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

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

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

84

85 ## Results

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

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

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

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

104 Site-specific glycosylation data contain significant associations between protein sequence, 3D 105 protein structures, and glycosylation

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

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

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

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

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

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

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

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

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

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

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

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

171 Many AA-substitutions are glycoimpactful on glycan biosynthesis, as exemplified by the F to W
172 substitution (**Figure** 3**d-e, Supplementary Results**). Upstream F is associated with bi- and tri-antennary
173 (>3-fold) while upstream W is associated with tetraantennary terminal Gal (>2-fold) implicating this
174 substitution in branching (**Figure** 3**d**). Additionally, spatially proximal F is strongly associated with
175 increased sialylation (>10-fold) implicating this amino acid in glycan maturation (**Figure** 3**e**). These
176 predictions suggest that W to F substitution has a large impact on glycoprofiles. Relevant substitution
177 events are highlighted as "glycoimpactful" (significantly high glycoimpact) and structurally ambivalent
178 (BLOSUM>0) (**Figure** 3**f**); those high-glycoimpact (red) and low structural impact (thick line) are
179 substitutions of particular note as potential avenues of non-structurally impactful glycan modulation.

180 Glycoimpact explains divergence of BLOSUM and PAM substitution matrices

181 We call the glycoimpact AA-substitution matrix the BLOSUM-PAM Orthology matrix, BLAMO x:y, where x
182 and y refer to the log odds ratio and FDR thresholds respectively. Sub-threshold odds ratios, those
183 insignificant or unsubstantial by the x:y threshold, are excluded from the glycoimpact calculation. **Figure**
184 3f displays a subset of BLAMO 0.5:0.1 relations. Comparisons can be made at multiple thresholds
185 (Supplementary **Figure** 5) but BLAMO 0.5:0.1 presents as a more representative, normative, and stable
186 threshold.

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

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

206 High glycoimpact residues are conserved around N-glycosites

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

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

223 We next aligned[15] glycosites permitting a tetraantennary N-glycan lacking fucose or sialic acids (**Figure**
224 **4b**). We examined the glycosite alignment for consistency with high-influence AA (**Figure 2c**) and
225 glycosite-coupled residues (**Figure 4c**). Of 20 glycosite-flanking AA, 16 show consistency between the
226 first or second most common AA and either the high-influence or glycosite coupled residues (see
227 **Supplementary Results**). At nearly every glycosite-flanking residue (N+/-10) there is consistency
228 between these three analyses, further corroborating that protein structure constrains glycosylation.

229

230 ## Glycoimpact correlates with discrepancy between functional variant predictions

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

246 ## Glycoimpact proposes mechanisms for pathogenic variants in ClinVar and PrP

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

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

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

269 We further analyzed glycosite proximity among causal variants in prion disease. We measured 3D
270 Euclidean min-distance from the two PrP glycosylation sites, N181 and N197, to all residues in human
271 prion protein (PrP) (including variants causing Creutzfeldt-Jakob disease (CJD)[66,67] and Gerstmann-
272 Sträussler-Scheinker disease (GSS))[68,69] (**Figure** 5a**,** Supplementary **Table** 2). CJD-causing variants were
273 approximately twice as close to glycosylation sites than the background distribution of all PrP sites (One-
274 sided Wilcoxon p=0.0003). GSS-causative variants also trend closer to glycosites (One-sided Wilcoxon
275 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

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

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

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

329 We next predicted differential glycosylation for the *ighg1* missense mutation (F299I) in the IgG1 heavy
330 chain.[78] IgG1 glycosylation impacts both adaptive humoral response[79–82] and monoclonal antibody (mAb)
331 response.[83–86] Experimentally, the IgG1:F299I substitution shows a strong glycoimpact. The IgG1 variant
332 expressed in C57BL/6 and BALB/c mouse strains shows less sialylation and digalactosylation compared
333 to similarly expressed *wt* IgG1.[87–89] Additionally, BALB/c-expressed IgG1:F299I glycosylation is more
334 similar to C57BL/6-expressed IgG1:F299I glycosylation than to glycans on *wt* IgG1 expressed in the same
335 BALB/c animals (**Figure** 7**b-c**). Fc N-glycans on IgG1:F299I expressed in both BALB/c and C57BL/6 animals
336 have more agalactosylation (Mann-Whitney p=1.02e-6), less digalactosylation, and less mono-, di- and
337 total sialylation (Mann-Whitney p<0.0073) compared to *wt* IgG1 expressed in the same animals (**Figure**
338 7**c**, Supplementary **Table** 3). The increase in galactosylation in *wt* IgG1:F299 is consistent with PGD
339 predicted IMR for upstream (N-terminal) F (**Figure** 7**a**). Upstream F is associated with increased di-

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

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

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

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

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

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

378 ### IMR-guided novel Rituximab variants are predictably and differentially glycosylated

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

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

394

395 ## Discussion

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

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

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

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

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

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

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

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

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

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

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

492

493 ## Acknowledgements

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

503

504 ## Conflicts

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

507

508 # Methods

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

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

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

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

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

544

545 ## Software and packages

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

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

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

557 ## Probability event space, information gain and conditional probability

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

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

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

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

597 ## Calculating glycoimpact from IMR and populating a BLAMO matrix

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

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

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

610 ## Comparison of SNP pathogenicity scores with glycoimpact

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

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

619 ## Estimation and Analysis of Evolutionary Couplings (EC)

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

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

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

648 ## Mouse breeding and Samples

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

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

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

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

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

674 ## Glycoproteomics for SARS-CoV-2 Spike glycoproteins

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

680 Briefly, glycoproteins were digested with Proteinase K (30 min or 4 h) or trypsin, sequentially
681 deglycosylated with Endo H (creating residual mass signature of +203 Da) to signify high
682 mannose/hybrid glycans, and then with PNGase F in $H_2{}^{18}O$ (creating residual mass signature of +3 Da) to
683 signify the remnant complex glycans on any sequon (NXS|T, where X is any amino acid except P)
684 asparagine. Unoccupied sequons will have no additional signature mass. Analysis of samples was done
685 on a Q Exactive HF-X mass spectrometer (Thermo), injecting directly onto a 25 cm, 100 µm ID column
686 packed with BEH 1.7 µm C18 resin (Waters). Liquid chromatography separation was achieved at a flow
687 rate of 300 nL min−1 on an EASY-nLC 1200 (Thermo). Buffers A and B were 0.1% formic acid in 5 and
688 80% acetonitrile, respectively. The gradient used was 1−25% B over 160 min, an increase to 40% B over
689 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
690 column was re-equilibrated with solution A before injecting sample. Peptides eluting from the tip of the
691 column were nanosprayed directly into the mass spectrometer by application of 2.8 kV at the back of
692 the column. The mass spectrometer was operated in a data-dependent mode. Full MS1 scans were
693 collected in the Orbitrap at 120000 resolution. The 10 most abundant ions per scan were selected for
694 HCD MS/MS at 25 NCE. Dynamic exclusion was enabled with a 10 s duration and +1 ions were excluded.
695 Peptides were identified with Integrated Proteomics Pipeline (IP2, Bruker Scientific LLC). Tandem mass
696 spectra were extracted from raw files using RawConverter[161] and searched with ProLuCID[162] against a
697 database comprising UniProt reviewed proteome for Homo sapiens (UP000005640), including additional
698 UniProt amino acid sequences for Endo H (P04067), PNGase F (Q9XBM8), and Proteinase K (P06873), the
699 amino acid sequences for the SARS-CoV-2 S1 subunits, and a list of general protein contaminants. The
700 search included no protease specificity (all fully tryptic and semitryptic peptide candidates when treated

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

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

724 Estimation of glycomotif preference from IMR

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

732 Glycomotif preference can be described in more precise terms. Given a substitution where X and Y are
733 WT and mutant amino acids respectively at an amino acid residue at index I, we may describe the
734 substitution as XIY. When I is glycosite-proximal, we may describe a substitution, XIY, as a point, p,
735 with a pair of IMR (ρ) relating each amino acid to a single glycomotif, $p_{XIY} = (\rho_X, \rho_Y)$. The IMR-pair,
736 $p_{XIY} = (\rho_X, \rho_Y)$, may also be described as a vector terminating on that point, $\vec{p}_{XIY}$. We may also
737 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
738 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 $\vec{n}$,
739 $\vec{p}_{\vec{n}} = \frac{\vec{p} \cdot \vec{n}}{\vec{n} \cdot \vec{n}} \cdot \vec{n}$. Therefore, $G = d(\vec{p}, \vec{n}) = \vec{p} - \vec{p}_{\vec{n}}$

740 With this definition of G, we can define G_i and G_c as the glycoimpact corresponding to the glycomotif of
741 interest and the contract glycomotif respectively. Then, the relative preference by first calculating the
742 perpendicular component of the difference in glycoimpact vectors projected onto $\vec{n}$, such that
743 $r_{\vec{n}} = (G_i - G_c)_{\vec{n}}$. 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
744 the sign, s, as positive if $y > x$ and negative if $x > y$. The sign is not defined if $x = y$.

745 ## Glycoprofiling of 25c4 monoclonal antibody using capillary electrophoresis

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

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

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

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

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

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

781 Captions

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

792 ***Figure 2*** **– N- and O-glycan substructures associated with glycosite-proximal protein structure. (a)** Dimensionality reduction
793 trained on UniCarbKB glycosite-proximal protein structures. When projected into the UniCarbKB-trained space, all human
794 UniProt annotated glycosites appear within the UniCarbKB sequence space implying that UniCarbKB glycosites are sufficiently
795 representative of most documented glycosites. The scatterplots show a two-dimensional projection of the FAMD using a
796 Uniform Manifold Approximation and Projection (UMAP).[40] Each point is a site on a glycoprotein, each color indicates the
797 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

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

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

871

872

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.,

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

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

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

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

921 17. Murray, A.N., Chen, W., Antonopoulos, A., Hanson, S.R., Wiseman, R.L., Dell, A., Haslam, S.M.,
922 Powers, D.L., Powers, E.T., and Kelly, J.W. (2015). Enhanced aromatic sequons increase
923 oligosaccharyltransferase glycosylation efficiency and glycan homogeneity. Chem. Biol. *22*, 1052–
924 1062.

