons can be made at multiple thresholds (Supplementary **Figure** 5) but BLAMO 0.5:0.1 presents as a more representative, normative, and stable threshold.

To further explore the relevance of glycoimpact, we compared it to conservation-based measures of amino acid substitution impact. The PAM[44] and BLOSUM[45] AA substitution matrices are popular but distinct estimates. PAM is based on global protein alignments and tends to reflect functional conservation, while BLOSUM relies on local protein alignment reflecting mostly structural conservation (e.g., domains).[46,47] Thus, PAM and BLOSUM diverge more when protein function is not fully described by protein structure.

Since changes in glycan structure can modify protein function in a protein structure-independent manner, we examined consistency between PAM and BLOSUM estimates across null and high glycoimpact substitutions. Comparing PAM and BLOSUM scores at multiple thresholds (RMSE($PAM_{i,j}$,$BLOSUM_{i,j}$)), we found that error in 4 of 5 PAM-BLOSUM comparisons was significantly correlated to glycoimpact (BLAMO 0.5:0.1) for high-glycoimpact (GI>2.5) substitutions (**Figure 4a,** Supplementary **Figure** 5). Acknowledging the limited effect size (R<0.4), these associations are strongly significant and consistent across multiple comparisons suggesting, as noted, a significant and global but noisy trend. The correlation between high-glycoimpact substitutions and PAM-BLOSUM inconsistency is maintained for most PAM and BLOSUM thresholds (Supplementary **Figure** 6). Both protein structure and glycosylation are necessary to fully explain protein function, and these results suggest a positive relationship between glycoimpact and the failure of structure (BLOSUM) to completely explain function (PAM). Given this relationship, we refer to the glycoimpact substitution matrix as the BLOSUM-PAM Orthology (BLAMO) matrix.

High glycoimpact residues are conserved around N-glycosites

If glycosite-proximal amino acids influence functional glycosylation, there should be evolutionary pressure imposed beyond the classical NX[S/T] N-glycan sequon. To map the broader glycosite structure, we aligned the surrounding sequences (five AA upstream and downstream) of N-glycosites (**Figure 4b-c**) and examined conservation and evolutionary coupling (EC). EC-derived from 2,005 glycoprotein alignments (see **Methods**, **Supplementary Results**),[48] substantiates enrichment between N-glycosites and flanking residues (Supplementary **Figure** 7a). We also observed several position-specific glycosite-coupled residues including S and T at N+2, F at N-2, and tyrosine (Y) at N-1 (pooled hypergeometric, FDR<0.1, Supplementary **Figure** 7c, **Figure 4c**). These findings are consistent with previous observations of the original sequon and enhanced aromatic sequon.[17] Glycosite-coupled residues are also consistent with IMR-observed high-impact residues (**Figure** 2**c**). Of the ten high-impact upstream residues, seven show enriched glycosite-coupling when they appear upstream. To further highlight global glycosite

structure, we clustered glycosites using coupling probabilities (**Figure 4d,** Supplementary **Figure** 8, see **Methods**). The N+2 aspartic acid (Asp) enriched in the univariate analysis (**Figure 4c**) co-occurs with an N-2 lysine (K) (**Figure 4d,** motif 1). Alternatively, glutamic acid (E) is more likely to co-occur with other E residues (N -4, +1, and +3) with an N+2 T-containing sequon (**Figure 4d,** motif 4). These couplings are reflective of evolutionary pressure surrounding the glycosylation sites.

We next aligned[15] glycosites permitting a tetraantennary N-glycan lacking fucose or sialic acids (**Figure 4b**). We examined the glycosite alignment for consistency with high-influence AA (**Figure 2c**) and glycosite-coupled residues (**Figure 4c**). Of 20 glycosite-flanking AA, 16 show consistency between the first or second most common AA and either the high-influence or glycosite coupled residues (see **Supplementary Results**). At nearly every glycosite-flanking residue (N+/-10) there is consistency between these three analyses, further corroborating that protein structure constrains glycosylation.

Glycoimpact correlates with discrepancy between functional variant predictions

Dozens of algorithms predict functional and pathogenic effects of genetic variants,[49–51] incorporating information ranging from sequences to protein structure, thus sometimes reporting different variants as deleterious. We hypothesized the differences in some pathogenicity scores between algorithms could be explained by glycoimpact (BLAMO 0.5:0.1, see **Methods**). Across 3,549,910 nonsynonymous mutations, we measured the disagreement (RMSE) between each of 27 rank-normalized functional impact prediction tools, precomputed with dbNSFP;[49] pathogenicity score divergence was then correlated with glycoimpact. After hierarchical clustering on the divergence-glycoimpact correlation coefficients, pathogenicity estimates separated into two major clusters: one containing nearly all (6/7) tools leveraging protein-structure and the other primarily containing conservation, sequence, and/or epigenetic-based tools (**Figure** 5**b**). Nearly all variant pathogenicity score differences across the two clusters correlated with glycoimpact. These correlations and clustering structure disappear when glycoimpact scores are shuffled (Supplementary **Figure** 10); thus, like BLOSUM-PAM discrepancies, glycoimpact correlates with discrepancies between conservation-based and protein structure-based pathogenicity estimates. These observations further implicate glycosylation as a potent functional regulator and glycoimpact as an appropriate proxy for the importance of glycosylation.

Glycoimpact proposes mechanisms for pathogenic variants in ClinVar and PrP

Because glycoimpact correlates with evolutionary, structural, and functional AA-substitution metrics, we hypothesized it is also pathologically relevant. Thus, we compared glycoimpact and clinical impact for ClinVar-annotated variants within 15Å, 20Å, and 30Å of UniProtKB-annotated glycosites. For all three distances tested, high glycoimpact (BLAMO 0.5:0.1) variants in ClinVar were robustly and significantly higher (Wilcoxon p=2.6e-4, 2.2e-7, and 1.6e-10) for pathogenic variants close to glycosylation sites compared to glycosite-proximal benign variants (**Figure** 5**c**). Similarly, glycoimpact is higher for likely-pathogenic variants and variants of unknown significance near glycosites.

Examining specific ClinVar-annotated variants, we identified multiple variants in glycosylation-related diseases. For example, Tyrosinase:A355V (P14679) is a high-glycoimpact and glycosite-proximal causal variant in oculocutaneous albinism.[52,53] While tyrosinase:A355V has not been examined for aberrant glycosylation, deglycosylation disrupts tyrosinase function consistent with type 1 oculocutaneous albinism.[53–57] Therefore, tyrosinase:A355V may act through aberrant glycosylation. We also observe high-glycoimpact, glycosite-proximal (<30Å), pathogenic (ClinVar) variants in multiple other glycan-modulated diseases including prion diseases,[58,59] lysosomal storage disorders,[60,61] and Gaucher's disease.[62–65]

More broadly, of 1,228 non-benign ClinVar annotated variants on glycoproteins, 340 are high-glycoimpact (GI>2.5) and closer than 30Å to a glycosylation site (**Table 1,** Supplementary **Table** 1**, Supplementary Dataset 1**). This includes major diseases not typically considered related to N-glycosylation including cystic fibrosis, long QT syndrome, renal cell carcinoma, acquired immunodeficiency syndrome, and multiple blood coagulation factor deficiencies. Notably, approximately 36% of the non-benign ClinVar-annotated glycoprotein variants we examined may be impacted by aberrant glycosylation, a potentially underappreciated mechanism of pathogenesis.

We further analyzed glycosite proximity among causal variants in prion disease. We measured 3D Euclidean min-distance from the two PrP glycosylation sites, N181 and N197, to all residues in human prion protein (PrP) (including variants causing Creutzfeldt-Jakob disease (CJD)[66,67] and Gerstmann-Sträussler-Scheinker disease (GSS))[68,69] (**Figure** 5a**,** Supplementary **Table** 2). CJD-causing variants were approximately twice as close to glycosylation sites than the background distribution of all PrP sites (One-sided Wilcoxon p=0.0003). GSS-causative variants also trend closer to glycosites (One-sided Wilcoxon p=0.07) (**Figure** 5a, Supplementary **Figure** 11a-b). Low expression mutants, an indication of possible

aberrant glycosylation, trend closer to site N180 (One-sided Wilcoxon p=0.16) and appeared further from N197 (One-sided Wilcoxon p=0.04) (Supplementary **Figure** 11c-d).

Table 1 – Summary of variants by disease type where variants are high glycoimpact, close to glycosylation sites, and annotated in ClinVar as non-benign. Each specific disease is listed followed by gene name(s) and an integer indicating the number (if larger than 1) of variants in that gene corresponding to each specific disease. "" indicates a multisystemic disorder.*

Cardiovascular Disorders	Ventricular tachycardia, polymorphic (*CASQ2* (2)); Arrhythmogenic right ventricular cardiomyopathy, dysplasia, hypertrophic (*DSG2* (4)); Long QT syndrome, Brugada syndrome (*KCNH2* (15) *KCNE1* (2) *KCNE2 TRHDE SCN1B* (3)); Primary pulmonary hypertension (*KCNK3*); Coronary artery disease, autosomal dominant (*LRP6*); Atrial fibrillation (*SCN3B*); Atrial septal defect 6 (*TLL1*); *Amyloidogenic transthyretin amyloidosis (*TTR* (6))
Coagulation & Hematologic Disorders	Upshaw-Schulman syndrome, Hereditary thrombotic thrombocytopenic purpura (*TTP ADAMTS13*); Platelet glycoprotein IV deficiency (*CD35*); Hereditary factor XI deficiency disease (*F11* (2)); Factor VII deficiency (*F7* (2)); Hemophilia (*F9*); Hereditary factor IX deficiency disease (*F9* (5)); Glanzmann's thrombasthenia (*ITGA2B*); Platelet-activating factor acetylhydrolase deficiency (*PLA2G7*); Thrombophilia (*PROC* (3)); Antithrombin III deficiency (*SERPINC1* (3))
Endocrine Disorders	*Serkal syndrome (*WNT4*); Familial Hypobetalipoproteinemia (*ANGPTL3*); *Hypocalciuric hypercalcemia (*CASR* (7)); Laron syndrome (*CGHR*); *Deficiency of ferroxidase (*CP* (3)); *Dopamine beta hydroxylase deficiency (*DBH*); Laron-type isolated somatotropin defect, short stature (*GHR* (3)); Isolated growth hormone deficiency (*GHRHR* (2)); Diabetes mellitus, insulin-resistant (*INSR*); Leprechaunism syndrome (*INSR*); Hypercholesterolaemia (*LDLR* (2)); Gonadotropin-independent familial sexual precocity, Leydig cell agenesis (*LHCGR* (2)); Hyperlipoproteinemia, type I (*LPL*); Hemochromatosis (*TFR2 HFE* (2)); *Deficiency of iodide peroxidase (*TPO* (3)); Hypothyroidism, congenital, nongoitrous (*TSHR*); *Ladd syndrome (*FGF10*)
Hearing Disorders	Nonsyndromic Hearing Loss and Deafness (*MET*)
Immune Disorders	Cyclical neutropenia (*ELANE* (5)); X-linked severe combined immunodeficiency (*IL2RG* (2)); Severe combined immunodeficiency disease (*IL7R* (2)); Myeloperoxidase deficiency (*MPO*); Complement deficiency, C1 esterase inhibitor deficiency (*SERPING1* (2)); Acquired immunodeficiency syndrome (AIDS) (*IL4R*)
Metabolic Disorders	*Aspartylglycosaminuria (*AGA*); Infantile hypophosphatasia (*ALPL* (2)); Farber's lipogranulomatosis (*ASAH*); Pseudocholinesterase deficiency, Bche, j variant (*BCHE*); *Biotinidase deficiency (*BTD* (13)); *Combined deficiency of sialidase and beta galactosidase, Galactosialidosis (*CTSA* (2)); Ceroid lipofuscinosis, neuronal 1, 10, and 13 (*CTSF PPT1* (5) *RP11-295K3.1 CTSD* (3)); Glycogen storage disease II (*GAA*); *Gaucher disease type 1, perinatal, lethal, subacute, acute, neuronopathic (*GBA* (13)); *Fabry disease (*GLA* (4)); Tay-Sachs disease, B1 variant (*HEXA* (4)); *Sandhoff disease (*HEXB* (2)); *Hurler syndrome, Gangliosidosis, Mucopolysaccharidosis, MPS-II, MPS-IV-B, MPS-I-H/S, MPS-III-A, MPS-III-B, MPS-IV-A, MPS-VII (*ARSB* (2) *IDS* (2) *GLB1* (5) *SGSH* (2) *NAGLU GALNS* (4) *MCOLN1 GUSB* (3)); *Danon disease (*LAMP*); *Sialidosis, type II (*NEU1*); *Niemann-Pick disease type A, B, C1, Sphingomyelin/cholesterol lipidosis (*NPC1* (7) *SMPD1* (2)); Hereditary acrodermatitis enteropathica (*SLC39A4*); *Congenital disorder of glycosylation type (*STT3A MOGS*); *Multiple sulfatase deficiency (*SUMF* (2)); Crigler-Najjar syndrome (*UGT1A8 UGT1A10* (2*) UGT1A9 UGT1A7 UGT1A6* (3) *UGT1A5 UGT1A4 UGT1A3* (2) *UGT1A1* (2))
Musculoskeletal Disorders	Geleophysic dysplasia (*ADAMTSL2*); Chondrodysplasia punctata 1, X-linked recessive (*ARSE*); Orofacial cleft (*BMP4*); Spondyloepiphyseal dysplasia with congenital joint dislocations (*CHST3*); Hypochondroplasia (*FGFR3*); Bruck syndrome (*FKBP10*); Brachydactyly, type A2, fibular hypoplasia, complex brachydactyly (*GDF5* (2)); Spondylocostal dysostosis (*LFNG*); Short stature with nonspecific skeletal abnormalities (*NPR2*); Hyperphosphatasemia with bone disease (*TNFRSF11B*); Osteogenesis imperfecta (*WNT1*); *Zimmermann-Laband syndrome (*KCNH1* (3))
Neurologic Disorders	Metachromatic leukodystrophy, late-onset, late infantile (*ARSA* (4)); Holoprosencephaly (*CDON*); Myasthenic syndrome, congenital, fast-channel, slow-channel (*CHRNA1* (5)); Epilepsy, nocturnal frontal lobe, type 3 (*CHRNB2*); *Congenital muscular dystrophy, Bethlem Myopathy (*COL6A2* (4)); Febrile seizures, familial, 11 (*CPA6*); *Lipid proteinosis (*ECM1* (1)); Familial febrile seizures 8|not provided (*GABRG2*); Focal epilepsy with speech disorder with or without mental retardation (*GRIN2A* (2)); Megalencephalic leukoencephalopathy (*HEPACAM*); Episodic ataxia type 1 (*KCNA1*); X-linked hydrocephalus syndrome (*L1CAM*); Charcot-Marie-Tooth disease (*MPZ* (2)); Early infantile epileptic encephalopathy 9, 19, 32 (*PCDH19 KCNA2 GABRA1*); *Congenital muscular dystrophy-dystroglycanopathy (*POMK*); Genetic prion diseases, Jakob-Creutzfeldt disease, Gerstmann-Straussler-Scheinker syndrome (*PRNP* (6)); Progressive myoclonic epilepsy (*SCARB2*); Infantile Parkinsonism-dystonia (*SLC6A3*); Hyperekplexia (*SLC6A5*); Dystonia (*TOR1A*); Temple-Baraister syndrome (*KCNH1*)
Oncogenic	*Hereditary diffuse gastric cancer, Endometrial carcinoma (*CDH1* (2)); *Lynch syndrome (*EPCAM*); Renal cell carcinoma, papillary (*MET* (2)); *Multiple endocrine neoplasia, type 2a, type 2 (*RET* (6))
Ophthalmological Disorders	Peters anomaly (*EPHB2*); Microphthalmia (*GDF3*); Retinitis pigmentosa 68, 12, 35, Cone-rod dystrophy 10 (*SLC7A14 CRB1 SEEMA4A* (2))
Pulmonary & Gastrointestinal	*Cystic fibrosis (*CFTR* (5)); *Alpha-1-antitrypsin deficiency (*SERPINA1* (2))
Renal Disorders	Renal dysplasia (*AGT*); Diabetes insipidus, nephrogenic (*AQP2*); Nephrotic syndrome (*LAMB2*); Finnish congenital nephrotic syndrome (*NPHS1* (5)); Cystinuria (*SLC3A1*)
Rheumatologic Disorders	*Polyarteritis nodosa (*CECR1* (2)); TNF receptor-associated periodic fever syndrome (*TRAPS TNFRSF1A*)

Skin Disorders	Diffuse palmoplantar keratoderma, Bothnian type (*AQPS*); Familial progressive hyperpigmentation with or without hypopigmentation (*KITLG*); Adult junctional epidermolysis bullosa (*LAMB3*); *Kanzaki disease (*NAGA*); *Ectodermal dysplasia-syndactyly syndrome 1 (*PVRL4*); *Oculocutaneous albinism, Waardenburg syndrome, Tyrosinase-negative, Oculocutaneous albinism type 3 (*TYR* (5) *TYRP1*); Odonto Onycho Dermal dysplasia (*WNT10A*)
Vascular Disorders	*Arterial calcification of infancy (*ENPP1*); Rienhoff syndrome (*TGFB3*)

Protein structure correlates with glycan complexity in HIV gp160

To further validate our predictions, we compared IMR to glycosylation on specific glycoproteins. We first found consistency between previously measured IMR (**Figure** 2**b**) and site-specific glycan complexity measurements in HIV ENV gp160 (**Figure** 6**b**).[70] PGD-measured GEE-estimated IMR suggest that downstream Q was significantly (FDR<1e-8; OR<0.5) predictive of complexity while spatially proximal P and K were weaker but significant distinguishers (FDR<1e-3, FDR<0.1 respectively, **Figure** 6**a**). As predicted, gp160 glycosites with proximal P (min distance < 6Å) present more oligomannose (Two-sided Wilcoxon p=0.0033), whereas C-terminus-proximal Q glycosites have more complex glycans (Two-sided Wilcoxon p=1e-4, **Figure** 6**c**). Spatially proximal K, while less significant (**Figure** 6**c**), had a nonlinear impact on glycan complexity in HIV gp160; first increasing with one proximal K then decreasing with two. The two most significant IMR predicted from PGD (spatially proximal P and C-terminal Q) were consistent with the site-specific glycosylation observed in HIV gp160.

