Protein structure, a genetic encoding for 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] 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,[12] Shlomi Ilan,[5] Isaac Shamie,[1,2,3] Jolene K. Diedrich,[9] Xiaoning Wang,[9] Esther van Woudenbergh,[5] Meghan Altman,[13] Anthony Aylward,[2,3] Bokan Bao,[1,2,3] Andrea Castro,[2,3] James Sorrentino,[1,2,3] Austin Chiang,[1,2] Matt Campbell,[14] Yannic Bartsch,[5] Patricia Aguilar-Calvo,[13,15,16] Christina Sigurdson,[13,15] Galit Alter,[5] Gordan Lauc,[7] John R. Yates III,[9] Debora Marks,[8,17] Frederique Lisacek,[18,19] Nathan E. Lewis[1,2,3,*]

[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, La Jolla, CA 92092
[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] Department of Pediatrics, Boston Medical Center, Boston, MA, USA
[13] Department of Chemistry and Biochemistry, University of California San Diego, 9500 Gilman Drive, La Jolla, CA, USA
[14] Institute for Glycomics, Griffith University, Queensland, Australia
[15] Department of Pathology, University of California San Diego, 9500 Gilman Drive, La Jolla, CA, USA
[16] Department of Neurobiology, The University of Alabama at Birmingham, Birmingham, AL 35243
[17] Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, United States
[18] Proteome Informatics Group, Swiss Institute of Bioinformatics, CH-1227 Geneva, Switzerland
[19] Computer Science Department & Section of Biology, University of Geneva, route de Drize 7, CH-1227, Geneva, Switzerland

Summary (150 words)

Unlike DNA, RNA, and proteins, the dogma describes glycosylation as metabolically determined and unconstrained by template molecules. Without template-based expectations for glycan structures, research is hampered, obscuring how these critical molecules impact the behavior in thousands of human glycoproteins. Here, we challenge the assertion of template-free glycosylation and discover protein-encoded rules for glycan biosynthesis, by quantifying associations between glycan and protein features, which we call "glycoimpact." We estimate 45-55% of amino acids substitutions will minimally change protein structure, but significantly impact glycosylation. We find that "glycoimpact" influences canonical substitution matrices and genetic variant pathogenicity. We identify thousands of high-glycoimpact pathogenic variants spanning hundreds of diseases, including several linked to aberrant glycosylation including Oculocutaneous Albinism, Prion, and Gaucher's Disease. We also successfully predict glycosylation in HIV, SARS-CoV-2, and immunoglobulins. Overall, we present rules defining a genetic encoding for glycosylation, enabling glycan prediction and discovery of glycoprotein functions in health and disease.

Abstract

DNA, RNA, and proteins are synthesized using template molecules, but glycosylation is not believed to be constrained by a template. However, if cellular environment is the only determinant of glycosylation, all sites should receive the same glycans on average. This template-free assertion is inconsistent with observations of microheterogeneity—wherein each site receives distinct and reproducible glycan structures. Here, we test the assumption of template-free glycan biosynthesis. Through structural analysis of site-specific glycosylation data, we find protein-sequence and structural features that predict specific glycan features. To quantify these relationships, we present a new amino acid substitution matrix that describes "glycoimpact" -- how glycosylation varies with protein structure. High-glycoimpact amino acids co-evolve with glycosites, and glycoimpact is high when estimates of amino acid conservation and variant pathogenicity diverge. We report hundreds 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, we validate glycoimpact quantification 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. In all, we show glycan biosynthesis is accurately guided by specific, genetically-encoded rules, and this presents a plausible refutation to the assumption of template-free glycosylation.

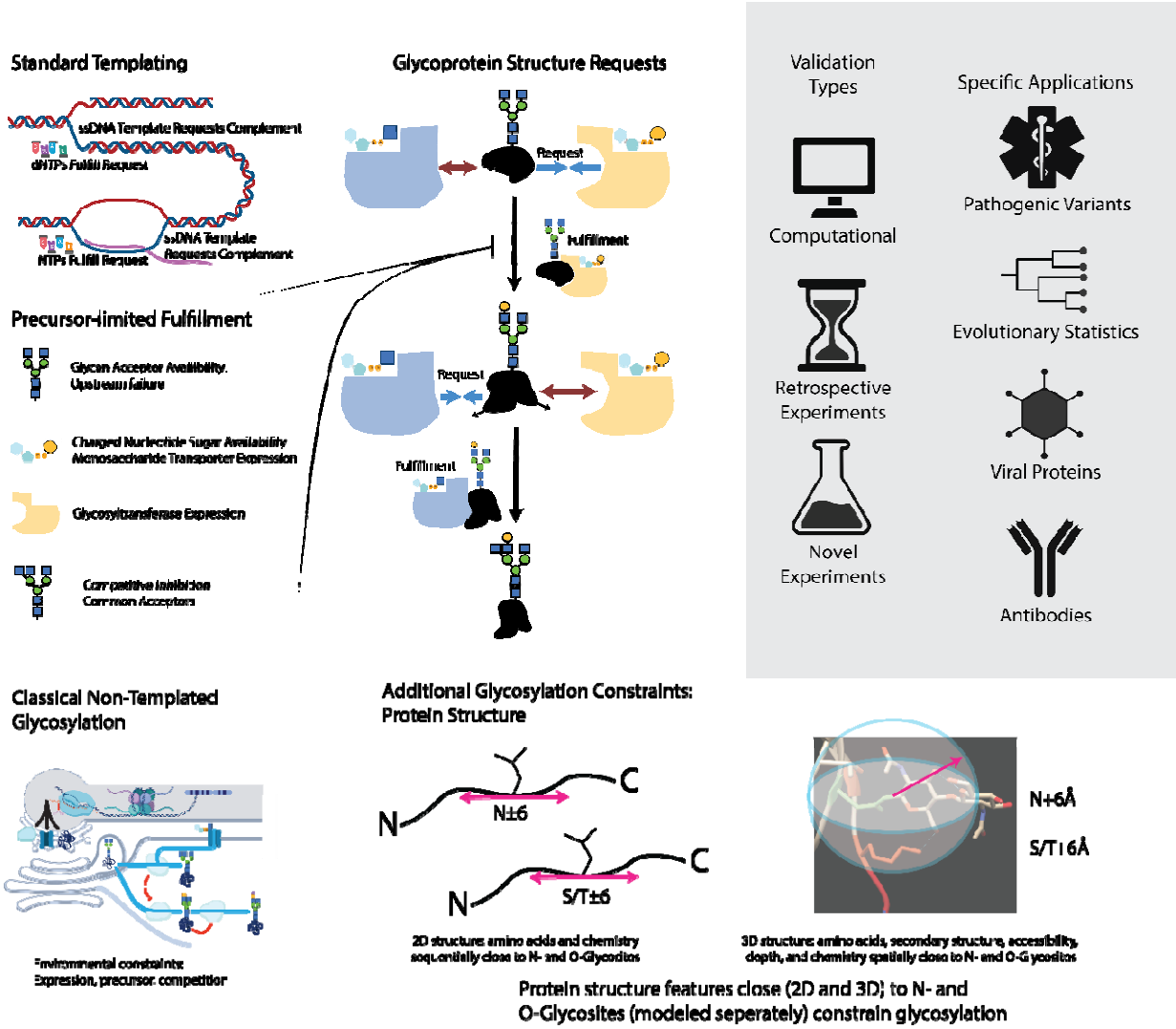


Figure 1 – DNA, RNA, and protein biosynthesis are template-driven processes. Glycan biosynthesis is believed to be non-templated and regulated by metabolic and enzymatic constraints. Here we explore protein structure influence on glycan biosynthesis. In our theory of "bounded biosynthesis", the glycoprotein (black) surface requests glycans by permitting specific glycosyltransferases (blue and yellow) to dock and the metabolic and enzymatic constraints govern the extent to which those requests are fulfilled. Therefore, we posit that glycan biosynthesis is accurately and reproducibly constrained by protein structure; an effect visible after accounting for metabolic influence. Glycans are represented using IUPAC-extension for glycans and the Symbol Nomenclature for Glycans (SNFG[1,2]); Mannose (Man), Galactose (Gal), Sialic Acid (Neu5Ac), N-Acetylglucosamine (GlcNAc), Fucose (Fuc). Here, we discover, validate, and demonstrate glycotemplating through comparison to established computational results and methods, comparison to published experimental results, and prediction of novel experimental results across pathogenic gene variants, evolution, viral proteins, and antibodies.

Introduction

DNA, RNA, and protein synthesis follow DNA and RNA templates. However, glycosylation is considered non-templated, assumed to be primarily regulated by the cellular environment.[3] This assertion leads to a dilemma. If glycan biosynthesis is template-free, all glycosylation sites on a protein should have similar glycans with some stochastic variation. Yet, each specific site on a protein can reproducibly host different glycans, a phenomenon termed "glycosite-specific microheterogeneity."[4]

Though the template-free hypothesis remains, many observations show protein structure influences glycosylation. For example, protein primary sequence influences glycan diversity and identity. In the N-glycosylation sequon,[5] glycans covalently attach 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[6] and accessibility impacts glycosite occupancy.[7] Glycosylation of sequons ending in threonine is ~40 times more efficient than those with serine.[8] Upstream of the glycosite, a phenylalanine to alanine substitution in human IgG3 can increase bi-galactose structures with a core-fucose.[9] Additionally, influenza evolves hemagglutinin glycosylation sites to facilitate immune evasion.[10] Glycoprotein-determined glycan evolution, used by HIV for immune evasion, has been leveraged to engineer better vaccine epitopes.[11] Tools like GlycoSiteAlign[12] 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[13] with a variable impact given the N+1 variation.[14] In prediction, several works since 2002 have shown variable success in predicting glycosites from sequons and hidden features using machine learning,[15–18] though none have predicted beyond glycosite presence or the root monosaccharide.

Beyond the primary sequence, secondary structures (β-sheets and α-helixes) can influence glycosylation[19] while concavity, accessibility and hydrophobicity also impact glycosite occupancy.[7] ERManI (MAN1B1) and Golgi Mannosidase IA (MAN1A1), crucial bottlenecks in glycan processing, have been co-crystalized with a Man9 glycan to study specific glycoconjugate features that permit or inhibit this interaction.[20–23] A study of >150 glycoproteins revealed differential glycosylation as a function of protein structure, wherein low accessibility sites impacted core fucosylation and branching;[24] similar accessibility constraints predicted oligomannose on the SARS-CoV-2 spike protein.[25] Also, a structural analysis found FUT8 activity is impacted by glycosite depth.[26] Site-specific kinetics for glycosylation of Protein Disulfide Isomerase[27–29] and five additional glycoproteins[30] further showed that protein structure influences occupancy. Recent studies have shown interest is grafting glycans to proteins.[17,31–33] These

89 approaches provide valuable instruments for exploring mechanistic and steric associations between
90 glycans and folded proteins. Yet, they do not explore generalizable associations between glycan and
91 protein structure features that can be mapped back to the genome.

92 These studies represent a growing refutation of non-templated glycan biosynthesis. However, they have
93 not been consolidated into an interoperable mapping from the genome to the glycome. Here, we define
94 rules governing glycosylation by analyzing trends in site-specific glycosylation[34] in the context of glycan
95 biosynthesis[35] and site-specific protein structure features.[36] We describe this genetic encoding for
96 glycosylation as a mapping from "glycoimpactful" protein structure to expected glycan structures. We
97 validate these glycoimpact relations by comparison to evolutionary substitution matrices, variant
98 pathogenicity scores, and glycosite co-evolution. Further, we find that glycosite-proximal pathogenic
99 variants correspond to higher glycoimpact substitutions. Finally, we use glycoimpact to accurately
100 predict changes in glycan complexity, galactosylation, sialylation and functional glycosylation. Taken
101 together, we present a comprehensive argument for updating our understanding of glycosylation from
102 template-free, to bounded biosynthesis. Under bounded glycan biosynthesis, a range of feasible glycans
103 are "requested" by local protein structure and are "fulfilled" contingent on precursor availability (**Figure**
104 1); biosynthetic constraints represent a genetic encoding, making glycans predictable for any biologist.

105

Results

Constructing a database containing paired protein and glycan structures

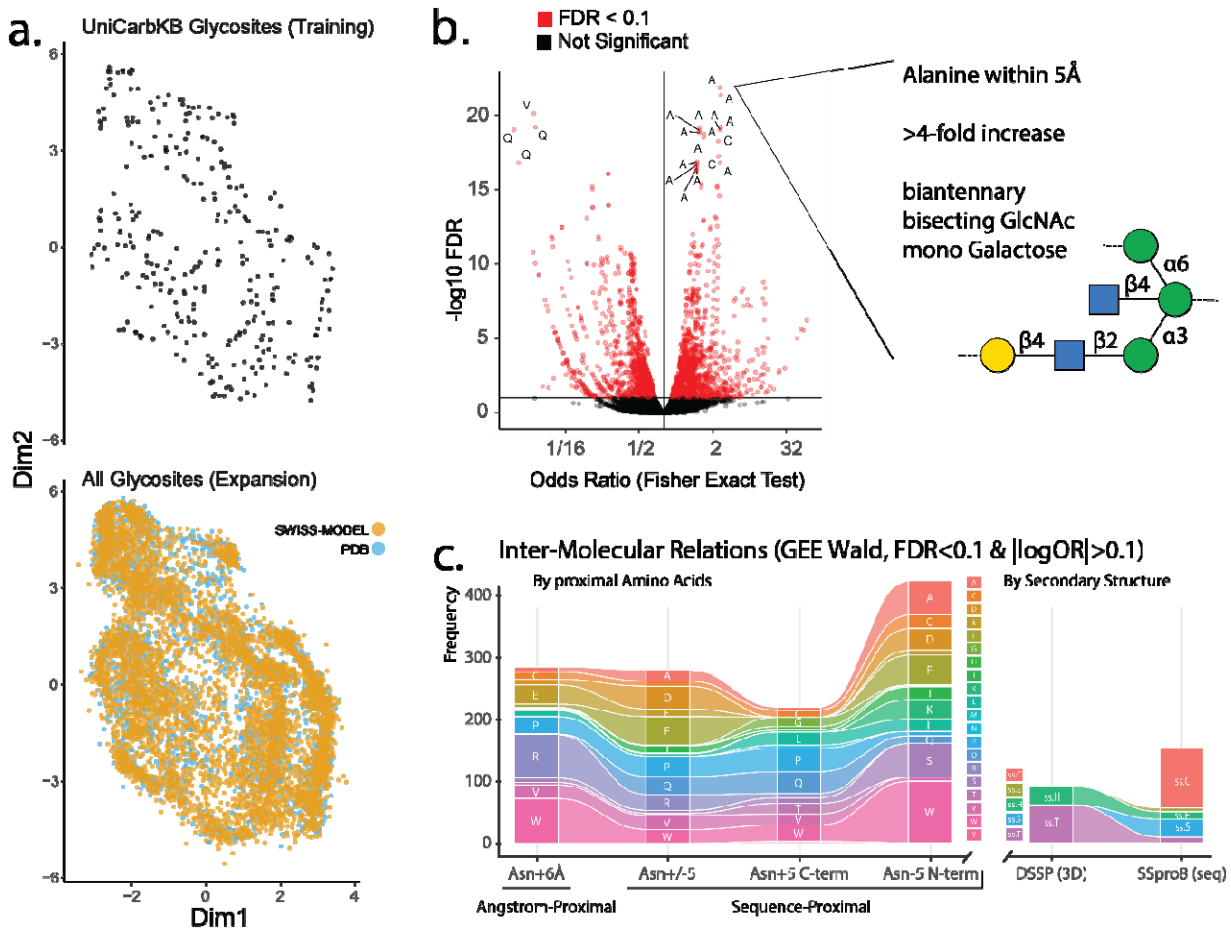


Figure 2 **– N- and O-glycan substructures associated with glycosite-proximal protein structure. (a)** Dimensionality reduction trained on UniCarbKB glycosite structures. When projected into the UniCarbKB-trained space, all human Uniprot annotated glycosites in appear within original density implying that UniCarbKB glycosites are sufficiently representative of most documented glycosites. The scatterplots show a two-dimensional projection of the FAMD using a Uniform Manifold Approximation and Projection (UMAP).[37] Each point is a site on a glycoprotein, each color indicates the source of that protein. (**b**) Volcano plot of the log odds ratio and False Discovery Rate adjusted p-values from a Fisher exact test between glycan and protein structure occurrence. We observe many large and significant Intermolecular Relations (IMR), with most significant relations associated with alanine, cysteine, valine, and glutamine (A, C, V, Q, respectively). (**c**) Inter-Molecular Relations (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).

We first tested if glycan structure correlates with protein structure. To do this, we built the Protein-Glycan Enriched Structure Database (PGES-DB), a compendium of experimentally measured and expert curated site-specific glycoprofiles on human glycoproteins collected from UniCarbKB[38] and GlyConnect[34] databases (**see Methods**). Glycan structures were decomposed into substructures (i.e., biosynthetic

125 intermediates) using GlyCompare. This maximizes statistical power by making explicit every permissible
126 precursor molecule at each site.[35] Glycosite-proximal protein structure features were extracted from
127 protein structure models, and proximity was defined by minimum distance—the minimum 3D Euclidean
128 distance between the nearest atoms in each residue. The resulting PDES database includes 98
129 glycoproteins with N-glycosylation sites and 38 glycoproteins with O-glycosylation sites and a total of
130 3,563 N-glycosylation and 700 O-glycosylation events (Supplementary **Figure** 1). In this work, we limited
131 training to human glycoproteins and limited predictions to glycoproteins expressed in humans.

132 We verified that the glycan-annotated glycosites in PGES-DB are representative of all glycosite
133 structures in the human secretome;[39] all secretome glycosites fall within a dimensionality reduction
134 trained only on PGES-DB glycosite structures (**Figure** 2a, see **Supplementary Results**, Supplementary
135 **Figure** 2). Thus, PGES-DB can provide generalizable conclusions about protein-glycan associations.
136 Furthermore, we examined the associations (termed "intermolecular relations (**IMR**)") between glycan
137 substructures (e.g., tri-mannose) and glycosite-proximal protein features (e.g., proximal tyrosine) within
138 PGES-DB, and we found that glycan substructures and protein features were significant and substantial
139 (Supplementary **Figure** 3a). Fisher exact test found 10,111 substantial and significant IMR with 9,296
140 positive and 815 negative associations (**Figure** 2b). Using multiple statistical tools (including Fisher's
141 Exact Test, conditional probability, and information gain, see **Supplementary Results**), we confirmed
142 that PGES-DB contains many notable IMR.