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

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

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

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

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

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

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

945 25. Hang, I., Lin, C.-W., Grant, O.C., Fleurkens, S., Villiger, T.K., Soos, M., Morbidelli, M., Woods, R.J.,
946 Gauss, R., and Aebi, M. (2015). Analysis of site-specific N-glycan remodeling in the endoplasmic
947 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.

985 39. Mih, N., Brunk, E., Chen, K., Catoiu, E., Sastry, A., Kavvas, E., Monk, J.M., Zhang, Z., and Palsson, B.O.
986 (2018). ssbio: a Python framework for structural systems biology. Bioinformatics *34*, 2155–2157.

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

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

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

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

998 44. Fitch, W.M., and Dayhoff, M.O. (1973). Atlas of protein sequence and structure, 1972. Syst. Zool.
999 *22*, 196.

1000 45. Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc.
1001 Natl. Acad. Sci. U. S. A. *89*, 10915–10919.

1002 46. Mount, D.W. (2008). Comparison of the PAM and BLOSUM amino acid substitution matrices. CSH
1003 Protoc. *2008*, db.ip59.

1004 47. Pearson, W.R. (2013). Selecting the right similarity-scoring matrix. Curr. Protoc. Bioinformatics *43*,
1005 3.5.1-3.5.9.

1006 48. Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011).
1007 Protein 3D structure computed from evolutionary sequence variation. PLoS One *6*, e28766.

1008 49. Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). DbNSFP v3.0: A one-stop database of functional
1009 predictions and annotations for human nonsynonymous and splice-site SNVs. Hum. Mutat. *37*,
1010 235–241.

1011 50. Thusberg, J., Olatubosun, A., and Vihinen, M. (2011). Performance of mutation pathogenicity
1012 prediction methods on missense variants. Hum. Mutat. *32*, 358–368.

1013 51. Cheng, N., Li, M., Zhao, L., Zhang, B., Yang, Y., Zheng, C.-H., and Xia, J. (2020). Comparison and
1014 integration of computational methods for deleterious synonymous mutation prediction. Brief.
1015 Bioinform. *21*, 970–981.

1016 52. Simeonov, D.R., Wang, X., Wang, C., Sergeev, Y., Dolinska, M., Bower, M., Fischer, R., Winer, D.,
1017 Dubrovsky, G., Balog, J.Z., et al. (2013). DNA variations in oculocutaneous albinism: an updated
1018 mutation list and current outstanding issues in molecular diagnostics. Hum. Mutat. *34*, 827–835.

1019 53. Spritz, R.A., Ho, L., Furumura, M., and Hearing, V.J., Jr (1997). Mutational analysis of copper binding
1020 by human tyrosinase. J. Invest. Dermatol. *109*, 207–212.

1021 54. Halaban, R., Svedine, S., Cheng, E., Smicun, Y., Aron, R., and Hebert, D.N. (2000). Endoplasmic
1022 reticulum retention is a common defect associated with tyrosinase-negative albinism. Proc. Natl.
1023 Acad. Sci. U. S. A. *97*, 5889–5894.

1024 55. Branza-Nichita, N., Negroiu, G., Petrescu, A.J., Garman, E.F., Platt, F.M., Wormald, M.R., Dwek, R.A.,
1025 and Petrescu, S.M. (2000). Mutations at critical N-glycosylation sites reduce tyrosinase activity by
1026 altering folding and quality control. J. Biol. Chem. *275*, 8169–8175.

1027 56. Dolinska, M.B., Kovaleva, E., Backlund, P., Wingfield, P.T., Brooks, B.P., and Sergeev, Y.V. (2014).
1028 Albinism-causing mutations in recombinant human tyrosinase alter intrinsic enzymatic activity.
1029 PLoS One *9*, e84494.

1030 57. Dolinska, M.B., and Sergeev, Y.V. (2017). The consequences of deglycosylation of recombinant
1031 intra-melanosomal domain of human tyrosinase. Biol. Chem. *399*, 73–77.

1032 58. Sevillano, A.M., Aguilar-Calvo, P., Kurt, T.D., Lawrence, J.A., Soldau, K., Nam, T.H., Schumann, T.,
1033 Pizzo, D.P., Nyström, S., Choudhury, B., et al. (2020). Prion protein glycans reduce intracerebral
1034 fibril formation and spongiosis in prion disease. J. Clin. Invest. *130*, 1350–1362.

1035 59. Yi, C.-W., Wang, L.-Q., Huang, J.-J., Pan, K., Chen, J., and Liang, Y. (2018). Glycosylation Significantly
1036 Inhibits the Aggregation of Human Prion Protein and Decreases Its Cytotoxicity. Preprint,
1037 https://doi.org/10.1038/s41598-018-30770-6 https://doi.org/10.1038/s41598-018-30770-6.

1038 60. Ng, B.G., and Freeze, H.H. (2018). Perspectives on Glycosylation and Its Congenital Disorders.
1039 Trends Genet. *34*, 466–476.

1040 61. Joshi, H.J., Hansen, L., Narimatsu, Y., Freeze, H.H., Henrissat, B., Bennett, E., Wandall, H.H., Clausen,
1041 H., and Schjoldager, K.T. (2018). Glycosyltransferase genes that cause monogenic congenital
1042 disorders of glycosylation are distinct from glycosyltransferase genes associated with complex
1043 diseases. Glycobiology *28*, 284–294.

1044 62. Berg-Fussman, A., Grace, M.E., Ioannou, Y., and Grabowski, G.A. (1993). Human acid beta-
1045 glucosidase. N-glycosylation site occupancy and the effect of glycosylation on enzymatic activity. J.
1046 Biol. Chem. *268*, 14861–14866.

1047 63. Pol-Fachin, L., Siebert, M., Verli, H., and Saraiva-Pereira, M.L. (2016). Glycosylation is crucial for a
1048 proper catalytic site organization in human glucocerebrosidase. Glycoconj. J. *33*, 237–244.

1049 64. Smith, L., Mullin, S., and Schapira, A.H.V. (2017). Insights into the structural biology of Gaucher
1050 disease. Exp. Neurol. *298*, 180–190.

1051 65. Souffrant, M.G., Yao, X.-Q., Momin, M., and Hamelberg, D. (2020). N-glycosylation and gaucher
1052 disease mutation allosterically alter active-site dynamics of acid-β-glucosidase. ACS Catal. *10*, 1810–
1053 1820.

1054 66. Di Fede, G., Catania, M., Atzori, C., Moda, F., Pasquali, C., Indaco, A., Grisoli, M., Zuffi, M., Guaita,
1055 M.C., Testi, R., et al. (2019). Clinical and neuropathological phenotype associated with the novel
1056 V189I mutation in the prion protein gene. Acta Neuropathol. Commun. *7*, 1.

1057 67. Ladogana, A., and Kovacs, G.G. (2018). Genetic Creutzfeldt–Jakob disease. In Human Prion Diseases

1058 Handbook of clinical neurology. (Elsevier), pp. 219–242.

1059 68. Liberski, P.P. (2012). Gerstmann-Sträussler-Scheinker disease. Adv. Exp. Med. Biol. *724*, 128–137.

1060 69. Cohen, M.L. (2014). Human Prion Diseases. In Pathobiology of Human Disease (Elsevier), pp. 2045–
1061 2054.

1062 70. Cao, L., Diedrich, J.K., Kulp, D.W., Pauthner, M., He, L., Park, S.-K.R., Sok, D., Su, C.Y., Delahunty,
1063 C.M., Menis, S., et al. (2017). Global site-specific N-glycosylation analysis of HIV envelope
1064 glycoprotein. Preprint, https://doi.org/10.1038/ncomms14954
1065 https://doi.org/10.1038/ncomms14954.

1066 71. Zhao, P., Praissman, J.L., Grant, O.C., Cai, Y., Xiao, T., Rosenbalm, K.E., Aoki, K., Kellman, B.P.,
1067 Bridger, R., Barouch, D.H., et al. (2020). Virus-Receptor Interactions of Glycosylated SARS-CoV-2
1068 Spike and Human ACE2 Receptor. Cell Host Microbe *28*, 586-601.e6.

1069 72. Watanabe, Y., Allen, J.D., Wrapp, D., McLellan, J.S., and Crispin, M. (2020). Site-specific glycan
1070 analysis of the SARS-CoV-2 spike. Science *369*, 330–333.

1071 73. Casalino, L., Gaieb, Z., Dommer, A.C., and Harbison, A.M. (2020). Shielding and Beyond: The Roles
1072 of Glycans in SARS-CoV-2 Spike Protein. bioRxiv.

1073 74. Grant, O.C., Montgomery, D., Ito, K., and Woods, R.J. (2020). 3D Models of glycosylated SARS-CoV-2
1074 spike protein suggest challenges and opportunities for vaccine development. bioRxiv.

1075 75. Wintjens, R., Bifani, A.M., and Bifani, P. (2020). Impact of glycan cloud on the B-cell epitope
1076 prediction of SARS-CoV-2 Spike protein. NPJ Vaccines *5*, 81.

1077 76. Zhang, L., Mann, M., Syed, Z.A., Reynolds, H.M., Tian, E., Samara, N.L., Zeldin, D.C., Tabak, L.A., and
1078 Ten Hagen, K.G. (2021). Furin cleavage of the SARS-CoV-2 spike is modulated by O-glycosylation.
1079 Proc. Natl. Acad. Sci. U. S. A. *118*, e2109905118.

1080 77. Baboo, S., Diedrich, J.K., Martínez-Bartolomé, S., Wang, X., Schiffner, T., Groschel, B., Schief, W.R.,
1081 Paulson, J.C., and Yates, J.R., 3rd (2021). DeGlyPHER: An ultrasensitive method for the analysis of
1082 viral spike N-glycoforms. Anal. Chem. *93*, 13651–13657.

1083 78. Krištić, J., Zaytseva, O.O., Ram, R., Nguyen, Q., Novokmet, M., Vučković, F., Vilaj, M., Trbojević-
1084 Akmačić, I., Pezer, M., Davern, K.M., et al. (2018). Profiling and genetic control of the murine
1085 immunoglobulin G glycome. Nat. Chem. Biol. *14*, 516–524.