IMR predict differential glycosylation on the SARS-CoV-2 spike glycoprotein

We also predicted glycosylation on SARS-CoV-2 spike protein S1 subunit in the ancestral strain to the Gamma and Delta variants. Several of the >20 glycosylation sites[71,72] have been implicated with stability, target engagement, furin cleavage, and immune evasion.[72–76] We found multiple glycosite-proximal mutations within 15Å (min-distance, cubic (3D) expansion of the 6Å IMR training threshold). Gamma spike S1 contains multiple mutations close to glycosylation sites including N17 (L18F, T20N, & D138Y), N61 (P26S & R190S), N122 (L18F, D138Y, & R190S), N616 (D614G), and N657 (H655Y). In the Delta S1, N17 and N122 have 5 and 4 proximal mutations (relative to ancestral S1), respectively, while N165 and N616 each have one high-proximity mutation. Of the glycosite proximal substitutions, only L18F in Gamma and F157V in Delta have a high predicted glycoimpact by IMR. L18F appears within 15Å of N17, N74, N122 in Gamma. Similarly, F157V appears within 15Å of N17, N122, and N165 in Delta. High-impact substitutions appear close to N17 and N122 in both variants.

We measured HEK293-expressed SARS-CoV-2 S1 variant glycan heterogeneity using DeGlyPHER, a mass spectrometry (MS)-based glycoproteomics method,[77] to determine glycan state and occupancy at each glycosite. We performed two independent technical replicate analyses of the S1 subunit comparing Ancestral to the Gamma and Delta variants and examined 11 of the 12 canonical S1 glycosites (N17 was

excluded due to variable signal peptide trimming). Site-specific unoccupied, complex, and oligomannose/hybrid proportions were compared between variants using a Mann-Whitney U test.[77] P-values were pooled across the two independent replicate analyses using the Fisher method (FDR-corrected). We observed three significant differential glycosylation events (**Figure** 6**d**). Complex glycans observed at N122 in Ancestral S1 were converted to more oligomannose/hybrid in both Delta (oligomannose/hybrid observations increased nearly 4-fold from 13.9% in S1 to 52.5%; FDR=3.3e-9) and Gamma (oligomannose/hybrid observations nearly doubled to 27.6%; FDR=7.9e-4) variants. Complex glycans at N331 increased marginally from Ancestral S1 in Delta variant (from 93.7% to 99.7%; FDR=0.031). Complex glycans seen at N657 in Ancestral S1 decreased in Gamma variant (by over 2-fold from 53.3% to 21.1%; FDR=1.27e-3). The Gamma S1 monomer was consistently expressed with two novel complex glycosites at N20 and N188.

Based on proximal high-glycoimpact substitutions, we predicted changes at N17, N74 (Gamma only), N122, and N165 (Delta only). Three of four (N122 in Gamma and Delta, N657 in Gamma) predicted differential glycosylation events were consistent with the four observed changes (sensitivity=0.75). Meanwhile, 15 sites where no change was predicted, were consistent with the 17 sites where no change was observed (specificity=0.88). This correct prediction of differential glycosylation is most substantial at N122 in the S1 monomer providing a proof-of-concept that motivates further inspection of differential glycosylation on the more physiological spike trimer and whole virus. This can provide further insights into how differential glycosylation may participate in immune evasion by SARS-CoV-2 and other viruses.

Glycosite-proximal variation in IgG sequence predicts differential Fc N-glycosylation

We next predicted differential glycosylation for the *ighg1* missense mutation (F299I) in the IgG1 heavy chain.[78] IgG1 glycosylation impacts both adaptive humoral response[79–82] and monoclonal antibody (mAb) response.[83–86] Experimentally, the IgG1:F299I substitution shows a strong glycoimpact. The IgG1 variant expressed in C57BL/6 and BALB/c mouse strains shows less sialylation and digalactosylation compared to similarly expressed *wt* IgG1.[87–89] Additionally, BALB/c-expressed IgG1:F299I glycosylation is more similar to C57BL/6-expressed IgG1:F299I glycosylation than to glycans on *wt* IgG1 expressed in the same BALB/c animals (**Figure** 7**b-c**). Fc N-glycans on IgG1:F299I expressed in both BALB/c and C57BL/6 animals have more agalactosylation (Mann-Whitney p=1.02e-6), less digalactosylation, and less mono-, di- and total sialylation (Mann-Whitney p<0.0073) compared to *wt* IgG1 expressed in the same animals (**Figure** 7**c**, Supplementary **Table** 3). The increase in galactosylation in *wt* IgG1:F299 is consistent with PGD predicted IMR for upstream (N-terminal) F (**Figure** 7**a**). Upstream F is associated with increased di-

galactosylated biantennary structures (OR>2), while upstream isoleucine (I) is associated with tetraantennary galactosylation. Since only biantennary structures are generally permitted on IgG, the galactose-promoting function of upstream isoleucine should be unrealized in this glycoprotein. The increased sialylation in *wt* IgG1:F299 is also consistent with PGD IMR which show an association between structurally proximal F and disialylated structures (OR>10). These results suggest that glycoimpact can accurately predict several specific glycan epitopes.

Core-fucose preference is associated with percent fucosylation and high-ADCC

To further demonstrate IMR accuracy and functional importance, we compared predicted core-fucose preference to observed fucosylation abundance and ADCC in a Fab-constant Fc-variant antibody panel[90] (**Figure** 7**d-e**). Core-fucose preference was calculated as the preference of variant compared to the wild type for N-glycan core motif with or without a core fucose, i.e., Man(b1-4)GlcNAc(b1-4)[Fuc(a1-6)]GlcNAc(b1-4)-Asn and Man(b1-4)GlcNAc(b1-4)GlcNAc(b1-4)-Asn, respectively (see **Methods**).

We compared core-fucose preference for Fc-variant glycosylation on a surface-binding *Mycobacterium tuberculosis*-specific monoclonal antibody (clone 24c5) to demonstrate the accuracy of IMR-based predictions. We determined core-fucose preference predicted from GEE-derived IMR (see **Methods**) for multiple Fc-variants. We then compared predicted core-fucose preference to the capillary electrophoresis measured relative abundance of fucosylated glycans. The Fc-variants naturally stratified into fucose-saturated (fucosylated glycans > 90%) and unsaturated variants (Supplementary **Figure** 16). Fucose saturated variants showed no significant association with predicted fucose preference. However, fucose-unsaturated variants alone (**Figure** 7**d**) show a strong correlation between predicted core-fucose preference and percent fucosylation (R=0.9, p=0.013). 24c5 Fc glycosylation profiles therefore demonstrate another instance of glycan predictability.

To determine if predicted changes in glycosylation can also predict glycan-modulated behaviors, we examined our ability to predict increases in antibody-dependent cellular cytotoxicity (ADCC) from predicted decreases in core-fucosylation, a well-characterized property of therapeutic antibodies.[86] We identified 10 glycosite-proximal variants (**Figure** 7**e**) from the REFORM Fc variant panel[90] and stratified the ADCC-enhancing variants from those associated with no change in ADCC. Of the five sequence-proximal variants, the non-enhancing variant (T307A) shows a positive preference for core-fucose while all four ADCC associated variants have a negative core-fucose preference. Among spatially proximal variants, several substitutions occur in multiple variants and are not uniquely associated with ADCC enhancement or non-enhancement (greyed out, **Figure** 7**e**). Of the two-remaining spatially proximal

variants, the non-enhancing variant (L235A) shows a positive preference for core-fucose while the ADCC-associated variant (G236A) has a negative core-fucose preference. These ADCC associations with core-fucose preference are consistent with percent fucosylation in 24c5 Fc. T307A and L235A are highly fucosylated in 24c5, predicted to prefer core-fucose, and do not enhance ADCC. Conversely, S298A is least fucosylated in 24c5, shows a negative preference for core-fucose, and is associated with ADCC enhancement. Our predictions therefore recapitulate the immunological implications of allotype-driven differential glycosylation.

IMR-guided novel Rituximab variants are predictably and differentially glycosylated

To demonstrate our genetic control over glycosylation, we designed 5 Rituximab variants. Fc-variants were created using IMR-designed single amino acid substitutions to modulate branching and fucosylation. Rituximab variants were transiently expressed in 1mL cultures of ExpiCHO cells, purified using affinity chromatography and glycoprofiled using LC-MC-backed UPLC. By creating novel Fc variants we can explore a maximally *ab initio* design space to avoid confounding influence from unrelated design objectives used to create pre-existing Fc variants.

We observed a profound change in glycoprofiles for 4 of the 5 Fc variants including a substantial increase in branching in AB1050 (**Figure 7f**); increased branching is a notable achievement as Fc glycans are almost exclusively biantennary. To better understand the robustness of IMR predictions we compared our structure- and sequence-based predictions to random predictions measured against a binomial distribution. Both structure- and sequence-based predictions performed significantly better than random at predicting fucosylation, and terminal GlcNAc or Gal (**Figure 7g**). Both sequence- and structure-based fucosylation predictions performed significantly better than random (structure- and sequence-based performance were not significantly different from each other). In all, IMR can design completely novel variants to substantially transform glycoprofiles by modifying single residues.

Discussion

Here, we identified constraints on glycan biosynthesis through the analysis of site-specific protein features[39] and glycan substructures[38] and then used these learned associations to effectively engineer glycosylation. With our Protein-Glycan Dataset (PGD), we enumerated and quantified intramolecular relations (IMR) between protein and glycan structure in human glycoproteins. We then computed the expected differential glycosylation, the “glycoimpact,” associated with specific protein structure

changes. We validated the importance of glycoimpact by comparison to substitution matrices, evolutionary couplings, and pathogenicity scores. We further examined glycoimpact sensitivity and specificity by comparing predictions with observed glycosylation on PrP, HIV gp160, and IgG glycoproteins. The existence and utility of glycoimpact suggests that glycosylation is constrained by protein structure. We anticipate these constraints depend on which glycosyltransferases can dock and continue glycosylation resulting in a range of feasible or preferred glycosylation for specific glycosites.

The relationship presented here between glycan and protein structures seen in PGD suggests that protein structure provides important information regarding glycan structure. Of course, in extreme expression systems, protein-based predictions of glycosylation and microheterogeneity breakdown as proteostatic stress increases, the unfolded protein response is triggered, and glycosylation machinery fails to meet opportunity for glycosylation.[91,92] Such low fulfillment is seen in high-yield expression systems[93] necessitating, in some cases, supplementation with maturing glycoenzymes.[94–96] Low fulfillment is not inconsistent with protein-based constraints, rather it is a condition which subverts those constraint.

Our quantified protein-glycan IMR describe an expanded glycosite structure, beyond the original sequon definition (NX[S/T]).[8] We discovered many new sequence-proximal AA IMR both upstream and downstream of traditional glycosites. The enhanced aromatic sequon (EAS)[17] corroborates one such sequon-expanding IMR. Upstream phenylalanine IMR predicts an increase in Man7 structures (i.e., N-glycan with seven-mannoses) and a decrease in Man6 structures, suggesting an increase in larger high-mannose structures (**Figure** 3**d**). The predicted difference in oligomannose is consistent with reported EAS glycosylation; upstream phenylalanine (N-2) can decrease glycan processing and increase homogeneity.[17] Structurally, W and Y are known to stabilize the chitobiose core through dispersion and increasing glycan accessibility for maturation; F lacks a dipole and therefore does not support similar maturation[97–101]. Both the EAS and selective aromatic stacking results are consistent with glycoimpact. We predict limited expected differential glycosylation following a W to Y substitution but a substantial decrease in processing following substitution from W or Y to F (**Figure** 3**f**, Supplementary **Figure** 5). The expanded sequon scaffolded by the IMR (**Figure** 2), glycoimpact substitution matrix (**Figure** 3**f**, Supplementary **Figure** 5), and glycosite-coupled residues (**Figure** 4**c-d**) together represent a portable and dynamic summary of our findings that can be easily applied to predict glycosylation following novel substitutions. We believe that the large number of both high glycoimpact and low structural impact

substitutions (around 50% of AA have one such substitution) will be useful in understanding pathogenic variants, viral evolution, and glycoprotein engineering.

Glycoimpact can be further used to inform the mechanism of pathogenesis when disease variants are close to glycosites that impact glycosylation. We can examine known annotated-pathogenic variants and propose novel mechanisms of pathogenesis. Unannotated variants can be automatically labelled for likelihood of glycan modulation. Though this study does not deeply interrogate causal associations with glycan-modulated pathogenesis, we found several prior works implicating high-glycoimpact glycosite proximal variants as pathogenic, including one thorough study of oculocutaneous albinism implicating variant-proximal glycosylation as a causal pathogenic modulator. The highest glycoimpact relation with a high BLOSUM score, a Valine-Isoleucine substitution, is known to have a negligible protein structural impact while dramatically changing glycosylation. As expected, pathogenic variant PrP:V180I (adjacent to glycosite N181) is a causal mutation in Creutzfeldt-Jakob disease.[102] Similarly, a glycosite-proximal V84I substitution in HIV-gp120 deactivates the virus; otherwise achieved by mutagenic glycosite-ablation.[103] Similarly, tyrosinase:A355V (P14679) is a sufficient cause of oculocutaneous albinism.[52,53] A355V is also a high glycoimpact variant and close to the function-critical N371 glycosite (<20Å; PDB:5M8N); tyrosinase glycosylation is critical for proper folding,[55,104] N371 glycosite ablation results in decreased protein abundance and activity,[105] and non-mutagenic post-expression tyrosinase deglycosylation also interrupts function.[57] Using glycoimpact, we can propose A335V as a glycan-based mechanism of pathogenicity. Beyond tyrosinase, we identified thousands of variants across hundreds of diseases that may be similarly explained by aberrant glycosylation. While aberrant glycosylation can also be caused by environmental changes and proteostatic stress,[91,92] glycoimpact opens a new an additional perspective on mechanisms by which aberrations in glycosylation may be transduced. There may be millions of high glycoimpact variants close to glycosites beyond the narrower scope of ClinVar annotation for which glycoimpact can suggest a possible pathogenic mechanism.

Just as amino acid substitution can assert pathogenic glycosylation on human proteins, similar substitutions can enable immune evasion for viruses. Consistent with glycoimpact predictions, we show that glycosite proximal proline and glutamine are associated with glycan complexity across HIV gp160. These results both validate our model and propose mechanisms for evolutionary control over the viral glycan shield.[14,28,106,107] Across three SARS-CoV-2 Variants of Concern, we show that high glycoimpact variations close to glycosites produce IMR-consistent differential spike glycosylation. If IMR can predict

differential glycosylation as viruses evolve, these predictions could help predict immune evasion and predict the viral evolutionary landscape for many viruses.[14,107,108]

Functional differential glycosylation is also critical in protein therapeutics. Glycoengineering for therapeutics is often costly, application-specific, and siloed from protein design. Glycans can be modified indirectly through genome editing of glycosyltransferases[109–111] or media optimization,[112–114] and such strategies can be guided by models of the biosynthetic pathways.[115–120] Here, we examined AA-specific differential glycosylation events on antibodies. On the previously-characterized and differentially-glycosylated mouse IgG1 allotype (I299F), we see IMR-predictable glycosylation encoded directly into the glycoprotein primary sequence. Similarly, in glycosite-proximal and high glycoimpact mAb-Fc-variants, both fucosylation and fucose-modulated functional response (ADCC) correlated with GEE-IMR core-fucose preference. In the most extreme challenge, we show that novel Rituximab variants can be designed with IMR to deliberately and effectively change glycoprofiles. In these examples, IMR is a biologically[79–81] and therapeutically[90,121–124] useful method for connecting allotype, glycosylation, and effector function. Therefore, IMR present an opportunity to embed glycoengineering directly into the glycoprotein sequence and combine the currently separate paradigms of glycoengineering and protein engineering to a unified practice of glycoprotein engineering: considering the structure of both protein and glycan and their mutual influence.

Our direct results focus on human-expressed glycoproteins and global trends, rather than specific mechanisms such as aromatic chemistry, or chaperone proteins. While the specific glycoimpact trends may not extend to other organisms, the chemistry of carbohydrate-protein interactions suggests that glycoimpact is foundational and therefore exists in some form across glycosylated proteins. Additionally, these observations speak to glycosylation potential and will be most accurate in nascently-expressed proteins under limited proteostatic stress when glycogenes are expressed and substrates are available. We do not suggest these predictions to describe the only glycoprofile but rather the set of plausible glycans that may comprise a site-specific glycoprofile.

The evidence presented demonstrates the constraint of glycosylation by protein structure and the relevance of glycoimpact. These findings are corroborated by multiple distinct analyses and datasets with which protein structure appears to bound glycosylation. In all, protein structure appears to inform glycan biosynthesis. Predictability in glycosylation from protein structure will unlock a wealth of theoretical, exploratory, and corroborative analysis making it inexpensive and easy to leverage glycobiological insights in fields throughout biology.

Acknowledgements

Thank you to Terry Platt who inspired me with tales of the *Trp* operon[131–135] and in doing so planted the seed of thought that substrate-limitation is not mutually exclusive with templated biosynthesis. Thanks also to Lorenzo Casalino, Christian Seitz, and Rommie Amaro for their early insights into the project in the context of influenza glycoprotein structure. We would like to acknowledge the invaluable contribution of Dr. Jasminka Krištić, Prof. Dr. Grant Morahan and Prof. Dr. Falk Nimmerjahn to the studies of IgG N-glycome that are referred to in this paper. EA is supported by ANID (DOCTORADO BECAS CHILE/2018 - 72190270), the Fulbright Chile Commission, and the Siebel Scholar Foundation. This work was supported by NIGMS (R35 GM119850, NEL) and the Novo Nordisk Foundation (NNF20SA0066621, NEL).

Conflicts

This work is associated with a provisional patent filed by the authors, and Augment Biologics, founded by BK and NEL.

Methods

Enrichment of glycan-protein site-matched data to generate the Protein-Glycan Enriched Structure Dataset (**PGD**)

Starting from site-specific glycosylation events, we extended the annotation of each glycosylation site and glycan to include detailed site-specific protein structural annotation and recorded the number of times each substructure (defined below) appeared in each glycan. Only human glycoproteins were analyzed. The final database includes 111 human glycoproteins (98 glycoproteins with N-glycosylation sites and 38 glycoproteins with O-glycosylation sites) 306 glycosylation sites and 4,263 glycans (3,563 N- and 700 O-linked glycans). We initially used site-specific glycosylation events documented in UniCarbKB (original deprecated, latest version archived at data.glygen.org/GLY_000040) and in the January 2022 release of GlyConnect[37] (glyconnect.expasy.org/)).[41] Later and current work was informed by glycosylation events documented in GlyConnect[136] with supplemental information from GlyGen.[137] Known glycosylation events from the UniCarbKB and GlyConnect were used to inform much of the core

analysis. Glycomes and glycoproteomes in UniCarbKB were collected from GlycosuiteDB,[138] EUROCarbDB,[139] and expanded meta-analysis data.[27] GlyConnect was built on the fall 2017 release of UniCarbKB and spans glycomes and glycoproteomes predominantly (70%) curated from experiments distinct from UniCarbKB.