143 Highly deterministic IMR are consistent with canonical templating

144 We found several highly deterministic amino acid (AA) specific intermolecular relations (IMR) ((Pr(G|P)
145 is 0.999-1 or 0-0.001). In these high-confidence IMR (hcIMR), the impact of precursor-limitation is
146 minimal, so the protein-imposed "requests" are more easily "fulfilled" and observed. Overall, confidence
147 in glycan presence increases when a specific protein structure feature is present (Supplementary **Figure**
148 4). For IMR with a spatially proximal AA (within 6A), 20.2% are highly deterministic of glycan
149 substructures. Additionally, 32.4% and 17.5% of down- and upstream AA IMR (+/-5aa) are highly
150 deterministic. The certainty with which an AA determines a glycan structure decreases substantially
151 when proximal AA are absent (Supplementary **Figure** 4**a**). These hcIMR are not dominated by small
152 numbers of motifs as the high-confidence IMR count is proportional to the number of unique
153 substructures (Supplementary **Figure** 4**b**). The decrease in hcIMR when a protein structure feature is
154 absent suggests bounded glycan biosynthesis is an active rather than passive process.

155 Among the highly deterministic protein-glycan relations, we observe (among 1,725 N-glycosylation
156 events) 1,553 glycans that contain a N-acetylglucosamine (GlcNAc) on the α-1,6-mannose branch
157 (Glc2NAc(β1-6)Man(α1-6)Man(β1-4)Glc2NAc); indicative of complex N-glycosylation. All 75 hcIMR with a
158 downstream tryptophan include glycans with the hybrid/complex substructure. These observations
159 suggest that tryptophan is sufficient to escape oligomannose structures. Similarly, in 454 O-glycosylation
160 hcIMR, 237 contain 6-sialyl-N-acetyllactosamine (Neu5Ac(α2-6)Gal(β1-3)GalNAc). Of 6 events
161 containing a sequence-proximal tryptophan, every event also contained a sialyl-T antigen. Thus, several
162 IMR link protein features, such as proximal tryptophan, with specific glycan features.

163 ## Glycosite-proximal structure and chemistry correlate with glycan structure

164 The numerous glycan-protein relations suggest the plausibility of bounded glycan biosynthesis. To
165 generate testable protein structural constraints on bounded glycan biosynthesis, we modeled IMR using
166 hierarchical regression to minimize relevant bias. We quantified specific IMR using univariate logistic
167 generalized estimation equations (GEE) to probe the site-matched glycan-protein co-occurrences in
168 PGES-DB and to control for protein identity effects (see **Methods**). A list of GEE odds ratios (OR) (each
169 describing the association between a protein structure and one of several glycan substructures)
170 describes the typical glycosylation near a given protein structure; the OR list indicates the "expected
171 substructure abundance" or an "expected glycoprofile" related to a glycosite-proximal protein structure.
172 When expected glycoprofiles are compared between two protein structures (e.g., the expected
173 glycoprofile difference between valine and isoleucine), we can estimate a glycosylation impact, or
174 "glycoimpact," of making amino acid substitutions or other protein feature changes. Glycoimpact is the
175 difference (z-score normalized Euclidean distance) in expected glycoprofiles between two protein
176 structures (**Figure** 3, see **Methods**).

177 We discovered 1,715 significant N-glycan IMR (FDR<0.1, |log(OR)|>0.1), many of which were associated
178 with spatially proximal (N+6Å) and sequence-proximal AA (N+/-5 residues). Stratifying sequence-
179 proximal effects, we observed almost twice as many IMR involving upstream (N-5) than downstream
180 (N+5) AA. Among the downstream AA effects, tryptophan, alanine, serine, and phenylalanine are most
181 impactful (99, 55, 55, and 48 IMR respectively). Tryptophan also has many IMR when downstream or
182 spatially proximal (26). Spatially proximal arginine (70) and downstream glutamine (35) are the largest
183 effectors. Finally, glycosylation sites on turns have many IMR (61) (**Figure** 2c).

184 Turn-associated IMR include >3-fold increases in di- and tri-sialylated tetra-antennary and >2-fold
185 increases in mono- and di-galactosylated structures with core fucose; all positively correlated structures

186 have at least one galactose while not all are core fucosylated. Structurally proximal glutamine is
187 associated with a >20-fold increase in monosialylated triantennary structures and a 10-fold decrease in
188 tetraantennary structures (**Figure** 3**a**). Histidine, threonine, and valine show correlation with increasing
189 GalNAc[4S] (**Figure** 3**b**). Expanding on **Figure** 3a-b, we further interrogated IMR for specific protein
190 structure features (i.e., proximal amino acid or local secondary structure) by comparing the IMR odds
191 ratio to monosaccharide count for several monosaccharides (**Figure** 3**c**). A biclustering highlights two
192 major types of protein structure influence, mirroring the N-glycan/O-glycan dichotomy in glycosylation.
193 Providing clues to the elusive O-glycosylation site, proximal alanine is negatively correlated with
194 galactose and GlcNAc but positively correlated with N-acetylgalactosamine (GalNAc). Conversely,
195 threonine and histidine are positively associated with GlcNAc and galactose but negatively correlated
196 with β-GalNAc. As expected, GlcNAc and complex-glycans follow similar trends to Neu5Ac. However,
197 Neu5Ac, GlcNAc and Gal trends diverge near proline, cysteine, and valine; these amino acids may act as
198 limiters of high-complexity (**Figure** 3**c**).

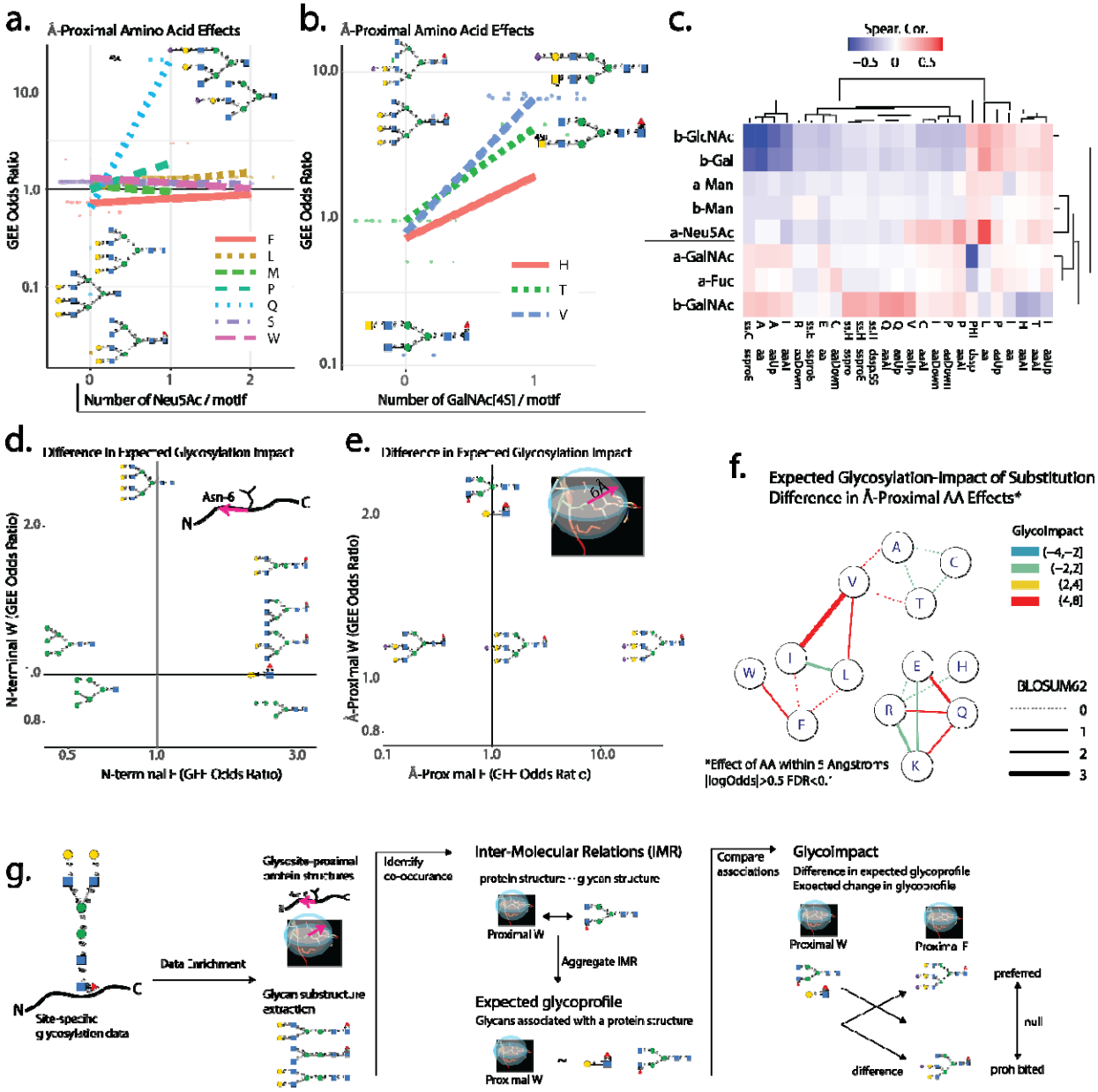


Figure 3 **– Glycosite-proximal amino acids impact glycosylation.** Specific Inter-Molecular Relations (IMR) were discovered by logistic GEE controlling for protein-bias. (**a-b**) IMR (FDR<0.1, |log(OR)|>0.1) relating structurally proximal amino acids to motifs stratified by the number of Sialic Acids (**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 (Blosum64) 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. (**g**) Overview of data flow and concepts examined in this manuscript.

213

214 Amino acid changes can predict glycan structural changes

215 We found that many AA-substitutions can have substantial "glycoimpact" (expected difference in
216 glycosylation) on glycan biosynthesis; exemplified by the phenylalanine-tryptophan substitution (**Figure**
217 3**d-e, Supplementary Results**). Upstream phenylalanine is associated with bi- and tri-antennary (>3-fold)
218 while upstream tryptophan is associated with tetraantennary terminal galactose (>2-fold) implicating
219 this substitution in branching (**Figure** 3**d**). Additionally, spatially proximal phenylalanine is strongly
220 associated with increased sialylation (>10-fold) implicating this amino acid in glycan maturation (**Figure**
221 3**e**). These predictions suggest that tryptophan-phenylalanine substitution has a large impact on
222 glycoprofiles.

223 We calculated all AA-substitution glycoimpact scores as the difference in expected glycoprofiles
224 between each AA-pair (see **Methods**, Supplementary **Figure** 5). Glycoimpact was calculated at multiple
225 IMR thresholds (Supplementary **Figure** 5). We highlight relevant substitution events as "glycoimpactful"
226 (significantly high glycoimpact) and structurally ambivalent (BLOSUM>0) (**Figure** 3**f**). As explained in the
227 next section, we call the glycoimpact AA-substitution matrix the BLOSUM-PAM Orthology matrix
228 (BLAMO X:Y); X and Y refer to the log odds ratio and FDR thresholds respectively. Sub-threshold odds
229 ratios, those insignificant or unsubstantial by the X:Y threshold, are excluded from the glycoimpact
230 calculation. **Figure** 3f displays a subset of BLAMO 0.5:0.1 relations.

231 Glycoimpact explains discrepancies between BLOSUM and PAM substitution matrixes

232 To further explore the relevance of glycoimpact, we compared it to conservation-based measures of
233 amino acid substitution impact. The PAM[41] and BLOSUM[42] AA substitution matrices are popular but
234 distinct estimates. PAM is based on global alignments within a protein and reflects more functional
235 conservation, while BLOSUM uses local alignment across proteins and highlights more structural
236 conservation.[43,44] Thus, these measures differ more when protein function is not fully described by
237 protein structure.

238 Since changes in glycan structure can modify protein function in a protein structure-independent
239 manner, we examined consistency between PAM and BLOSUM estimates across null and high
240 glycoimpact substitutions. Comparing PAM and BLOSUM scores at multiple thresholds
241 (RMSE($PAM_{i,j}$,$BLOSUM_{i,j}$)), we found that error in 4 of 5 PAM-BLOSUM comparisons was significantly
242 correlated to glycoimpact (BLAMO 0.5:0.1) for high-glycoimpact (GI>2.5) substitutions (**Figure 4a,**

243 Supplementary **Figure** 5). The correlation between high-glycoimpact substitutions and PAM-BLOSUM
244 inconsistency is maintained for most PAM and BLOSUM thresholds (Supplementary **Figure** 6). Both
245 protein structure and glycosylation are necessary to fully explain protein function, and these results
246 suggest a positive relationship between glycoimpact and the failure of structure (BLOSUM) to
247 completely explain function (PAM). Given this relationship, we refer to the glycoimpact substitution
248 matrix as the BLOSUM-PAM Orthology (BLAMO) matrix.

249 High glycoimpact residues are conserved around glycosites

250 If glycosite-proximal amino acids influence functional glycosylation, there should be evolutionary
251 pressure imposed beyond the classical NX[S/T] N-glycan sequon. To map the broader glycosite structure,
252 we aligned glycosites (**Figure 4b-c**) and examine conservation and evolutionary coupling (EC). Using EC
253 derived from 2,005 glycoprotein alignments (see **Methods**, **Supplementary Results**),[45] we observed
254 enriched coupling between N-glycosites and flanking residues (Supplementary **Figure** 7a). We also
255 observed several position-specific glycosite-coupled residues including serine and threonine at N+2,
256 phenylalanine at N-2, and tyrosine at N-1 (pooled hypergeometric, FDR<0.1, Supplementary **Figure** 7c,
257 **Figure 4c**). These findings are consistent with previous observations of the original sequon and
258 enhanced aromatic sequon.[14] Glycosite-coupled residues are also consistent with IMR-observed high-
259 impact residues (**Figure** 2**c**). Of the ten high-impact upstream residues, seven show enriched glycosite-
260 coupling when they appear upstream. To further highlight global glycosite structure, we clustered
261 glycosites using coupling probabilities (**Figure 4d,** Supplementary **Figure** 8, see **Methods**). The N+2
262 aspartic acid enriched in the univariate analysis (**Figure 4c**) co-occurs with an N-2 lysine (**Figure 4d,** motif
263 1). Alternatively, glutamic acid is more likely to co-occur with other glutamic acid residues (N -4, +1, and
264 +3) with an N+2 threonine sequon (**Figure 4d,** motif 4). These couplings are reflective of evolutionary
265 pressures surrounding the glycosylation sites.

266 We next aligned[12] glycosites permitting a tetraantennary N-glycan lacking fucose or sialic acids (**Figure**
267 **4b**). We examined the glycosite alignment for consistency with high-influence AA (**Figure 2c**) and
268 glycosite-coupled residues (**Figure 4c**). Of 20 glycosite-flanking AA, 16 show consistency between the
269 first or second most common AA and either the high-influence or glycosite coupled residues (see
270 **Supplementary Results**). At nearly every glycosite-flanking residue (N+/-10) there is consistency
271 between these three analyses, further corroborating that protein structure constrains glycosylation.

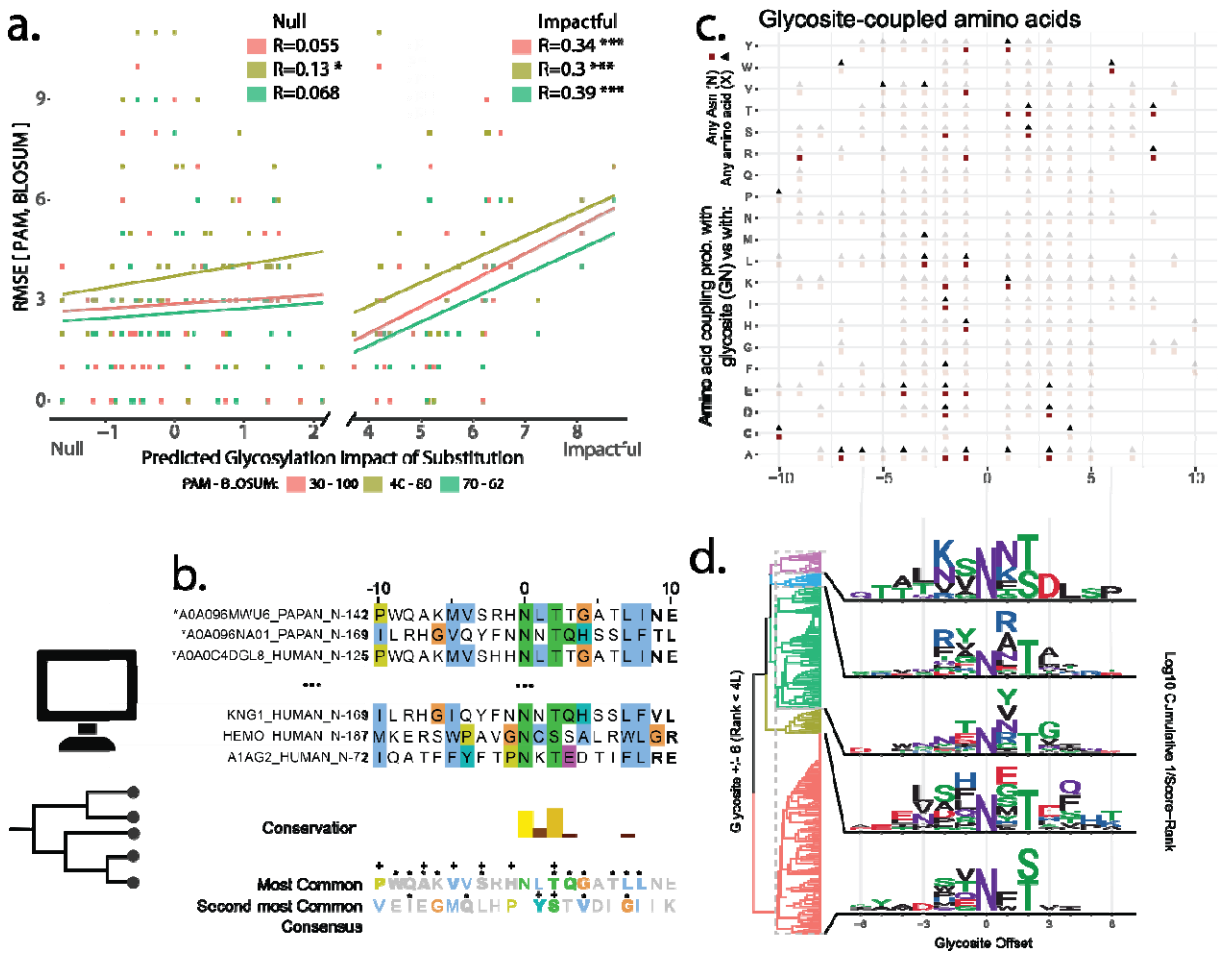


Figure 4 **- Glycoimpact of amino acid substitution correlate with evolution metrics and conservation.** (**a**) Comparison of the error between the PAM and BLOSUM substitution matrices and the glycoimpact for corresponding substitutions. Linear regressions are split into null glycoimpact (<2.5) and 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. Correlation (R) significance indicated as <0.001 (***), <0.01 (**), and <0.05 (*). (**b**) Glycosite Alignments [12] 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***).