1086 79. Cheng, H.D., Tirosh, I., de Haan, N., Stöckmann, H., Adamczyk, B., McManus, C.A., O'Flaherty, R.,
1087 Greville, G., Saldova, R., Bonilla, F.A., et al. (2020). IgG Fc glycosylation as an axis of humoral
1088 immunity in childhood. J. Allergy Clin. Immunol. *145*, 710-713.e9.

1089 80. Lofano, G., Gorman, M.J., Yousif, A.S., Yu, W.-H., Fox, J.M., Dugast, A.-S., Ackerman, M.E.,
1090 Suscovich, T.J., Weiner, J., Barouch, D., et al. (2018). Antigen-specific antibody Fc glycosylation
1091 enhances humoral immunity via the recruitment of complement. Sci. Immunol. *3*, eaat7796.

1092 81. Atyeo, C., Pullen, K.M., Bordt, E.A., Fischinger, S., Burke, J., Michell, A., Slein, M.D., Loos, C., Shook,
1093 L.L., Boatin, A.A., et al. (2021). Compromised SARS-CoV-2-specific placental antibody transfer. Cell

1094 *184*, 628-642.e10.

1095 82. Grace, P.S., Dolatshahi, S., Lu, L.L., Cain, A., Palmieri, F., Petrone, L., Fortune, S.M., Ottenhoff,
1096 T.H.M., Lauffenburger, D.A., Goletti, D., et al. (2021). Antibody subclass and glycosylation shift
1097 following effective TB treatment. Front. Immunol. *12*, 679973.

1098 83. Majewska, N.I., Tejada, M.L., Betenbaugh, M.J., and Agarwal, N. (2020). N-glycosylation of IgG and
1099 IgG-like recombinant therapeutic proteins: Why is it important and how can we control it? Annu.
1100 Rev. Chem. Biomol. Eng. *11*, 311–338.

1101 84. Li, W., Zhu, Z., Chen, W., Feng, Y., and Dimitrov, D.S. (2017). Crystallizable fragment
1102 glycoengineering for therapeutic antibodies development. Front. Immunol. *8*.
1103 https://doi.org/10.3389/fimmu.2017.01554.

1104 85. Zhang, P., Woen, S., Wang, T., Liau, B., Zhao, S., Chen, C., Yang, Y., Song, Z., Wormald, M.R., Yu, C.,
1105 et al. (2016). Challenges of glycosylation analysis and control: an integrated approach to producing
1106 optimal and consistent therapeutic drugs. Drug Discov. Today *21*, 740–765.

1107 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.
1108 (2002). Lack of Fucose on Human IgG1 N-Linked Oligosaccharide Improves Binding to Human FcγRIII
1109 and Antibody-dependent Cellular Toxicity. J. Biol. Chem. *277*, 26733–26740.

1110 87. de Haan, N., Reiding, K.R., Krištić, J., Hipgrave Ederveen, A.L., Lauc, G., and Wuhrer, M. (2017). The
1111 N-Glycosylation of Mouse Immunoglobulin G (IgG)-Fragment Crystallizable Differs Between IgG
1112 Subclasses and Strains. Front. Immunol. *8*, 608.

1113 88. Zaytseva, O.O., Jansen, B.C., Hanić, M., Mrčela, M., Razdorov, G., Stojković, R., Erhardt, J., Brizić, I.,
1114 Jonjić, S., Pezer, M., et al. (2018). MIgGGly (mouse IgG glycosylation analysis) - a high-throughput
1115 method for studying Fc-linked IgG N-glycosylation in mice with nanoUPLC-ESI-MS. Sci. Rep. *8*,
1116 13688.

1117 89. Zaytseva, O.O., Seeling, M., Krištić, J., Lauc, G., Pezer, M., and Nimmerjahn, F. (2020). Fc-Linked IgG
1118 N-Glycosylation in FcγR Knock-Out Mice. Preprint, https://doi.org/10.3389/fcell.2020.00067
1119 https://doi.org/10.3389/fcell.2020.00067.

1120 90. Gunn, B.M., Lu, R., Slein, M.D., Ilinykh, P.A., Huang, K., Atyeo, C., Schendel, S.L., Kim, J., Cain, C.,
1121 Roy, V., et al. (2021). A Fc engineering approach to define functional humoral correlates of
1122 immunity against Ebola virus. Immunity *54*, 815-828.e5.

1123 91. Wong, M.Y., Chen, K., Antonopoulos, A., Kasper, B.T., Dewal, M.B., Taylor, R.J., Whittaker, C.A.,
1124 Hein, P.P., Dell, A., Genereux, J.C., et al. (2018). XBP1s activation can globally remodel N-glycan
1125 structure distribution patterns. Proc. Natl. Acad. Sci. U. S. A. *115*, E10089–E10098.

1126 92. Chen, K., and Shoulders, M.D. (2024). Protein glycosylation patterns shaped by the IRE1-XBP1s arm
1127 of the unfolded protein response. Isr. J. Chem. *64*. https://doi.org/10.1002/ijch.202300162.

1128 93. Karch, C.P., Paquin-Proulx, D., Eller, M.A., Matyas, G.R., Burkhard, P., and Beck, Z. (2020). Impact of
1129 the expression system on the immune responses to self-assembling protein nanoparticles (SAPNs)
1130 displaying HIV-1 V1V2 loop. Nanomedicine *29*, 102255.

1131 94. Hombu, R., Neelamegham, S., and Park, S. (2021). Cellular and molecular engineering of glycan
1132 sialylation in heterologous systems. Molecules *26*, 5950.

1133 95. Zhong, X., Ma, W., Meade, C.L., Tam, A.S., Llewellyn, E., Cornell, R., Cote, K., Scarcelli, J.J., Marshall,
1134 J.K., Tzvetkova, B., et al. (2019). Transient CHO expression platform for robust antibody production
1135 and its enhanced N-glycan sialylation on therapeutic glycoproteins. Biotechnol. Prog. *35*, e2724.

1136 96. Zhong, X., Schwab, A., Ma, W., Meade, C.L., Zhou, J., D'Antona, A.M., Somers, W., and Lin, L. (2022).
1137 Large-scale transient production in ExpiCHO-S™ with enhanced N-galactosylation-sialylation and
1138 PEI-based transfection. Methods Mol. Biol. *2313*, 143–150.

1139 97. Hudson, K.L., Bartlett, G.J., Diehl, R.C., Agirre, J., Gallagher, T., Kiessling, L.L., and Woolfson, D.N.
1140 (2015). Carbohydrate-aromatic interactions in proteins. J. Am. Chem. Soc. *137*, 15152–15160.

1141 98. Chen, W., Enck, S., Price, J.L., Powers, D.L., Powers, E.T., Wong, C.-H., Dyson, H.J., and Kelly, J.W.
1142 (2013). Structural and energetic basis of carbohydrate-aromatic packing interactions in proteins. J.
1143 Am. Chem. Soc. *135*, 9877–9884.

1144 99. Asensio, J.L., Ardá, A., Cañada, F.J., and Jiménez-Barbero, J. (2013). Carbohydrate-aromatic
1145 interactions. Acc. Chem. Res. *46*, 946–954.

1146 100. Montalvillo-Jiménez, L., Santana, A.G., Corzana, F., Jiménez-Osés, G., Jiménez-Barbero, J., Gómez,
1147 A.M., and Asensio, J.L. (2019). Impact of aromatic stacking on glycoside reactivity: Balancing CH/π
1148 and cation/π interactions for the stabilization of glycosyl-oxocarbenium ions. J. Am. Chem. Soc. *141*,
1149 13372–13384.

1150 101. Ardejani, M.S., Noodleman, L., Powers, E.T., and Kelly, J.W. (2021). Stereoelectronic effects in
1151 stabilizing protein-N-glycan interactions revealed by experiment and machine learning. Nat. Chem.
1152 *13*, 480–487.

1153 102. Goldfarb, L.G., Petersen, R.B., Tabaton, M., Brown, P., LeBlanc, A.C., Montagna, P., Cortelli, P.,
1154 Julien, J., Vital, C., and Pendelbury, W.W. (1992). Fatal familial insomnia and familial Creutzfeldt-
1155 Jakob disease: disease phenotype determined by a DNA polymorphism. Science *258*, 806–808.

1156 103. Wang, W.K., Essex, M., and Lee, T.H. (1996). Single amino acid substitution in constant region 1 or 4
1157 of gp120 causes the phenotype of a human immunodeficiency virus type 1 variant with mutations
1158 in hypervariable regions 1 and 2 to revert. J. Virol. *70*, 607–611.

1159 104. Branza-Nichita, N., Petrescu, A.J., Negroiu, G., Dwek, R.A., and Petrescu, S.M. (2000). N-
1160 glycosylation processing and glycoprotein folding-lessons from the tyrosinase-related proteins.
1161 Chem. Rev. *100*, 4697–4712.

1162 105. Dolinska, M.B., Kus, N.J., Farney, S.K., Wingfield, P.T., Brooks, B.P., and Sergeev, Y.V. (2017).
1163 Oculocutaneous albinism type 1: link between mutations, tyrosinase conformational stability, and
1164 enzymatic activity. Pigment Cell Melanoma Res. *30*, 41–52.

1165 106. Grant, O.C., Montgomery, D., Ito, K., and Woods, R.J. (2020). Analysis of the SARS-CoV-2 spike
1166 protein glycan shield reveals implications for immune recognition. Sci. Rep. *10*, 14991.

1167 107. Peng, W., Rayaprolu, V., Parvate, A.D., Pronker, M.F., Hui, S., Parekh, D., Shaffer, K., Yu, X., Saphire,

1168 E.O., and Snijder, J. (2022). Glycan shield of the ebolavirus envelope glycoprotein GP. Commun.
1169 Biol. *5*, 785.

1170 108. Wei, C.-J., Boyington, J.C., Dai, K., Houser, K.V., Pearce, M.B., Kong, W.-P., Yang, Z.-Y., Tumpey,
1171 T.M., and Nabel, G.J. (2010). Cross-neutralization of 1918 and 2009 influenza viruses: role of
1172 glycans in viral evolution and vaccine design. Sci. Transl. Med. *2*, 24ra21.