The protein structure annotation was done using the Structural Systems Biology (ssbio) package in python.[39] The package uses several tools to perform a variety of annotations. For each human protein, empirical and homology modeled structures were collected from the Protein Data Bank (PDB)[140] and SWISMOD,[141] respectively. Proteins without existing models were modelled using I-TASSER.[142] Protein structures and chemistry close to the glycosylation sites were annotated multiple software packages through ssbio: sequence properties (EMBOS:*pepstats*),[143] sequence alignment (EMBOS:*needle*),[143] secondary structure (DSSP,[144] SCRATCH::SSpro[145] and SCRATCH::SSpro8), solvent accessibility (DSSP and FreeSASA), and residue depth (MSMS). Additional amino acid aggregate features were calculated using R::seqinr. Spatial proximity was defined using "min-distance" between two amino acids; the minimum distance between any pair of atoms spanning the amino acids.

Glycan structures were used to represent shared structures between specific observations. Several authors have relied on the representation of glycans as graphs starting with early computer encodings such as KCF[146] or GlycoCT[147] and all the way to latest glycan language models[148]. Grounding our work in this trend, we consider a substructure as a subgraph of the full graph that represents a whole structure. Substructures were annotated using a combination of glypy[149] and GlyCompare[38] for structure parsing and comparison respectively. All glycan *substructures*, a connected subset of monosaccharides with and without linkage information, were extracted from each glycan, merged to make a superset of substructures, then mapped to each glycan. Thus, resulting in a mapping from every glycan in the input database to shared substructures. We define *motifs* as notable glycan substructures.

Software and packages

Protein structure analysis was performed in Python v2.7.15 using ssbio v0.9.9.8 to retrieve and calculate: existing empirical and homology models from PDB and SWISSMOD (PDBe SIFTS),[150] *de novo* homology models (I-TASSER v5.1), sequence properties (EMBOS v6.6.0.0 pepstats), sequence alignment(EMBOS v6.6.0.0 needle), secondary structure (DSSP v3.0.0, SCRATCHv1.1::SSpro and

SCRATCHv1.1::SSpro8), solvent accessibility (DSSPv3.0.0 and FreeSASAv2.0.2), and residue depth (MSMSv2.2.6.1). Additional amino acid aggregate features were calculated using R::seqinr.

Statistical analysis was performed in R v3.6.1. R::entropy v1.2.1 was used for entropy, Kullback-Leibler divergence and other information-theoretic calculations. Generalized Estimating Equations (GEE) were fit using R::geepack v1.3.1. Gaussian Mixture Models were used to z-score normalize the glycoimpact using R::mixtools v1.1.0. BLOSUM and PAM substitution matrixes were accessed from R::Biostrings v2.52.

Probability event space, information gain and conditional probability

Here we define an event (a row in our enriched glycosylation-glycosite database) as "the observation of a glycan at a glycosylation site in an experiment." If two separate experiments in the input database both report the same glycan at the same site on the same protein, we consider that event to have occurred twice. Within each event, we ask if the glycan structure random variable (the presence or absence of a specific glycan substructure) is present or absent in the observed glycosylation event and if the protein structure random variable (a proximal amino acid, a secondary structure or another discrete protein structural feature). A Fisher exact test (R::base::fisher.test) was used to estimate the odds ratio (OR) and significance (p) of each intra-molecular relation (IMR). P-values were corrected for False Discovery Rate (FDR, q) permitting 10% false discovery ($q<0.1$); a common threshold for systems-level analyses and distinct from p-values. Conditional probability was calculated by dividing joint probability by the marginal probability of protein and glycan structure presence. Kullback-Leibler divergence (KLd, R::entropy::KL.Dirichlet, pseudo count=1/6) was calculated by comparing the conditional probability distribution to the marginal probability distributions. In summary, we quantify an individual IMR as the Fischer exact test odds ratio (OR), significance (p), that were corrected for by the false discovery rate (FDR), which were later analyzed by clustering.

Quantitative characterization of Intra-Molecular Relations (IMR) using Generalized Estimation Equations (GEE)

To characterize the IMR in the PGD while controlling for protein-specific confounding effects and handle nonlinear relations we used a population-averaging approach; logistic Generalized Estimating Equations (GEE) with glycoprotein identity as the cluster identifier.[151] We used an exchangeable correlation structure to describe and balance the in-protein similarity. Models were fit to predict glycan substructure binary (presence or absence) from either z-score-normalized continuous or binary (presence or absence) protein structures. For each model, the data from PGD was isolated for one

glycan-type (N-glycan or O-glycan), one glycan substructure and one protein structure. Incomplete observations (events/rows) were removed and then several checks on each data-slice were run to minimize overfitting. Glycan substructures were excluded from modelling if standard deviation was less than 1e-6 or if there were fewer than 5 observations of the structure within the pertinent data-slice. Discrete protein structure features were excluded if there were fewer than 4 observations within the data-slice. Models were excluded if there were fewer than 4 instances in any cell (of the 2x2 absence/occurrence matrix) or if the chi-squared expected value of any cell was less than or equal to 5. Observations were weighted by the reciprocal-count of the corresponding label type to balance label contributions to the model and scaled by exponentiated c-score to maximize the contribution of high-quality protein structure models ($w_i = 2^c / n$); c is the c-score given by I-TASSER and n is the number of times a structure is present (1) or absent (0). Models with $|\log(OR)| > 50$ were excluded as likely overfit. Quasi-likelihood under independent model criterion (QIC) and the Wald tests were used to evaluate the significance and magnitude of the estimated IMR. We also ran this analysis using publication identifiers as a group/cluster identity variable to account for researcher and group biases; this produced similar results likely because protein identity is strongly correlated with the publications in which the proteins appear.

Calculating glycoimpact from IMR and populating a BLAMO matrix

Glycoimpact is the z-score normalized Euclidean distance between the significant log(OR) for all motifs associated with a protein structure; the null distribution of Euclidean distances was determined using Gaussian mixture models (Supplementary **Figure** 5)

Glycoimpact is calculated for every pair of AAs as the Euclidean distance between significant and substantial log odds ratios for each AA; the Euclidean distance between expected glycoprofiles for each AA. The substantial (log(OR)>X) and significant (FDR<Y) log(OR) values are retained while insignificant or unsubstantial log(OR) values are set to zero. The resulting matrix describes the expected glycoimpact due to each AA-substitution, termed the BLAMO XY matrix where X and Y denote the log(OR) and FDR thresholds respectively.

Glycoimpact values from a BLAMO XY matrix can then be z-score normalized to a Gaussian Mixture Model[152] estimated null distribution. We use z=2.5 as a heuristic but stringent cutoff between "impactful" (z>2.5) and "null" (z<2.5) substitutions.

Comparison of SNP pathogenicity scores with glycoimpact

Functional prediction rank normalized scores were obtained from dbNSFP (v3.2) for the following 27 tools: SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, 2x PhyloP, MetaSVM, MetaLR, CADD, VEST3, PROVEAN, 4× fitCons scores, fathmm-MKL, DANN, 2× phastCons, GenoCanyon, Eigen and Eigen-PC.[49] Variants were excluded from the analysis if they had more than 3 missing functional score predictions, did not result in an amino acid change, or not on proteins that had known glycosylation sites.

Assignments of "prediction-type" and "structure-usage" (**Figure** 5b) were adapted from classifications provided by dbNSFP.[49]

Estimation and Analysis of Evolutionary Couplings (EC)

For EVCouplings calculation, hits composed of more than 50% gaps were filtered from the alignment, and sequences with homologs more than 80% identical were down-weighted to compute N_{eff}, the effective number of sequences.[48] ECs were calculated using pseudo-likelihood maximization,[153,154] as implemented previously.[155] The λ_J term was scaled by the number of amino acids minus one times the number of sites in the model minus one. Pre- and post-processing was performed using the EVCouplings Python package.[156]

High-ranking EC events are generally considered those ranking less than *L* a measure of sequence length where only residues with fewer than 30% gaps were counted.[48] Specifically, the gap threshold (minimum column coverage parameter) was set to 70%. Therefore, residue positions with more than 30% gaps disregarded. We also used a fragment filter of 70%, meaning that individual sequences had to have 30% non-gap characters in each row as well. We explored multiple high-rank thresholds between *L/5* and *3L*. To explore the increased coupling with glycosylation sites, we examined couplings between each amino acid with glycosites (GN), asparagines (N) and any amino acid (AA). We compared the number of high-ranking coupling events, the distributions of EC probabilities and the relative numbers of high and low-ranking ECs for each group with various amino acids at relative positions N+/-6. Distributions were compared with a one-sided Wilcoxon test and high/low-ranking counts were compared with hypergeometric enrichment. The hypergeometric enrichment of glycosite-coupling was performed at multiple high-rank thresholds (*L/3, L/2, L, 2L, 3L*) and p-values were pooled for each amino acid at each relative position across ranks using Fisher's method. Finally, the pooled p-values were corrected for multiple tests using the Benjamini-Hochberg method.

To examine larger structures in ECs, we used EC rank to mask extended sequons (N+/-6) then clustered the sequons and extracted motifs. For each sequon, the residues were retained if the residue-glycosite coupling rank was less than *L4*. The extended and masked sequons were distinguished using a hamming distance (DECIPHERv2.18.1)[157] then clustered using agglomerative hierarchical clustering (factoextra::hcut v1.0.7). Motif logos were generated using custom-scaled position-specific scoring matrices[158] reflecting the cumulative rank of amino acids at each glycosite relative position. Specifically, the aggregate score, *S*, for each amino acid, *a*, at each position, *p*, was aggregated over EC score ranks, *r*, within each extended-masked sequons, *s*, in a cluster, *c*, such that $S_{a,p} = log_{10}\left(\sum_{s \in c} \quad L/r\right)$

Mouse breeding and Samples

The Collaborative Cross (CC) recombinant inbred mouse strains (N = 333, 95 strains, age 20-117 weeks) were produced by Geniad Pty Ltd and housed at Animal Resources Centre (Murdoch, WA, Australia).[159] The CC strains were genotyped using the MegaMUGA platform (GeneSeek; Lincoln, NE). C57BL/6 (N = 10) and BALB/c mice (N = 10), sex- and age-matched (10 weeks old, 1:1 male:female) were obtained from Elevage Janvier (Le Genest-Saint-Isle, France).[89] The studies received appropriate ethics approvals from the Animal Ethics Committee of the Animal Resources Centre[78] and the Ethical Committee of the District Government of Lower Franconia.[89]

Liquid Chromatography – Mass Spectrometry (LC-MS), Normalization and Statistical Analysis of Mouse Fc-linked IgG N-glycopeptides

Immunoglobulin G was isolated from 100-500 µl of mouse serum on 96-well Protein G monolithic plates (BIA Separations) as described previously.[78,88] LC-MS analysis of tryptic Fc-glycopeptides was performed as described in.[78,88] In brief, approximately 10–20 µg of isolated IgG was digested with 200 ng trypsin (Worthington, USA). The resulting glycopeptides were purified by reverse-phase solid phase extraction using Chromabond C18ec beads (Marcherey-Nagel, Germany) as described in.[88] Tryptic digests were analyzed on a nanoACQUITY UPLC system (Waters, USA) coupled to a Compact mass spectrometer (Bruker Daltonics, Germany). Peak areas were calculated by summing areas for doubly and triply charged ions determined with LaCyTools v 1.0.1 b.7 software[160] and normalized to the total integrated area per IgG subclass.

Batch correction was performed on the log-transformed values using the ComBat method (R package "sva") to remove possible experimental variations due to LC-MS analysis having been performed on several 96-well plates within each cohort. Derived glycosylation traits describing relative abundance of

N-glycans sharing specific structural features (agalactosylated, galactosylated, sialylated, monogalactosylated, digalactosylated, monosialylated, disialylated structures, structures with bisecting GlcNAc) were calculated in a subclass-specific manner.[89] Statistical analysis and data visualization were performed using R programming language v 4.0.3.

Glycoproteomics for SARS-CoV-2 Spike glycoproteins

His-tagged recombinant SARS-CoV-2 Spike S1 protein, expressed in HEK293 cells were purchased from Sino Biologics (Wayne, PA). Lyophilized glycoproteins corresponding to the original 2019 strain (40591-V08H), the Gamma variant (40591-V08H14 with L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, D614G and H655Y mutations), and the Delta variant (40591-V08H23 with T19R, G142D, E156G, 157-158 deletion, L452R, T478K, D614G and P681R mutations). Samples were analyzed using DeGlyPHER.[77]

Briefly, glycoproteins were digested with Proteinase K (30 min or 4 h) or trypsin, sequentially deglycosylated with Endo H (creating residual mass signature of +203 Da) to signify high mannose/hybrid glycans, and then with PNGase F in $H_2{}^{18}O$ (creating residual mass signature of +3 Da) to signify the remnant complex glycans on any sequon (NXS|T, where X is any amino acid except P) asparagine. Unoccupied sequons will have no additional signature mass. Analysis of samples was done on a Q Exactive HF-X mass spectrometer (Thermo), injecting directly onto a 25 cm, 100 µm ID column packed with BEH 1.7 µm C18 resin (Waters). Liquid chromatography separation was achieved at a flow rate of 300 nL min−1 on an EASY-nLC 1200 (Thermo). Buffers A and B were 0.1% formic acid in 5 and 80% acetonitrile, respectively. The gradient used was 1−25% B over 160 min, an increase to 40% B over 40 min, an increase to 90% B over another 10 and 30 min at 90% B for a total run time of 240 min. The column was re-equilibrated with solution A before injecting sample. Peptides eluting from the tip of the column were nanosprayed directly into the mass spectrometer by application of 2.8 kV at the back of the column. The mass spectrometer was operated in a data-dependent mode. Full MS1 scans were collected in the Orbitrap at 120000 resolution. The 10 most abundant ions per scan were selected for HCD MS/MS at 25 NCE. Dynamic exclusion was enabled with a 10 s duration and +1 ions were excluded. Peptides were identified with Integrated Proteomics Pipeline (IP2, Bruker Scientific LLC). Tandem mass spectra were extracted from raw files using RawConverter[161] and searched with ProLuCID[162] against a database comprising UniProt reviewed proteome for Homo sapiens (UP000005640), including additional UniProt amino acid sequences for Endo H (P04067), PNGase F (Q9XBM8), and Proteinase K (P06873), the amino acid sequences for the SARS-CoV-2 S1 subunits, and a list of general protein contaminants. The search included no protease specificity (all fully tryptic and semitryptic peptide candidates when treated

with trypsin). Carbamidomethylation (+57.02146 C) was used as a static modification. Deamidation in the presence of $H_2{}^{18}O$ (+2.988261 N), GlcNAc (+203.079373 N), oxidation (+15.994915 M), and N-terminal pyroglutamate formation (−17.026549 Q) were used as differential modifications. Data were searched with 50 ppm parent mass tolerance and 50 ppm fragment mass tolerance. Identified proteins were filtered using DTASelect2[163] while using a target-decoy database search strategy to limit the false discovery rate to 1%, at the spectrum level.[164] At least one peptide per protein and no tryptic end (or one tryptic end when treated with trypsin) per peptide were necessary, and precursor delta mass cutoff was fixed at 10 ppm. Statistical models for peptide mass modification (modstat) were applied (trypstat was additionally applied for trypsin-treated samples). Semi-quantitative label-free analysis was performed based on the precursor peak area, with a 10 ppm parent mass tolerance and 0.1 min retention time tolerance, Census2.[165] "Match between runs" was used to find missing peptides between runs. GlycoMSQuant (v.1.4.1, https://github.com/proteomicsyates/GlycoMSQuant)[77] was used, summing precursor peak areas across the 3 conditions – 30 min and 4 h Proteinase K, and trypsin, discarded peptides without glycosites, and discarded misidentified peptides when N-glycan remnant mass modifications were localized to non-glycosite asparagines and corrected/fixed N-glycan mis-localization where appropriate, to finally calculate proportions of 3 glycosylation states at each glycosite, with +/-SEM (standard error of mean). Pairwise-statistical comparison between variants was performed using Mann-Whitney U Test.

The entire DeGlyPHER pipeline and measurement were run twice to collect two independent technical replicates. Using the Fisher's method for pooling independent p-values, we combined the Mann-Whitney U-derived p-values for each glycosite comparison. Pooled p-values were adjusted for multiple-testing using FDR. The N17 glycosite was inconsistently cleaved with the signal peptide precluding stable measurements needed for robust comparison.

Estimation of glycomotif preference from IMR

Here, we define glycomotif (e.g., core-fucose) preference. Briefly, this is the preference for a glycomotif following a substitution compared to the wildtype (WT) and a close (within one monosaccharide) precursor of the glycomotif of interest. When the IMR relating the lost (WT, x-axis) and gained (variant, y-axis) amino acids association (IMR) to the glycomotif and glycomotif precursor, the variant preference can be calculated as the distance between the glycomotif and glycomotif-precursor points; distance is the component perpendicular to the line of equity (y=x). Supplementary **Figure** 15 demonstrates this calculation visually.

Glycomotif preference can be described in more precise terms. Given a substitution where X and Y are WT and mutant amino acids respectively at an amino acid residue at index I, we may describe the substitution as XIY. When I is glycosite-proximal, we may describe a substitution, XIY, as a point, p, with a pair of IMR (ρ) relating each amino acid to a single glycomotif, $p_{XIY} = (\rho_X, \rho_Y)$. The IMR-pair, $p_{XIY} = (\rho_X, \rho_Y)$, may also be described as a vector terminating on that point, $\vec{p}_{XIY}$. We may also represent the line of equality ($y = x$) by the normalized vector $\vec{n} = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$. Here, the glycoimpact is the minimum distance, $G = d(\vec{p}, \vec{n})$. To calculate the component of $\vec{p}$ parallel to $\vec{n}$, we project $\vec{p}$ onto $\vec{n}$, $\vec{p}_{\vec{n}} = \frac{\vec{p} \cdot \vec{n}}{\vec{n} \cdot \vec{n}} \cdot \vec{n}$. Therefore, $G = d(\vec{p}, \vec{n}) = \vec{p} - \vec{p}_{\vec{n}}$

With this definition of G, we can define G_i and G_c as the glycoimpact corresponding to the glycomotif of interest and the contract glycomotif respectively. Then, the relative preference by first calculating the perpendicular component of the difference in glycoimpact vectors projected onto $\vec{n}$, such that $r_{\vec{n}} = (G_i - G_c)_{\vec{n}}$. Then R is the l^2-norm of the vector difference, $R = s|r_{\vec{n}}|$. For $r_{\vec{n}} = \langle x, y \rangle$, we define the sign, s, as positive if $y > x$ and negative if $x > y$. The sign is not defined if $x = y$.