293 ## Glycoimpact correlates with discrepancy between functional variant predictions

294 Dozens of algorithms predict functional and pathogenic effects of genetic variants,[40,46,47] incorporating
295 information ranging from sequences to protein structure, thus sometimes reporting different variants as
296 deleterious. We hypothesized the differences in some pathogenicity scores between algorithms could
297 be explained by glycoimpact (BLAMO 0.5:0.1, see **Methods**). Across 3,549,910 nonsynonymous
298 mutations, we measured the disagreement (RMSE) between each of 27 rank-normalized functional
299 impact prediction tools, precomputed with dbNSFP;[40] pathogenicity score divergence was then
300 correlated with glycoimpact. After hierarchical clustering on the divergence-glycoimpact correlation
301 coefficients, pathogenicity estimates separated into two major clusters: one containing nearly all (6/7)
302 tools leveraging protein-structure and the other primarily containing conservation, sequence, and/or
303 epigenetic-based tools (**Figure** 5**b**). Nearly all variant pathogenicity score differences across the two
304 clusters correlated with glycoimpact. These correlations and clustering structure disappear when
305 glycoimpact scores are shuffled (Supplementary **Figure** 10); thus, like BLOSUM-PAM discrepancies,
306 glycoimpact correlates with discrepancies between conservation-based and protein structure-based
307 pathogenicity estimates. These observations further implicate glycosylation as a potent functional
308 regulator and glycoimpact as an appropriate proxy for the importance of glycosylation.

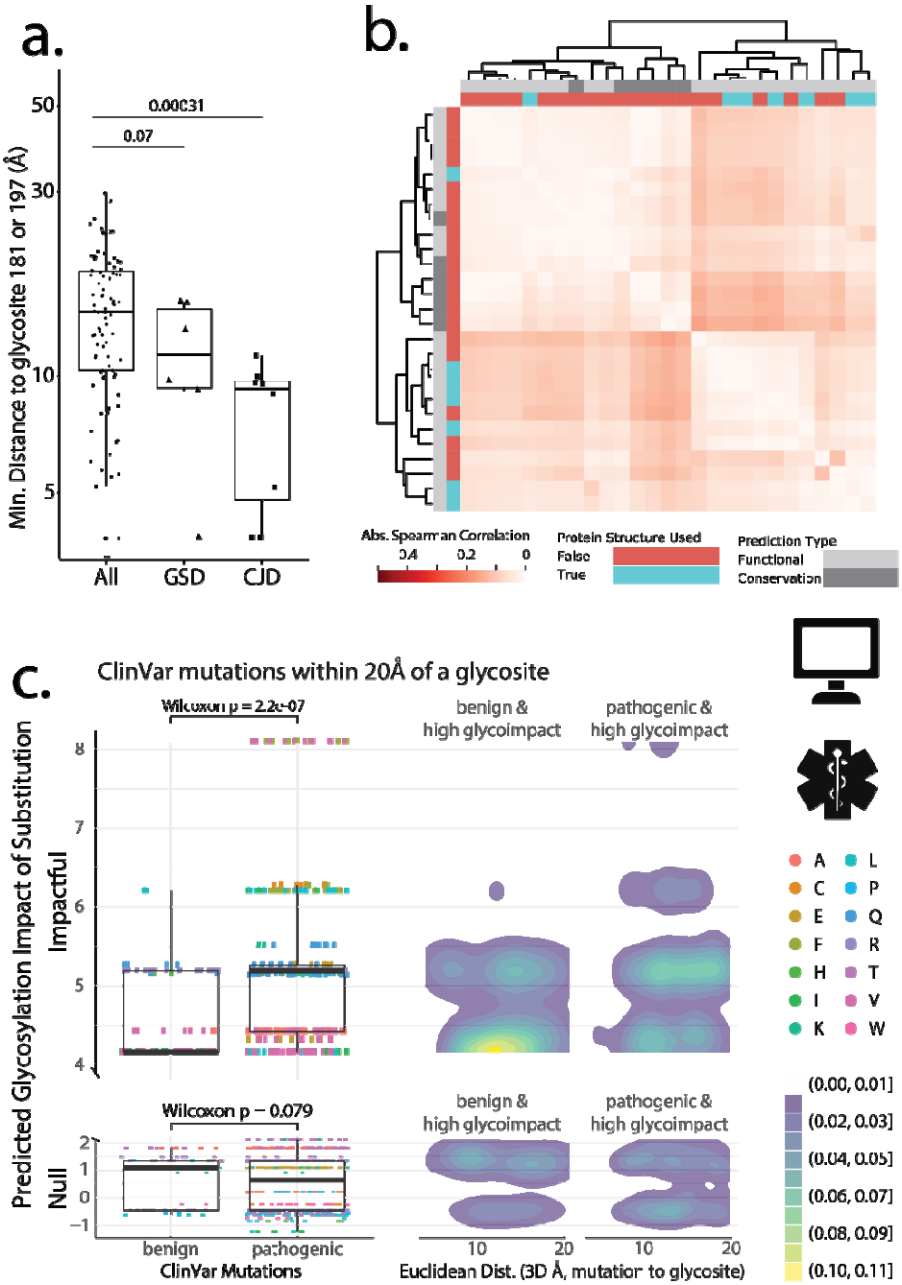


Figure 5 **- Substitution glycoimpact predicts pathogenicity.** (**a**) Boxplots show the min-distance from all residues within human PrP to the N197 or N181 glycosylation sites. Residues are stratified by all sites (All) and causative mutations of prion disease including Creutzfeldt-Jakob disease (CJD) and Gerstmann-Straussler disease (GSS), (one-sided Wilcoxon test). For glycositesite-specific proximity, see ***Supplementary Figure 11***. (**b**) A hierarchical clustered heatmap (average-linkage with Euclidean distance) of Spearman correlations 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 [40]. 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.

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

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

328 Examining specific ClinVar-annotated variants, we identified multiple variants in glycosylation-related
329 diseases For example, Tyrosinase:A355V (P14679) is a high-glycoimpact and glycosite-proximal causal
330 variant in oculocutaneous albinism.[48,49] While tyrosinase:A355V has not been examined for aberrant
331 glycosylation, deglycosylation disrupts tyrosinase function consistent with type 1 oculocutaneous
332 albinism.[49–53] Therefore, tyrosinase:A355V may act through aberrant glycosylation. We also observe
333 high-glycoimpact, glycosite-proximal (<30Å), pathogenic (ClinVar) variants in multiple other glycan-
334 modulated diseases including prion diseases,[54,55] lysosomal storage disorders,[56,57] and Gaucher's
335 disease.[58–61]

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

343 We further analyzed glycosite proximity among causal variants in prion disease. We measured 3D
344 Euclidean min-distance from the two PrP glycosylation sites, N181 and N197, to all residues in human
345 prion protein (PrP) (including variants causing Creutzfeldt-Jakob disease (CJD)[62,63] and Gerstmann-
346 Sträussler-Scheinker disease (GSS))[64,65] (**Figure** 5a**,** Supplementary **Table** 2). CJD-causing variants were
347 approximately twice as close to glycosylation sites than the background distribution of all PrP sites (One-
348 sided Wilcoxon p=0.0003). GSS-causative variants also trend closer to glycosites (One-sided Wilcoxon
349 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 Deafnes (*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 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**).[66] PGES-DB-measured GEE-estimated IMR suggest that downstream glutamine was significantly (FDR<1e-8; OR<0.5) predictive of complexity while spatially proximal proline and lysine were weaker but significant distinguishers (FDR<1e-3, FDR<0.1 respectively, **Figure** 6**a**). As predicted, proline-proximal (within 6Å, min-distance) gp160 glycosites present more oligomannose (Two-sided Wilcoxon p=0.0033), whereas C-terminus-proximal glutamine glycosites have more complex glycans (Two-sided Wilcoxon p=1e-4, **Figure** 6**c**). Spatially proximal lysine, while less significant (**Figure** 6**c**), had a nonlinear impact on glycan complexity in HIV gp160; first increasing with one proximal lysine then decreasing with two. The two most significant IMR predicted from PGES-DB (spatially proximal proline and C-terminal glutamine) 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[67,68] have been implicated with stability, target engagement, furin cleavage, and immune evasion.[68–72] 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,[73] to determine glycan state and occupancy at each glycosylation site. We performed two independent technical replicate analyses of the S1 subunit

383 comparing Ancestral to the Gamma and Delta variants and examined 11 of the 12 canonical S1
384 glycosites (N17 was excluded due to variable signal peptide trimming). Site-specific unoccupied,
385 complex, and oligomannose/hybrid proportions were compared between variants using a Mann-
386 Whitney U test.[73] P-values were pooled across the two independent replicate analyses using the Fisher
387 method and FDR-corrected. We observed three significant differential glycosylation events (**Figure** 6**d**).
388 Complex glycans observed at N122 in Ancestral S1 were converted to more oligomannose/hybrid in
389 both Delta (oligomannose/hybrid observations increased nearly 4-fold from 13.9% in S1 to 52.5%;
390 FDR=3.3e-9) and Gamma (oligomannose/hybrid observations nearly doubled to 27.6%; FDR=7.9e-4)
391 variants. Complex glycans at N331 increased marginally from Ancestral S1 in Delta variant (from 93.7%
392 to 99.7%; FDR=0.031). Complex glycans seen at N657 in Ancestral S1 decreased in Gamma variant (by
393 over 2-fold from 53.3% to 21.1%; FDR=1.27e3). The Gamma S1 monomer was consistently expressed
394 with two novel complex glycosylation sites at N20 and N188.

395 Based on proximal high-glycoimpact substitutions, we predicted changes at N17, N74 (Gamma only),
396 N122, and N165 (Delta only). Two of four (N122 in Gamma and Delta) predicted differential
397 glycosylation events were consistent with the four observed changes (sensitivity=0.5), while 15 sites
398 where no change was predicted were consistent with the 17 sites where no change was observed
399 (specificity=0.88). This proof of concept, wherein we correctly predicted the most substantial differential
400 glycosylation event at N122 in the S1 monomer, will motivate further inspection of differential
401 glycosylation on the more physiological spike trimer and whole virus, to elucidate how differential
402 glycosylation may participate in immune evasion by SARS-CoV-2 and other viruses.

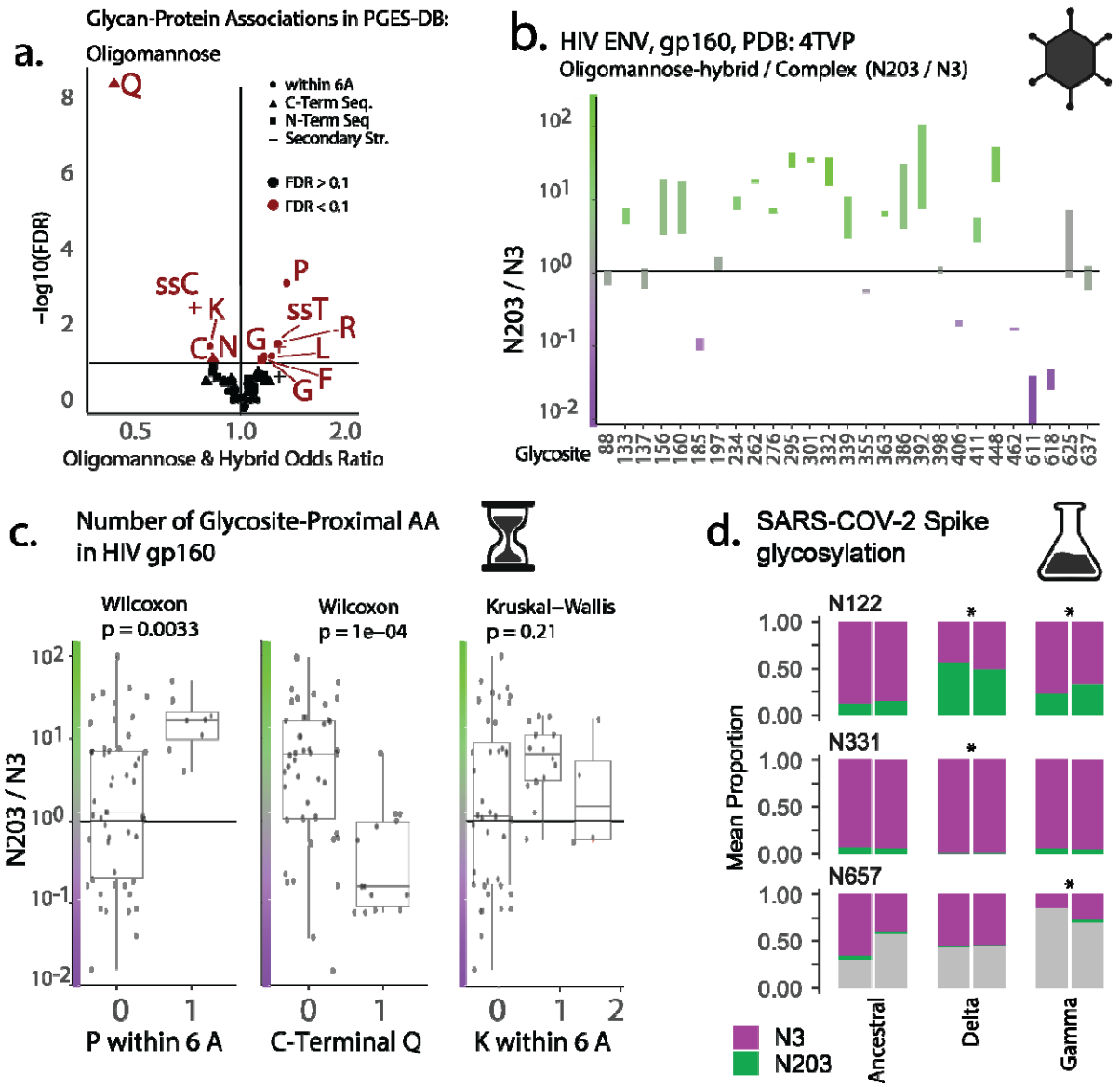


Figure 6 – Changes in high-mannose, hybrid, and complex glycosylation are predicted by glycosite-proximal high-glycoimpact substitutions in HIV, SARS-CoV2. (**a**) GEE-calculated IMR from PGES-DB denoting relations between sequence (triangle & square) or structural protein features (circle & plus) and 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).[66] Relative abundance is represented as a ratio of oligomannose-hybrid to complex glycan peaks (N203/N3). (**c**) Distributions of oligomannose-hybrid to complexity (N203/N3, panel **b**) glycan peaks stratified by proximal protein structure features selected in panel **a**. For the N203/N3 distribution of select IMR, see *Supplementary **Figure** 12*. (**d**) Mean proportion of mass-spectrometry-observed peptide with mass offsets corresponding to complex glycan peaks (+3, purple), oligomannose or hybrid glycan peaks (+203, green), or no glycosylation (+0, grey) in the SARS-CoV-2 S1 subunit at 3 sites with significant glycosylation differences between the original ancestral strain, and the Delta and Gamma variants (see *Supplementary **Figure** 13* for all sites). Peak count for each glycosylation offset type (+3, +203, or +0) for each glycosylation site is divided by the total number of peaks for that site to determine the site-specific proportion of each glycosylation type. Significant differential glycosylation (FDR<0.05) in VOC compared to ancestral is indicated by "*".

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

419 We next predicted differential glycosylation across *ighg1* missense mutation (F299I) in the IgG1 heavy
420 chain.[75] IgG1 glycosylation impacts both adaptive humoral response[76–79] and monoclonal antibody (mAb)
421 response.[80–83] Experimentally, the IgG1:F299I substitution shows a strong glycoimpact. The IgG1 variant
422 expressed in C57BL/6 and BALB/c mouse strains shows less sialylation and digalactosylation compared
423 to similarly expressed *wt* IgG1.[84–86] Additionally, BALB/c-expressed IgG1:F299I glycosylation is more
424 similar to C57BL/6-expressed IgG1:F299I glycosylation than to glycans on *wt* IgG1 expressed in the same
425 BALB/c animals (**Figure** 7**b-c**). Fc N-glycans on IgG1:F299I expressed in both BALB/c and C57BL/6 animals
426 have more agalactosylation (Mann-Whitney p=1.02e-6), less digalactosylation, and less mono-, di- and
427 total sialylation (Mann-Whitney p<0.0073) compared to *wt* IgG1 expressed in the same animals (**Figure**
428 7**c**, Supplementary **Table** 3). The increase in galactosylation in *wt* IgG1:F299 is consistent with PGES-DB
429 predicted IMR for upstream (N-terminal) phenylalanine (**Figure** 7**a**). Upstream phenylalanine is
430 associated with increased di-galactosylated biantennary structures (OR>2), while upstream isoleucine is
431 associated with tetraantennary galactosylation. Since only biantennary structures are generally
432 permitted on IgG, the galactose-promoting function of upstream isoleucine should be unrealized in this
433 glycoprotein. The increased sialylation in *wt* IgG1:F299 is also consistent with PGES-DB IMR which show
434 an association between structurally proximal phenylalanine and disialylated structures (OR>10). These
435 results suggest that glycoimpact can accurately predict several specific glycan epitopes.