1173 109. Schulz, M.A., Tian, W., Mao, Y., Van Coillie, J., Sun, L., Larsen, J.S., Chen, Y.-H., Kristensen, C.,
1174 Vakhrushev, S.Y., Clausen, H., et al. (2018). Glycoengineering design options for IgG1 in CHO cells
1175 using precise gene editing. Glycobiology *28*, 542–549.

1176 110. Narimatsu, Y., Joshi, H.J., Nason, R., Van Coillie, J., Karlsson, R., Sun, L., Ye, Z., Chen, Y.-H.,
1177 Schjoldager, K.T., Steentoft, C., et al. (2019). An atlas of human glycosylation pathways enables
1178 display of the human glycome by gene engineered cells. Mol. Cell *75*, 394-407.e5.

1179 111. Narimatsu, Y., Büll, C., Chen, Y.-H., Wandall, H.H., Yang, Z., and Clausen, H. (2021). Genetic
1180 glycoengineering in mammalian cells. J. Biol. Chem. *296*, 100448.

1181 112. Sumit, M., Dolatshahi, S., Chu, A.-H.A., Cote, K., Scarcelli, J.J., Marshall, J.K., Cornell, R.J., Weiss, R.,
1182 Lauffenburger, D.A., Mulukutla, B.C., et al. (2019). Dissecting N-Glycosylation Dynamics in Chinese
1183 Hamster Ovary Cells Fed-batch Cultures using Time Course Omics Analyses. iScience *12*, 102–120.

1184 113. Ehret, J., Zimmermann, M., Eichhorn, T., and Zimmer, A. (2019). Impact of cell culture media
1185 additives on IgG glycosylation produced in Chinese hamster ovary cells. Biotechnol. Bioeng. *116*,
1186 816–830.

1187 114. Fan, Y., Kildegaard, H.F., and Andersen, M.R. (2017). Engineer medium and feed for modulating N-
1188 glycosylation of recombinant protein production in CHO cell culture. Methods Mol. Biol. *1603*, 209–
1189 226.

1190 115. Kotidis, P., and Kontoravdi, C. (2020). Harnessing the potential of artificial neural networks for
1191 predicting protein glycosylation. Metabolic Engineering Communications, e00131.

1192 116. Jimenez del Val, I., Nagy, J.M., and Kontoravdi, C. (2011). A dynamic mathematical model for
1193 monoclonal antibody N-linked glycosylation and nucleotide sugar donor transport within a
1194 maturing Golgi apparatus. Biotechnol. Prog. *27*, 1730–1743.

1195 117. Sou, S.N., Jedrzejewski, P.M., Lee, K., Sellick, C., Polizzi, K.M., and Kontoravdi, C. (2017). Model-
1196 based investigation of intracellular processes determining antibody Fc-glycosylation under mild
1197 hypothermia. Biotechnol. Bioeng. *114*, 1570–1582.

1198 118. Spahn, P.N., Hansen, A.H., Hansen, H.G., Arnsdorf, J., Kildegaard, H.F., and Lewis, N.E. (2016). A
1199 Markov chain model for N-linked protein glycosylation – towards a low-parameter tool for model-
1200 driven glycoengineering. Preprint, https://doi.org/10.1016/j.ymben.2015.10.007
1201 https://doi.org/10.1016/j.ymben.2015.10.007.

1202 119. Spahn, P.N., Hansen, A.H., Kol, S., Voldborg, B.G., and Lewis, N.E. (2017). Predictive
1203 glycoengineering of biosimilars using a Markov chain glycosylation model. Biotechnol. J. *12*.
1204 https://doi.org/10.1002/biot.201600489.

1205 120. Liang, C., Chiang, A.W.T., Hansen, A.H., Arnsdorf, J., Schoffelen, S., Sorrentino, J.T., Kellman, B.P.,
1206 Bao, B., Voldborg, B.G., and Lewis, N.E. (2020). A Markov model of glycosylation elucidates isozyme
1207 specificity and glycosyltransferase interactions for glycoengineering. Curr Res Biotechnol *2*, 22–36.

1208 121. Michaeli, Y., and Reiter, Y. (2006). Optimised Fc variants with enhanced effector function. Expert
1209 Opin. Ther. Pat. *16*, 1449–1452.

1210 122. Liu, R., Oldham, R.J., Teal, E., Beers, S.A., and Cragg, M.S. (2020). Fc-engineering for modulated
1211 effector functions-improving antibodies for cancer treatment. Antibodies (Basel) *9*, 64.

1212 123. van Erp, E.A., Luytjes, W., Ferwerda, G., and van Kasteren, P.B. (2019). Fc-mediated antibody
1213 effector functions during respiratory syncytial virus infection and disease. Front. Immunol. *10*, 548.

1214 124. Leon, P.E., He, W., Mullarkey, C.E., Bailey, M.J., Miller, M.S., Krammer, F., Palese, P., and Tan, G.S.
1215 (2016). Optimal activation of Fc-mediated effector functions by influenza virus hemagglutinin
1216 antibodies requires two points of contact. Proc. Natl. Acad. Sci. U. S. A. *113*, E5944–E5951.

1217 125. Zhou, M., Palanca, A.M.S., and Law, J.A. (2018). Locus-specific control of the de novo DNA
1218 methylation pathway in Arabidopsis by the CLASSY family. Nat. Genet. *50*, 865–873.

1219 126. Nováček, V., McGauran, G., Matallanas, D., Vallejo Blanco, A., Conca, P., Muñoz, E., Costabello, L.,
1220 Kanakaraj, K., Nawaz, Z., Walsh, B., et al. (2020). Accurate prediction of kinase-substrate networks
1221 using knowledge graphs. PLoS Comput. Biol. *16*, e1007578.

1222 127. Malaker, S.A., Riley, N.M., Shon, D.J., Pedram, K., Krishnan, V., Dorigo, O., and Bertozzi, C.R. (2022).
1223 Revealing the human mucinome. Nat. Commun. *13*, 3542.

1224 128. Madhani, H.D., and Guthrie, C. (1994). Dynamic RNA-RNA interactions in the spliceosome. Annu.
1225 Rev. Genet. *28*, 1–26.

1226 129. Bao, P., Boon, K.-L., Will, C.L., Hartmuth, K., and Lührmann, R. (2018). Multiple RNA–RNA tertiary
1227 interactions are dispensable for formation of a functional U2/U6 RNA catalytic core in the
1228 spliceosome. Preprint, https://doi.org/10.1093/nar/gky966 https://doi.org/10.1093/nar/gky966.

1229 130. Corley, M., Burns, M.C., and Yeo, G.W. (2020). How RNA-Binding Proteins Interact with RNA:
1230 Molecules and Mechanisms. Mol. Cell *78*, 9–29.

1231 131. Farnham, P.J., and Platt, T. (1982). Effects of DNA base analogs on transcription termination at the
1232 tryptophan operon attenuator of EScherichia coli. Proc. Natl. Acad. Sci. U. S. A. *79*, 998–1002.

1233 132. Lee, F., and Yanofsky, C. (1977). Transcription termination at the trp operon attenuators of
1234 Escherichia coli and Salmonella typhimurium: RNA secondary structure and regulation of
1235 termination. Proc. Natl. Acad. Sci. U. S. A. *74*, 4365–4369.

1236 133. Oxender, D.L., Zurawski, G., and Yanofsky, C. (1979). Attenuation in the Escherichia coli tryptophan
1237 operon: role of RNA secondary structure involving the tryptophan codon region. Proc. Natl. Acad.
1238 Sci. U. S. A. *76*, 5524–5528.

1239 134. Farnham, P.J., and Platt, T. (1980). A model for transcription termination suggested by studies on
1240 the trp attenuator in vitro using base analogs. Cell *20*, 739–748.

1241 135. Wu, A.M., and Platt, T. (1978). Transcription termination: nucleotide sequence at 3' end of
1242 tryptophan operon in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. *75*, 5442–5446.

1243 136. Mariethoz, J., Alocci, D., Gastaldello, A., Horlacher, O., Gasteiger, E., Rojas-Macias, M., Karlsson,
1244 N.G., Packer, N.H., and Lisacek, F. (2018). Glycomics@ExPASy: Bridging the Gap. Preprint,
1245 https://doi.org/10.1074/mcp.ra118.000799 https://doi.org/10.1074/mcp.ra118.000799.

1246 137. York, W.S., Mazumder, R., Ranzinger, R., Edwards, N., Kahsay, R., Aoki-Kinoshita, K.F., Campbell,
1247 M.P., Cummings, R.D., Feizi, T., Martin, M., et al. (2020). GlyGen: Computational and Informatics
1248 Resources for Glycoscience. Glycobiology *30*, 72–73.

1249 138. Cooper, C.A., Joshi, H.J., Harrison, M.J., Wilkins, M.R., and Packer, N.H. (2003). GlycoSuiteDB: a
1250 curated relational database of glycoprotein glycan structures and their biological sources. 2003
1251 update. Nucleic Acids Res. *31*, 511–513.

1252 139. Al Jadda, K., Porterfield, M.P., Bridger, R., Heiss, C., Tiemeyer, M., Wells, L., Miller, J.A., York, W.S.,
1253 and Ranzinger, R. (2015). EUROCarbDB(CCRC): a EUROCarbDB node for storing glycomics standard
1254 data. Bioinformatics *31*, 242–245.

1255 140. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and
1256 Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

1257 141. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer,
1258 T.A.P., Rempfer, C., Bordoli, L., et al. (2018). SWISS-MODEL: homology modelling of protein
1259 structures and complexes. Nucleic Acids Res. *46*, W296–W303.

1260 142. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER Suite: protein
1261 structure and function prediction. Nat. Methods *12*, 7–8.

1262 143. Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N.,
1263 Potter, S.C., Finn, R.D., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019.
1264 Nucleic Acids Res. *47*, W636–W641.

1265 144. Zhang, Y., and Sagui, C. (2015). Secondary structure assignment for conformationally irregular
1266 peptides: comparison between DSSP, STRIDE and KAKSI. J. Mol. Graph. Model. *55*, 72–84.

1267 145. Magnan, C.N., and Baldi, P. (2014). SSpro/ACCpro 5: almost perfect prediction of protein secondary
1268 structure and relative solvent accessibility using profiles, machine learning and structural similarity.
1269 Bioinformatics *30*, 2592–2597.