Glycoprofiling of 25c4 monoclonal antibody using capillary electrophoresis

As was recently described,[166] 54 Fc variants of 25c4 were cloned from the REFORM Fc variant panel.[90] The REFORM plasmid library[90] is comprised of golden gate cloning plasmids[167] with BsaI restriction sites flanking distinct antibody domains and a furin 2A cleavage site to enable self-cleavage and successful assembly of a complete antibody from a single open reading frame.[168] Variable Light chains containing REFORM variants and 25c4 Variable Heavy chains were simultaneously transfected at a 1:1 ratio in CHO cells, purified using a Protein A chromatography resin, then dialyzed and concentrated with Phosphate Saline Buffer (PBS).

Fc glycosylation was then measured on purified 25c4 antibodies. Purified antibodies were incubated with magnetic G protein beads (Millipore) and separated from Fab fragments following enzymatic digestion (IdeZ, NEB). Fc glycans were released and labelled using the GlycanAssure ATPS kit (Thermo Fisher Scientific) then separated using 3500xL Genetic Analyzer (Thermo Fisher Scientific). As previously described,[169] retention times (RT) were matched to glycan standards using Glycan Acquisition Software Version 3500 v1.0.3 and Glycan Analysis Software v1.1. Abundance (area under peaks) was normalized to total area per sample to calculate relative abundance of each glycan. Non-uniquely determined glycans were excluded (where the difference in RT was below the detection threshold).

Expression, purification and glycoprofiling of Rituximab variants using LC-MS-backed UPLC

Rituximab titers were determined in triplicate on an Octet® Red96 biolayer interferometry (BLI) instrument. Binding to ProA biosensors was recorded for 120 s at 30 °C. Binding rates were converted to concentrations based on a standard curve generated using wildtype Rituximab, produced and purified in-house.

The Rituximab variants were purified by affinity chromatography using a 1-mL MAb Select Sure column (Cytiva) mounted on an Äkta Pure instrument. Equilibration and washing steps were performed using 20 mM sodium phosphate, 0.15 M NaCl, pH 7.2. The antibody was eluted with 0.1M Sodium citrate, pH 3. Elution fractions were neutralized with 0.2 V of 1 M Tris, pH 9. Next, the protein solutions were desalted using 5-mL Zeba™ Spin desalting columns (7K MWCO, Thermo Fisher) and dPBS as eluent. Finally, the desalted solutions were concentrated on 4-mL Amicon centrifugal filter units (50K MWCO, Millipore) aiming for a concentration of approximately 0.5 mg/mL. The final concentrations were determined by measuring absorbance at 280 nm on a Nanodrop 2000 spectrophotometer using an extinction coefficient of 1.46 $(mg/mL)^{-1}cm^{-1}$.

Purified and concentrated protein extract were fluorescently labelled (N-glycan labeling using the GlycoWorks RapiFluor-MS N-Glycan Kit, Waters, Milford, MA). Fluorescent glycans were stratified and measured using a HILIC-FLR with ACQUITY UPLC Glycan BEH Amide column (2.1 x 150 mm, 1.7 µm, Waters, Milford, MA) mounted on an Ultimate 3000 UPLC system and a Fusion Orbitrap mass spectrometer (Thermo Scientific). Acetonitrile (100%) and ammonium formate (50 mM, pH 4.4) were used as mobile phases.

Captions

Figure 1 – (**a**) DNA, RNA, and protein biosynthesis are template-driven processes. Glycan biosynthesis is described as metabolically and enzymatically constrained. (**b**) Glycoprotein 2D and 3D structural features are considered in relation to proximal glycosylation. (**c**) Site-specific glycosylation data (left) are used to estimate associations between site-specific protein glycan occurrence and proximal sequence (b, left) or structure (b, right), as quantified with our Intra-Molecular Relations (IMR) metric (middle). Following amino acid changes, IMR reflect the agreement between expected glycosylation and a glycoprotein structure, while disagreement is singled out and called "glycoimpact" to capture glycan sequence changes (right). Glycans are represented using IUPAC-extension for glycans and the Symbol Nomenclature for Glycans (SNFG[1,2]); Mannose (Man), Galactose (Gal), Sialic Acid (Neu5Ac), N-Acetylglucosamine (GlcNAc), Fucose (Fuc). (**d**) We validated predictions of glycoprotein structure constraining glycosylation with computational results, published experimental data, and novel experimental results across pathogenic gene variants, evolution, viral proteins, and antibodies.

Figure 2 **– N- and O-glycan substructures associated with glycosite-proximal protein structure. (a)** Dimensionality reduction trained on UniCarbKB glycosite-proximal protein structures. When projected into the UniCarbKB-trained space, all human UniProt annotated glycosites appear within the UniCarbKB sequence space implying that UniCarbKB glycosites are sufficiently representative of most documented glycosites. The scatterplots show a two-dimensional projection of the FAMD using a Uniform Manifold Approximation and Projection (UMAP).[40] Each point is a site on a glycoprotein, each color indicates the source of that protein. (**b**) Volcano plot of the log odds ratio and False Discovery Rate adjusted p-values from a Fisher exact test

between co-occurring glycan and protein structures. We observe IMR, with most significant relations associated with alanine, cysteine, valine, and glutamine (A, C, V, Q, respectively). (**c**) IMR were also estimated by logistic GEE controlling for protein-bias. The number of significant (FDR<0.1, |log(OR)|>0.1) IMR relating to structurally proximal AAs (N+6Å), sequence proximal AAs C-terminal (N+5), N-terminal (N-5) or either direction (N+/-5), predicted secondary structure from sequence (SSpro8) and structure (DSSP): alpha-helix (ss.H), extended strand (ss.E), beta-bridge (ss.B), turn (ss.T) bend (ss.S), other (ss.C).

Figure 3 **– Glycosite-proximal amino acids impact glycosylation.** Specific IMRs were discovered by logistic GEE controlling for protein-bias. **(a-b)** IMR (FDR<0.1, |log(OR)|>0.1) relating structurally-proximal amino acids to motifs stratified by the number of Sialic Acids (**a**) and 4-Sulfated GalNAc (**b**). **(c)** Spearman correlation between the monosaccharide count of glycan substructures from protein-structure features; protein structure features with an average absolute correlation > 0.2 were retained. Terms used here are consistent with panel **a** but "aa" denotes a structurally proximal amino acid, "aaUp," "aaDown," and "aaAll" denotes N-terminus, C-terminus, or any sequence proximal amino acid. (**d-e**) IMR (FDR<0.1, | log(OR)|>0.1) compared across two sequence-proximal (d) and spatially proximal (**e**) amino acids, phenylalanine (F) and tryptophan (W). The direct comparison of proximal-amino acid effects visualized the expected change in glycosylation associated with that substitution. (**f**) Network depicting the glycoimpact (see ***Methods***) of spatially proximal (within 5Å) substitutions for structurally low impact (BLOSUM62) substitutions. Panel **f** shows glycoimpact predicted from BLAMO 0.5:0.1 (|log(OR)| > 0.5, FDR<0.1), see *Supplementary* ***Figure 5*** for additional thresholds, raw scores, and sequence-proximal substitution predictions. Note that plots including glycan substructures were manually curated to summarize the selected glycan substructures.

Figure 4 **- Glycoimpact of amino acid substitution correlates with evolution metrics and conservation.** (**a**) Comparison of the error between the PAM and BLOSUM substitution matrices and the glycoimpact for corresponding substitutions. Linear regressions are split into null glycoimpact (<2.5) and glycoimpactful (>2.5). Glycoimpact scores from BLAMO 0.5:0.1 were used; those computed from strong IMR (|log(OR)| > 0.5, FDR<0.1). Error (y-axis) was calculated as the root mean square error between PAM and BLOSUM scores. Pearson's Correlation (*r*) significance indicated as <0.001 (***), <0.01 (**), and <0.05 (*). (**b**) Glycosite Alignments [15] corresponding to GlyConnect-documented tetraantennary structures (Hex:7 HexNAc:6) with no sialic acid or fucose. See *Supplementary* ***Figure 9*** for full alignment. The first (top line) and second (bottom line) most popular amino acids are displayed for each position N+/-30. Consensus AAs consistent with other analyses are highlighted in bold and marked with a "+" (Glycosite-coupled, ***Figure 4***h) or "*" (high-influence AA, ***Figure 2***c). (**c**) An aggregation of all residue-glycosite enrichments at N+/-10 (hypergeometric enrichment illustrated in panel *Supplementary* ***Figure 7***c). The proportion of high-ranking Evolutionary Couplings (EC) for each amino acid (rows) at the column-specified relative position was compared with glycosites (GN), asparagines (N), or any residue (X). An opaque red circle indicates that for the residue (row) in the given position (column), high-EC proportion is higher with GN than N. An opaque black triangle indicates high-EC proportion is higher with GN than X. A transparent circle or square indicates GN was not significantly more coupled (hypergeometric test). Significance was assessed at multiple EC-rank thresholds between L/3 and 3L (see Methods) and pooled using Fisher's method (FDR <0.1). (**d**) Hierarchical clustering of coupling-masked (Rank<4L) amino acids surrounding a glycosite (+/-6 AA). Each of 5 clusters was summarized as a motif. Height is the log of cumulative reciprocal EC-Rank with a pseudo-count of 0.25. The asparagine at the center was fixed at 2 for context. Residues are colored by chemical properties. A more granular clustering, 25 clusters is included in the supplement (*Supplementary* ***Figure 8***).

Figure 5 **- Substitution glycoimpact predicts pathogenicity.** (**a**) Boxplots show the min-distance from all residues within human PrP to the N197 or N181 glycosylation sites. Residues are stratified by all sites (All) and causative mutations of prion disease including Creutzfeldt-Jakob disease (CJD) and Gerstmann-Straussler disease (GSS), (one-sided Wilcoxon test). For glycosite-specific proximity, see *Supplementary* ***Figure 11***. (**b**) A hierarchical clustered heatmap (average-linkage with Euclidean distance) of Spearman correlation coefficients between glycoimpact (BLAMO 0.5:0.1) and error between variant impact prediction scores. Prediction-type and protein structure indicate the training data used to build various tools as described in dbNSFP [49]. Each row and column refer to a variant function prediction tool. (**c**) Null and impactful glycoimpact (BLAMO 0.5:0.1) stratified by variant pathogenicity in ClinVar within 20Å (min-distance) of an N-glycosylation site. Two-dimensional density plots compare glycoimpact and the glycosite-mutation distance.

Figure 6 – Changes in high-mannose, hybrid, and complex glycosylation are predicted by glycosite-proximal high-glycoimpact substitutions in HIV, SARS-CoV-2. (**a**) GEE-calculated IMR from PGD denoting relations between sequence (triangle & square) or structural protein features (circle & plus) and motifs containing >3 mannose (hybrid & high-mannose). (**b**) The range of mass spectrometry peaks consistent with either oligomannose or hybrid glycans (N203) or complex glycan peaks (N3) at each site on the HIV envelope gp160 (BG505 SOSIP.664, PDB:4TVP).[70] Relative abundance is represented as a ratio of oligomannose-hybrid to complex glycan peaks (N203/N3). (**c**) Distributions of oligomannose-hybrid to complexity (N203/N3, panel **b**) glycan peaks stratified by proximal protein structure features selected in panel **a**. For the N203/N3 distribution of select IMR, see *Supplementary* ***Figure 12***. (**d**) Mean proportion of mass-spectrometry-observed peptide with mass offsets corresponding to complex glycan peaks (+3, purple), oligomannose or hybrid glycan peaks (+203, green), or no glycosylation (+0, grey) in the SARS-CoV-2 S1 subunit at 3 sites with significant glycosylation differences between the original ancestral strain, and the Delta

and Gamma variants (see *Supplementary **Figure** 13* for all sites). Peak count for each glycosylation offset type (+3, +203, or +0) for each glycosylation site is divided by the total number of peaks for that site to determine the site-specific proportion of each glycosylation type. Significant differential glycosylation (FDR<0.05) in VOC compared to ancestral is indicated by "*".

***Figure** 7* –Differential glycosylation is predicted by high-glycoimpact events near glycosylation sites in IgG (**a**) GEE-learned IMR relating to sequence-proximal (upstream/N-terminal) effects of I and F. IgG allotypes, F299 and I299, segregated by Principal Component Analysis of relative abundance (**b**) separate by allotype, not strain. (**c**) Galactose (Gal), Sialylation (Neu5Ac) and Bisecting GlcNAc abundance distributions for IgG1 F299 and I299 allotypes across BALB/c and C57BL/6 mice (see *Supplementary **Figure** 14*). (**d**-**e**) Core-fucose preference (see **Methods,** *Supplementary **Figure** 15*) for each glycosite-proximal variant in the REFORM Fc variant panel.[90] Variants are colored by index. (**d**) Variants are stratified continuously (top) by percent of fucosylated Fc-cleaved glycans observed on the 24C5 monoclonal antibody. (**e**) Variants are stratified categorically (bottom) by association with enhanced ADCC.[90] Variants appearing in Fc regions associated with both no-enhancement and enhanced ADCC are shown as transparent. Only L235A and G236A are uniquely associated with ADCC. (**f**) *UPLC chromatogram of PNGaseF-released mAb Fc glycans. The expected Rituximab glycoprofile (black) and GlycoTemplated Rituximab variant AB1050 (red) are shown to visualize the changes to the glycoprofile (glycans verified via LC-MS).* (**g**) *Number (n) of Rituximab variants (of 5) correctly predicted (n) to increase, decrease, or not-change fucosylation, terminal galactose (gal) or terminal N-acetylglucosamine (GlcNAc). Predictions were either from Random estimation, sequence-based IMR-guided predictions (Seq Pred), or protein structure-based IMR-guided predictions (Struct Pred). Significance of consistency between experimental observations and predictions was calculated by a binomial distribution (* < 0.05, ** < 0.01, *** < 0.001)*

References

1. Varki, A., Cummings, R.D., Aebi, M., Packer, N.H., Seeberger, P.H., Esko, J.D., Stanley, P., Hart, G., Darvill, A., Kinoshita, T., et al. (2015). Symbol Nomenclature for Graphical Representations of Glycans. Glycobiology *25*, 1323–1324.

2. Neelamegham, S., Aoki-Kinoshita, K., Bolton, E., Frank, M., Lisacek, F., Lütteke, T., O'Boyle, N., Packer, N.H., Stanley, P., Toukach, P., et al. (2019). Updates to the Symbol Nomenclature for Glycans guidelines. Glycobiology *29*, 620–624.

3. Pothukuchi, P., Agliarulo, I., Russo, D., Rizzo, R., Russo, F., and Parashuraman, S. (2019). Translation of genome to glycome: role of the Golgi apparatus. FEBS Lett. *593*, 2390–2411.

4. Bagdonaite, I., Malaker, S.A., Polasky, D.A., Riley, N.M., Schjoldager, K., Vakhrushev, S.Y., Halim, A., Aoki-Kinoshita, K.F., Nesvizhskii, A.I., Bertozzi, C.R., et al. (2022). Glycoproteomics. Nat. Rev. Methods Primers *2*. https://doi.org/10.1038/s43586-022-00128-4.

5. Reily, C., Stewart, T.J., Renfrow, M.B., and Novak, J. (2019). Glycosylation in health and disease. Nat. Rev. Nephrol. *15*, 346–366.

6. Makrydaki, E., Donini, R., Krueger, A., Royle, K., Moya Ramirez, I., Kuntz, D.A., Rose, D.R., Haslam, S.M., Polizzi, K.M., and Kontoravdi, C. (2024). Immobilized enzyme cascade for targeted glycosylation. Nat. Chem. Biol. https://doi.org/10.1038/s41589-023-01539-4.

7. Johnson, R.L., and Deutsch, H.F. (1970). Preparation and studies of myeloma Fab subfractions. Immunochemistry *7*, 207–215.

8. Marshall, R.D. (1974). The nature and metabolism of the carbohydrate-peptide linkages of glycoproteins. Biochem. Soc. Symp., 17–26.

9. Shakin-Eshleman, S.H., Spitalnik, S.L., and Kasturi, L. (1996). The amino acid at the X position of an Asn-X-Ser sequon is an important determinant of N-linked core-glycosylation efficiency. J. Biol. Chem. *271*, 6363–6366.

10. Petrescu, A.-J., Milac, A.-L., Petrescu, S.M., Dwek, R.A., and Wormald, M.R. (2004). Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding. Glycobiology *14*, 103–114.

11. Kasturi, L., Eshleman, J.R., Wunner, W.H., and Shakin-Eshleman, S.H. (1995). The hydroxy amino acid in an Asn-X-Ser/Thr sequon can influence N-linked core glycosylation efficiency and the level of expression of a cell surface glycoprotein. J. Biol. Chem. *270*, 14756–14761.

12. Lund, J., Takahashi, N., Pound, J.D., Goodall, M., and Jefferis, R. (1996). Multiple interactions of IgG with its core oligosaccharide can modulate recognition by complement and human Fc gamma receptor I and influence the synthesis of its oligosaccharide chains. The Journal of Immunology *157*, 4963–4969.

13. Altman, M.O., Angel, M., Košík, I., Trovão, N.S., Zost, S.J., Gibbs, J.S., Casalino, L., Amaro, R.E.,

Hensley, S.E., Nelson, M.I., et al. (2019). Human Influenza A Virus Hemagglutinin Glycan Evolution Follows a Temporal Pattern to a Glycan Limit. Preprint, https://doi.org/10.1128/mbio.00204-19 https://doi.org/10.1128/mbio.00204-19.

14. Yu, W.-H., Zhao, P., Draghi, M., Arevalo, C., Karsten, C.B., Suscovich, T.J., Gunn, B., Streeck, H., Brass, A.L., Tiemeyer, M., et al. (2018). Exploiting glycan topography for computational design of Env glycoprotein antigenicity. PLoS Comput. Biol. *14*, e1006093.

15. Gastaldello, A., Alocci, D., Baeriswyl, J.-L., Mariethoz, J., and Lisacek, F. (2016). GlycoSiteAlign: Glycosite Alignment Based on Glycan Structure. J. Proteome Res. *15*, 3916–3928.

16. Huang, Y.-W., Yang, H.-I., Wu, Y.-T., Hsu, T.-L., Lin, T.-W., Kelly, J.W., and Wong, C.-H. (2017). Residues comprising the enhanced aromatic sequon influence protein N-glycosylation efficiency. J. Am. Chem. Soc. *139*, 12947–12955.