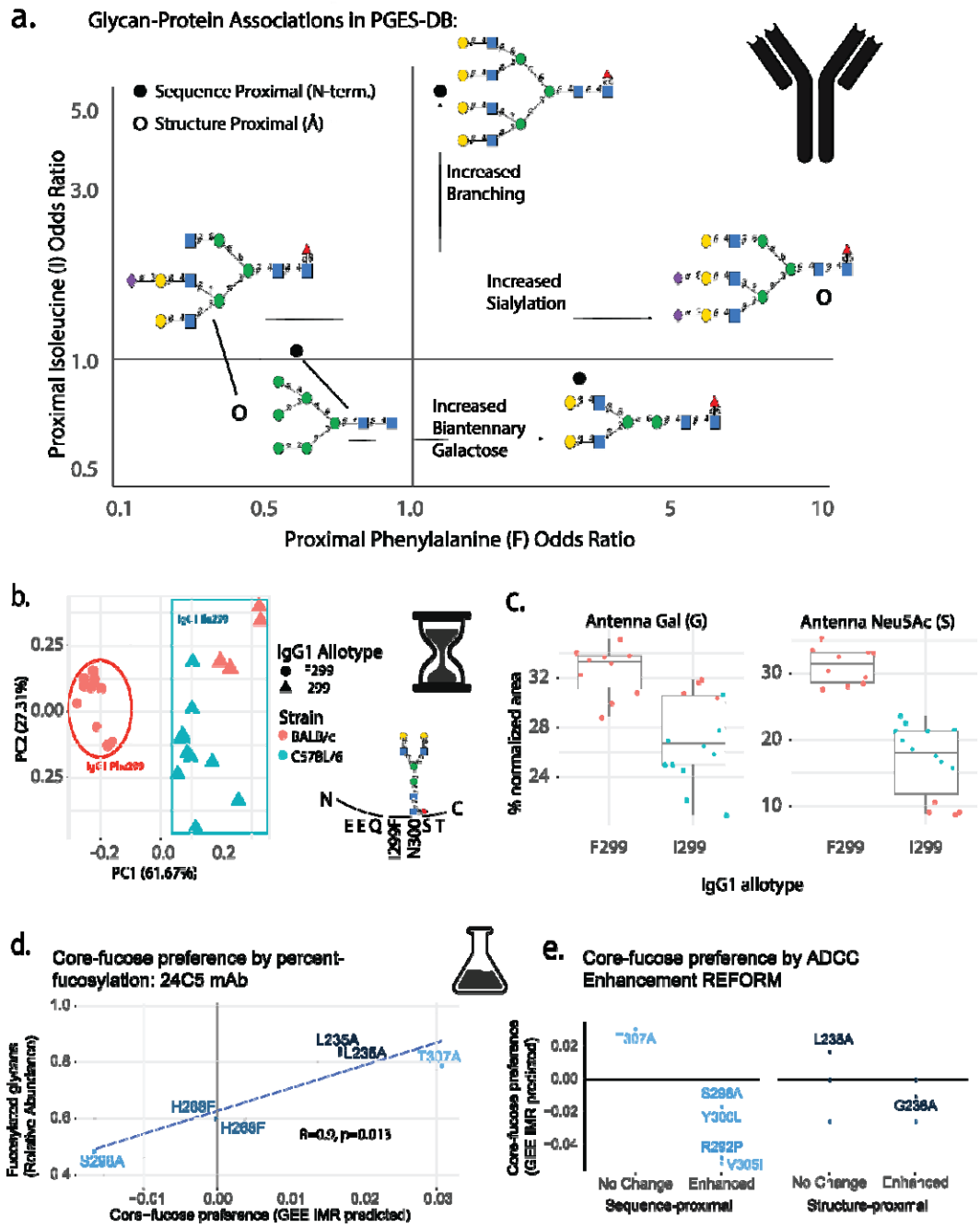


Figure 7 –Differential glycosylation is predicted by high-glycoimpact events near glycosylation sites in IgG (**a**) GEE-learned IMR relating to sequence-proximal (upstream/N-terminal) effects of I and F. IgG allotypes, F299 and I299, segregated by Principal Component Analysis of relative abundance (**b**) separate by allotype, not strain. (**c**) Galactose (Gal), Sialylation (Neu5Ac) and Bisecting GlcNAc abundance distributions for IgG1 F299 and I299 allotypes across BALB/c and C57BL/6 mice (see *Supplementary Figure 14*). (**d-e**) Core-fucose preference (see **Methods**, *Supplementary Figure 15*) for each glycosite-proximal variant in the REFORM Fc variant panel.[74] Variants are colored by index. (**d**) Variants are stratified continuously (top) by percent of fucosylated Fc-cleaved glycans observed on the 24C5 monoclonal antibody. (**e**) Variants are stratified categorically (bottom) by association with enhanced ADCC.[74] Variants appearing in Fc regions associated with both no-enhancement and enhanced ADCC are shown as transparent. Only L235A and G236A are uniquely associated with ADCC.

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

448 To further demonstrate IMR accuracy and functional importance, we compared predicted core-fucose
449 preference to observed fucosylation abundance and ADCC in a Fab-constant Fc-variant antibody panel[74]
450 (**Figure** 7**d-e**). Core-fucose preference was calculated as the preference of variant compared to the wild
451 type for N-glycan core motif with or without a core fucose, i.e., Man(b1-4)GlcNAc(b1-4)[Fuc(a1-
452 6)]GlcNAc(b1-4)-Asn and Man(b1-4)GlcNAc(b1-4)GlcNAc(b1-4)-Asn, respectively (see **Methods**).

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

463 To determine if predicted changes in glycosylation can also predict glycan-modulated behaviors, we
464 examined our ability to predict increases in antibody-dependent cellular cytotoxicity (ADCC) from
465 predicted decreases in core-fucosylation, a well-characterized property of therapeutic antibodies.[83] We
466 identified 10 glycosite-proximal variants (**Figure** 7**e**) from the REFORM Fc variant panel[74] and stratified
467 the ADCC-enhancing variants from those associated with no change in ADCC. Of the five sequence-
468 proximal variants, the non-enhancing variant (T307A) shows a positive preference for core-fucose while
469 all four ADCC associated variants have a negative core-fucose preference. Among spatially proximal
470 variants, several substitutions occur in multiple variants and are not uniquely associated with ADCC
471 enhancement or non-enhancement (greyed out, **Figure** 7**e**). Of the two-remaining spatially proximal
472 variants, the non-enhancing variant (L235A) shows a positive preference for core-fucose while the ADCC
473 associated variant (G236A) has a negative core-fucose preference. These ADCC associations with core-
474 fucose preference are consistent with percent fucosylation in 24c5 Fc. T307A and L235A are highly
475 fucosylated in 24c5, predicted to prefer core-fucose, and do not enhance ADCC. Conversely, S298A is
476 least fucosylated in 24c5, shows a negative preference for core-fucose, and is associated with ADCC

477 enhancement. Our predictions therefore recapitulate the immunological implications of allotype-driven
478 differential glycosylation.

479 Discussion

480 Here, we have identified constraints on glycan biosynthesis through the analysis of site-specific protein
481 features[36] and glycan substructures.[35] With our Protein-Glycan Enriched Structure Database (PGES-DB),
482 we enumerated and quantified intermolecular relations (IMR) between protein and glycan structure in
483 human glycoproteins. We then computed the expected differential glycosylation, the "glycoimpact,"
484 associated with specific protein structure changes. We validated glycoimpact importance by comparison
485 to substitution matrices, evolutionary couplings, and pathogenicity scores. We further examined
486 glycoimpact sensitivity and specificity comparing predictions with observed glycosylation on PrP, HIV
487 gp160, and IgG glycoproteins. The existence and utility of glycoimpact suggests that glycans are not
488 synthesized template-free. Rather, the glycoprotein surface constrains glycan biosynthesis, deciding
489 which glycosyltransferases may dock and continue glycosylation. Thus, glycosylation is a bounded
490 biosynthesis wherein a range of glycans within biosynthetic pathways are requested through local
491 glycosite structural constraints (likely dictated by glycoprotein-glycosyltransferase docking feasibility)
492 and those requests are variably fulfilled given metabolic precursor-availability.

493 The significant associations between glycan and protein structures seen in PGES-DB is not consistent
494 with the expected independence if glycan biosynthesis was template-free. Instead, it suggests protein
495 structure provides substantial information regarding glycan structure. However, the observation that
496 larger glycans are less determined by protein structure supports precursor-limitations of bounded
497 biosynthesis, wherein larger glycans require more precursors, any of which can limit fulfillment.
498 Similarly, low fulfillment is seen in high-yield expression systems[87] necessitating, in some cases,
499 supplementation with maturing glycoenzymes.[88–90]

500 Our quantified protein-glycan IMR describe an expanded glycosite structure, beyond the original sequon
501 definition (NX[S/T]).[5] We discovered many new sequence-proximal AA IMR both upstream and
502 downstream of traditional glycosites. The enhanced aromatic sequon (EAS)[14] corroborates one such
503 sequon-expanding IMR. Upstream phenylalanine IMR predicts an increase in Man7 structures (i.e., N-
504 glycan with seven-mannoses) and a decrease in Man6 structures, suggesting an increase in larger high-
505 mannose structures (**Figure** 3**d**). The predicted difference in oligomannose is consistent with reported
506 EAS glycosylation; upstream phenylalanine (N-2) can decrease glycan processing and increase
507 homogeneity.[14] Structurally, tryptophan and tyrosine are known to stabilize the chitobiose core through

508 dispersion and increasing glycan accessibility for maturation; phenylalanine lacks a dipole and therefore
509 does not support similar maturation[91–95]. Both the EAS and selective aromatic stacking results are
510 consistent with glycoimpact. We predict limited expected differential glycosylation following a
511 tryptophan-tyrosine substitution but a substantial decrease in processing following substitution from
512 tryptophan or tyrosine to phenylalanine (**Figure** 3**f**, Supplementary **Figure** 5). The expanded sequon
513 scaffolded by the IMR (**Figure** 2), glycoimpact substitution matrix (**Figure** 3**f**, Supplementary **Figure** 5),
514 and glycosite-coupled residues (**Figure** 4**c-d**) together represent a portable and dynamic summary of our
515 findings that can be easily applied to predict glycosylation following novel substitutions. We believe that
516 the large number of both high glycoimpact and low structural impact substitutions (around 50% of AA
517 have one such substitution) will be useful in understanding pathogenic variants, viral evolution, and
518 glycoprotein engineering.

519 Glycoimpact can be further used to inform the mechanism of pathogenesis when disease variants are
520 close to glycosites that impact glycosylation. We can examine known annotated-pathogenic variants and
521 propose novel mechanisms of pathogenesis. Unannotated variants can be automatically labelled for
522 likelihood of glycan modulation. Though this work does not deeply interrogate causal associations with
523 glycan-modulated pathogenesis, we found several prior works implicating high-glycoimpact glycosite
524 proximal variants as pathogenic, including one thorough study of oculocutaneous albinism implicating
525 variant-proximal glycosylation as a causal pathogenic modulator. Examining the highest glycoimpact
526 relation with a high BLOSUM score, Ile/Val; the substitution should have a negligible structural impact
527 while dramatically changing glycosylation. As expected, pathogenic variant PrP:V180I (adjacent to
528 glycosite N181) is a causal mutation in Creutzfeldt-Jakob disease.[96] Similarly, a glycosite-proximal V84I
529 substitution in HIV-gp120 deactivates the virus; otherwise achieved by mutagenic glycosite-ablation.[97]
530 Similarly, tyrosinase:A355V (P14679) is a sufficient cause of oculocutaneous albinism.[48,49] A355V is also a
531 high glycoimpact variant and close to the function-critical N371 glycosite (<20Å; PDB:5M8N); tyrosinase
532 glycosylation is critical for proper folding,[51,98] N371 glycosite ablation results in decreased protein
533 abundance and activity,[99] and non-mutagenic post-expression tyrosinase deglycosylation also interrupts
534 function.[53] Using glycoimpact, we can propose A335V as a glycan-based mechanism of pathogenicity.
535 Beyond tyrosinase, we identified thousands of variants across hundreds of diseases that may be
536 similarly explained by aberrant glycosylation. There may be millions of high glycoimpact variants close to
537 glycosites beyond the narrow scope of ClinVar annotation for which glycoimpact can suggest a possible
538 pathogenic mechanism.

539 Just as amino acid substitution can assert pathogenic glycosylation on human proteins, similar
540 substitutions can enable immune evasion for viruses. Consistent with glycoimpact predictions, we show
541 that glycosite proximal proline and glutamine are associated with glycan complexity across HIV gp160.
542 These results both validate our model and propose mechanisms for evolutionary control over the viral
543 glycan shield.[11,25,100,101] Across three SARS-CoV-2 Variants of Concern, we show that high glycoimpact
544 variations close to glycosites produce IMR-consistent differential spike glycosylation. If IMR can predict
545 differential glycosylation as viruses evolve, these predictions could help predict immune evasion and
546 predict the viral evolutionary landscape for many viruses.[11,101,102]

547 Functional differential glycosylation is also critical in protein therapeutics. Glycoengineering for
548 therapeutics is often costly, application specific, and siloed from protein design. Glycans can be modified
549 indirectly through genome editing of glycosyltransferases[103–105] or media optimization,[106–108] and such
550 strategies can be guided by models of the biosynthetic pathways.[109–114] Here, we examined AA-specific
551 differential glycosylation events on antibodies. On the previously characterized and differentially
552 glycosylated mouse IgG1 allotype (I299F), we see IMR-predictable glycosylation encoded directly into
553 the glycoprotein primary sequence. Similarly, in glycosite-proximal and high glycoimpact mAb-Fc-
554 variants both fucosylation and fucose-modulated functional response (ADCC) correlated with GEE-IMR
555 core-fucose preference. In these examples, IMR is a biologically[76–78] and therapeutically[74,115–118] useful
556 method for connecting allotype, glycosylation, and effector function. Therefore, IMR represent an
557 opportunity to embed glycoengineering directly into the glycoprotein sequence and combine the
558 currently separate paradigms of glycoengineering and protein engineering to a unified practice of
559 glycoprotein engineering: considering the structure of both protein and glycan and their mutual
560 influence.

561 Our direct results focus on human-expressed glycoproteins and global trends rather than specific
562 mechanisms such as aromatic chemistry, or chaperone proteins. While the specific glycoimpact trends
563 may not extend to other organisms, the chemistry of carbohydrate-protein interactions suggest that
564 glycoimpact is foundational and therefore exists in some form across glycosylated proteins.

565 We believe the evidence presented demonstrates the utility of glycoimpact. Further, the predictability
566 of glycosylation from protein structure demonstrates that glycosylation is a genetically-encoded
567 bounded biosynthesis. Due to the high degree of accuracy, precision, and reproducibility with which
568 protein structure appears to bound glycosylation, we believe protein structure may be described as a
569 template for glycan biosynthesis. Predictability in glycobiology from such templates will unlock a wealth

570 of theoretical, exploratory, and corroborative analysis making it inexpensive and easy to leverage
571 glycobiological insights in fields throughout biology.

572

573 Conclusions

574 Our results are incompatible with the template-free principle of glycan biosynthesis. Glycoimpact,
575 defined and validated here, presents a metric for estimating the influence of a protein structure on
576 glycosylation. The evidence presented asserts that once the precursor-limitations of biosynthesis are
577 considered, protein structure influence on glycans becomes visible. Because precursor-limitations are a
578 substantial force in glycan biosynthesis, protein structure describes a less obvious mapping than
579 canonical templates such as DNA, RNA, or protein biosynthesis. It is nonetheless predictable, accurate,
580 and reproducible within a requested range of permissible glycans.

581 Notably, bounded biosynthesis, imposed through local protein structural constraints, is not a behavior
582 unique to glycans. Within biology there are likely other cases of biosynthesis or modification constrained
583 or directed by local protein structure. Because molecules exist in physical space and possess chemical
584 properties, both the target molecules and the modifying proteins may influence each other by
585 generalizable steric and chemical interactions between their surfaces. Other target-modifier relations
586 likely exist in similar interactions such as chromatin and epigenetic modification (e.g. neither histone
587 deacetylases nor chromatin demethylases uniformly impact accessible chromatin),[119] post-translational
588 modification,[120,121] and RNA-binding protein actuated splicing.[122–124] Any instance involving a target
589 molecule placed in distinct local protein structures modified by a small set of enzymes could be an
590 instance of bounded biosynthesis.

591

592 Acknowledgements

593 Thank you to Terry Platt who inspired me with tales of the *Trp* operon[125–129] and in doing so planted the
594 seed of thought that substrate-limitation is not mutually exclusive with templated biosynthesis. Thanks
595 also to Lorenzo Casalino, Christian Seitz, and Rommie Amaro for their early insights into the project in
596 the context of influenza glycoprotein structure. We would like to acknowledge the invaluable
597 contribution of Dr. Jasminka Krištić, Prof. Dr. Grant Morahan and Prof. Dr. Falk Nimmerjahn to the
598 studies of IgG N-glycome that are referred to in this paper. EA is supported by ANID (DOCTORADO
599 BECAS CHILE/2018 - 72190270), the Fulbright Chile Commission, and the Siebel Scholar Foundation. This

work was supported by NIGMS (R35 GM119850, NEL) and the Novo Nordisk Foundation (NNF20SA0066621, NEL).

Conflicts

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

Methods

Enrichment of glycan-protein site-matched data to generate the Protein-Glycan Enriched Structure Database (**PGES-DB**)

Starting from site-specific glycosylation events, we extended the annotation of each glycosylation site and glycan to include detailed site-specific protein structural annotation and recorded the number of times each substructure appeared in each glycan. Only human glycoproteins were analyzed. The final database includes 111 proteins, 306 glycosylation sites and 4,263 glycans. We initially used site-specific glycosylation events documented in UniCarbKB.[38] Later and current work was informed by glycosylation events documented in GlyConnect[130] with supplemental information from GlyGen.[131] Empirical and expert-curated site-specific glycosylation events from the UniCarbKB and GlyConnect were used to inform much of the core analysis. Glycoprofiles from UniCarbKB were collected from GlycosuiteDB,[132] EUROCarbDB,[133] and expanded meta-analysis data.[24] GlyConnect glycoprofiles were predominantly (70%) curated from novel experiments but includes GlycosuiteDB[132], and select large scale studies.

The protein structure annotation was done using the Structural Systems Biology (ssbio) package in python.[36] The package uses several tools to perform a variety of annotations. For each human protein, empirical and homology modeled structures were collected from the Protein Data Bank (PDB)[134] and SWISMOD,[135] respectively. Proteins without existing models were modelled using I-TASSER.[136] Protein structures and chemistry close to the glycosylation sites were annotated multiple software packages through ssbio: sequence properties (EMBOS:*pepstats*),[137] sequence alignment (EMBOS:*needle*),[137] secondary structure (DSSP,[138] SCRATCH::SSpro[139] and SCRATCH::SSpro8), solvent accessibility (DSSP and FreeSASA), and residue depth (MSMS). Additional amino acid aggregate features were calculated using

628 R::seqinr. Spatial proximity was defined using "min-distance" between two amino acids; the minimum
629 distance between any pair of atoms spanning the amino acids.

630 Glycan structures were annotated using a combination of glypy[140] and GlyCompare[35] for structure
631 parsing and comparison respectively. All glycan substructures, a connected subset of monosaccharides
632 with and without linkage information, were extracted from each glycan, merged to make a superset of
633 substructures, then mapped to each glycan. Thus, resulting in a mapping from every glycan in the input
634 database to shared substructures.

635

636 Software and packages

637 Protein structure analysis was performed in Python v2.7.15 using ssbio v0.9.9.8 to retrieve and
638 calculate: existing empirical and homology models from PDB and SWISSMOD (PDBe SIFTS),[141] *de novo*
639 homology models (I-TASSER v5.1), sequence properties (EMBOS v6.6.0.0 pepstats), sequence
640 alignment(EMBOS v6.6.0.0 needle), secondary structure (DSSP v3.0.0, SCRATCHv1.1::SSpro and
641 SCRATCHv1.1::SSpro8), solvent accessibility (DSSPv3.0.0 and FreeSASAv2.0.2), and residue depth
642 (MSMSv2.2.6.1). Additional amino acid aggregate features were calculated using R::seqinr.