1270 146. Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K.F., Ueda, N., Hamajima, M., Kawasaki, T., and
1271 Kanehisa, M. (2006). KEGG as a glycome informatics resource. Glycobiology *16*, 63R-70R.

1272 147. Herget, S., Ranzinger, R., Maass, K., and Lieth, C.-W.V.D. (2008). GlycoCT-a unifying sequence
1273 format for carbohydrates. Carbohydr. Res. *343*, 2162–2171.

1274 148. Burkholz, R., Quackenbush, J., and Bojar, D. (2021). Using graph convolutional neural networks to
1275 learn a representation for glycans. Cell Rep. *35*, 109251.

1276 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.

1312 164. Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J., and Gygi, S.P. (2003). Evaluation of
1313 multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for
1314 large-scale protein analysis: the yeast proteome. J. Proteome Res. *2*, 43–50.

1315 165. Park, S.K., Venable, J.D., Xu, T., and Yates, J.R., 3rd (2008). A quantitative analysis software tool for
1316 mass spectrometry-based proteomics. Nat. Methods *5*, 319–322.

1317 166. Irvine, E.B., Peters, J.M., Lu, R., Grace, P.S., Sixsmith, J., Wallace, A., Schneider, M., Shin, S.,
1318 Karpinski, W., Hsiao, J.C., et al. (2022). Fc-engineered antibodies leverage neutrophils to drive
1319 control of *Mycobacterium tuberculosis*. bioRxiv. https://doi.org/10.1101/2022.05.01.490220.

1320 167. Engler, C., and Marillonnet, S. (2014). Golden Gate cloning. Methods Mol. Biol. *1116*, 119–131.

1321 168. Fang, J., Qian, J.-J., Yi, S., Harding, T.C., Tu, G.H., VanRoey, M., and Jooss, K. (2005). Stable antibody
1322 expression at therapeutic levels using the 2A peptide. Nat. Biotechnol. *23*, 584–590.

1323 169. Mahan, A.E., Tedesco, J., Dionne, K., Baruah, K., Cheng, H.D., De Jager, P.L., Barouch, D.H.,
1324 Suscovich, T., Ackerman, M., Crispin, M., et al. (2015). A method for high-throughput, sensitive
1325 analysis of IgG Fc and Fab glycosylation by capillary electrophoresis. J. Immunol. Methods *417*, 34–
1326 44.

1327 170. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and
1328 Powerful Approach to Multiple Testing. J. R. Stat. Soc. Series B Stat. Methodol. *57*, 289–300.

1329 171. Xu, C., and Ng, D.T.W. (2015). Glycosylation-directed quality control of protein folding. Nat. Rev.
1330 Mol. Cell Biol. *16*, 742–752.

1331 172. Shental-Bechor, D., and Levy, Y. (2008). Effect of glycosylation on protein folding: a close look at
1332 thermodynamic stabilization. Proc. Natl. Acad. Sci. U. S. A. *105*, 8256–8261.

1333 173. Shihab, H.A., Gough, J., Mort, M., Cooper, D.N., Day, I.N.M., and Gaunt, T.R. (2014). Ranking non-
1334 synonymous single nucleotide polymorphisms based on disease concepts. Hum. Genomics *8*, 11.

1335 174. Shihab, H.A., Gough, J., Cooper, D.N., Day, I.N.M., and Gaunt, T.R. (2013). Predicting the functional
1336 consequences of cancer-associated amino acid substitutions. Bioinformatics *29*, 1504–1510.

1337 175. Shihab, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L.A., Edwards, K.J., Day, I.N.M., and
1338 Gaunt, T.R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino
1339 acid substitutions using hidden Markov models. Hum. Mutat. *34*, 57–65.

1340 176. Baković, M.P., Selman, M.H.J., Hoffmann, M., Rudan, I., Campbell, H., Deelder, A.M., Lauc, G., and
1341 Wuhrer, M. (2013). High-throughput IgG Fc N-glycosylation profiling by mass spectrometry of
1342 glycopeptides. J. Proteome Res. *12*, 821–831.

1343 177. Pucić, M., Knezević, A., Vidic, J., Adamczyk, B., Novokmet, M., Polasek, O., Gornik, O., Supraha-
1344 Goreta, S., Wormald, M.R., Redzić, I., et al. (2011). High throughput isolation and glycosylation
1345 analysis of IgG-variability and heritability of the IgG glycome in three isolated human populations.
1346 Mol. Cell. Proteomics *10*, M111.010090.

1347 178. Klarić, L., Tsepilov, Y.A., Stanton, C.M., Mangino, M., Sikka, T.T., Esko, T., Pakhomov, E., Salo, P.,

1348 Deelen, J., McGurnaghan, S.J., et al. (2020). Glycosylation of immunoglobulin G is regulated by a
1349 large network of genes pleiotropic with inflammatory diseases. Sci. Adv. *6*, eaax0301.

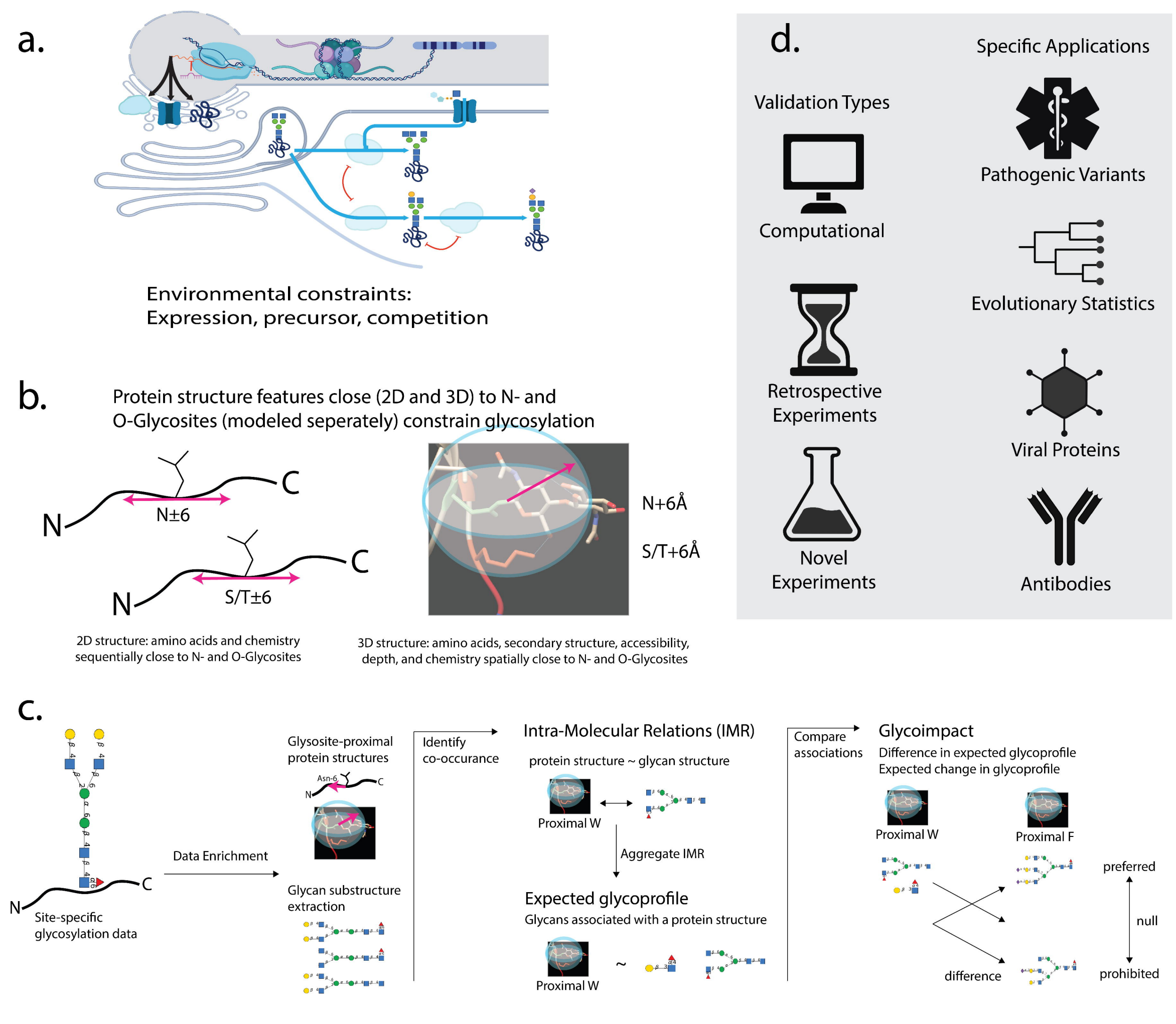
1350 179. Rose, R.J., van Berkel, P.H.C., van den Bremer, E.T.J., Labrijn, A.F., Vink, T., Schuurman, J., Heck,
1351 A.J.R., and Parren, P.W.H.I. (2013). Mutation of Y407 in the CH3 domain dramatically alters
1352 glycosylation and structure of human IgG. MAbs *5*, 219–228.