17. Murray, A.N., Chen, W., Antonopoulos, A., Hanson, S.R., Wiseman, R.L., Dell, A., Haslam, S.M., Powers, D.L., Powers, E.T., and Kelly, J.W. (2015). Enhanced aromatic sequons increase oligosaccharyltransferase glycosylation efficiency and glycan homogeneity. Chem. Biol. *22*, 1052–1062.

18. Gupta, R., and Brunak, S. (2002). Prediction of glycosylation across the human proteome and the correlation to protein function. Pac. Symp. Biocomput., 310–322.

19. Steentoft, C., Vakhrushev, S.Y., Joshi, H.J., Kong, Y., Vester-Christensen, M.B., Schjoldager, K.T.-B.G., Lavrsen, K., Dabelsteen, S., Pedersen, N.B., Marcos-Silva, L., et al. (2013). Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. EMBO J. *32*, 1478–1488.

20. Ives, C.M., Singh, O., D'Andrea, S., Fogarty, C.A., Harbison, A.M., Satheesan, A., Tropea, B., and Fadda, E. (2023). Restoring Protein Glycosylation with GlycoShape. bioRxiv. https://doi.org/10.1101/2023.12.11.571101.

21. Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., and Brunak, S. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics *4*, 1633–1649.

22. Silverman, J.M., and Imperiali, B. (2016). Bacterial N-Glycosylation Efficiency Is Dependent on the Structural Context of Target Sequons. J. Biol. Chem. *291*, 22001–22010.

23. Kawatkar, S.P., Kuntz, D.A., Woods, R.J., Rose, D.R., and Boons, G.-J. (2006). Structural Basis of the Inhibition of Golgi α-Mannosidase II by Mannostatin A and the Role of the Thiomethyl Moiety in Ligand–Protein Interactions. Preprint, https://doi.org/10.1021/ja061216p https://doi.org/10.1021/ja061216p.

24. Li, B., Kawatkar, S.P., George, S., Strachan, H., Woods, R.J., Siriwardena, A., Moremen, K.W., and Boons, G.-J. (2004). Inhibition of Golgi Mannosidase II with Mannostatin A Analogues: Synthesis, Biological Evaluation, and Structure-Activity Relationship Studies. Chembiochem *5*, 1220–1227.

25. Hang, I., Lin, C.-W., Grant, O.C., Fleurkens, S., Villiger, T.K., Soos, M., Morbidelli, M., Woods, R.J., Gauss, R., and Aebi, M. (2015). Analysis of site-specific N-glycan remodeling in the endoplasmic reticulum and the Golgi. Glycobiology *25*, 1335–1349.

948 26. Xiang, Y., Karaveg, K., and Moremen, K.W. (2016). Substrate recognition and catalysis by GH47 α-
949 mannosidases involved in Asn-linked glycan maturation in the mammalian secretory pathway. Proc.
950 Natl. Acad. Sci. U. S. A. *113*, E7890–E7899.

951 27. Thaysen-Andersen, M., and Packer, N.H. (2012). Site-specific glycoproteomics confirms that protein
952 structure dictates formation of N-glycan type, core fucosylation and branching. Glycobiology *22*,
953 1440–1452.

954 28. Allen, J.D., Chawla, H., Samsudin, F., Zuzic, L., Shivgan, A.T., Watanabe, Y., He, W.-T., Callaghan, S.,
955 Song, G., Yong, P., et al. (2021). Site-specific steric control of SARS-CoV-2 spike glycosylation. Cold
956 Spring Harbor Laboratory, 2021.03.08.433764. https://doi.org/10.1101/2021.03.08.433764.

957 29. García-García, A., Serna, S., Yang, Z., Delso, I., Taleb, V., Hicks, T., Artschwager, R., Vakhrushev, S.Y.,
958 Clausen, H., Angulo, J., et al. (2021). FUT8-directed core fucosylation of N-glycans is regulated by
959 the glycan structure and protein environment. ACS Catal. *11*, 9052–9065.

960 30. Losfeld, M.-E., Scibona, E., Lin, C.-W., Villiger, T.K., Gauss, R., Morbidelli, M., and Aebi, M. (2017).
961 Influence of protein/glycan interaction on site-specific glycan heterogeneity. FASEB J. *31*, 4623–
962 4635.

963 31. Losfeld, M.-E., Scibona, E., Lin, C.-W., and Aebi, M. (2022). Glycosylation network mapping and site-
964 specific glycan maturation in vivo. iScience, 105417.

965 32. Mathew, C., Weiß, R.G., Giese, C., Lin, C.-W., Losfeld, M.-E., Glockshuber, R., Riniker, S., and Aebi,
966 M. (2021). Glycan-protein interactions determine kinetics of N-glycan remodeling. RSC Chem Biol *2*,
967 917–931.

968 33. Adams, T.M., Zhao, P., Chapla, D., Moremen, K.W., and Wells, L. (2022). Sequential in vitro
969 enzymatic N-glycoprotein modification reveals site-specific rates of glycoenzyme processing. J. Biol.
970 Chem., 102474.

971 34. Bagdonas, H., Fogarty, C.A., Fadda, E., and Agirre, J. (2021). The case for post-predictional
972 modifications in the AlphaFold Protein Structure Database. Nat. Struct. Mol. Biol. *28*, 869–870.

973 35. Tsai, Y.-X., Chang, N.-E., Reuter, K., Chang, H.-T., Yang, T.-J., von Bülow, S., Sehrawat, V., Zerrouki,
974 N., Tuffery, M., Gecht, M., et al. (2024). Rapid simulation of glycoprotein structures by grafting and
975 steric exclusion of glycan conformer libraries. Cell *187*, 1296-1311.e26.

976 36. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L.,
977 Ballard, A.J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions
978 with AlphaFold 3. Nature. https://doi.org/10.1038/s41586-024-07487-w.

979 37. Alocci, D., Mariethoz, J., Gastaldello, A., Gasteiger, E., Karlsson, N.G., Kolarich, D., Packer, N.H., and
980 Lisacek, F. (2019). GlyConnect: Glycoproteomics Goes Visual, Interactive, and Analytical. J.
981 Proteome Res. *18*, 664–677.

982 38. Bao, B., Kellman, B.P., Chiang, A.W.T., Zhang, Y., Sorrentino, J.T., York, A.K., Mohammad, M.A.,
983 Haymond, M.W., Bode, L., and Lewis, N.E. (2021). Correcting for sparsity and interdependence in
984 glycomics by accounting for glycan biosynthesis. Nat. Commun. *12*, 4988.

39. Mih, N., Brunk, E., Chen, K., Catoiu, E., Sastry, A., Kavvas, E., Monk, J.M., Zhang, Z., and Palsson, B.O. (2018). ssbio: a Python framework for structural systems biology. Bioinformatics *34*, 2155–2157.

40. McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv [stat.ML].

41. Campbell, M.P., Peterson, R., Mariethoz, J., Gasteiger, E., Akune, Y., Aoki-Kinoshita, K.F., Lisacek, F., and Packer, N.H. (2014). UniCarbKB: building a knowledge platform for glycoproteomics. Nucleic Acids Res. *42*, D215-21.

42. Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science *347*, 1260419.

43. Allen, J.D., Chawla, H., Samsudin, F., Zuzic, L., Shivgan, A.T., Watanabe, Y., He, W.-T., Callaghan, S., Song, G., Yong, P., et al. (2021). Site-specific steric control of SARS-CoV-2 spike glycosylation. Biochemistry *60*, 2153–2169.

44. Fitch, W.M., and Dayhoff, M.O. (1973). Atlas of protein sequence and structure, 1972. Syst. Zool. *22*, 196.

45. Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U. S. A. *89*, 10915–10919.

46. Mount, D.W. (2008). Comparison of the PAM and BLOSUM amino acid substitution matrices. CSH Protoc. *2008*, db.ip59.

47. Pearson, W.R. (2013). Selecting the right similarity-scoring matrix. Curr. Protoc. Bioinformatics *43*, 3.5.1-3.5.9.

48. Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. PLoS One *6*, e28766.

49. Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). DbNSFP v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. Hum. Mutat. *37*, 235–241.

50. Thusberg, J., Olatubosun, A., and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. Hum. Mutat. *32*, 358–368.

51. Cheng, N., Li, M., Zhao, L., Zhang, B., Yang, Y., Zheng, C.-H., and Xia, J. (2020). Comparison and integration of computational methods for deleterious synonymous mutation prediction. Brief. Bioinform. *21*, 970–981.

52. Simeonov, D.R., Wang, X., Wang, C., Sergeev, Y., Dolinska, M., Bower, M., Fischer, R., Winer, D., Dubrovsky, G., Balog, J.Z., et al. (2013). DNA variations in oculocutaneous albinism: an updated mutation list and current outstanding issues in molecular diagnostics. Hum. Mutat. *34*, 827–835.

53. Spritz, R.A., Ho, L., Furumura, M., and Hearing, V.J., Jr (1997). Mutational analysis of copper binding by human tyrosinase. J. Invest. Dermatol. *109*, 207–212.

54. Halaban, R., Svedine, S., Cheng, E., Smicun, Y., Aron, R., and Hebert, D.N. (2000). Endoplasmic reticulum retention is a common defect associated with tyrosinase-negative albinism. Proc. Natl. Acad. Sci. U. S. A. *97*, 5889–5894.

55. Branza-Nichita, N., Negroiu, G., Petrescu, A.J., Garman, E.F., Platt, F.M., Wormald, M.R., Dwek, R.A., and Petrescu, S.M. (2000). Mutations at critical N-glycosylation sites reduce tyrosinase activity by altering folding and quality control. J. Biol. Chem. *275*, 8169–8175.

56. Dolinska, M.B., Kovaleva, E., Backlund, P., Wingfield, P.T., Brooks, B.P., and Sergeev, Y.V. (2014). Albinism-causing mutations in recombinant human tyrosinase alter intrinsic enzymatic activity. PLoS One *9*, e84494.

57. Dolinska, M.B., and Sergeev, Y.V. (2017). The consequences of deglycosylation of recombinant intra-melanosomal domain of human tyrosinase. Biol. Chem. *399*, 73–77.

58. Sevillano, A.M., Aguilar-Calvo, P., Kurt, T.D., Lawrence, J.A., Soldau, K., Nam, T.H., Schumann, T., Pizzo, D.P., Nyström, S., Choudhury, B., et al. (2020). Prion protein glycans reduce intracerebral fibril formation and spongiosis in prion disease. J. Clin. Invest. *130*, 1350–1362.

59. Yi, C.-W., Wang, L.-Q., Huang, J.-J., Pan, K., Chen, J., and Liang, Y. (2018). Glycosylation Significantly Inhibits the Aggregation of Human Prion Protein and Decreases Its Cytotoxicity. Preprint, https://doi.org/10.1038/s41598-018-30770-6 https://doi.org/10.1038/s41598-018-30770-6.

60. Ng, B.G., and Freeze, H.H. (2018). Perspectives on Glycosylation and Its Congenital Disorders. Trends Genet. *34*, 466–476.

61. Joshi, H.J., Hansen, L., Narimatsu, Y., Freeze, H.H., Henrissat, B., Bennett, E., Wandall, H.H., Clausen, H., and Schjoldager, K.T. (2018). Glycosyltransferase genes that cause monogenic congenital disorders of glycosylation are distinct from glycosyltransferase genes associated with complex diseases. Glycobiology *28*, 284–294.

62. Berg-Fussman, A., Grace, M.E., Ioannou, Y., and Grabowski, G.A. (1993). Human acid beta-glucosidase. N-glycosylation site occupancy and the effect of glycosylation on enzymatic activity. J. Biol. Chem. *268*, 14861–14866.

63. Pol-Fachin, L., Siebert, M., Verli, H., and Saraiva-Pereira, M.L. (2016). Glycosylation is crucial for a proper catalytic site organization in human glucocerebrosidase. Glycoconj. J. *33*, 237–244.

64. Smith, L., Mullin, S., and Schapira, A.H.V. (2017). Insights into the structural biology of Gaucher disease. Exp. Neurol. *298*, 180–190.

65. Souffrant, M.G., Yao, X.-Q., Momin, M., and Hamelberg, D. (2020). N-glycosylation and gaucher disease mutation allosterically alter active-site dynamics of acid-β-glucosidase. ACS Catal. *10*, 1810–1820.

66. Di Fede, G., Catania, M., Atzori, C., Moda, F., Pasquali, C., Indaco, A., Grisoli, M., Zuffi, M., Guaita, M.C., Testi, R., et al. (2019). Clinical and neuropathological phenotype associated with the novel V189I mutation in the prion protein gene. Acta Neuropathol. Commun. *7*, 1.

67. Ladogana, A., and Kovacs, G.G. (2018). Genetic Creutzfeldt–Jakob disease. In Human Prion Diseases

Handbook of clinical neurology. (Elsevier), pp. 219–242.

68. Liberski, P.P. (2012). Gerstmann-Sträussler-Scheinker disease. Adv. Exp. Med. Biol. *724*, 128–137.

69. Cohen, M.L. (2014). Human Prion Diseases. In Pathobiology of Human Disease (Elsevier), pp. 2045–2054.

70. Cao, L., Diedrich, J.K., Kulp, D.W., Pauthner, M., He, L., Park, S.-K.R., Sok, D., Su, C.Y., Delahunty, C.M., Menis, S., et al. (2017). Global site-specific N-glycosylation analysis of HIV envelope glycoprotein. Preprint, https://doi.org/10.1038/ncomms14954 https://doi.org/10.1038/ncomms14954.

71. Zhao, P., Praissman, J.L., Grant, O.C., Cai, Y., Xiao, T., Rosenbalm, K.E., Aoki, K., Kellman, B.P., Bridger, R., Barouch, D.H., et al. (2020). Virus-Receptor Interactions of Glycosylated SARS-CoV-2 Spike and Human ACE2 Receptor. Cell Host Microbe *28*, 586-601.e6.

72. Watanabe, Y., Allen, J.D., Wrapp, D., McLellan, J.S., and Crispin, M. (2020). Site-specific glycan analysis of the SARS-CoV-2 spike. Science *369*, 330–333.

73. Casalino, L., Gaieb, Z., Dommer, A.C., and Harbison, A.M. (2020). Shielding and Beyond: The Roles of Glycans in SARS-CoV-2 Spike Protein. bioRxiv.

74. Grant, O.C., Montgomery, D., Ito, K., and Woods, R.J. (2020). 3D Models of glycosylated SARS-CoV-2 spike protein suggest challenges and opportunities for vaccine development. bioRxiv.

75. Wintjens, R., Bifani, A.M., and Bifani, P. (2020). Impact of glycan cloud on the B-cell epitope prediction of SARS-CoV-2 Spike protein. NPJ Vaccines *5*, 81.

76. Zhang, L., Mann, M., Syed, Z.A., Reynolds, H.M., Tian, E., Samara, N.L., Zeldin, D.C., Tabak, L.A., and Ten Hagen, K.G. (2021). Furin cleavage of the SARS-CoV-2 spike is modulated by O-glycosylation. Proc. Natl. Acad. Sci. U. S. A. *118*, e2109905118.

77. Baboo, S., Diedrich, J.K., Martínez-Bartolomé, S., Wang, X., Schiffner, T., Groschel, B., Schief, W.R., Paulson, J.C., and Yates, J.R., 3rd (2021). DeGlyPHER: An ultrasensitive method for the analysis of viral spike N-glycoforms. Anal. Chem. *93*, 13651–13657.

78. Krištić, J., Zaytseva, O.O., Ram, R., Nguyen, Q., Novokmet, M., Vučković, F., Vilaj, M., Trbojević-Akmačić, I., Pezer, M., Davern, K.M., et al. (2018). Profiling and genetic control of the murine immunoglobulin G glycome. Nat. Chem. Biol. *14*, 516–524.

79. Cheng, H.D., Tirosh, I., de Haan, N., Stöckmann, H., Adamczyk, B., McManus, C.A., O'Flaherty, R., Greville, G., Saldova, R., Bonilla, F.A., et al. (2020). IgG Fc glycosylation as an axis of humoral immunity in childhood. J. Allergy Clin. Immunol. *145*, 710-713.e9.

80. Lofano, G., Gorman, M.J., Yousif, A.S., Yu, W.-H., Fox, J.M., Dugast, A.-S., Ackerman, M.E., Suscovich, T.J., Weiner, J., Barouch, D., et al. (2018). Antigen-specific antibody Fc glycosylation enhances humoral immunity via the recruitment of complement. Sci. Immunol. *3*, eaat7796.

81. Atyeo, C., Pullen, K.M., Bordt, E.A., Fischinger, S., Burke, J., Michell, A., Slein, M.D., Loos, C., Shook, L.L., Boatin, A.A., et al. (2021). Compromised SARS-CoV-2-specific placental antibody transfer. Cell

184, 628-642.e10.

82. Grace, P.S., Dolatshahi, S., Lu, L.L., Cain, A., Palmieri, F., Petrone, L., Fortune, S.M., Ottenhoff, T.H.M., Lauffenburger, D.A., Goletti, D., et al. (2021). Antibody subclass and glycosylation shift following effective TB treatment. Front. Immunol. *12*, 679973.

83. Majewska, N.I., Tejada, M.L., Betenbaugh, M.J., and Agarwal, N. (2020). N-glycosylation of IgG and IgG-like recombinant therapeutic proteins: Why is it important and how can we control it? Annu. Rev. Chem. Biomol. Eng. *11*, 311–338.

84. Li, W., Zhu, Z., Chen, W., Feng, Y., and Dimitrov, D.S. (2017). Crystallizable fragment glycoengineering for therapeutic antibodies development. Front. Immunol. *8*. https://doi.org/10.3389/fimmu.2017.01554.

85. Zhang, P., Woen, S., Wang, T., Liau, B., Zhao, S., Chen, C., Yang, Y., Song, Z., Wormald, M.R., Yu, C., et al. (2016). Challenges of glycosylation analysis and control: an integrated approach to producing optimal and consistent therapeutic drugs. Drug Discov. Today *21*, 740–765.

86. Shields, R.L., Lai, J., Keck, R., O'Connell, L.Y., Hong, K., Meng, Y.G., Weikert, S.H.A., and Presta, L.G. (2002). Lack of Fucose on Human IgG1 N-Linked Oligosaccharide Improves Binding to Human FcγRIII and Antibody-dependent Cellular Toxicity. J. Biol. Chem. *277*, 26733–26740.