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

648 Probability event space, information gain and conditional probability

649 Here we define an event (a row in our enriched glycosylation-glycosite database) as "the observation of
650 a glycan at a glycosylation site in an experiment." If two separate experiments in the input database
651 both observe the same glycan at the same site on the same protein, we consider that event to have
652 occurred twice. Within each event, we ask if the glycan structure random variable (the presence or
653 absence of a specific glycan substructure) is present or absent in the observed glycosylation event and if
654 the protein structure random variable (a proximal amino acid, a secondary structure or another discrete
655 protein structure). A Fisher exact test (R::base::fisher.test) was used to estimate the odds ratio (OR) and
656 significance (p) of each inter-molecular relation (IMR). P-values were corrected for False Discovery Rate
657 (FDR, q) permitting 10% false discovery ($q<0.1$). Conditional probability was calculated by dividing joint

658 probability by the marginal probability of protein and glycan structure presence. Kullback-Leibler
659 divergence (KLd, R::entropy::KL.Dirichlet, pseudo count=1/6) was calculated by comparing the
660 conditional probability distribution to the marginal probability distributions.

661 Quantitative characterization of Inter-Molecular Relations (IMR) using Generalized Estimation 662 Equations (GEE)

663 To characterize the IMR in the PGES-DB while controlling for protein-specific confounding effects and
664 handle nonlinear relations we used a population-averaging approach; logistic Generalized Estimating
665 Equations (GEE) with glycoprotein identity as the cluster identifier.[142] We used an exchangeable
666 correlation structure to describe and balance the in-protein similarity. Models were fit to predict glycan
667 substructure binary (presence or absence) from either z-score normalized continuous or binary
668 (presence or absence) protein structures. For each model, the data from PGES-DB was isolated for one
669 glycan-type (N-glycan or O-glycan), one glycan substructure and one protein structure. Incomplete
670 observations (events/rows) were removed and then several checks on each data-slice were run to
671 minimize overfitting. Glycan substructures were excluded from modelling if standard deviation was less
672 than 1e-6 or if there were fewer than 5 observations of the structure within the pertinent data-slice.
673 Discrete protein structure features were excluded if there were fewer than 4 observations within the
674 data-slice. Models were excluded if there were fewer than 4 instances in any cell (of the 2x2
675 absence/occurrence matrix) or if the chi-squared expected value of any cell was less than or equal to 5.
676 Observations were weighted by the reciprocal-count of the corresponding label type to balance label
677 contributions to the model and scaled by exponentiated cscore to maximize the contribution of high-
678 quality protein structure models ($w_i = 2^c / n$); c is the c-score given by I-TASSER and n is the number of
679 times a structure is present (1) or absent (0). Models with $|\log(OR)| > 50$ were excluded as likely overfit.
680 Quasi-likelihood under independent model criterion (QIC) and the Wald tests were used to evaluate the
681 significance and magnitude of the estimated IMR. We also ran this analysis using publication identifiers
682 as a control variable to account for researcher and group biases; this produced similar results likely
683 because protein identity is strongly correlated with the publications in which they appear.

684 Calculating glycoimpact from IMR and populating a BLAMO matrix

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

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

694 Glycoimpact values from a BLAMO XY matrix can then be z-score normalized to a Gaussian Mixture
695 Model[143] estimated null distribution. We use z=2.5 as a heuristic but stringent cutoff between
696 "impactful" (z>2.5) and "null" (z<2.5) substitutions.

697 Comparison of SNP pathogenicity scores with glycoimpact

698 Functional prediction rank normalized scores were obtained from dbNSFP (v3.2) for the following 27
699 tools: SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM,
700 LRT, SiPhy, 2x PhyloP, MetaSVM, MetaLR, CADD, VEST3, PROVEAN, 4× fitCons scores, fathmm-MKL,
701 DANN, 2× phastCons, GenoCanyon, Eigen and Eigen-PC.[40] Variants were excluded from the analysis if
702 they had more than 3 missing functional score predictions, did not result in an amino acid change, or not
703 on proteins that had known glycosylation sites.

704 Assignments of "prediction-type" and "structure-usage" (**Figure** 5b) were adapted from classifications
705 provided by dbNSFP.[40]

706 Estimation and Analysis of Evolutionary Couplings (EC)

707 For EVCouplings calculation, hits composed of more than 50% gaps were filtered from the alignment,
708 and sequences with homologs more than 80% identical were down-weighted to compute N_{eff}, the
709 effective number of sequences.[45] ECs were calculated using pseudo-likelihood maximization,[144,145] as
710 implemented previously.[146] The λ_J term was scaled by the number of amino acids minus one times the
711 number of sites in the model minus one. Pre- and post-processing was performed using the EVCouplings
712 Python package.[147]

713 High-ranking EC events are generally considered those ranking less than *L* a measure of sequence length
714 where only residues with fewer than 30% gaps were counted.[45] Specifically, the gap threshold (minimum
715 column coverage parameter) was set to 70%. Therefore, residue positions with more than 30% gaps
716 disregarded. We also used a fragment filter of 70%, meaning that individual sequences had to have 30%
717 non-gap characters in each row as well. We explored multiple high-rank thresholds between *L/5* and *3L*.

718 To explore the increased coupling with glycosylation sites, we examined couplings between each amino
719 acid with glycosites (GN), asparagines (N) and any amino acid (AA). We compared the number of high-
720 ranking coupling events, the distributions of EC probabilities and the relative numbers of high and low-
721 ranking ECs for each group with various amino acids at relative positions N+/-6. Distributions were
722 compared with a one-sided Wilcoxon test and high/low-ranking counts were compared with
723 hypergeometric enrichment. The hypergeometric enrichment of glycosite-coupling was performed at
724 multiple high-rank thresholds (*L/3, L/2, L, 2L, 3L*) and p-values were pooled for each amino acid at each
725 relative position across ranks using Fisher's method. Finally, the pooled p-values were corrected for
726 multiple tests using the Benjamini-Hotchberg method.

727 To examine larger structures in ECs, we used EC rank to mask extended sequons (N+/-6) then clustered
728 the sequons and extracted motifs. For each sequon, the residues were retained if the residue-glycosite
729 coupling rank was less than *L4*. The extended and masked sequons were distinguished using a hamming
730 distance (DECIPHERv2.18.1)[148] then clustered using agglomerative hierarchical clustering
731 (factoextra::hcut v1.0.7). Motif logos were generated using custom-scaled position-specific scoring
732 matrices[149] reflecting the cumulative rank of amino acids at each glyco-site relative position. Specifically,
733 the aggregate score, *S,* for each amino acid, *a*, at each position, *p*, was aggregated over EC score ranks, *r,*
734 within each extended-masked sequons, *s*, in a cluster, *c*, such that $S_{a,p} = log_{10}\left(\sum_{s \in c} \quad L/r\right)$

735 Mouse breeding and Samples

736 The Collaborative Cross (CC) recombinant inbred mouse strains (N = 333, 95 strains, age 20-117 weeks)
737 were produced by Geniad Pty Ltd and housed at Animal Resources Centre (Murdoch, WA, Australia).[150]
738 The CC strains were genotyped using the MegaMUGA platform (GeneSeek; Lincoln, NE). C57BL/6 (N =
739 10) and BALB/c mice (N = 10), sex- and age-matched (10 weeks old, 1:1 male:female) were obtained
740 from Elevage Janvier (Le Genest-Saint-Isle, France).[86] The studies received appropriate ethics approvals
741 from the Animal Ethics Committee of the Animal Resources Centre[75] and the Ethical Committee of the
742 District Government of Lower Franconia.[86]

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

745 Immunoglobulin G was isolated from 100-500 µl of mouse serum on 96-well Protein G monolithic plates
746 (BIA Separations) as described previously.[75,85] LC-MS analysis of tryptic Fc-glycopeptides was performed
747 as described in.[75,85] In brief, approximately 10–20 µg of isolated IgG was digested with 200 ng trypsin

748 (Worthington, USA). The resulting glycopeptides were purified by reverse-phase solid phase extraction
749 using Chromabond C18ec beads (Marcherey-Nagel, Germany) as described in.[85] Tryptic digests were
750 analyzed on a nanoACQUITY UPLC system (Waters, USA) coupled to a Compact mass spectrometer
751 (Bruker Daltonics, Germany). Peak areas were calculated by summing areas for doubly and triply
752 charged ions determined with LaCyTools v 1.0.1 b.7 software[151] and normalized to the total integrated
753 area per IgG subclass.

754 Batch correction was performed on the log-transformed values using the ComBat method (R package
755 "sva") to remove possible experimental variations due to LC-MS analysis having been performed on
756 several 96-well plates within each cohort. Derived glycosylation traits describing relative abundance of
757 N-glycans sharing specific structural features (agalactosylated, galactosylated, sialylated,
758 monogalactosylated, digalactosylated, monosialylated, disialylated structures, structures with bisecting
759 GlcNAc) were calculated in a subclass-specific manner.[86] Statistical analysis and data visualization were
760 performed using R programming language v 4.0.3.

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

762 His-tagged recombinant SARS-CoV-2 Spike S1 protein, expressed in HEK293 cells were purchased from
763 Sino Biologics (Wayne, PA). Lyophilized glycoproteins corresponding to the original 2019 strain (40591-
764 V08H), the Gamma variant (40591-V08H14 with L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y,
765 D614G and H655Y mutations), and the Delta variant (40591-V08H23 with T19R, G142D, E156G, 157-158
766 deletion, L452R, T478K, D614G and P681R mutations). Samples were analyzed using DeGlyPHER.[73]

767 Briefly, glycoproteins were digested with Proteinase K (30 min or 4 h) or trypsin, sequentially
768 deglycosylated with Endo H (creating residual mass signature of +203 Da) to signify high
769 mannose/hybrid glycans, and then with PNGase F in $H_2{}^{18}O$ (creating residual mass signature of +3 Da) to
770 signify the remnant complex glycans on any sequon (NXS|T, where X is any amino acid except P)
771 asparagine. Unoccupied sequons will have no additional signature mass. Analysis of samples was done
772 on a Q Exactive HF-X mass spectrometer (Thermo), injecting directly onto a 25 cm, 100 µm ID column
773 packed with BEH 1.7 µm C18 resin (Waters). Liquid chromatography separation was achieved at a flow
774 rate of 300 nL min−1 on an EASY-nLC 1200 (Thermo). Buffers A and B were 0.1% formic acid in 5 and
775 80% acetonitrile, respectively. The gradient used was 1–25% B over 160 min, an increase to 40% B over
776 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
777 column was re-equilibrated with solution A before injecting sample. Peptides eluting from the tip of the
778 column were nanosprayed directly into the mass spectrometer by application of 2.8 kV at the back of

779 the column. The mass spectrometer was operated in a data-dependent mode. Full MS1 scans were
780 collected in the Orbitrap at 120000 resolution. The 10 most abundant ions per scan were selected for
781 HCD MS/MS at 25 NCE. Dynamic exclusion was enabled with a 10 s duration and +1 ions were excluded.
782 Peptides were identified with Integrated Proteomics Pipeline (IP2, Bruker Scientific LLC). Tandem mass
783 spectra were extracted from raw files using RawConverter[152] and searched with ProLuCID[153] against a
784 database comprising UniProt reviewed proteome for Homo sapiens (UP000005640), including additional
785 UniProt amino acid sequences for Endo H (P04067), PNGase F (Q9XBM8), and Proteinase K (P06873), the
786 amino acid sequences for the SARS-CoV-2 S1 subunits, and a list of general protein contaminants. The
787 search included no protease specificity (all fully tryptic and semitryptic peptide candidates when treated
788 with trypsin). Carbamidomethylation (+57.02146 C) was used as a static modification. Deamidation in
789 the presence of $H_2{}^{18}O$ (+2.988261 N), GlcNAc (+203.079373 N), oxidation (+15.994915 M), and N-
790 terminal pyroglutamate formation (−17.026549 Q) were used as differential modifications. Data were
791 searched with 50 ppm parent mass tolerance and 50 ppm fragment mass tolerance. Identified proteins
792 were filtered using DTASelect2[154] while using a target-decoy database search strategy to limit the false
793 discovery rate to 1%, at the spectrum level.[155] At least one peptide per protein and no tryptic end (or
794 one tryptic end when treated with trypsin) per peptide were necessary, and precursor delta mass cutoff
795 was fixed at 10 ppm. Statistical models for peptide mass modification (modstat) were applied (trypstat
796 was additionally applied for trypsin-treated samples). Semi-quantitative label-free analysis was
797 performed based on the precursor peak area, with a 10 ppm parent mass tolerance and 0.1 min
798 retention time tolerance, Census2.[156] "Match between runs" was used to find missing peptides between
799 runs. GlycoMSQuant (v.1.4.1, https://github.com/proteomicsyates/GlycoMSQuant)[73] was used,
800 summing precursor peak areas across the 3 conditions – 30 min and 4 h Proteinase K, and trypsin,
801 discarded peptides without glycosites, and discarded misidentified peptides when N-glycan remnant
802 mass modifications were localized to non-glycosite asparagines and corrected/fixed N-glycan
803 mislocalization where appropriate, to finally calculate proportions of 3 glycosylation states at each
804 glycosite, with +/-SEM (standard error of mean). Pairwise statistical comparison between variants was
805 performed using Mann-Whitney U Test.

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

811 ## Estimation of glycomotif preference from IMR

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

819 Glycomotif preference can be described in more precise terms. Given a substitution where X and $Y$are
820 WT and mutant amino acids respectively at an amino acid residue at index I, we may describe the
821 substitution as XIY. When I is glycosite-proximal, we may describe a substitution, XIY, as a point, p,
822 with a pair of IMR (ρ) relating each amino acid to a single glycomotif, $p_{XIY} = (\rho_X, \rho_Y)$. The IMR-pair,
823 $p_{XIY} = (\rho_X, \rho_Y)$, may also be described as a vector terminating on that point, $\vec{p}_{XIY}$. We may also
824 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
825 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 $\vec{n}$,
826 $\vec{p}_{\vec{n}} = \frac{\vec{p} \cdot \vec{n}}{\vec{n} \cdot \vec{n}} \cdot \vec{n}$. Therefore, $G = d(\vec{p}, \vec{n}) = \vec{p} - \vec{p}_{\vec{n}}$

827 With this definition of G, we can define G_i and G_c as the glycoimpact corresponding to the glycomotif of
828 interest and the contract glycomotif respectively. Then, the relative preference by first calculating the
829 perpendicular component of the difference in glycoimpact vectors projected onto $\vec{n}$, such that
830 $r_{\vec{n}} = (G_i - G_c)_{\vec{n}}$. 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
831 the sign, s, as positive if $y > x$ and negative if $x > y$. The sign is not defined if $x = y$.

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

833 As was recently described,[157] 54 Fc variants of 25c4 were cloned from the REFORM Fc variant panel.[74]
834 The REFORM plasmid library[74] is comprised of golden gate cloning plasmids[158] with BsaI restriction sites
835 flanking distinct antibody domains and a furin 2A cleavage site to enable self-cleavage and successful
836 assembly of a complete antibody from a single open reading frame.[159] Variable Light chains containing
837 REFORM variants and 25c4 Variable Heavy chains were simultaneously transfected at a 1:1 ratio in CHO
838 cells, purified using a Protein A chromatography resin, then dialyzed and concentrated with Phosphate
839 Saline Buffer (PBS).

840 Fc glycosylation was then measured on purified 25c4 antibodies. Purified antibodies were incubated
841 with magnetic G protein beads (Millipore) and separated from Fab fragments following enzymatic
842 digestion (IdeZ, NEB). Fc glycans were released and labelled using the GlycanAssure ATPS kit (Thermo
843 Fisher Scientific) then separated using 3500xL Genetic Analyzer (Thermo Fisher Scientific). As previously
844 described,[160] retention times (RT) were matched to glycan standards using Glycan Acquisition Software
845 Version 3500 v1.0.3 and Glycan Analysis Software v1.1. Abundance (area under peaks) was normalized
846 to total area per sample to calculate relative abundance of each glycan. Non-uniquely determined
847 glycans were excluded (where the difference in RT was below the detection threshold).

848

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

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

6. 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.

7. 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.

8. 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.

9. 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.

10. Altman, M.O., Angel, M., Košík, I., Trovão, N.S., Zost, S.J., Gibbs, J.S., Casalino, L., Amaro, R.E., Hensley, S.E., Nelson, M.I., et al. (2019). Human Influenza A Virus Hemagglutinin Glycan Evolution Follows a Temporal Pattern to a Glycan Limit. Preprint, 10.1128/mbio.00204-19 10.1128/mbio.00204-19.

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

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

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

889 14. Murray, A.N., Chen, W., Antonopoulos, A., Hanson, S.R., Wiseman, R.L., Dell, A., Haslam, S.M.,
890 Powers, D.L., Powers, E.T., and Kelly, J.W. (2015). Enhanced aromatic sequons increase
891 oligosaccharyltransferase glycosylation efficiency and glycan homogeneity. Chem. Biol. *22*, 1052–
892 1062.

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

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

898 17. Ives, C.M., Singh, O., D'Andrea, S., Fogarty, C.A., Harbison, A.M., Satheesan, A., Tropea, B., and
899 Fadda, E. (2023). Restoring Protein Glycosylation with GlycoShape. bioRxiv.
900 10.1101/2023.12.11.571101.

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

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

906 20. Kawatkar, S.P., Kuntz, D.A., Woods, R.J., Rose, D.R., and Boons, G.-J. (2006). Structural Basis of the
907 Inhibition of Golgi α-Mannosidase II by Mannostatin A and the Role of the Thiomethyl Moiety in
908 Ligand−Protein Interactions. Preprint, 10.1021/ja061216p 10.1021/ja061216p.

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

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

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

918 24. Thaysen-Andersen, M., and Packer, N.H. (2012). Site-specific glycoproteomics confirms that protein
919 structure dictates formation of N-glycan type, core fucosylation and branching. Glycobiology *22*,
920 1440–1452.

921 25. Allen, J.D., Chawla, H., Samsudin, F., Zuzic, L., Shivgan, A.T., Watanabe, Y., He, W.-T., Callaghan, S.,
922 Song, G., Yong, P., et al. (2021). Site-specific steric control of SARS-CoV-2 spike glycosylation. Cold
923 Spring Harbor Laboratory, 2021.03.08.433764. 10.1101/2021.03.08.433764.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

961 proteome. Science *347*, 1260419.

962 40. Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). DbNSFP v3.0: A one-stop database of functional
963 predictions and annotations for human nonsynonymous and splice-site SNVs. Hum. Mutat. *37*,
964 235–241.

965 41. Fitch, W.M., and Dayhoff, M.O. (1973). Atlas of protein sequence and structure, 1972. Syst. Zool.
966 *22*, 196.

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

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

971 44. Pearson, W.R. (2013). Selecting the right similarity-scoring matrix. Curr. Protoc. Bioinformatics *43*,
972 3.5.1-3.5.9.

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

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

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

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

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

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

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

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

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

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

999 55. Yi, C.-W., Wang, L.-Q., Huang, J.-J., Pan, K., Chen, J., and Liang, Y. (2018). Glycosylation Significantly
1000 Inhibits the Aggregation of Human Prion Protein and Decreases Its Cytotoxicity. Preprint,
1001 10.1038/s41598-018-30770-6 10.1038/s41598-018-30770-6.