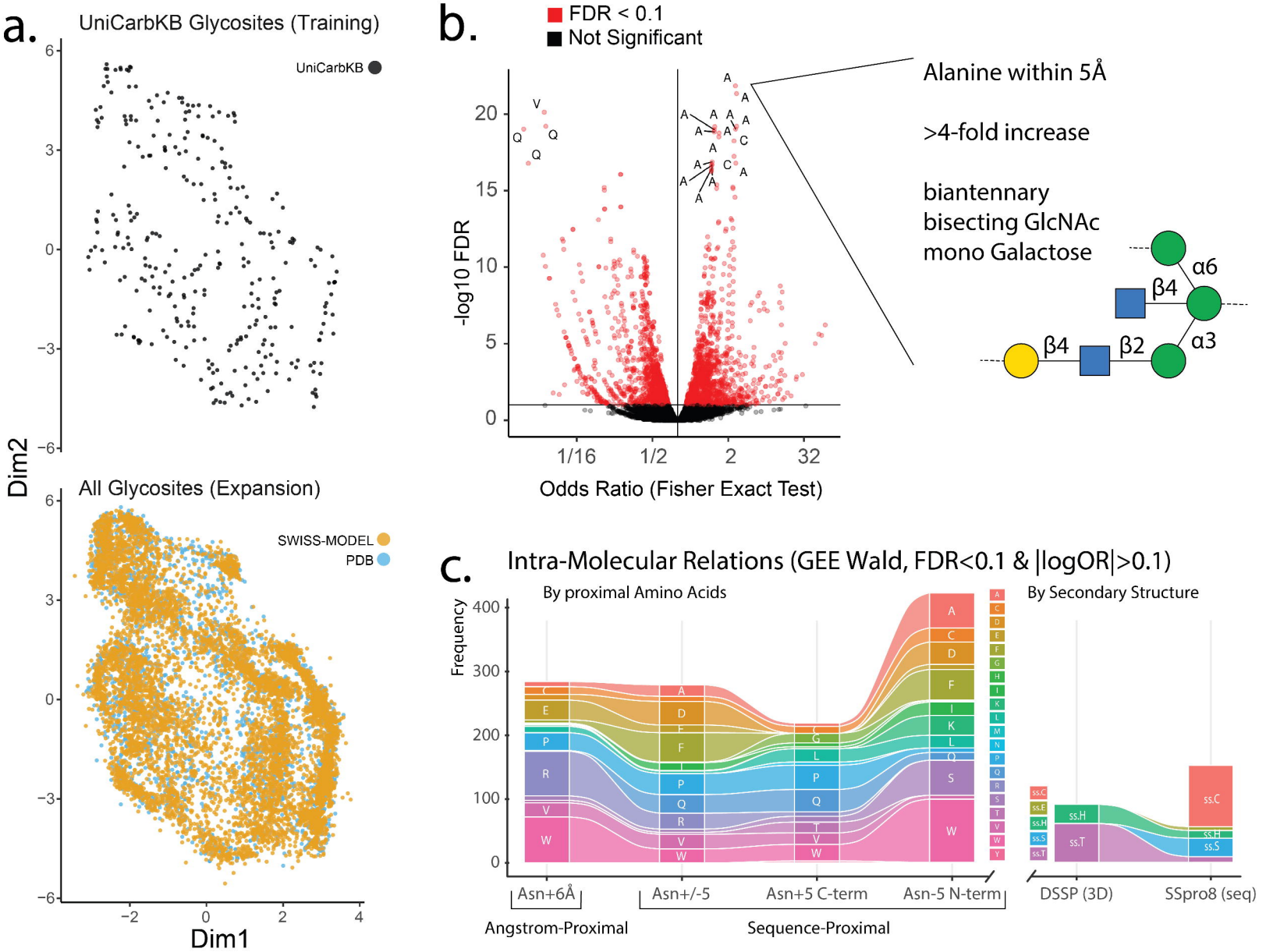
1353 180. Vidarsson, G., Dekkers, G., and Rispens, T. (2014). IgG subclasses and allotypes: from structure to
1354 effector functions. Front. Immunol. *5*, 520.

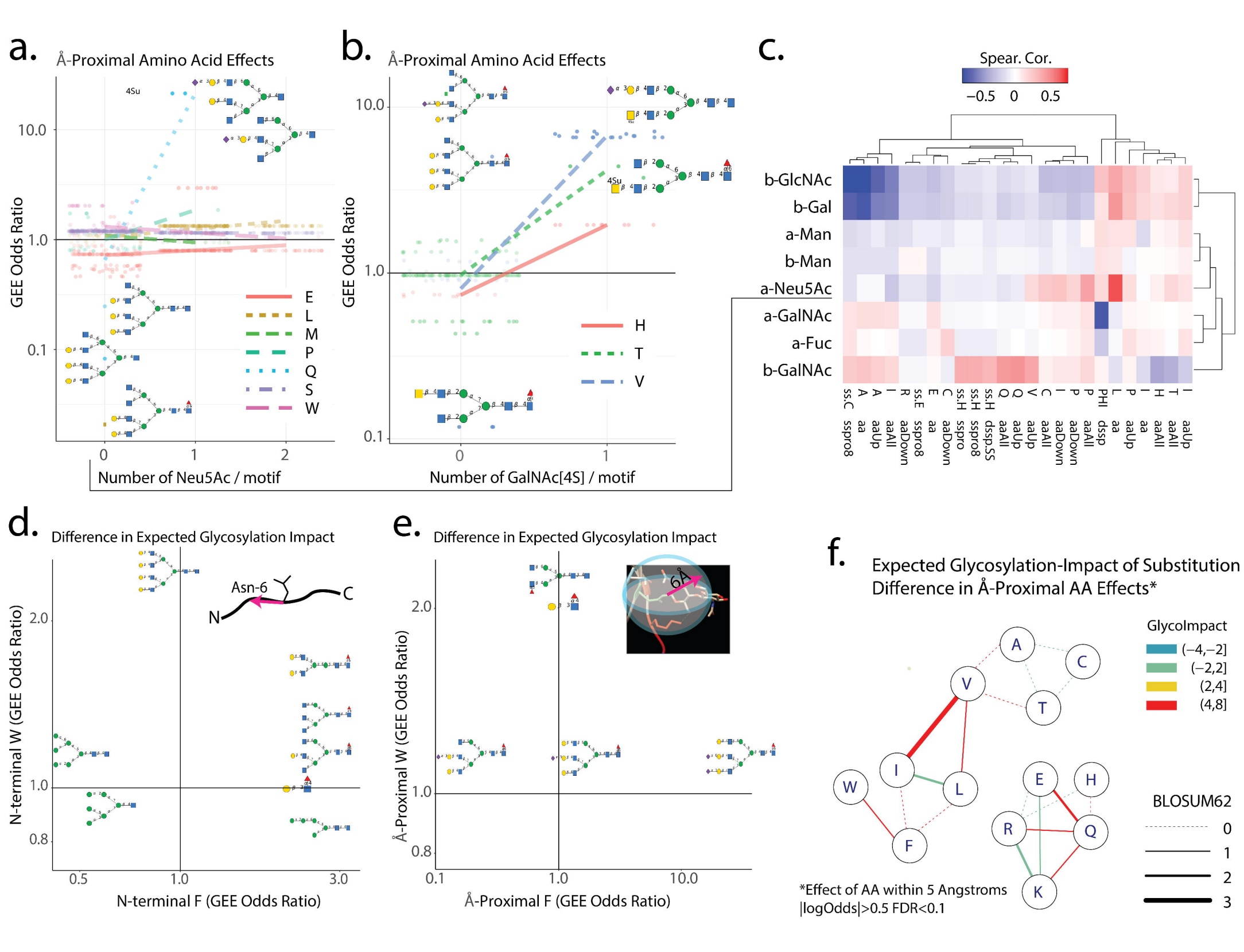
1355 181. Hutton, S.M., and Spritz, R.A. (2008). Comprehensive analysis of oculocutaneous albinism among
1356 non-Hispanic caucasians shows that OCA1 is the most prevalent OCA type. J. Invest. Dermatol. *128*,
1357 2442–2450.

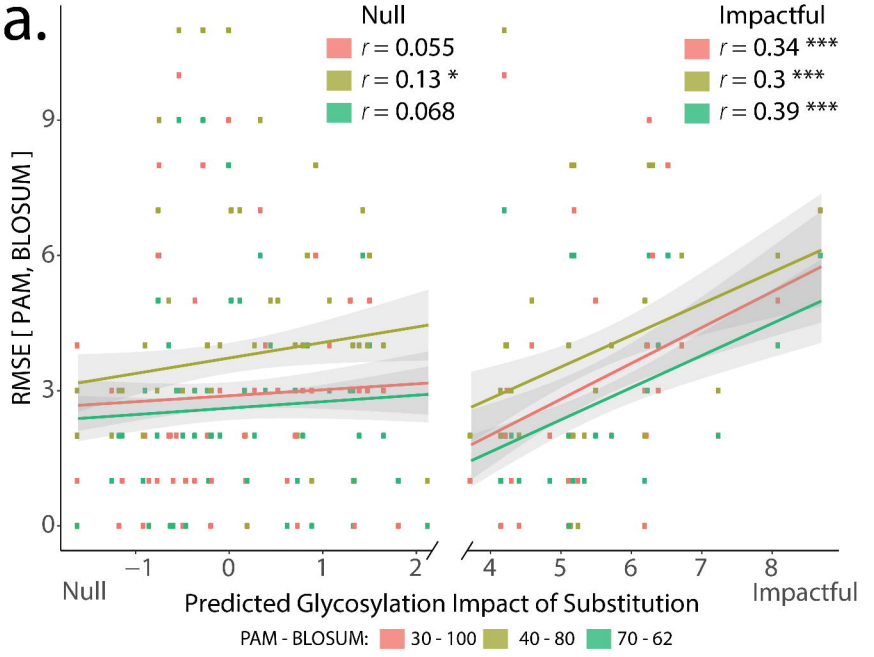
1358 182. Rademacher, C., Bru, T., McBride, R., Robison, E., Nycholat, C.M., Kremer, E.J., and Paulson, J.C.
1359 (2012). A Siglec-like sialic-acid-binding motif revealed in an adenovirus capsid protein. Glycobiology
1360 *22*, 1086–1091.

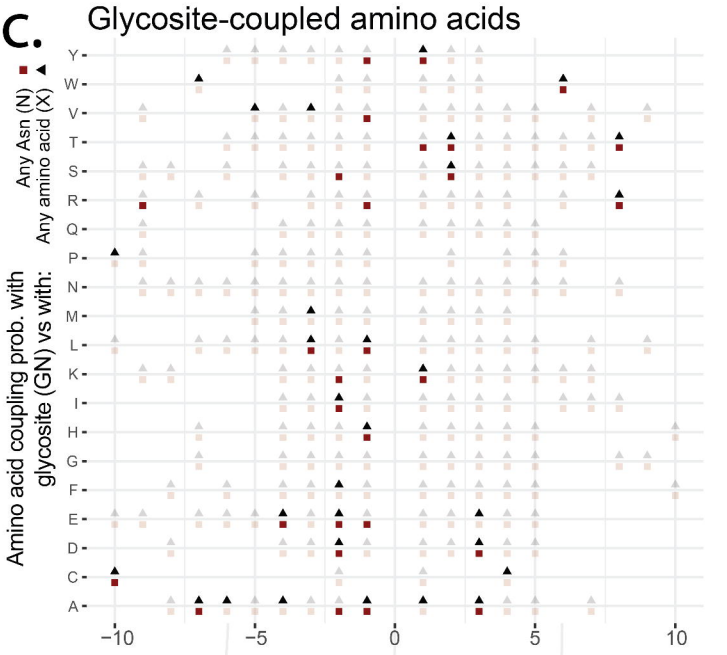
1361 183. Chavent, M., Kuentz-Simonet, V., Labenne, A., and Saracco, J. (2014). Multivariate analysis of mixed
1362 data: The R package PCAmixdata. arXiv [stat.CO].