87. de Haan, N., Reiding, K.R., Krištić, J., Hipgrave Ederveen, A.L., Lauc, G., and Wuhrer, M. (2017). The N-Glycosylation of Mouse Immunoglobulin G (IgG)-Fragment Crystallizable Differs Between IgG Subclasses and Strains. Front. Immunol. *8*, 608.

88. Zaytseva, O.O., Jansen, B.C., Hanić, M., Mrčela, M., Razdorov, G., Stojković, R., Erhardt, J., Brizić, I., Jonjić, S., Pezer, M., et al. (2018). MIgGGly (mouse IgG glycosylation analysis) - a high-throughput method for studying Fc-linked IgG N-glycosylation in mice with nanoUPLC-ESI-MS. Sci. Rep. *8*, 13688.

89. Zaytseva, O.O., Seeling, M., Krištić, J., Lauc, G., Pezer, M., and Nimmerjahn, F. (2020). Fc-Linked IgG N-Glycosylation in FcγR Knock-Out Mice. Preprint, https://doi.org/10.3389/fcell.2020.00067 https://doi.org/10.3389/fcell.2020.00067.

90. Gunn, B.M., Lu, R., Slein, M.D., Ilinykh, P.A., Huang, K., Atyeo, C., Schendel, S.L., Kim, J., Cain, C., Roy, V., et al. (2021). A Fc engineering approach to define functional humoral correlates of immunity against Ebola virus. Immunity *54*, 815-828.e5.

91. Wong, M.Y., Chen, K., Antonopoulos, A., Kasper, B.T., Dewal, M.B., Taylor, R.J., Whittaker, C.A., Hein, P.P., Dell, A., Genereux, J.C., et al. (2018). XBP1s activation can globally remodel N-glycan structure distribution patterns. Proc. Natl. Acad. Sci. U. S. A. *115*, E10089–E10098.

92. Chen, K., and Shoulders, M.D. (2024). Protein glycosylation patterns shaped by the IRE1-XBP1s arm of the unfolded protein response. Isr. J. Chem. *64*. https://doi.org/10.1002/ijch.202300162.

93. Karch, C.P., Paquin-Proulx, D., Eller, M.A., Matyas, G.R., Burkhard, P., and Beck, Z. (2020). Impact of the expression system on the immune responses to self-assembling protein nanoparticles (SAPNs) displaying HIV-1 V1V2 loop. Nanomedicine *29*, 102255.

94. Hombu, R., Neelamegham, S., and Park, S. (2021). Cellular and molecular engineering of glycan sialylation in heterologous systems. Molecules *26*, 5950.

95. Zhong, X., Ma, W., Meade, C.L., Tam, A.S., Llewellyn, E., Cornell, R., Cote, K., Scarcelli, J.J., Marshall, J.K., Tzvetkova, B., et al. (2019). Transient CHO expression platform for robust antibody production and its enhanced N-glycan sialylation on therapeutic glycoproteins. Biotechnol. Prog. *35*, e2724.

96. Zhong, X., Schwab, A., Ma, W., Meade, C.L., Zhou, J., D'Antona, A.M., Somers, W., and Lin, L. (2022). Large-scale transient production in ExpiCHO-S^{TM} with enhanced N-galactosylation-sialylation and PEI-based transfection. Methods Mol. Biol. *2313*, 143–150.

97. Hudson, K.L., Bartlett, G.J., Diehl, R.C., Agirre, J., Gallagher, T., Kiessling, L.L., and Woolfson, D.N. (2015). Carbohydrate-aromatic interactions in proteins. J. Am. Chem. Soc. *137*, 15152–15160.

98. Chen, W., Enck, S., Price, J.L., Powers, D.L., Powers, E.T., Wong, C.-H., Dyson, H.J., and Kelly, J.W. (2013). Structural and energetic basis of carbohydrate-aromatic packing interactions in proteins. J. Am. Chem. Soc. *135*, 9877–9884.

99. Asensio, J.L., Ardá, A., Cañada, F.J., and Jiménez-Barbero, J. (2013). Carbohydrate-aromatic interactions. Acc. Chem. Res. *46*, 946–954.

100. Montalvillo-Jiménez, L., Santana, A.G., Corzana, F., Jiménez-Osés, G., Jiménez-Barbero, J., Gómez, A.M., and Asensio, J.L. (2019). Impact of aromatic stacking on glycoside reactivity: Balancing CH/π and cation/π interactions for the stabilization of glycosyl-oxocarbenium ions. J. Am. Chem. Soc. *141*, 13372–13384.

101. Ardejani, M.S., Noodleman, L., Powers, E.T., and Kelly, J.W. (2021). Stereoelectronic effects in stabilizing protein-N-glycan interactions revealed by experiment and machine learning. Nat. Chem. *13*, 480–487.

102. Goldfarb, L.G., Petersen, R.B., Tabaton, M., Brown, P., LeBlanc, A.C., Montagna, P., Cortelli, P., Julien, J., Vital, C., and Pendelbury, W.W. (1992). Fatal familial insomnia and familial Creutzfeldt-Jakob disease: disease phenotype determined by a DNA polymorphism. Science *258*, 806–808.

103. Wang, W.K., Essex, M., and Lee, T.H. (1996). Single amino acid substitution in constant region 1 or 4 of gp120 causes the phenotype of a human immunodeficiency virus type 1 variant with mutations in hypervariable regions 1 and 2 to revert. J. Virol. *70*, 607–611.

104. Branza-Nichita, N., Petrescu, A.J., Negroiu, G., Dwek, R.A., and Petrescu, S.M. (2000). N-glycosylation processing and glycoprotein folding-lessons from the tyrosinase-related proteins. Chem. Rev. *100*, 4697–4712.

105. Dolinska, M.B., Kus, N.J., Farney, S.K., Wingfield, P.T., Brooks, B.P., and Sergeev, Y.V. (2017). Oculocutaneous albinism type 1: link between mutations, tyrosinase conformational stability, and enzymatic activity. Pigment Cell Melanoma Res. *30*, 41–52.

106. Grant, O.C., Montgomery, D., Ito, K., and Woods, R.J. (2020). Analysis of the SARS-CoV-2 spike protein glycan shield reveals implications for immune recognition. Sci. Rep. *10*, 14991.

107. Peng, W., Rayaprolu, V., Parvate, A.D., Pronker, M.F., Hui, S., Parekh, D., Shaffer, K., Yu, X., Saphire,

E.O., and Snijder, J. (2022). Glycan shield of the ebolavirus envelope glycoprotein GP. Commun. Biol. *5*, 785.

108. Wei, C.-J., Boyington, J.C., Dai, K., Houser, K.V., Pearce, M.B., Kong, W.-P., Yang, Z.-Y., Tumpey, T.M., and Nabel, G.J. (2010). Cross-neutralization of 1918 and 2009 influenza viruses: role of glycans in viral evolution and vaccine design. Sci. Transl. Med. *2*, 24ra21.

109. Schulz, M.A., Tian, W., Mao, Y., Van Coillie, J., Sun, L., Larsen, J.S., Chen, Y.-H., Kristensen, C., Vakhrushev, S.Y., Clausen, H., et al. (2018). Glycoengineering design options for IgG1 in CHO cells using precise gene editing. Glycobiology *28*, 542–549.

110. Narimatsu, Y., Joshi, H.J., Nason, R., Van Coillie, J., Karlsson, R., Sun, L., Ye, Z., Chen, Y.-H., Schjoldager, K.T., Steentoft, C., et al. (2019). An atlas of human glycosylation pathways enables display of the human glycome by gene engineered cells. Mol. Cell *75*, 394-407.e5.

111. Narimatsu, Y., Büll, C., Chen, Y.-H., Wandall, H.H., Yang, Z., and Clausen, H. (2021). Genetic glycoengineering in mammalian cells. J. Biol. Chem. *296*, 100448.

112. Sumit, M., Dolatshahi, S., Chu, A.-H.A., Cote, K., Scarcelli, J.J., Marshall, J.K., Cornell, R.J., Weiss, R., Lauffenburger, D.A., Mulukutla, B.C., et al. (2019). Dissecting N-Glycosylation Dynamics in Chinese Hamster Ovary Cells Fed-batch Cultures using Time Course Omics Analyses. iScience *12*, 102–120.

113. Ehret, J., Zimmermann, M., Eichhorn, T., and Zimmer, A. (2019). Impact of cell culture media additives on IgG glycosylation produced in Chinese hamster ovary cells. Biotechnol. Bioeng. *116*, 816–830.

114. Fan, Y., Kildegaard, H.F., and Andersen, M.R. (2017). Engineer medium and feed for modulating N-glycosylation of recombinant protein production in CHO cell culture. Methods Mol. Biol. *1603*, 209–226.

115. Kotidis, P., and Kontoravdi, C. (2020). Harnessing the potential of artificial neural networks for predicting protein glycosylation. Metabolic Engineering Communications, e00131.

116. Jimenez del Val, I., Nagy, J.M., and Kontoravdi, C. (2011). A dynamic mathematical model for monoclonal antibody N-linked glycosylation and nucleotide sugar donor transport within a maturing Golgi apparatus. Biotechnol. Prog. *27*, 1730–1743.

117. Sou, S.N., Jedrzejewski, P.M., Lee, K., Sellick, C., Polizzi, K.M., and Kontoravdi, C. (2017). Model-based investigation of intracellular processes determining antibody Fc-glycosylation under mild hypothermia. Biotechnol. Bioeng. *114*, 1570–1582.

118. Spahn, P.N., Hansen, A.H., Hansen, H.G., Arnsdorf, J., Kildegaard, H.F., and Lewis, N.E. (2016). A Markov chain model for N-linked protein glycosylation – towards a low-parameter tool for model-driven glycoengineering. Preprint, https://doi.org/10.1016/j.ymben.2015.10.007 https://doi.org/10.1016/j.ymben.2015.10.007.

119. Spahn, P.N., Hansen, A.H., Kol, S., Voldborg, B.G., and Lewis, N.E. (2017). Predictive glycoengineering of biosimilars using a Markov chain glycosylation model. Biotechnol. J. *12*. https://doi.org/10.1002/biot.201600489.

120. Liang, C., Chiang, A.W.T., Hansen, A.H., Arnsdorf, J., Schoffelen, S., Sorrentino, J.T., Kellman, B.P., Bao, B., Voldborg, B.G., and Lewis, N.E. (2020). A Markov model of glycosylation elucidates isozyme specificity and glycosyltransferase interactions for glycoengineering. Curr Res Biotechnol *2*, 22–36.

121. Michaeli, Y., and Reiter, Y. (2006). Optimised Fc variants with enhanced effector function. Expert Opin. Ther. Pat. *16*, 1449–1452.

122. Liu, R., Oldham, R.J., Teal, E., Beers, S.A., and Cragg, M.S. (2020). Fc-engineering for modulated effector functions-improving antibodies for cancer treatment. Antibodies (Basel) *9*, 64.

123. van Erp, E.A., Luytjes, W., Ferwerda, G., and van Kasteren, P.B. (2019). Fc-mediated antibody effector functions during respiratory syncytial virus infection and disease. Front. Immunol. *10*, 548.

124. Leon, P.E., He, W., Mullarkey, C.E., Bailey, M.J., Miller, M.S., Krammer, F., Palese, P., and Tan, G.S. (2016). Optimal activation of Fc-mediated effector functions by influenza virus hemagglutinin antibodies requires two points of contact. Proc. Natl. Acad. Sci. U. S. A. *113*, E5944–E5951.

125. Zhou, M., Palanca, A.M.S., and Law, J.A. (2018). Locus-specific control of the de novo DNA methylation pathway in Arabidopsis by the CLASSY family. Nat. Genet. *50*, 865–873.

126. Nováček, V., McGauran, G., Matallanas, D., Vallejo Blanco, A., Conca, P., Muñoz, E., Costabello, L., Kanakaraj, K., Nawaz, Z., Walsh, B., et al. (2020). Accurate prediction of kinase-substrate networks using knowledge graphs. PLoS Comput. Biol. *16*, e1007578.

127. Malaker, S.A., Riley, N.M., Shon, D.J., Pedram, K., Krishnan, V., Dorigo, O., and Bertozzi, C.R. (2022). Revealing the human mucinome. Nat. Commun. *13*, 3542.

128. Madhani, H.D., and Guthrie, C. (1994). Dynamic RNA-RNA interactions in the spliceosome. Annu. Rev. Genet. *28*, 1–26.

129. Bao, P., Boon, K.-L., Will, C.L., Hartmuth, K., and Lührmann, R. (2018). Multiple RNA–RNA tertiary interactions are dispensable for formation of a functional U2/U6 RNA catalytic core in the spliceosome. Preprint, https://doi.org/10.1093/nar/gky966 https://doi.org/10.1093/nar/gky966.

130. Corley, M., Burns, M.C., and Yeo, G.W. (2020). How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. Mol. Cell *78*, 9–29.

131. Farnham, P.J., and Platt, T. (1982). Effects of DNA base analogs on transcription termination at the tryptophan operon attenuator of EScherichia coli. Proc. Natl. Acad. Sci. U. S. A. *79*, 998–1002.

132. Lee, F., and Yanofsky, C. (1977). Transcription termination at the trp operon attenuators of Escherichia coli and Salmonella typhimurium: RNA secondary structure and regulation of termination. Proc. Natl. Acad. Sci. U. S. A. *74*, 4365–4369.

133. Oxender, D.L., Zurawski, G., and Yanofsky, C. (1979). Attenuation in the Escherichia coli tryptophan operon: role of RNA secondary structure involving the tryptophan codon region. Proc. Natl. Acad. Sci. U. S. A. *76*, 5524–5528.

134. Farnham, P.J., and Platt, T. (1980). A model for transcription termination suggested by studies on the trp attenuator in vitro using base analogs. Cell *20*, 739–748.

135. Wu, A.M., and Platt, T. (1978). Transcription termination: nucleotide sequence at 3' end of tryptophan operon in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. *75*, 5442–5446.

136. Mariethoz, J., Alocci, D., Gastaldello, A., Horlacher, O., Gasteiger, E., Rojas-Macias, M., Karlsson, N.G., Packer, N.H., and Lisacek, F. (2018). Glycomics@ExPASy: Bridging the Gap. Preprint, https://doi.org/10.1074/mcp.ra118.000799 https://doi.org/10.1074/mcp.ra118.000799.

137. York, W.S., Mazumder, R., Ranzinger, R., Edwards, N., Kahsay, R., Aoki-Kinoshita, K.F., Campbell, M.P., Cummings, R.D., Feizi, T., Martin, M., et al. (2020). GlyGen: Computational and Informatics Resources for Glycoscience. Glycobiology *30*, 72–73.

138. Cooper, C.A., Joshi, H.J., Harrison, M.J., Wilkins, M.R., and Packer, N.H. (2003). GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. Nucleic Acids Res. *31*, 511–513.

139. Al Jadda, K., Porterfield, M.P., Bridger, R., Heiss, C., Tiemeyer, M., Wells, L., Miller, J.A., York, W.S., and Ranzinger, R. (2015). EUROCarbDB(CCRC): a EUROCarbDB node for storing glycomics standard data. Bioinformatics *31*, 242–245.

140. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

141. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. *46*, W296–W303.

142. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. Nat. Methods *12*, 7–8.

143. Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N., Potter, S.C., Finn, R.D., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. *47*, W636–W641.

144. Zhang, Y., and Sagui, C. (2015). Secondary structure assignment for conformationally irregular peptides: comparison between DSSP, STRIDE and KAKSI. J. Mol. Graph. Model. *55*, 72–84.

145. Magnan, C.N., and Baldi, P. (2014). SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. Bioinformatics *30*, 2592–2597.

146. Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K.F., Ueda, N., Hamajima, M., Kawasaki, T., and Kanehisa, M. (2006). KEGG as a glycome informatics resource. Glycobiology *16*, 63R-70R.

147. Herget, S., Ranzinger, R., Maass, K., and Lieth, C.-W.V.D. (2008). GlycoCT-a unifying sequence format for carbohydrates. Carbohydr. Res. *343*, 2162–2171.

148. Burkholz, R., Quackenbush, J., and Bojar, D. (2021). Using graph convolutional neural networks to learn a representation for glycans. Cell Rep. *35*, 109251.

149. Klein, J., and Zaia, J. (2019). glypy: An Open Source Glycoinformatics Library. J. Proteome Res. *18*,

3532–3537.

150. Velankar, S., Dana, J.M., Jacobsen, J., van Ginkel, G., Gane, P.J., Luo, J., Oldfield, T.J., O'Donovan, C., Martin, M.-J., and Kleywegt, G.J. (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res. *41*, D483-9.

151. Zeger, S.L., and Liang, K.-Y. (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. Preprint, https://doi.org/10.2307/2531248 https://doi.org/10.2307/2531248.

152. Benaglia, T., Chauveau, D., Hunter, D.R., and Young, D. (2009). mixtools: AnRPackage for Analyzing Finite Mixture Models. Preprint, https://doi.org/10.18637/jss.v032.i06 https://doi.org/10.18637/jss.v032.i06.

153. Balakrishnan, S., Kamisetty, H., Carbonell, J.G., Lee, S.-I., and Langmead, C.J. (2011). Learning generative models for protein fold families. Proteins *79*, 1061–1078.

154. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., and Aurell, E. (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. *87*, 012707.

155. Weinreb, C., Riesselman, A.J., Ingraham, J.B., Gross, T., Sander, C., and Marks, D.S. (2016). 3D RNA and Functional Interactions from Evolutionary Couplings. Cell *165*, 963–975.

156. Hopf, T.A., Green, A.G., Schubert, B., Mersmann, S., Schärfe, C.P.I., Ingraham, J.B., Toth-Petroczy, A., Brock, K., Riesselman, A.J., Palmedo, P., et al. (2019). The EVcouplings Python framework for coevolutionary sequence analysis. Bioinformatics *35*, 1582–1584.

157. Wright, E.S. (2016). Using DECIPHER v2. 0 to analyze big biological sequence data in R. R J. *8*.

158. Wagih, O. (2017). ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics *33*, 3645–3647.

159. Morahan, G., Balmer, L., and Monley, D. (2008). Establishment of "The Gene Mine": a resource for rapid identification of complex trait genes. Mamm. Genome *19*, 390–393.

160. Jansen, B.C., Falck, D., de Haan, N., Hipgrave Ederveen, A.L., Razdorov, G., Lauc, G., and Wuhrer, M. (2016). LaCyTools: A Targeted Liquid Chromatography–Mass Spectrometry Data Processing Package for Relative Quantitation of Glycopeptides. J. Proteome Res. *15*, 2198–2210.

161. He, L., Diedrich, J., Chu, Y.-Y., and Yates, J.R., 3rd (2015). Extracting accurate precursor information for tandem mass spectra by RawConverter. Anal. Chem. *87*, 11361–11367.