1002 56. Ng, B.G., and Freeze, H.H. (2018). Perspectives on Glycosylation and Its Congenital Disorders.
1003 Trends Genet. *34*, 466–476.

1004 57. Joshi, H.J., Hansen, L., Narimatsu, Y., Freeze, H.H., Henrissat, B., Bennett, E., Wandall, H.H., Clausen,
1005 H., and Schjoldager, K.T. (2018). Glycosyltransferase genes that cause monogenic congenital
1006 disorders of glycosylation are distinct from glycosyltransferase genes associated with complex
1007 diseases. Glycobiology *28*, 284–294.

1008 58. Berg-Fussman, A., Grace, M.E., Ioannou, Y., and Grabowski, G.A. (1993). Human acid beta-
1009 glucosidase. N-glycosylation site occupancy and the effect of glycosylation on enzymatic activity. J.
1010 Biol. Chem. *268*, 14861–14866.

1011 59. Pol-Fachin, L., Siebert, M., Verli, H., and Saraiva-Pereira, M.L. (2016). Glycosylation is crucial for a
1012 proper catalytic site organization in human glucocerebrosidase. Glycoconj. J. *33*, 237–244.

1013 60. Smith, L., Mullin, S., and Schapira, A.H.V. (2017). Insights into the structural biology of Gaucher
1014 disease. Exp. Neurol. *298*, 180–190.

1015 61. Souffrant, M.G., Yao, X.-Q., Momin, M., and Hamelberg, D. (2020). N-glycosylation and gaucher
1016 disease mutation allosterically alter active-site dynamics of acid-β-glucosidase. ACS Catal. *10*, 1810–
1017 1820.

1018 62. Di Fede, G., Catania, M., Atzori, C., Moda, F., Pasquali, C., Indaco, A., Grisoli, M., Zuffi, M., Guaita,
1019 M.C., Testi, R., et al. (2019). Clinical and neuropathological phenotype associated with the novel
1020 V189I mutation in the prion protein gene. Acta Neuropathol. Commun. *7*, 1.

1021 63. Ladogana, A., and Kovacs, G.G. (2018). Genetic Creutzfeldt–Jakob disease. In Human Prion Diseases
1022 Handbook of clinical neurology. (Elsevier), pp. 219–242.

1023 64. Liberski, P.P. (2012). Gerstmann-Sträussler-Scheinker disease. Adv. Exp. Med. Biol. *724*, 128–137.

1024 65. Cohen, M.L. (2014). Human Prion Diseases. In Pathobiology of Human Disease (Elsevier), pp. 2045–
1025 2054.

1026 66. Cao, L., Diedrich, J.K., Kulp, D.W., Pauthner, M., He, L., Park, S.-K.R., Sok, D., Su, C.Y., Delahunty,
1027 C.M., Menis, S., et al. (2017). Global site-specific N-glycosylation analysis of HIV envelope
1028 glycoprotein. Preprint, 10.1038/ncomms14954 10.1038/ncomms14954.

1029 67. Zhao, P., Praissman, J.L., Grant, O.C., Cai, Y., Xiao, T., Rosenbalm, K.E., Aoki, K., Kellman, B.P.,
1030 Bridger, R., Barouch, D.H., et al. (2020). Virus-Receptor Interactions of Glycosylated SARS-CoV-2
1031 Spike and Human ACE2 Receptor. Cell Host Microbe *28*, 586-601.e6.

1032 68. Watanabe, Y., Allen, J.D., Wrapp, D., McLellan, J.S., and Crispin, M. (2020). Site-specific glycan
1033 analysis of the SARS-CoV-2 spike. Science *369*, 330–333.

1034 69. Casalino, L., Gaieb, Z., Dommer, A.C., and Harbison, A.M. (2020). Shielding and Beyond: The Roles
1035 of Glycans in SARS-CoV-2 Spike Protein. bioRxiv.

1036 70. Grant, O.C., Montgomery, D., Ito, K., and Woods, R.J. (2020). 3D Models of glycosylated SARS-CoV-2
1037 spike protein suggest challenges and opportunities for vaccine development. bioRxiv.

1038 71. Wintjens, R., Bifani, A.M., and Bifani, P. (2020). Impact of glycan cloud on the B-cell epitope
1039 prediction of SARS-CoV-2 Spike protein. NPJ Vaccines *5*, 81.

1040 72. Zhang, L., Mann, M., Syed, Z.A., Reynolds, H.M., Tian, E., Samara, N.L., Zeldin, D.C., Tabak, L.A., and
1041 Ten Hagen, K.G. (2021). Furin cleavage of the SARS-CoV-2 spike is modulated by O-glycosylation.
1042 Proc. Natl. Acad. Sci. U. S. A. *118*, e2109905118.

1043 73. Baboo, S., Diedrich, J.K., Martínez-Bartolomé, S., Wang, X., Schiffner, T., Groschel, B., Schief, W.R.,
1044 Paulson, J.C., and Yates, J.R., 3rd (2021). DeGlyPHER: An ultrasensitive method for the analysis of
1045 viral spike N-glycoforms. Anal. Chem. *93*, 13651–13657.

1046 74. Gunn, B.M., Lu, R., Slein, M.D., Ilinykh, P.A., Huang, K., Atyeo, C., Schendel, S.L., Kim, J., Cain, C.,
1047 Roy, V., et al. (2021). A Fc engineering approach to define functional humoral correlates of
1048 immunity against Ebola virus. Immunity *54*, 815-828.e5.

1049 75. Krištić, J., Zaytseva, O.O., Ram, R., Nguyen, Q., Novokmet, M., Vučković, F., Vilaj, M., Trbojević-
1050 Akmačić, I., Pezer, M., Davern, K.M., et al. (2018). Profiling and genetic control of the murine
1051 immunoglobulin G glycome. Nat. Chem. Biol. *14*, 516–524.

1052 76. Cheng, H.D., Tirosh, I., de Haan, N., Stöckmann, H., Adamczyk, B., McManus, C.A., O'Flaherty, R.,
1053 Greville, G., Saldova, R., Bonilla, F.A., et al. (2020). IgG Fc glycosylation as an axis of humoral
1054 immunity in childhood. J. Allergy Clin. Immunol. *145*, 710-713.e9.

1055 77. Lofano, G., Gorman, M.J., Yousif, A.S., Yu, W.-H., Fox, J.M., Dugast, A.-S., Ackerman, M.E.,
1056 Suscovich, T.J., Weiner, J., Barouch, D., et al. (2018). Antigen-specific antibody Fc glycosylation
1057 enhances humoral immunity via the recruitment of complement. Sci. Immunol. *3*, eaat7796.

1058 78. Atyeo, C., Pullen, K.M., Bordt, E.A., Fischinger, S., Burke, J., Michell, A., Slein, M.D., Loos, C., Shook,
1059 L.L., Boatin, A.A., et al. (2021). Compromised SARS-CoV-2-specific placental antibody transfer. Cell
1060 *184*, 628-642.e10.

1061 79. Grace, P.S., Dolatshahi, S., Lu, L.L., Cain, A., Palmieri, F., Petrone, L., Fortune, S.M., Ottenhoff,
1062 T.H.M., Lauffenburger, D.A., Goletti, D., et al. (2021). Antibody subclass and glycosylation shift
1063 following effective TB treatment. Front. Immunol. *12*, 679973.

1064 80. Majewska, N.I., Tejada, M.L., Betenbaugh, M.J., and Agarwal, N. (2020). N-glycosylation of IgG and
1065 IgG-like recombinant therapeutic proteins: Why is it important and how can we control it? Annu.
1066 Rev. Chem. Biomol. Eng. *11*, 311–338.

1067 81. Li, W., Zhu, Z., Chen, W., Feng, Y., and Dimitrov, D.S. (2017). Crystallizable fragment
1068 glycoengineering for therapeutic antibodies development. Front. Immunol. *8*.

1069 10.3389/fimmu.2017.01554.

1070 82. Zhang, P., Woen, S., Wang, T., Liau, B., Zhao, S., Chen, C., Yang, Y., Song, Z., Wormald, M.R., Yu, C.,
1071 et al. (2016). Challenges of glycosylation analysis and control: an integrated approach to producing
1072 optimal and consistent therapeutic drugs. Drug Discov. Today *21*, 740–765.

1073 83. Shields, R.L., Lai, J., Keck, R., O'Connell, L.Y., Hong, K., Meng, Y.G., Weikert, S.H.A., and Presta, L.G.
1074 (2002). Lack of Fucose on Human IgG1 N-Linked Oligosaccharide Improves Binding to Human FcγRIII
1075 and Antibody-dependent Cellular Toxicity. J. Biol. Chem. *277*, 26733–26740.

1076 84. de Haan, N., Reiding, K.R., Krištić, J., Hipgrave Ederveen, A.L., Lauc, G., and Wuhrer, M. (2017). The
1077 N-Glycosylation of Mouse Immunoglobulin G (IgG)-Fragment Crystallizable Differs Between IgG
1078 Subclasses and Strains. Front. Immunol. *8*, 608.

1079 85. Zaytseva, O.O., Jansen, B.C., Hanić, M., Mrčela, M., Razdorov, G., Stojković, R., Erhardt, J., Brizić, I.,
1080 Jonjić, S., Pezer, M., et al. (2018). MIgGGly (mouse IgG glycosylation analysis) - a high-throughput
1081 method for studying Fc-linked IgG N-glycosylation in mice with nanoUPLC-ESI-MS. Sci. Rep. *8*,
1082 13688.

1083 86. Zaytseva, O.O., Seeling, M., Krištić, J., Lauc, G., Pezer, M., and Nimmerjahn, F. (2020). Fc-Linked IgG
1084 N-Glycosylation in FcγR Knock-Out Mice. Preprint, 10.3389/fcell.2020.00067
1085 10.3389/fcell.2020.00067.

1086 87. Karch, C.P., Paquin-Proulx, D., Eller, M.A., Matyas, G.R., Burkhard, P., and Beck, Z. (2020). Impact of
1087 the expression system on the immune responses to self-assembling protein nanoparticles (SAPNs)
1088 displaying HIV-1 V1V2 loop. Nanomedicine *29*, 102255.

1089 88. Hombu, R., Neelamegham, S., and Park, S. (2021). Cellular and molecular engineering of glycan
1090 sialylation in heterologous systems. Molecules *26*, 5950.

1091 89. Zhong, X., Ma, W., Meade, C.L., Tam, A.S., Llewellyn, E., Cornell, R., Cote, K., Scarcelli, J.J., Marshall,
1092 J.K., Tzvetkova, B., et al. (2019). Transient CHO expression platform for robust antibody production
1093 and its enhanced N-glycan sialylation on therapeutic glycoproteins. Biotechnol. Prog. *35*, e2724.

1094 90. Zhong, X., Schwab, A., Ma, W., Meade, C.L., Zhou, J., D'Antona, A.M., Somers, W., and Lin, L. (2022).
1095 Large-scale transient production in ExpiCHO-S^{TM} with enhanced N-galactosylation-sialylation and
1096 PEI-based transfection. Methods Mol. Biol. *2313*, 143–150.

1097 91. Hudson, K.L., Bartlett, G.J., Diehl, R.C., Agirre, J., Gallagher, T., Kiessling, L.L., and Woolfson, D.N.
1098 (2015). Carbohydrate-aromatic interactions in proteins. J. Am. Chem. Soc. *137*, 15152–15160.

1099 92. Chen, W., Enck, S., Price, J.L., Powers, D.L., Powers, E.T., Wong, C.-H., Dyson, H.J., and Kelly, J.W.
1100 (2013). Structural and energetic basis of carbohydrate-aromatic packing interactions in proteins. J.
1101 Am. Chem. Soc. *135*, 9877–9884.

1102 93. Asensio, J.L., Ardá, A., Cañada, F.J., and Jiménez-Barbero, J. (2013). Carbohydrate-aromatic
1103 interactions. Acc. Chem. Res. *46*, 946–954.

1104 94. Montalvillo-Jiménez, L., Santana, A.G., Corzana, F., Jiménez-Osés, G., Jiménez-Barbero, J., Gómez,
1105 A.M., and Asensio, J.L. (2019). Impact of aromatic stacking on glycoside reactivity: Balancing CH/π

1106 and cation/π interactions for the stabilization of glycosyl-oxocarbenium ions. J. Am. Chem. Soc. *141*,
1107 13372–13384.

1108 95. Ardejani, M.S., Noodleman, L., Powers, E.T., and Kelly, J.W. (2021). Stereoelectronic effects in
1109 stabilizing protein-N-glycan interactions revealed by experiment and machine learning. Nat. Chem.
1110 *13*, 480–487.

1111 96. Goldfarb, L.G., Petersen, R.B., Tabaton, M., Brown, P., LeBlanc, A.C., Montagna, P., Cortelli, P.,
1112 Julien, J., Vital, C., and Pendelbury, W.W. (1992). Fatal familial insomnia and familial Creutzfeldt-
1113 Jakob disease: disease phenotype determined by a DNA polymorphism. Science *258*, 806–808.

1114 97. Wang, W.K., Essex, M., and Lee, T.H. (1996). Single amino acid substitution in constant region 1 or 4
1115 of gp120 causes the phenotype of a human immunodeficiency virus type 1 variant with mutations
1116 in hypervariable regions 1 and 2 to revert. J. Virol. *70*, 607–611.

1117 98. Branza-Nichita, N., Petrescu, A.J., Negroiu, G., Dwek, R.A., and Petrescu, S.M. (2000). N-
1118 glycosylation processing and glycoprotein folding-lessons from the tyrosinase-related proteins.
1119 Chem. Rev. *100*, 4697–4712.

1120 99. Dolinska, M.B., Kus, N.J., Farney, S.K., Wingfield, P.T., Brooks, B.P., and Sergeev, Y.V. (2017).
1121 Oculocutaneous albinism type 1: link between mutations, tyrosinase conformational stability, and
1122 enzymatic activity. Pigment Cell Melanoma Res. *30*, 41–52.

1123 100. Grant, O.C., Montgomery, D., Ito, K., and Woods, R.J. (2020). Analysis of the SARS-CoV-2 spike
1124 protein glycan shield reveals implications for immune recognition. Sci. Rep. *10*, 14991.

1125 101. Peng, W., Rayaprolu, V., Parvate, A.D., Pronker, M.F., Hui, S., Parekh, D., Shaffer, K., Yu, X., Saphire,
1126 E.O., and Snijder, J. (2022). Glycan shield of the ebolavirus envelope glycoprotein GP. Commun.
1127 Biol. *5*, 785.

1128 102. Wei, C.-J., Boyington, J.C., Dai, K., Houser, K.V., Pearce, M.B., Kong, W.-P., Yang, Z.-Y., Tumpey,
1129 T.M., and Nabel, G.J. (2010). Cross-neutralization of 1918 and 2009 influenza viruses: role of
1130 glycans in viral evolution and vaccine design. Sci. Transl. Med. *2*, 24ra21.

1131 103. Schulz, M.A., Tian, W., Mao, Y., Van Coillie, J., Sun, L., Larsen, J.S., Chen, Y.-H., Kristensen, C.,
1132 Vakhrushev, S.Y., Clausen, H., et al. (2018). Glycoengineering design options for IgG1 in CHO cells
1133 using precise gene editing. Glycobiology *28*, 542–549.

1134 104. Narimatsu, Y., Joshi, H.J., Nason, R., Van Coillie, J., Karlsson, R., Sun, L., Ye, Z., Chen, Y.-H.,
1135 Schjoldager, K.T., Steentoft, C., et al. (2019). An atlas of human glycosylation pathways enables
1136 display of the human glycome by gene engineered cells. Mol. Cell *75*, 394-407.e5.

1137 105. Narimatsu, Y., Büll, C., Chen, Y.-H., Wandall, H.H., Yang, Z., and Clausen, H. (2021). Genetic
1138 glycoengineering in mammalian cells. J. Biol. Chem. *296*, 100448.

1139 106. Sumit, M., Dolatshahi, S., Chu, A.-H.A., Cote, K., Scarcelli, J.J., Marshall, J.K., Cornell, R.J., Weiss, R.,
1140 Lauffenburger, D.A., Mulukutla, B.C., et al. (2019). Dissecting N-Glycosylation Dynamics in Chinese
1141 Hamster Ovary Cells Fed-batch Cultures using Time Course Omics Analyses. iScience *12*, 102–120.

1142 107. Ehret, J., Zimmermann, M., Eichhorn, T., and Zimmer, A. (2019). Impact of cell culture media

1143 additives on IgG glycosylation produced in Chinese hamster ovary cells. Biotechnol. Bioeng. *116*,
1144 816–830.

1145 108. Fan, Y., Kildegaard, H.F., and Andersen, M.R. (2017). Engineer medium and feed for modulating N-
1146 glycosylation of recombinant protein production in CHO cell culture. Methods Mol. Biol. *1603*, 209–
1147 226.

1148 109. Kotidis, P., and Kontoravdi, C. (2020). Harnessing the potential of artificial neural networks for
1149 predicting protein glycosylation. Metabolic Engineering Communications, e00131.

1150 110. Jimenez del Val, I., Nagy, J.M., and Kontoravdi, C. (2011). A dynamic mathematical model for
1151 monoclonal antibody N-linked glycosylation and nucleotide sugar donor transport within a
1152 maturing Golgi apparatus. Biotechnol. Prog. *27*, 1730–1743.

1153 111. Sou, S.N., Jedrzejewski, P.M., Lee, K., Sellick, C., Polizzi, K.M., and Kontoravdi, C. (2017). Model-
1154 based investigation of intracellular processes determining antibody Fc-glycosylation under mild
1155 hypothermia. Biotechnol. Bioeng. *114*, 1570–1582.

1156 112. Spahn, P.N., Hansen, A.H., Hansen, H.G., Arnsdorf, J., Kildegaard, H.F., and Lewis, N.E. (2016). A
1157 Markov chain model for N-linked protein glycosylation – towards a low-parameter tool for model-
1158 driven glycoengineering. Preprint, 10.1016/j.ymben.2015.10.007 10.1016/j.ymben.2015.10.007.

1159 113. Spahn, P.N., Hansen, A.H., Kol, S., Voldborg, B.G., and Lewis, N.E. (2017). Predictive
1160 glycoengineering of biosimilars using a Markov chain glycosylation model. Biotechnol. J. *12*.
1161 10.1002/biot.201600489.

1162 114. Liang, C., Chiang, A.W.T., Hansen, A.H., Arnsdorf, J., Schoffelen, S., Sorrentino, J.T., Kellman, B.P.,
1163 Bao, B., Voldborg, B.G., and Lewis, N.E. (2020). A Markov model of glycosylation elucidates isozyme
1164 specificity and glycosyltransferase interactions for glycoengineering. Curr Res Biotechnol *2*, 22–36.