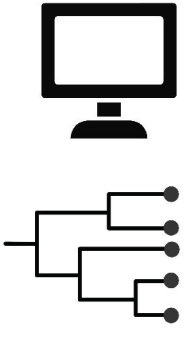

a.
Environmental constraints:
Expression, precursor, competition
b.
Protein structure features close (2D and 3D) to N- and
O-Glycosites (modeled seperately) constrain glycosylation
N
C
N±6
N
C
S/T±6
N+6Å
S/T+6Å
2D structure: amino acids and chemistry
sequentially close to N- and O-Glycosites
3D structure: 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.
N
C
Site-specific
glycosylation data
Data Enrichment
Glysosite-proximal
protein structures
Asn-6
N
C
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
Proximal W
Proximal F
difference
preferred
null
prohibited

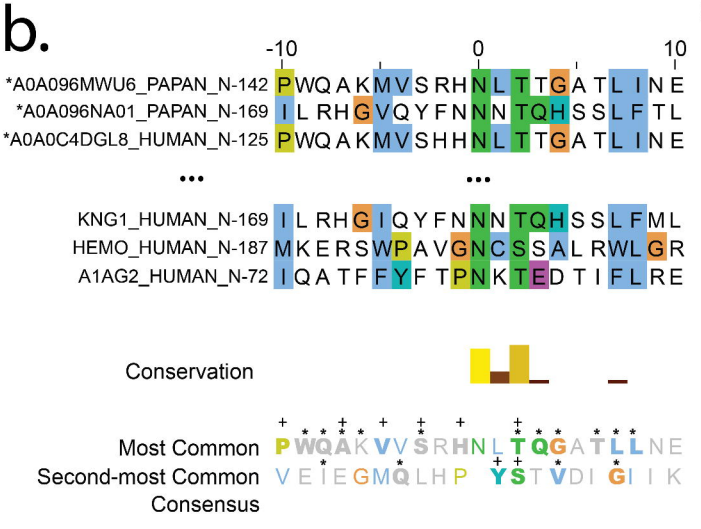

a.
UniCarbKB Glycosites (Training)
UniCarbKB
All Glycosites (Expansion)
SWISS-MODEL
PDB
Dim1
Dim2
b.
FDR < 0.1
Not Significant
-log10 FDR
Odds Ratio (Fisher Exact Test)
1/16
1/2
2
32
Alanine within 5Å
>4-fold increase
biantennary
bisecting GlcNAc
mono Galactose
α6
β4
β4
β2
α3
c.
Intra-Molecular Relations (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
DSSP (3D)
SSpro8 (seq)
Angstrom-Proximal
Sequence-Proximal

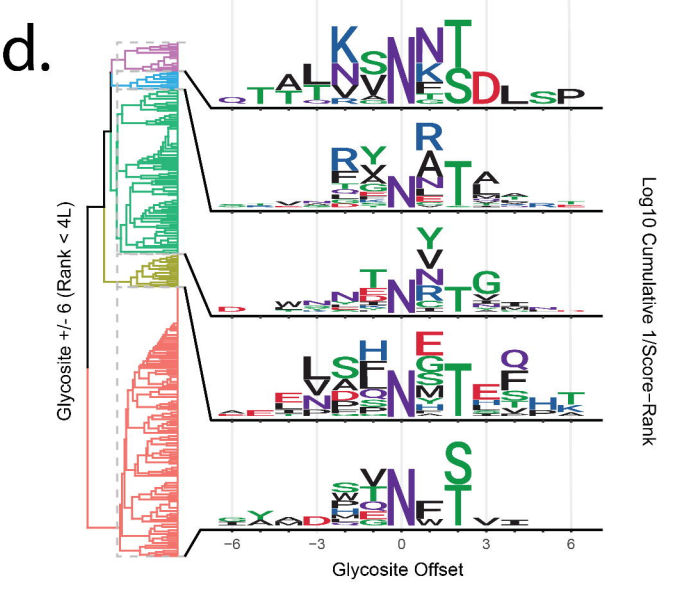

a.
Å-Proximal Amino Acid Effects
GEE Odds Ratio
Number of Neu5Ac / motif
E
L
M
P
Q
S
W
b.
Å-Proximal Amino Acid Effects
GEE Odds Ratio
Number of GalNAc[4S] / motif
H
T
V
c.
Spear. Cor.
−0.5 0 0.5
b-GlcNAc
b-Gal
a-Man
b-Man
a-Neu5Ac
a-GalNAc
a-Fuc
b-GalNAc
d.
Difference in Expected Glycosylation Impact
Asn-6
N
C
N-terminal W (GEE Odds Ratio)
N-terminal F (GEE Odds Ratio)
e.
Difference in Expected Glycosylation Impact
6Å
Å-Proximal W (GEE Odds Ratio)
Å-Proximal F (GEE Odds Ratio)
f.
Expected Glycosylation-Impact of Substitution
Difference in Å-Proximal AA Effects*
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

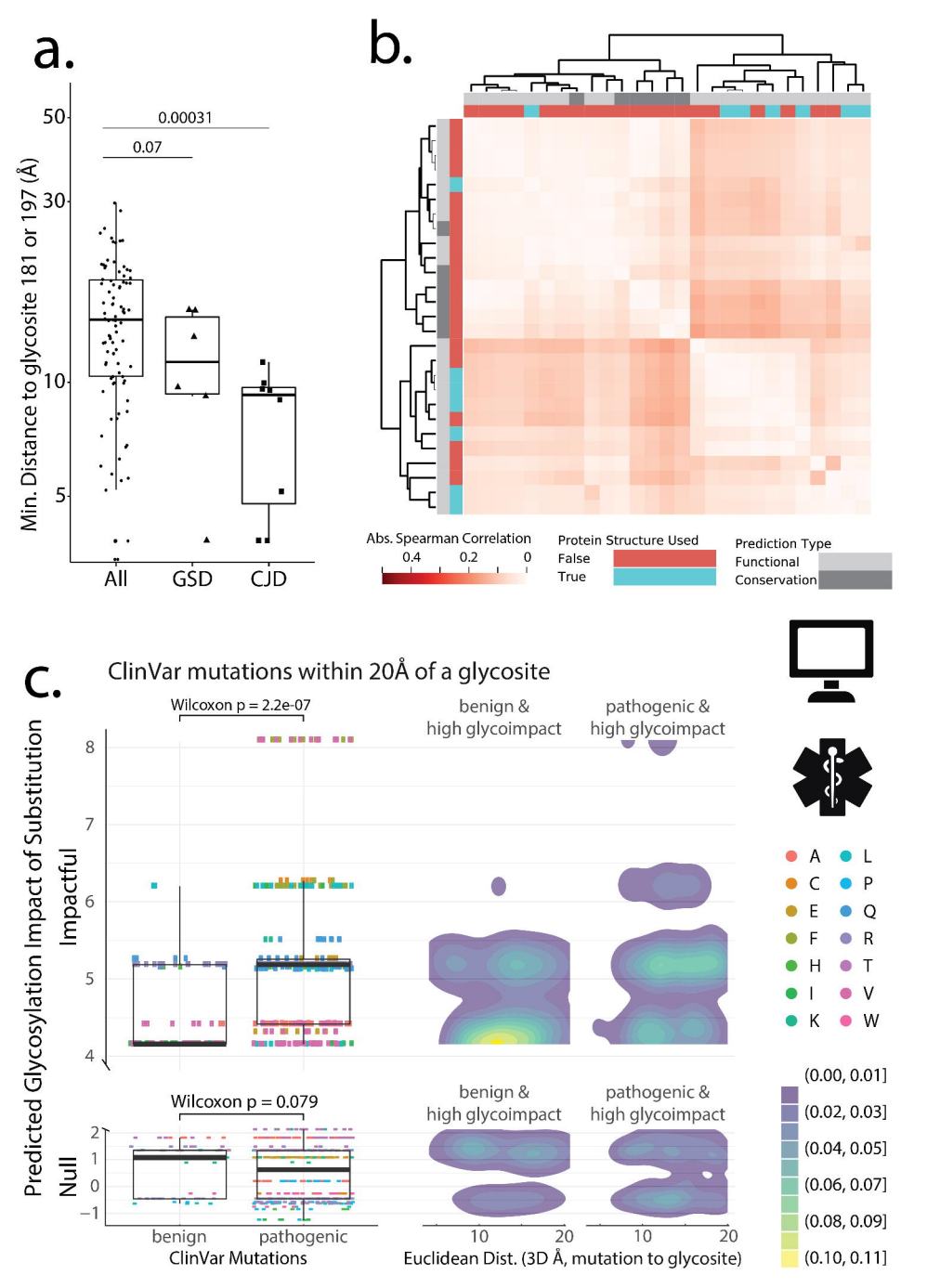

a.
Null
r = 0.055
r = 0.13 *
r = 0.068
Impactful
r = 0.34 ***
r = 0.3 ***
r = 0.39 ***
RMSE [PAM, BLOSUM]
Null
Impactful
Predicted Glycosylation Impact of Substitution
PAM - BLOSUM: 30 - 100 40 - 80 70 - 62