162. Xu, T., Park, S.K., Venable, J.D., Wohlschlegel, J.A., Diedrich, J.K., Cociorva, D., Lu, B., Liao, L., Hewel, J., Han, X., et al. (2015). ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity and specificity. J. Proteomics *129*, 16–24.

163. Tabb, D.L., McDonald, W.H., and Yates, J.R., 3rd (2002). DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. J. Proteome Res. *1*, 21–26.

164. Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J., and Gygi, S.P. (2003). Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J. Proteome Res. *2*, 43–50.

165. Park, S.K., Venable, J.D., Xu, T., and Yates, J.R., 3rd (2008). A quantitative analysis software tool for mass spectrometry-based proteomics. Nat. Methods *5*, 319–322.

166. Irvine, E.B., Peters, J.M., Lu, R., Grace, P.S., Sixsmith, J., Wallace, A., Schneider, M., Shin, S., Karpinski, W., Hsiao, J.C., et al. (2022). Fc-engineered antibodies leverage neutrophils to drive control of *Mycobacterium tuberculosis*. bioRxiv. https://doi.org/10.1101/2022.05.01.490220.

167. Engler, C., and Marillonnet, S. (2014). Golden Gate cloning. Methods Mol. Biol. *1116*, 119–131.

168. Fang, J., Qian, J.-J., Yi, S., Harding, T.C., Tu, G.H., VanRoey, M., and Jooss, K. (2005). Stable antibody expression at therapeutic levels using the 2A peptide. Nat. Biotechnol. *23*, 584–590.

169. Mahan, A.E., Tedesco, J., Dionne, K., Baruah, K., Cheng, H.D., De Jager, P.L., Barouch, D.H., Suscovich, T., Ackerman, M., Crispin, M., et al. (2015). A method for high-throughput, sensitive analysis of IgG Fc and Fab glycosylation by capillary electrophoresis. J. Immunol. Methods *417*, 34–44.

170. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Series B Stat. Methodol. *57*, 289–300.

171. Xu, C., and Ng, D.T.W. (2015). Glycosylation-directed quality control of protein folding. Nat. Rev. Mol. Cell Biol. *16*, 742–752.

172. Shental-Bechor, D., and Levy, Y. (2008). Effect of glycosylation on protein folding: a close look at thermodynamic stabilization. Proc. Natl. Acad. Sci. U. S. A. *105*, 8256–8261.

173. Shihab, H.A., Gough, J., Mort, M., Cooper, D.N., Day, I.N.M., and Gaunt, T.R. (2014). Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. Hum. Genomics *8*, 11.

174. Shihab, H.A., Gough, J., Cooper, D.N., Day, I.N.M., and Gaunt, T.R. (2013). Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics *29*, 1504–1510.

175. Shihab, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L.A., Edwards, K.J., Day, I.N.M., and Gaunt, T.R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum. Mutat. *34*, 57–65.

176. Baković, M.P., Selman, M.H.J., Hoffmann, M., Rudan, I., Campbell, H., Deelder, A.M., Lauc, G., and Wuhrer, M. (2013). High-throughput IgG Fc N-glycosylation profiling by mass spectrometry of glycopeptides. J. Proteome Res. *12*, 821–831.

177. Pucić, M., Knezević, A., Vidic, J., Adamczyk, B., Novokmet, M., Polasek, O., Gornik, O., Supraha-Goreta, S., Wormald, M.R., Redzić, I., et al. (2011). High throughput isolation and glycosylation analysis of IgG-variability and heritability of the IgG glycome in three isolated human populations. Mol. Cell. Proteomics *10*, M111.010090.

178. Klarić, L., Tsepilov, Y.A., Stanton, C.M., Mangino, M., Sikka, T.T., Esko, T., Pakhomov, E., Salo, P.,

Deelen, J., McGurnaghan, S.J., et al. (2020). Glycosylation of immunoglobulin G is regulated by a large network of genes pleiotropic with inflammatory diseases. Sci. Adv. *6*, eaax0301.

179. Rose, R.J., van Berkel, P.H.C., van den Bremer, E.T.J., Labrijn, A.F., Vink, T., Schuurman, J., Heck, A.J.R., and Parren, P.W.H.I. (2013). Mutation of Y407 in the CH3 domain dramatically alters glycosylation and structure of human IgG. MAbs *5*, 219–228.

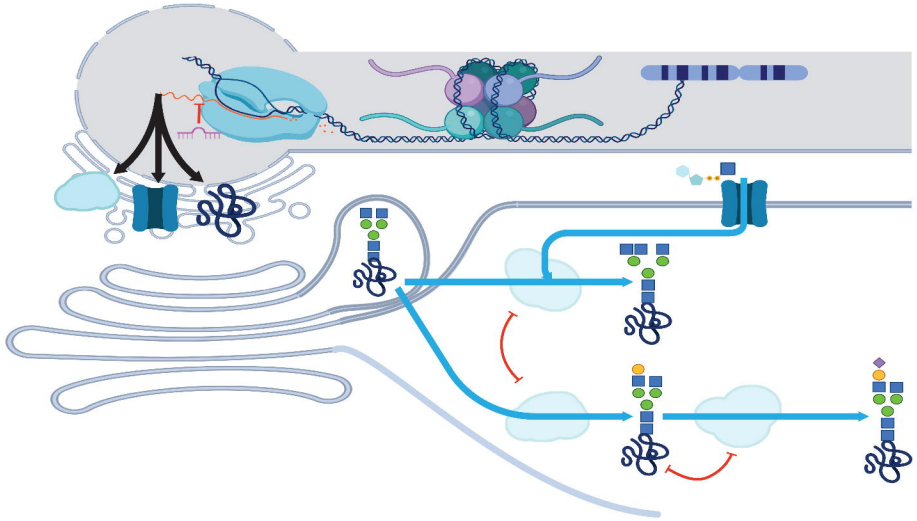
180. Vidarsson, G., Dekkers, G., and Rispens, T. (2014). IgG subclasses and allotypes: from structure to effector functions. Front. Immunol. *5*, 520.

181. Hutton, S.M., and Spritz, R.A. (2008). Comprehensive analysis of oculocutaneous albinism among non-Hispanic caucasians shows that OCA1 is the most prevalent OCA type. J. Invest. Dermatol. *128*, 2442–2450.

182. Rademacher, C., Bru, T., McBride, R., Robison, E., Nycholat, C.M., Kremer, E.J., and Paulson, J.C. (2012). A Siglec-like sialic-acid-binding motif revealed in an adenovirus capsid protein. Glycobiology *22*, 1086–1091.

183. Chavent, M., Kuentz-Simonet, V., Labenne, A., and Saracco, J. (2014). Multivariate analysis of mixed data: The R package PCAmixdata. arXiv [stat.CO].

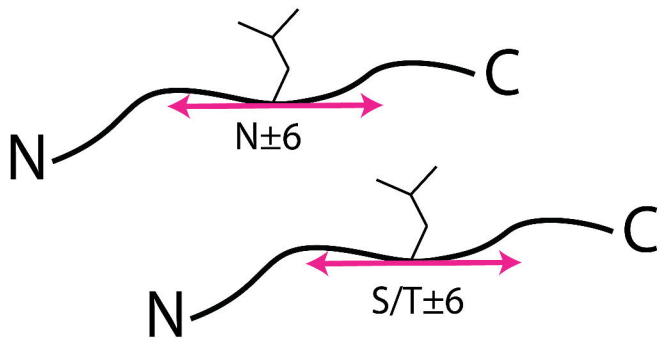
a.



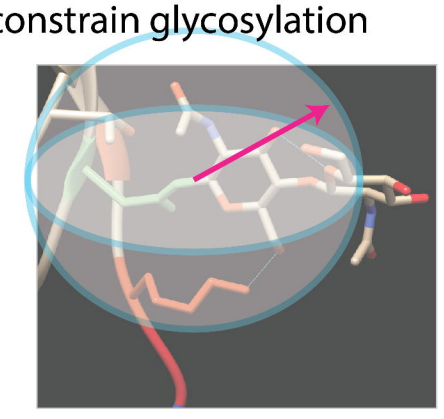
Environmental constraints:
Expression, precursor, competition

b.

Protein structure features close (2D and 3D) to N- and O-Glycosites (modeled seperately) constrain glycosylation



2D structure: amino acids and chemistry sequentially close to N- and O-Glycosites



3D structure: amino acids, secondary structure, accessibility, depth, and chemistry spatially close to N- and O-Glycosites

d.

Validation Types

Computational

Retrospective Experiments

Novel Experiments

Specific Applications

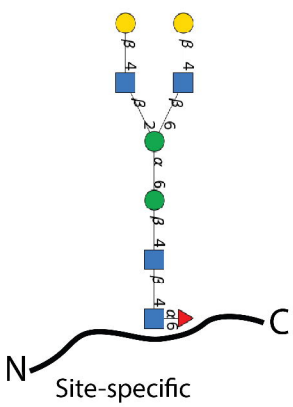
Pathogenic Variants

Evolutionary Statistics

Viral Proteins

Antibodies

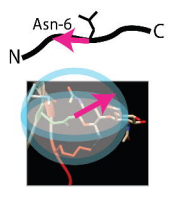
C.



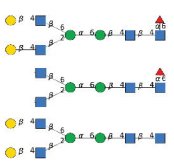
Site-specific glycosylation data

Data Enrichment

Glysosite-proximal protein structures



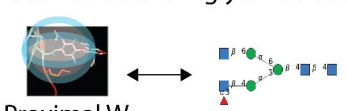
Glycan substructure extraction



Identify co-occurance

Intra-Molecular Relations (IMR)

protein structure ~ glycan structure

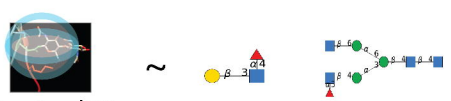


Proximal W

Aggregate IMR

Expected glycoprofile

Glycans associated with a protein structure

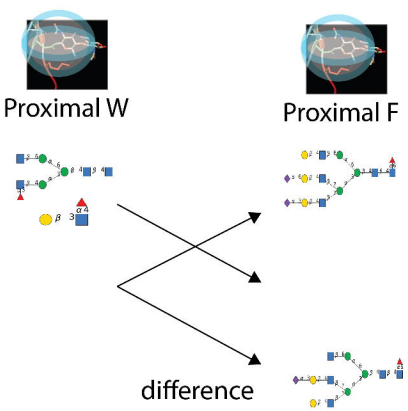


Proximal W

Compare associations

Glycoimpact

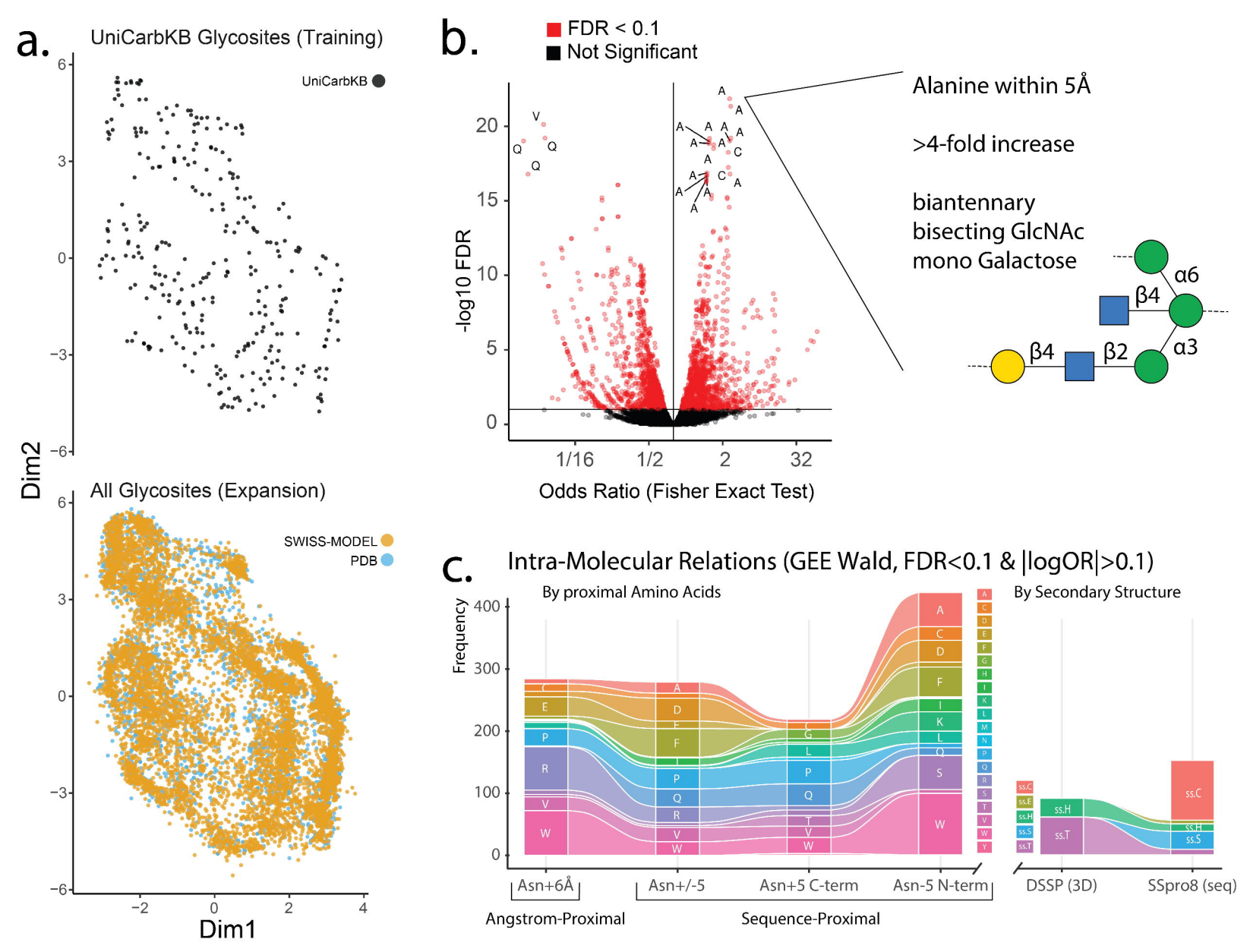
Difference in expected glycoprofile
Expected change in glycoprofile