1165 115. Michaeli, Y., and Reiter, Y. (2006). Optimised Fc variants with enhanced effector function. Expert
1166 Opin. Ther. Pat. *16*, 1449–1452.

1167 116. Liu, R., Oldham, R.J., Teal, E., Beers, S.A., and Cragg, M.S. (2020). Fc-engineering for modulated
1168 effector functions-improving antibodies for cancer treatment. Antibodies (Basel) *9*, 64.

1169 117. van Erp, E.A., Luytjes, W., Ferwerda, G., and van Kasteren, P.B. (2019). Fc-mediated antibody
1170 effector functions during respiratory syncytial virus infection and disease. Front. Immunol. *10*, 548.

1171 118. Leon, P.E., He, W., Mullarkey, C.E., Bailey, M.J., Miller, M.S., Krammer, F., Palese, P., and Tan, G.S.
1172 (2016). Optimal activation of Fc-mediated effector functions by influenza virus hemagglutinin
1173 antibodies requires two points of contact. Proc. Natl. Acad. Sci. U. S. A. *113*, E5944–E5951.

1174 119. Zhou, M., Palanca, A.M.S., and Law, J.A. (2018). Locus-specific control of the de novo DNA
1175 methylation pathway in Arabidopsis by the CLASSY family. Nat. Genet. *50*, 865–873.

1176 120. Nováček, V., McGauran, G., Matallanas, D., Vallejo Blanco, A., Conca, P., Muñoz, E., Costabello, L.,
1177 Kanakaraj, K., Nawaz, Z., Walsh, B., et al. (2020). Accurate prediction of kinase-substrate networks
1178 using knowledge graphs. PLoS Comput. Biol. *16*, e1007578.

1179 121. Malaker, S.A., Riley, N.M., Shon, D.J., Pedram, K., Krishnan, V., Dorigo, O., and Bertozzi, C.R. (2022).
1180 Revealing the human mucinome. Nat. Commun. *13*, 3542.

1181 122. Madhani, H.D., and Guthrie, C. (1994). Dynamic RNA-RNA interactions in the spliceosome. Annu.
1182 Rev. Genet. *28*, 1–26.

1183 123. Bao, P., Boon, K.-L., Will, C.L., Hartmuth, K., and Lührmann, R. (2018). Multiple RNA–RNA tertiary
1184 interactions are dispensable for formation of a functional U2/U6 RNA catalytic core in the
1185 spliceosome. Preprint, 10.1093/nar/gky966 10.1093/nar/gky966.

1186 124. Corley, M., Burns, M.C., and Yeo, G.W. (2020). How RNA-Binding Proteins Interact with RNA:
1187 Molecules and Mechanisms. Mol. Cell *78*, 9–29.

1188 125. Farnham, P.J., and Platt, T. (1982). Effects of DNA base analogs on transcription termination at the
1189 tryptophan operon attenuator of EScherichia coli. Proc. Natl. Acad. Sci. U. S. A. *79*, 998–1002.

1190 126. Lee, F., and Yanofsky, C. (1977). Transcription termination at the trp operon attenuators of
1191 Escherichia coli and Salmonella typhimurium: RNA secondary structure and regulation of
1192 termination. Proc. Natl. Acad. Sci. U. S. A. *74*, 4365–4369.

1193 127. Oxender, D.L., Zurawski, G., and Yanofsky, C. (1979). Attenuation in the Escherichia coli tryptophan
1194 operon: role of RNA secondary structure involving the tryptophan codon region. Proc. Natl. Acad.
1195 Sci. U. S. A. *76*, 5524–5528.

1196 128. Farnham, P.J., and Platt, T. (1980). A model for transcription termination suggested by studies on
1197 the trp attenuator in vitro using base analogs. Cell *20*, 739–748.

1198 129. Wu, A.M., and Platt, T. (1978). Transcription termination: nucleotide sequence at 3' end of
1199 tryptophan operon in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. *75*, 5442–5446.

1200 130. Mariethoz, J., Alocci, D., Gastaldello, A., Horlacher, O., Gasteiger, E., Rojas-Macias, M., Karlsson,
1201 N.G., Packer, N.H., and Lisacek, F. (2018). Glycomics@ExPASy: Bridging the Gap. Preprint,
1202 10.1074/mcp.ra118.000799 10.1074/mcp.ra118.000799.

1203 131. York, W.S., Mazumder, R., Ranzinger, R., Edwards, N., Kahsay, R., Aoki-Kinoshita, K.F., Campbell,
1204 M.P., Cummings, R.D., Feizi, T., Martin, M., et al. (2020). GlyGen: Computational and Informatics
1205 Resources for Glycoscience. Glycobiology *30*, 72–73.

1206 132. Cooper, C.A., Joshi, H.J., Harrison, M.J., Wilkins, M.R., and Packer, N.H. (2003). GlycoSuiteDB: a
1207 curated relational database of glycoprotein glycan structures and their biological sources. 2003
1208 update. Nucleic Acids Res. *31*, 511–513.

1209 133. Al Jadda, K., Porterfield, M.P., Bridger, R., Heiss, C., Tiemeyer, M., Wells, L., Miller, J.A., York, W.S.,
1210 and Ranzinger, R. (2015). EUROCarbDB(CCRC): a EUROCarbDB node for storing glycomics standard
1211 data. Bioinformatics *31*, 242–245.

1212 134. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and
1213 Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

1214 135. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer,

1215 T.A.P., Rempfer, C., Bordoli, L., et al. (2018). SWISS-MODEL: homology modelling of protein
1216 structures and complexes. Nucleic Acids Res. *46*, W296–W303.

1217 136. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER Suite: protein
1218 structure and function prediction. Nat. Methods *12*, 7–8.

1219 137. Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N.,
1220 Potter, S.C., Finn, R.D., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019.
1221 Nucleic Acids Res. *47*, W636–W641.

1222 138. Zhang, Y., and Sagui, C. (2015). Secondary structure assignment for conformationally irregular
1223 peptides: comparison between DSSP, STRIDE and KAKSI. J. Mol. Graph. Model. *55*, 72–84.

1224 139. Magnan, C.N., and Baldi, P. (2014). SSpro/ACCpro 5: almost perfect prediction of protein secondary
1225 structure and relative solvent accessibility using profiles, machine learning and structural similarity.
1226 Bioinformatics *30*, 2592–2597.

1227 140. Klein, J., and Zaia, J. (2019). glypy: An Open Source Glycoinformatics Library. J. Proteome Res. *18*,
1228 3532–3537.

1229 141. Velankar, S., Dana, J.M., Jacobsen, J., van Ginkel, G., Gane, P.J., Luo, J., Oldfield, T.J., O'Donovan, C.,
1230 Martin, M.-J., and Kleywegt, G.J. (2013). SIFTS: Structure Integration with Function, Taxonomy and
1231 Sequences resource. Nucleic Acids Res. *41*, D483-9.

1232 142. Zeger, S.L., and Liang, K.-Y. (1986). Longitudinal Data Analysis for Discrete and Continuous
1233 Outcomes. Preprint, 10.2307/2531248 10.2307/2531248.

1234 143. Benaglia, T., Chauveau, D., Hunter, D.R., and Young, D. (2009). mixtools: AnRPackage for Analyzing
1235 Finite Mixture Models. Preprint, 10.18637/jss.v032.i06 10.18637/jss.v032.i06.

1236 144. Balakrishnan, S., Kamisetty, H., Carbonell, J.G., Lee, S.-I., and Langmead, C.J. (2011). Learning
1237 generative models for protein fold families. Proteins *79*, 1061–1078.

1238 145. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., and Aurell, E. (2013). Improved contact prediction in
1239 proteins: using pseudolikelihoods to infer Potts models. Phys. Rev. E Stat. Nonlin. Soft Matter Phys.
1240 *87*, 012707.

1241 146. Weinreb, C., Riesselman, A.J., Ingraham, J.B., Gross, T., Sander, C., and Marks, D.S. (2016). 3D RNA
1242 and Functional Interactions from Evolutionary Couplings. Cell *165*, 963–975.

1243 147. Hopf, T.A., Green, A.G., Schubert, B., Mersmann, S., Schärfe, C.P.I., Ingraham, J.B., Toth-Petroczy,
1244 A., Brock, K., Riesselman, A.J., Palmedo, P., et al. (2019). The EVcouplings Python framework for
1245 coevolutionary sequence analysis. Bioinformatics *35*, 1582–1584.

1246 148. Wright, E.S. (2016). Using DECIPHER v2. 0 to analyze big biological sequence data in R. R J. *8*.

1247 149. Wagih, O. (2017). ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics *33*,
1248 3645–3647.

1249 150. Morahan, G., Balmer, L., and Monley, D. (2008). Establishment of "The Gene Mine": a resource for

1250 rapid identification of complex trait genes. Mamm. Genome *19*, 390–393.

1251 151. Jansen, B.C., Falck, D., de Haan, N., Hipgrave Ederveen, A.L., Razdorov, G., Lauc, G., and Wuhrer, M.
1252 (2016). LaCyTools: A Targeted Liquid Chromatography–Mass Spectrometry Data Processing Package
1253 for Relative Quantitation of Glycopeptides. J. Proteome Res. *15*, 2198–2210.

1254 152. He, L., Diedrich, J., Chu, Y.-Y., and Yates, J.R., 3rd (2015). Extracting accurate precursor information
1255 for tandem mass spectra by RawConverter. Anal. Chem. *87*, 11361–11367.

1256 153. Xu, T., Park, S.K., Venable, J.D., Wohlschlegel, J.A., Diedrich, J.K., Cociorva, D., Lu, B., Liao, L., Hewel,
1257 J., Han, X., et al. (2015). ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity
1258 and specificity. J. Proteomics *129*, 16–24.

1259 154. Tabb, D.L., McDonald, W.H., and Yates, J.R., 3rd (2002). DTASelect and Contrast: tools for
1260 assembling and comparing protein identifications from shotgun proteomics. J. Proteome Res. *1*,
1261 21–26.

1262 155. Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J., and Gygi, S.P. (2003). Evaluation of
1263 multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for
1264 large-scale protein analysis: the yeast proteome. J. Proteome Res. *2*, 43–50.

1265 156. Park, S.K., Venable, J.D., Xu, T., and Yates, J.R., 3rd (2008). A quantitative analysis software tool for
1266 mass spectrometry-based proteomics. Nat. Methods *5*, 319–322.

1267 157. Irvine, E.B., Peters, J.M., Lu, R., Grace, P.S., Sixsmith, J., Wallace, A., Schneider, M., Shin, S.,
1268 Karpinski, W., Hsiao, J.C., et al. (2022). Fc-engineered antibodies leverage neutrophils to drive
1269 control of *Mycobacterium tuberculosis*. bioRxiv. 10.1101/2022.05.01.490220.

1270 158. Engler, C., and Marillonnet, S. (2014). Golden Gate cloning. Methods Mol. Biol. *1116*, 119–131.

1271 159. Fang, J., Qian, J.-J., Yi, S., Harding, T.C., Tu, G.H., VanRoey, M., and Jooss, K. (2005). Stable antibody
1272 expression at therapeutic levels using the 2A peptide. Nat. Biotechnol. *23*, 584–590.

1273 160. Mahan, A.E., Tedesco, J., Dionne, K., Baruah, K., Cheng, H.D., De Jager, P.L., Barouch, D.H.,
1274 Suscovich, T., Ackerman, M., Crispin, M., et al. (2015). A method for high-throughput, sensitive
1275 analysis of IgG Fc and Fab glycosylation by capillary electrophoresis. J. Immunol. Methods *417*, 34–
1276 44.

1277 161. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and
1278 Powerful Approach to Multiple Testing. J. R. Stat. Soc. Series B Stat. Methodol. *57*, 289–300.

1279 162. Xu, C., and Ng, D.T.W. (2015). Glycosylation-directed quality control of protein folding. Nat. Rev.
1280 Mol. Cell Biol. *16*, 742–752.

1281 163. Shental-Bechor, D., and Levy, Y. (2008). Effect of glycosylation on protein folding: a close look at
1282 thermodynamic stabilization. Proc. Natl. Acad. Sci. U. S. A. *105*, 8256–8261.

1283 164. Shihab, H.A., Gough, J., Mort, M., Cooper, D.N., Day, I.N.M., and Gaunt, T.R. (2014). Ranking non-
1284 synonymous single nucleotide polymorphisms based on disease concepts. Hum. Genomics *8*, 11.

1285 165. Shihab, H.A., Gough, J., Cooper, D.N., Day, I.N.M., and Gaunt, T.R. (2013). Predicting the functional
1286 consequences of cancer-associated amino acid substitutions. Bioinformatics *29*, 1504–1510.

1287 166. Shihab, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L.A., Edwards, K.J., Day, I.N.M., and
1288 Gaunt, T.R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino
1289 acid substitutions using hidden Markov models. Hum. Mutat. *34*, 57–65.

1290 167. Baković, M.P., Selman, M.H.J., Hoffmann, M., Rudan, I., Campbell, H., Deelder, A.M., Lauc, G., and
1291 Wuhrer, M. (2013). High-throughput IgG Fc N-glycosylation profiling by mass spectrometry of
1292 glycopeptides. J. Proteome Res. *12*, 821–831.

1293 168. Pucić, M., Knezević, A., Vidic, J., Adamczyk, B., Novokmet, M., Polasek, O., Gornik, O., Supraha-
1294 Goreta, S., Wormald, M.R., Redzić, I., et al. (2011). High throughput isolation and glycosylation
1295 analysis of IgG-variability and heritability of the IgG glycome in three isolated human populations.
1296 Mol. Cell. Proteomics *10*, M111.010090.

1297 169. Klarić, L., Tsepilov, Y.A., Stanton, C.M., Mangino, M., Sikka, T.T., Esko, T., Pakhomov, E., Salo, P.,
1298 Deelen, J., McGurnaghan, S.J., et al. (2020). Glycosylation of immunoglobulin G is regulated by a
1299 large network of genes pleiotropic with inflammatory diseases. Sci. Adv. *6*, eaax0301.

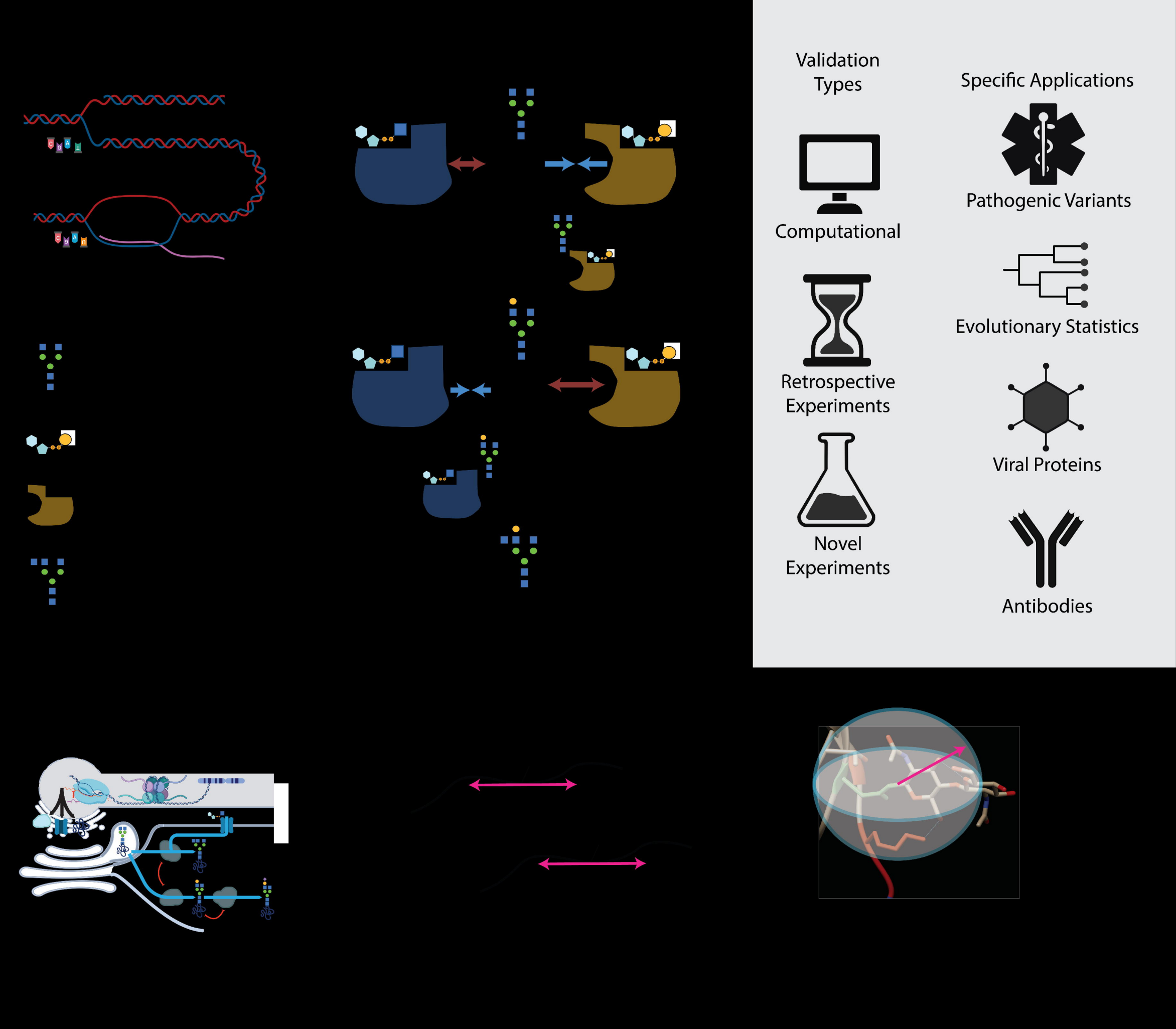
1300 170. Rose, R.J., van Berkel, P.H.C., van den Bremer, E.T.J., Labrijn, A.F., Vink, T., Schuurman, J., Heck,
1301 A.J.R., and Parren, P.W.H.I. (2013). Mutation of Y407 in the CH3 domain dramatically alters
1302 glycosylation and structure of human IgG. MAbs *5*, 219–228.