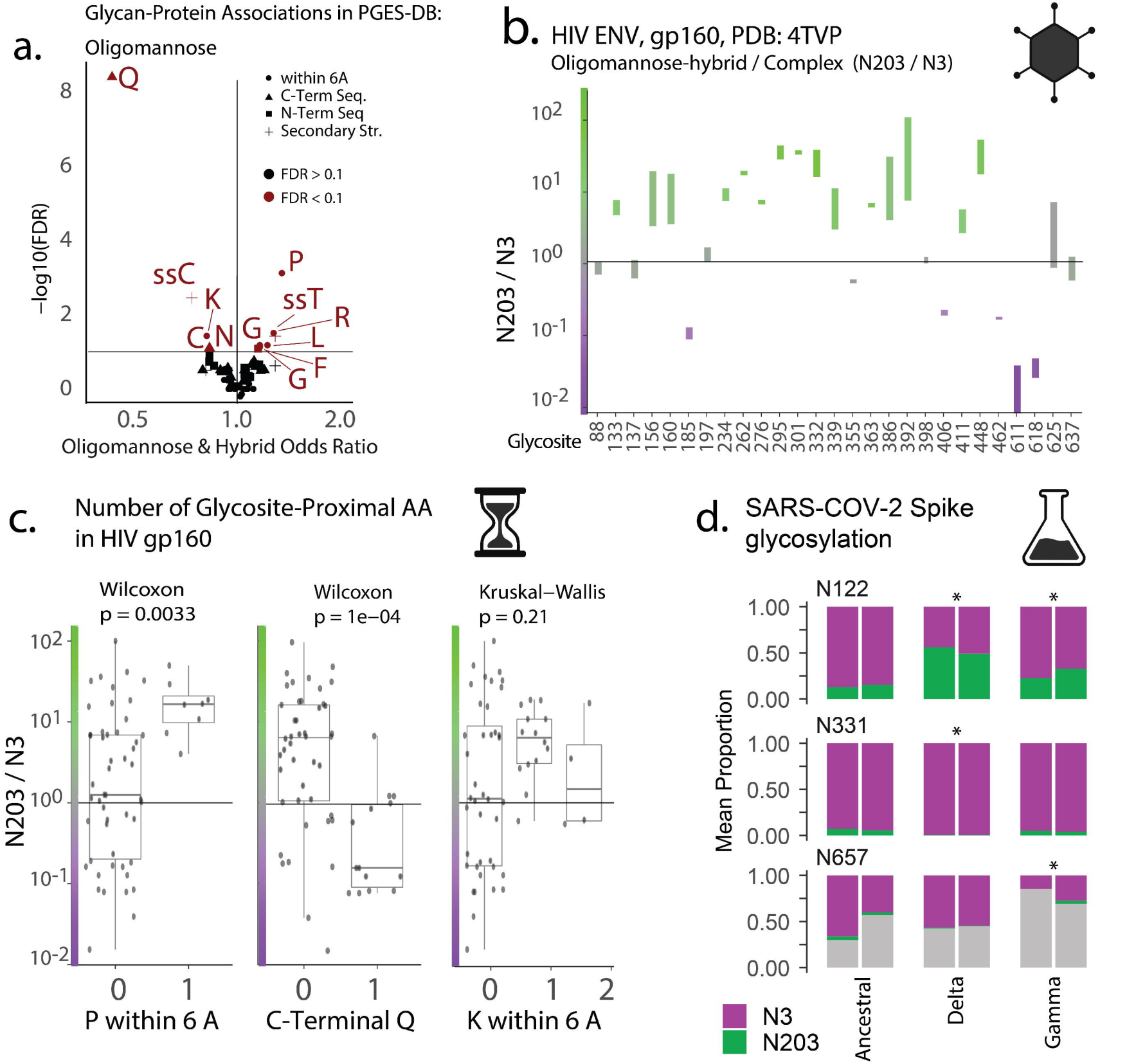

c.
Glycosite-coupled amino acids
Any Asn (N)
Any amino acid (X)
Amino acid coupling prob. with glycosite (GN) vs with:

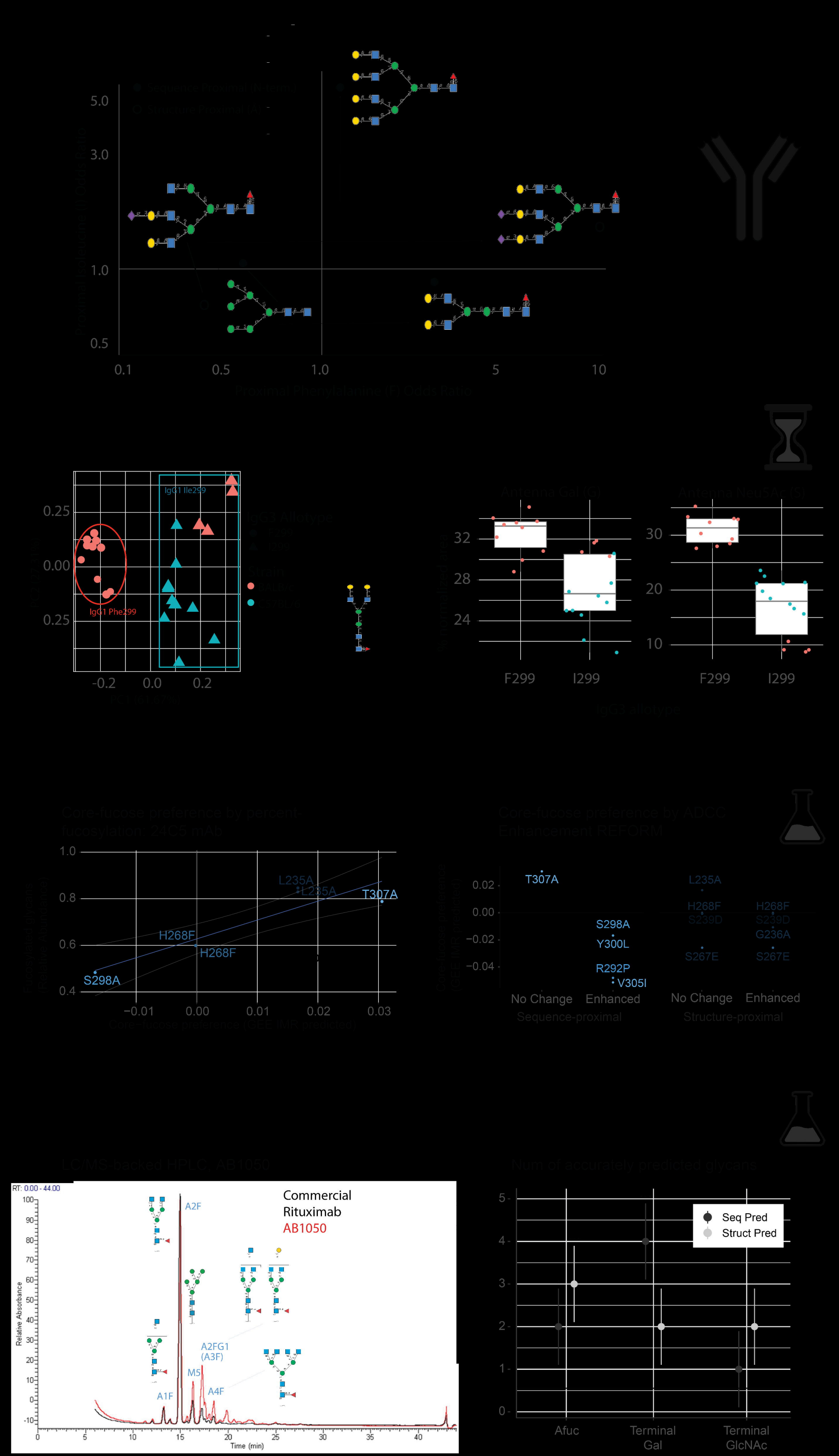



b.
*A0A096MWU6_PAPAN_N-142
*A0A096NA01_PAPAN_N-169
*A0A0C4DGL8_HUMAN_N-125
KNG1_HUMAN_N-169
HEMO_HUMAN_N-187
A1AG2_HUMAN_N-72
Conservation
Most Common
Second-most Common
Consensus


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


a.
0.00031
0.07
Min. Distance to glycosite 181 or 197 (Å)
50
30
10
5
All
GSD
CJD
b.
Abs. Spearman Correlation
0.4
0.2
0
Protein Structure Used
False
True
Prediction Type
Functional
Conservation
c.
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
Q
ssC
K
C
N
G
P
ssT
R
L
F
G
−log10(FDR)
Oligomannose & Hybrid Odds Ratio
b.
HIV ENV, gp160, PDB: 4TVP
Oligomannose-hybrid / Complex (N203 / N3)
N203 / N3
Glycosite
88 133 137 156 160 185 197 234 262 276 295 301 332 339 355 363 386 392 398 406 411 448 462 611 618 625 637
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


Sequence Proximal (N-term.)
Structure Proximal (A)
Proximal Isoleucine (I) Odds Ratio
Proximal Phenylalanine (F) Odds Ratio
5.0
3.0
1.0
0.5
0.1
0.5
1.0
5
10
IgG1 Ile299
IgG1 Phe299
IgG3 Allotype
F299
I299
Strain
BALB/c
C57BL/6
PC1 (61.67%)
0.25
0.00
0.25
-0.2
0.0
0.2
Antenna Gal (G)
Antenna Neu5Ac (S)
Normalized Area
32
28
24
30
20
10
F299
I299
IgG3 allotype
Core-fucose preference by percent-fucosylation: 24C5 mAb
Fucosylated glycans (Relative Abundance)
Core-fucose preference (GEE IMR predicted)
1.0
0.8
0.6
0.4
−0.01
0.00
0.01
0.02
0.03
L235A
L235A
T307A
H268F
H268F
S298A
Core-fucose preference by ADCC Enhancement REFORM
Core-fucose preference (GEE IMR predicted)
0.02
0.00
−0.02
−0.04
T307A
S298A
Y300L
R292P
V305I
L235A
H268F
S239D
S267E
H268F
S239D
G236A
S267E
No Change
Enhanced
No Change
Enhanced
Sequence-proximal
Structure-proximal
LC/MS-backed HPLC: AB1050
RT: 0.00 - 44.00
Relative Absorbance
Time (min)
Commercial
Rituximab
AB1050
A2F
A2FG1
(A3F)
M5
A4F
A1F
Num of accurately predicted glycans
Seq Pred
Struct Pred
Afuc
Terminal
Gal
Terminal
GlcNAc