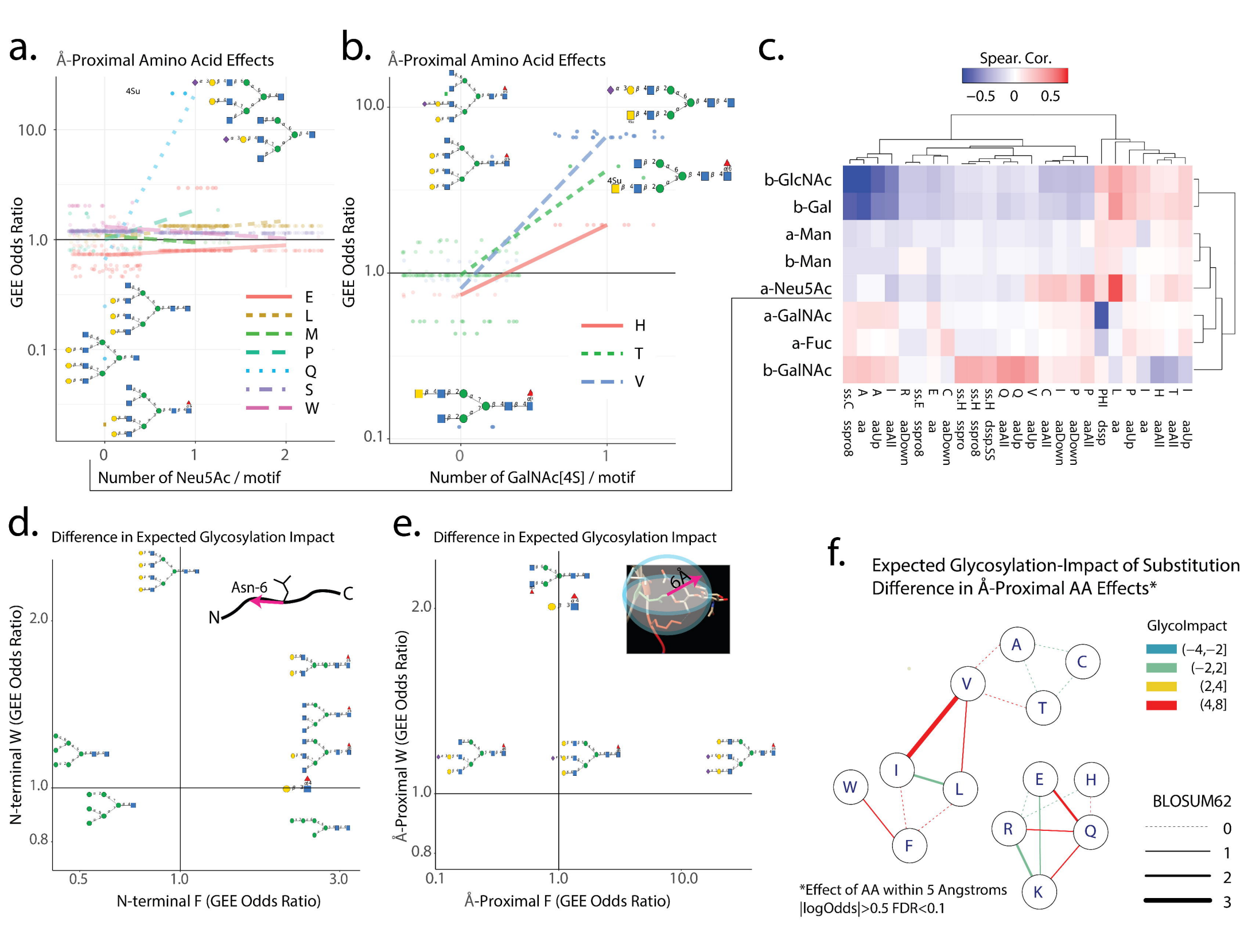


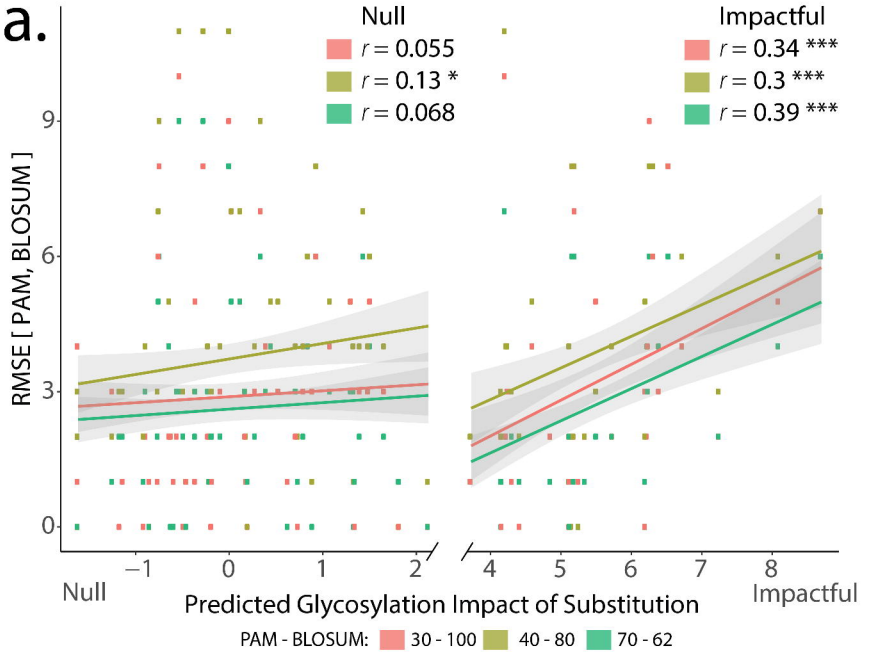
preferred

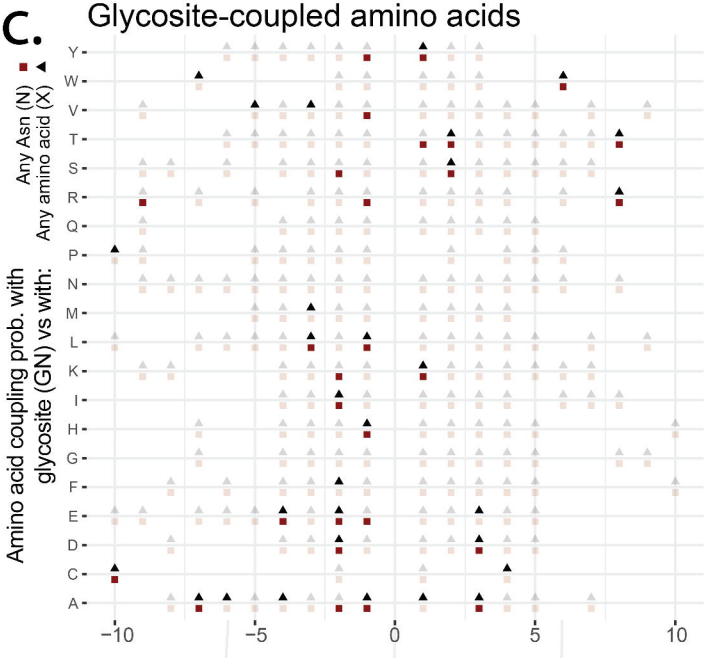
null

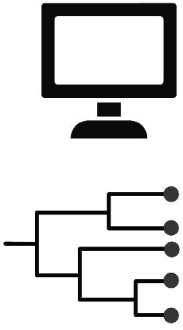
prohibited

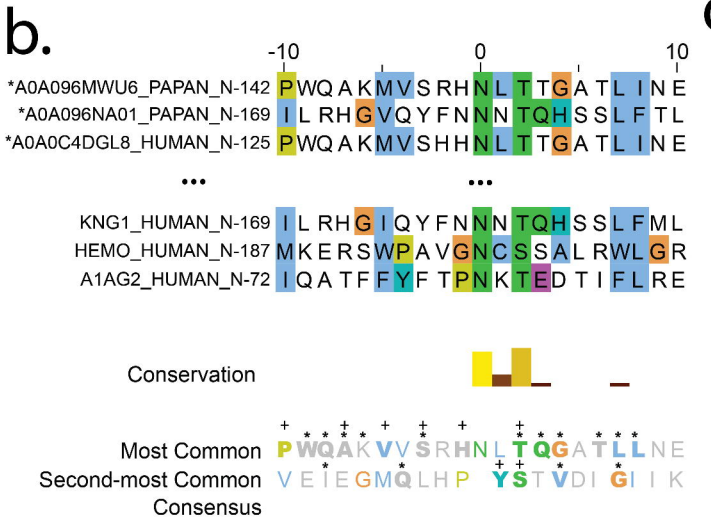

a.
UniCarbKB Glycosites (Training)
UniCarbKB
All Glycosites (Expansion)
SWISS-MODEL
PDB
Dim1
Dim2
b.
FDR < 0.1
Not Significant
-log10 FDR
Odds Ratio (Fisher Exact Test)
1/16
1/2
2
32
Alanine within 5Å
>4-fold increase
biantennary
bisecting GlcNAc
mono Galactose
α6
β4
α3
β4
β2
c.
Intra-Molecular Relations (GEE Wald, FDR<0.1 & |logOR|>0.1)
By proximal Amino Acids
By Secondary Structure
Frequency
Asn+6Å
Asn+/-5
Asn+5 C-term
Asn-5 N-term
Angstrom-Proximal
Sequence-Proximal
DSSP (3D)
SSpro8 (seq)

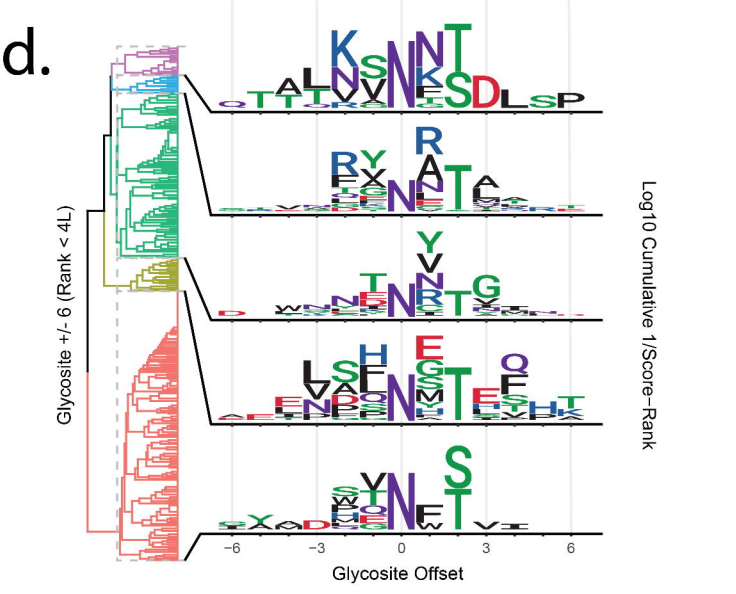

a.
Å-Proximal Amino Acid Effects
4Su
GEE Odds Ratio
10.0
1.0
0.1
E
L
M
P
Q
S
W
0
1
2
Number of Neu5Ac / motif
b.
Å-Proximal Amino Acid Effects
GEE Odds Ratio
10.0
1.0
0.1
4Su
H
T
V
0
1
Number of GalNAc[4S] / motif
c.
Spear. Cor.
−0.5 0 0.5
b-GlcNAc
b-Gal
a-Man
b-Man
a-Neu5Ac
a-GalNAc
a-Fuc
b-GalNAc
ss.C sspro8
A aa
A aaUp
I aaAll
R aaDown
ss.E sspro8
E aa
C aaDown
ss.H sspro
ss.H sspro8
ss.H dssp.SS
Q aaAll
Q aaUp
V aaUp
C aaAll
I aaDown
P aaDown
P aaAll
PHI dssp
L aa
P aaUp
I aa
H aaAll
T aaAll
I aaUp
d.
Difference in Expected Glycosylation Impact
Asn-6
N
C
N-terminal W (GEE Odds Ratio)
2.0
1.0
0.8
0.5
1.0
3.0
N-terminal F (GEE Odds Ratio)
e.
Difference in Expected Glycosylation Impact
6Å
Å-Proximal W (GEE Odds Ratio)
2.0
1.0
0.8
0.1
1.0
10.0
Å-Proximal F (GEE Odds Ratio)
f.
Expected Glycosylation-Impact of Substitution
Difference in Å-Proximal AA Effects*
A
C
V
T
I
L
W
F
E
H
R
Q
K
GlycoImpact
(−4,−2]
(−2,2]
(2,4]
(4,8]
BLOSUM62
0
1
2
3
*Effect of AA within 5 Angstroms
|logOdds|>0.5 FDR<0.1

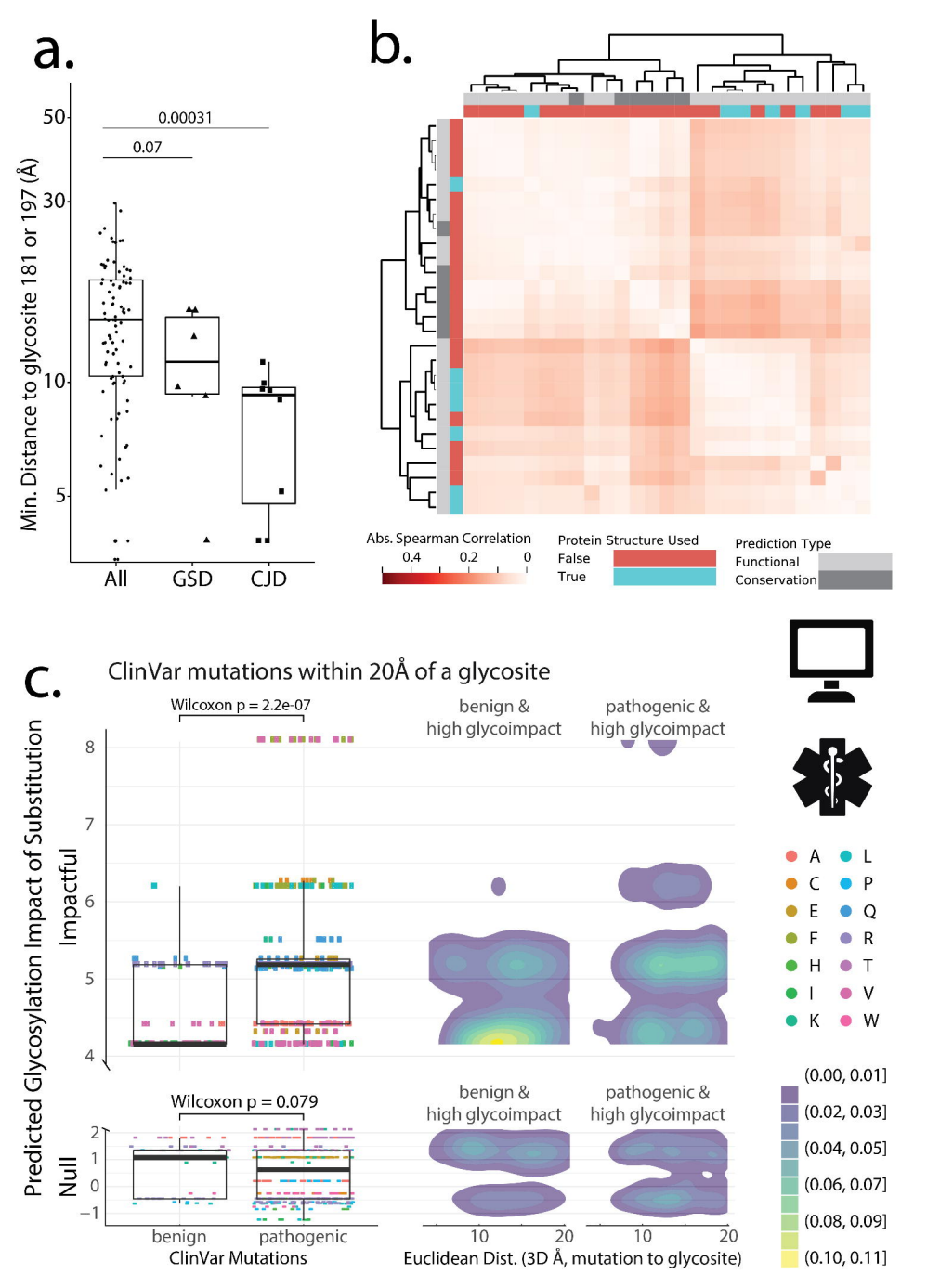

a.
Null
r = 0.055
r = 0.13 *
r = 0.068
Impactful
r = 0.34 ***
r = 0.3 ***
r = 0.39 ***
RMSE [PAM, BLOSUM]
Null
Impactful
Predicted Glycosylation Impact of Substitution
PAM - BLOSUM: 30 - 100 40 - 80 70 - 62

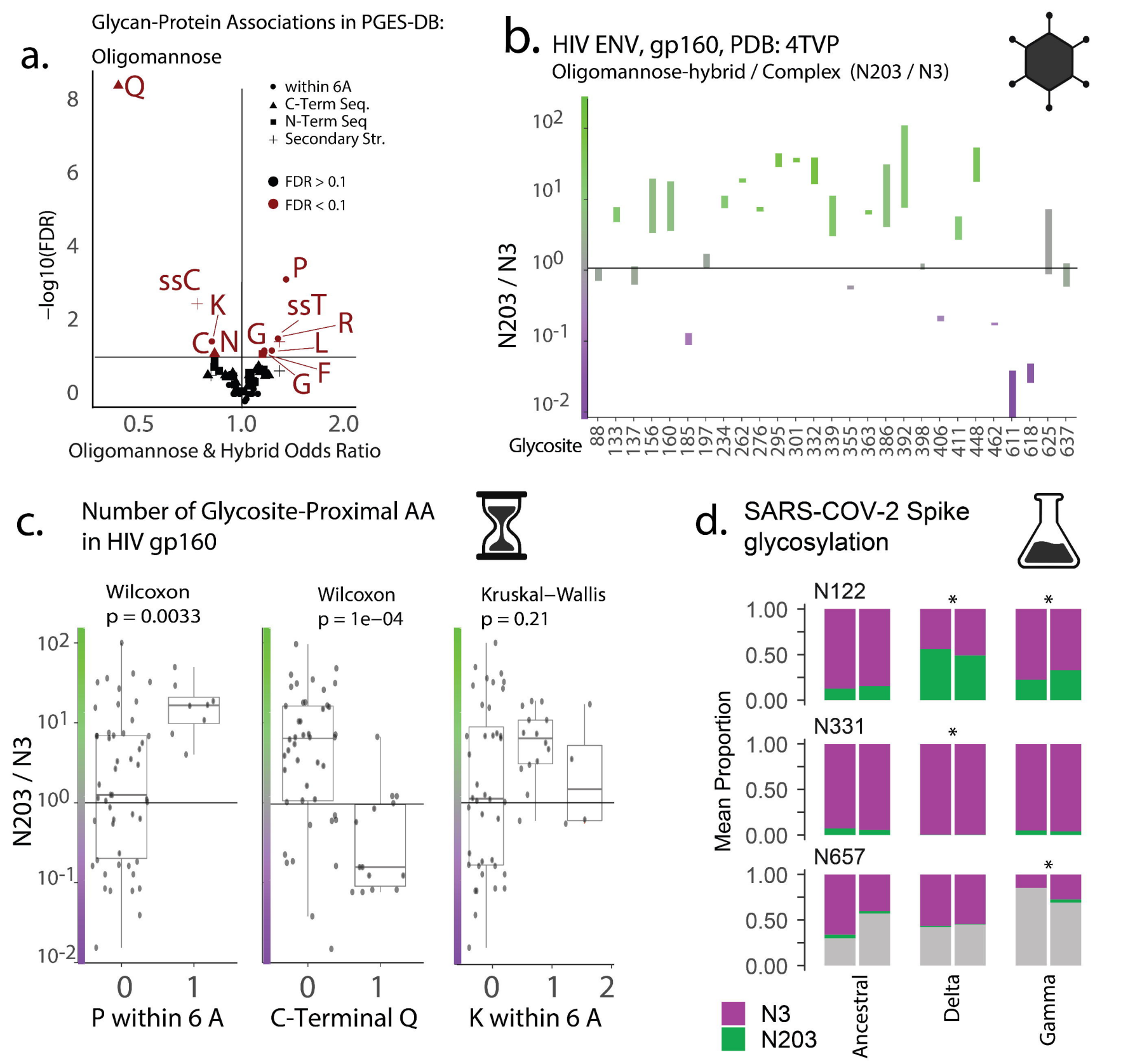

c.
Glycosite-coupled amino acids
Any Asn (N)
Any amino acid (X)
Amino acid coupling prob. with glycosite (GN) vs with:




b.
*A0A096MWU6_PAPAN_N-142
*A0A096NA01_PAPAN_N-169
*A0A0C4DGL8_HUMAN_N-125
KNG1_HUMAN_N-169
HEMO_HUMAN_N-187
A1AG2_HUMAN_N-72
Conservation
Most Common
Second-most Common
Consensus


d.
Glycosite +/- 6 (Rank < 4L)
Log10 Cumulative 1/Score−Rank
Glycosite Offset


a.
b.
c.
Min. Distance to glycosite 181 or 197 (Å)
50
30
10
5
0.00031
0.07
All
GSD
CJD
Abs. Spearman Correlation
0.4
0.2
0
Protein Structure Used
False
True
Prediction Type
Functional
Conservation
ClinVar mutations within 20Å of a glycosite
Wilcoxon p = 2.2e-07
benign &
high glycoimpact
pathogenic &
high glycoimpact
Predicted Glycosylation Impact of Substitution
Impactful
Null
Wilcoxon p = 0.079
benign
pathogenic
ClinVar Mutations
Euclidean Dist. (3D Å, mutation to glycosite)
A
C
E
F
H
I
K
L
P
Q
R
T
V
W
(0.00, 0.01]
(0.02, 0.03]
(0.04, 0.05]
(0.06, 0.07]
(0.08, 0.09]
(0.10, 0.11]


a.
Glycan-Protein Associations in PGES-DB:
Oligomannose
within 6A
C-Term Seq.
N-Term Seq
Secondary Str.
FDR > 0.1
FDR < 0.1
−log10(FDR)
Oligomannose & Hybrid Odds Ratio
Q
P
ssC
K
ssT
R
C
N
G
L
F
G
b.
HIV ENV, gp160, PDB: 4TVP
Oligomannose-hybrid / Complex (N203 / N3)
N203 / N3
Glycosite
c.
Number of Glycosite-Proximal AA in HIV gp160
Wilcoxon p = 0.0033
Wilcoxon p = 1e−04
Kruskal–Wallis p = 0.21
N203 / N3
P within 6 A
C-Terminal Q
K within 6 A
d.
SARS-COV-2 Spike glycosylation
N122
N331
N657
Mean Proportion
N3
N203
Ancestral
Delta
Gamma