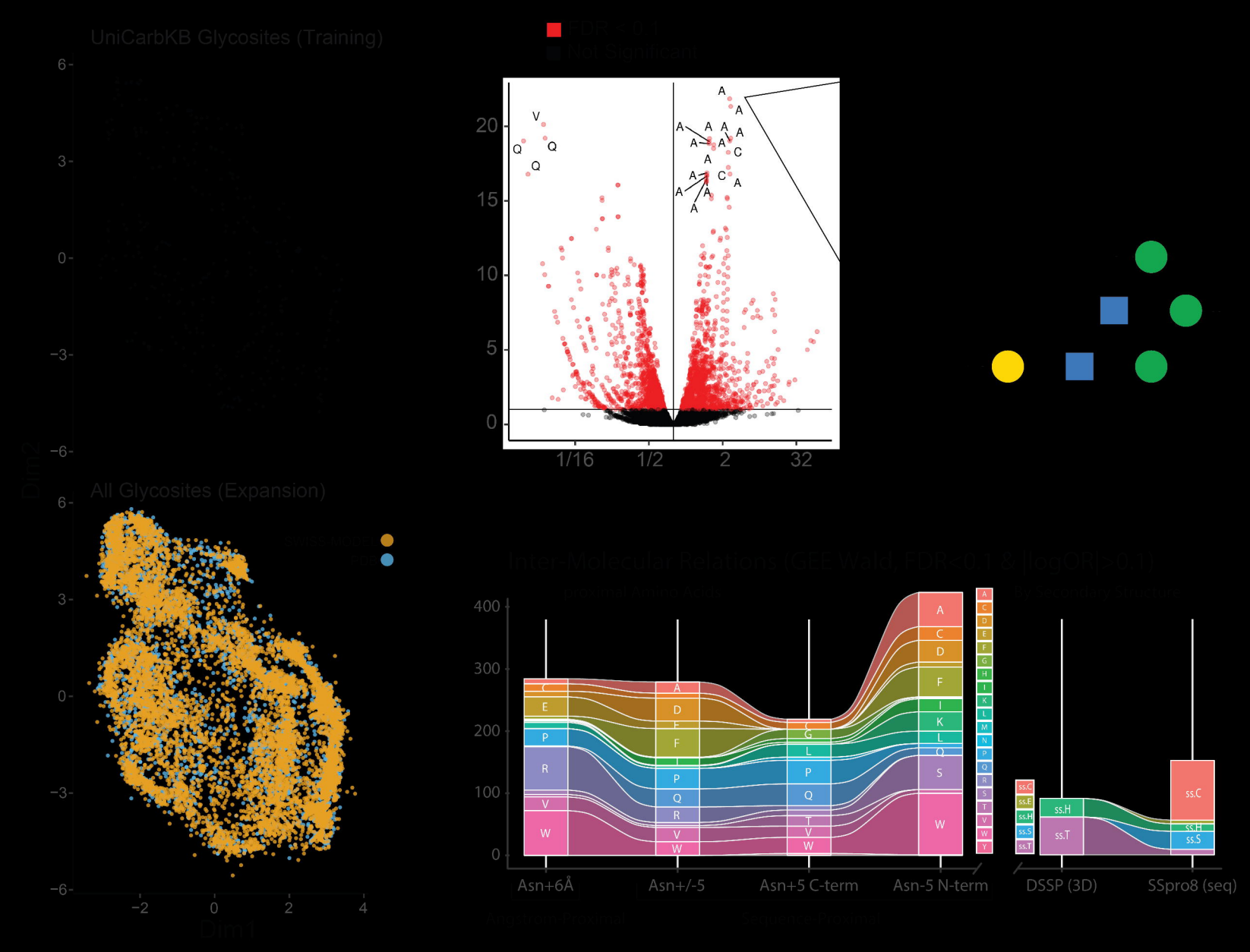
1303 171. Vidarsson, G., Dekkers, G., and Rispens, T. (2014). IgG subclasses and allotypes: from structure to
1304 effector functions. Front. Immunol. *5*, 520.

1305 172. Hutton, S.M., and Spritz, R.A. (2008). Comprehensive analysis of oculocutaneous albinism among
1306 non-Hispanic caucasians shows that OCA1 is the most prevalent OCA type. J. Invest. Dermatol. *128*,
1307 2442–2450.

1308 173. Rademacher, C., Bru, T., McBride, R., Robison, E., Nycholat, C.M., Kremer, E.J., and Paulson, J.C.
1309 (2012). A Siglec-like sialic-acid-binding motif revealed in an adenovirus capsid protein. Glycobiology
1310 *22*, 1086–1091.

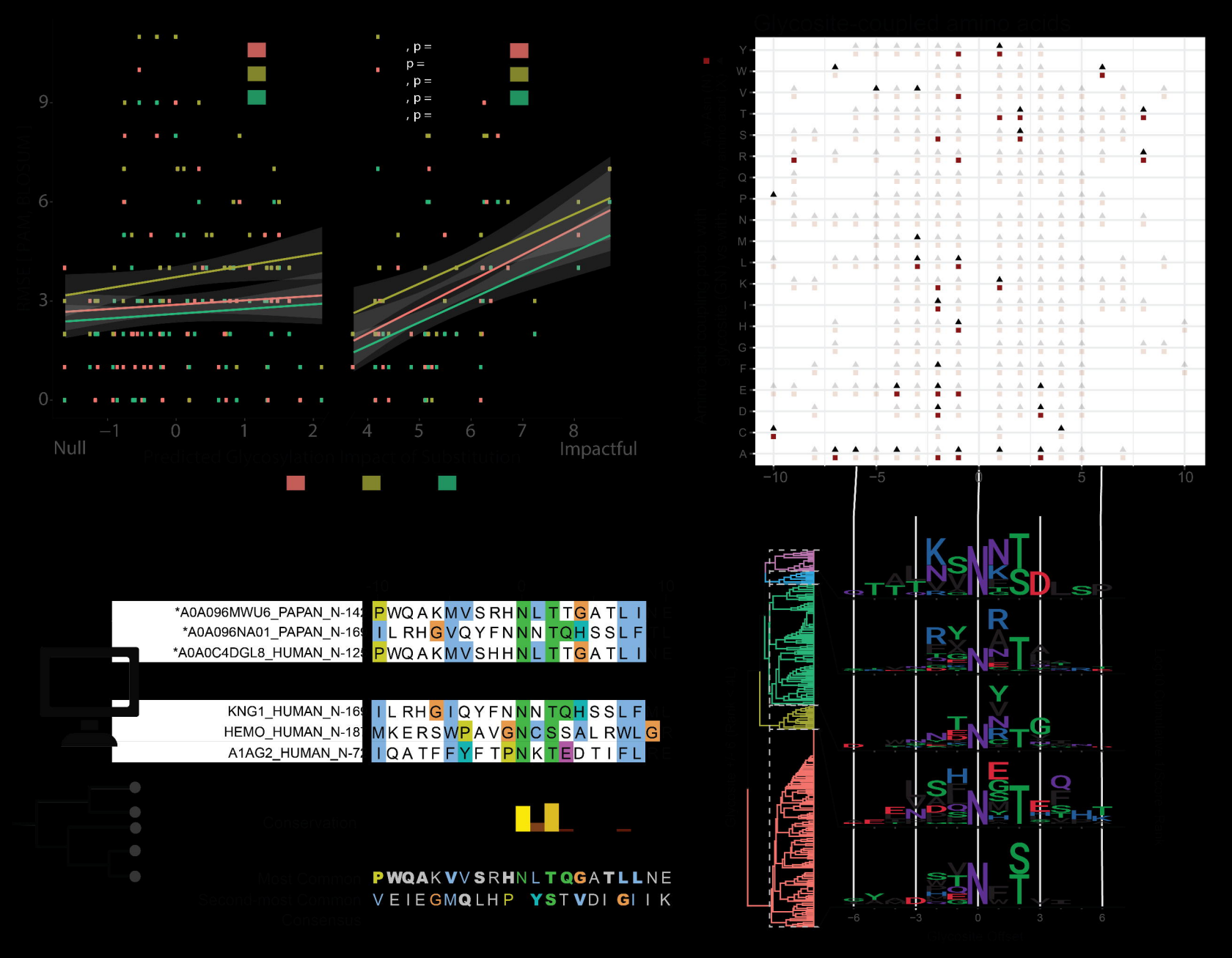
1311 174. Chavent, M., Kuentz-Simonet, V., Labenne, A., and Saracco, J. (2014). Multivariate analysis of mixed
1312 data: The R package PCAmixdata. arXiv [stat.CO].

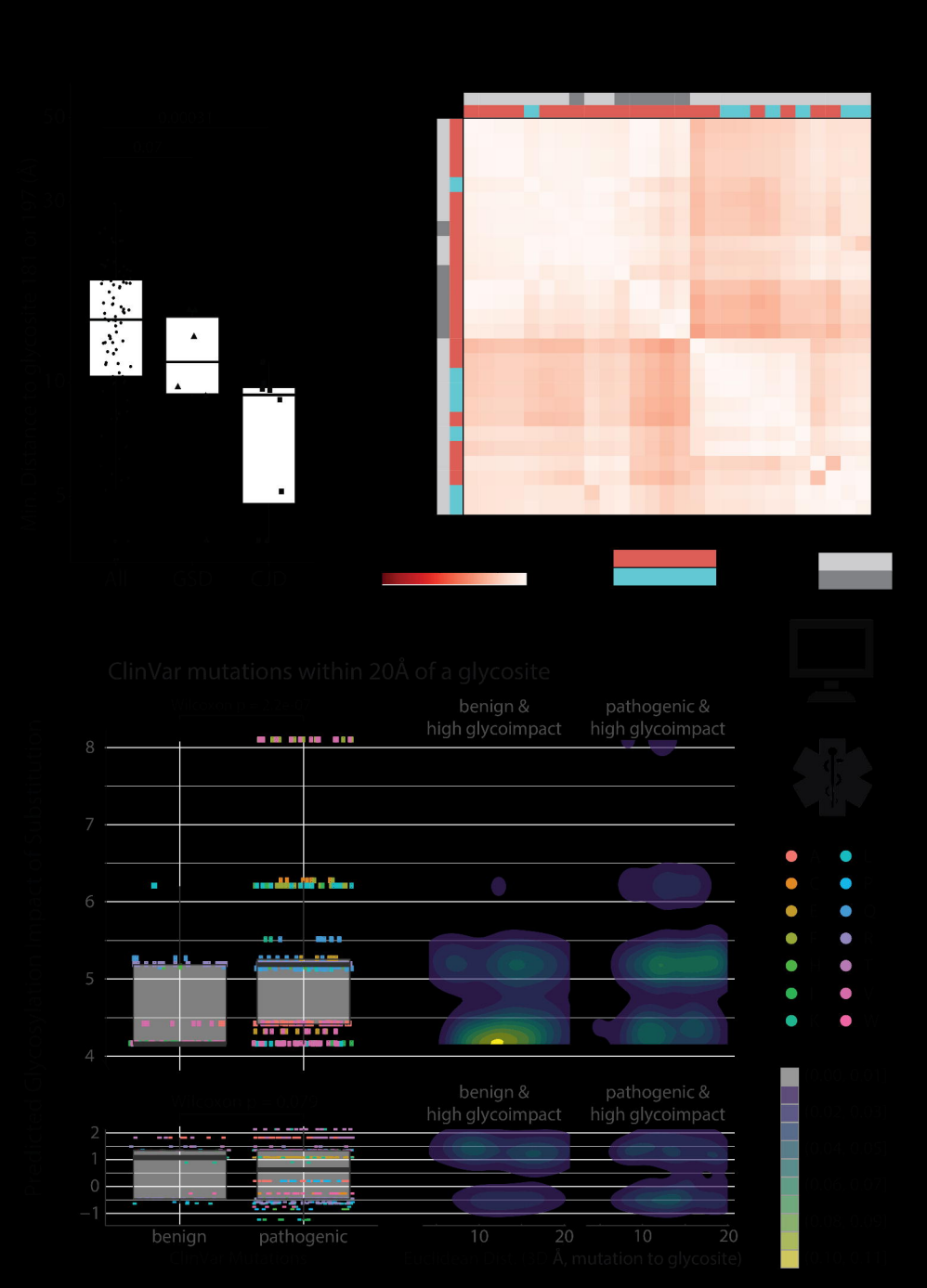

Validation Types
Computational
Retrospective Experiments
Novel Experiments
Specific Applications
Pathogenic Variants
Evolutionary Statistics
Viral Proteins
Antibodies



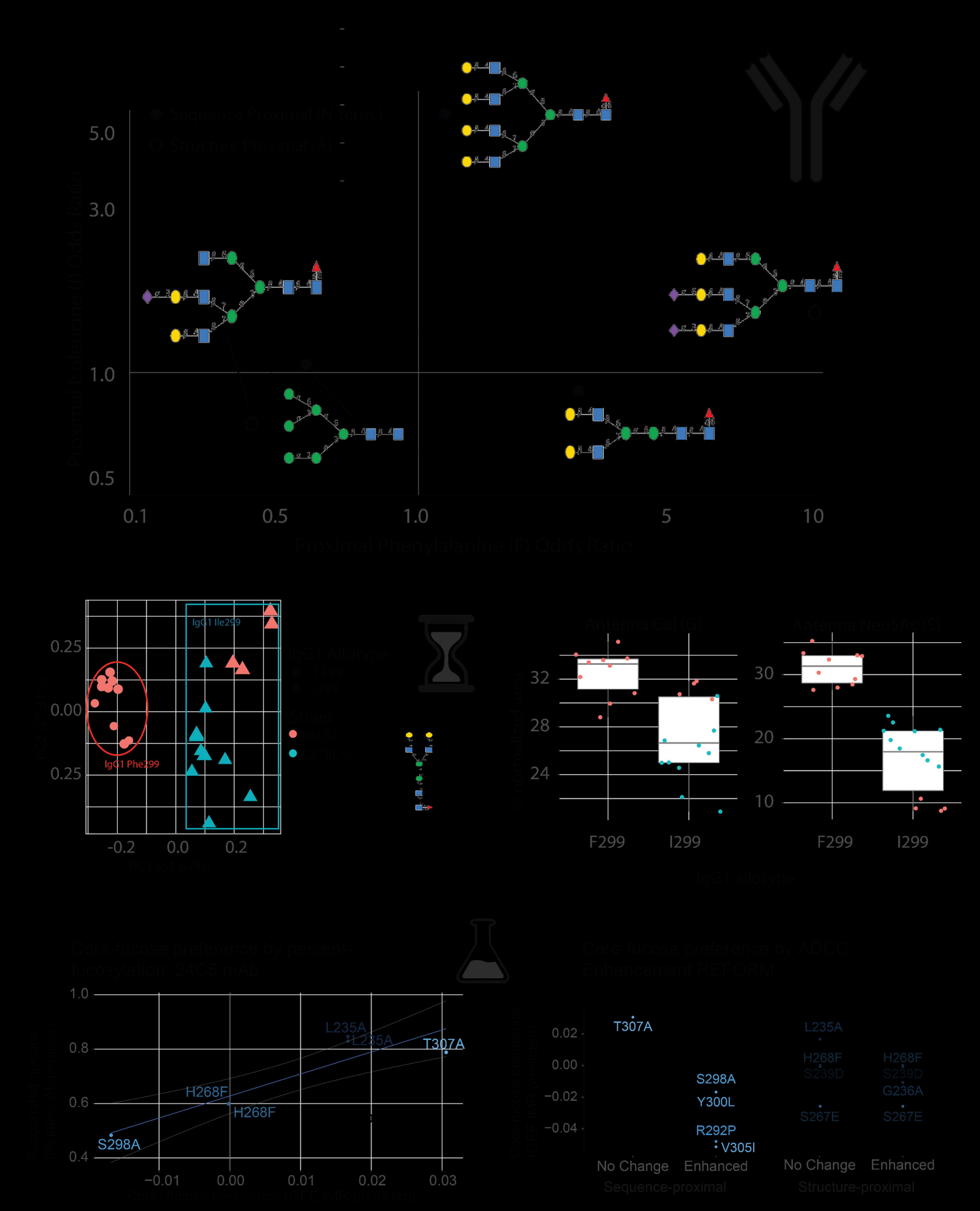
UniCarbKB Glycosites (Training)
All Glycosites (Expansion)
Inter-Molecular Relations (GEE Wald, FDR<0.1 & |logOR|>0.1)
Asn+6Å
Asn+/-5
Asn+5 C-term
Asn-5 N-term
DSSP (3D)
SSpro8 (seq)

Å-Proximal Amino Acid Effects
10.0
1.0
0.1
0
1
2
E
L
M
P
Q
S
W
Å Proximal Amino Acid Effects
H
T
V
Spear Corr
b-GlcNAc
b-Gal
a-Man
b-Man
a-NeuNAc
a-GalNAc
a-Fuc
b-GalNAc
2.0
1.0
0.8
0.5
3.0
0.1
N-terminal F (GEE Odds Ratio)
Å Proximal F (GEE Odds Ratio)
6Å
Expected Glycosylation Impact of Substitution
Difference in Å Proximal AA Effects
Glycoimpact
A
C
V
T
I
L
W
F
E
H
R
Q
K
BLOSUM62


, p =
p =
, p =
, p =
, p =
Null
Impactful
Predicted Glycosylation Impact of Substitution
Glycosite-coupled amino acids
*A0A096MWU6_PAPAN_N-14
*A0A096NA01_PAPAN_N-16
*A0A0C4DGL8_HUMAN_N-12
KNG1_HUMAN_N-16
HEMO_HUMAN_N-18
A1AG2_HUMAN_N-7
PWQAKMVSRHNLTTGATLI
ILRHGVQYFNNNTQHSSLF
PWQAKMVSHHNLTTGATLI
ILRHGIQYFNNNTQHSSLF
MKERSWPAVGNCSSALRWLG
IQATFFYFTPNKTEDTIFL
Conservation
Most Common
Second-most Common
Consensus
PWQAKVVSRHNLTQGATLLNE
VEIEGMQLHP YSTVDI GIIK


ClinVar mutations within 20Å of a glycosite
benign & high glycoimpact
pathogenic & high glycoimpact
benign
pathogenic
10
20

Q
P
ssC
K
ssT
R
C
N
G
L
F
G
8
6
4
2
0
0.5
1.0
2.0
Oligomannose & Hybrid Odds Ratio
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
Number of Glycosite-Proximal AA in HIV gp160
Wilcoxon
p = 0.0033
Wilcoxon
p = 1e-04
Kruskal-Wallis
p = 0.21
P within 6 A
C-Terminal Q
K within 6 A
SARS-COV-2 Spike glycosylation
N122
N331
N657
Mean Proportion
N3
N203


Sequence Proximal (N-term)
Structure Proximal (Å)
Proximal Isoleucine (I) Odds Ratio
5.0
3.0
1.0
0.5
0.1
0.5
1.0
5
10
Proximal Phenylalanine (F) Odds Ratio
IgG1 Ile299
IgG1 Phe299
IgG1 Allotype
F299
I299
Strain
BALB/c
C57BL/6
0.25
0.00
0.25
-0.2
0.0
0.2
Antenna Gal (G)
Antenna Neu5Ac (S)
32
28
24
30
20
10
F299
I299
IgG1 allotype
Core fucose preference by percent fucosylation: 2A10.5 mAb
Core fucose preference by ADCC Enhancement: REFORM
1.0
0.8
0.6
0.4
L235A
L235A
T307A
H268F
H268F
S298A
−0.01
0.00
0.01
0.02
0.03
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