MOLECULAR BIOLOGY

N-glycosylation as a eukaryotic protective mechanism against protein aggregation

Ramon Duran-Romaña[1,2], Bert Houben[1,2], Matthias De Vleeschouwer[1,2], Nikolaos Louros[1,2], Matthew P. Wilson[3], Gert Matthijs[3], Joost Schymkowitz[1,2]*, Frederic Rousseau[1,2]*

The tendency for proteins to form aggregates is an inherent part of every proteome and arises from the self-assembly of short protein segments called aggregation-prone regions (APRs). While posttranslational modifications (PTMs) have been implicated in modulating protein aggregation, their direct role in APRs remains poorly understood. In this study, we used a combination of proteome-wide computational analyses and biophysical techniques to investigate the potential involvement of PTMs in aggregation regulation. Our findings reveal that while most PTM types are disfavored near APRs, N-glycosylation is enriched and evolutionarily selected, especially in proteins prone to misfolding. Experimentally, we show that N-glycosylation inhibits the aggregation of peptides in vitro through steric hindrance. Moreover, mining existing proteomics data, we find that the loss of N-glycans at the flanks of APRs leads to specific protein aggregation in Neuro2a cells. Our findings indicate that, among its many molecular functions, N-glycosylation directly prevents protein aggregation in higher eukaryotes.

INTRODUCTION

The conversion of soluble functional proteins into structured aggregates is triggered by short, generally hydrophobic, amino acid stretches known as aggregation-prone regions (APRs) (*1*). Most proteins contain one and usually several APRs. Around 20% of all residues in globular proteins are predicted to reside within these regions (*2*). In globular proteins, APRs are mostly buried inside the hydrophobic core, preventing them from initiating aggregation (*3*). However, under physiological stress or during translation and translocation, APRs are exposed to the solvent and are prone to aggregate, requiring rigorous regulation by the cellular proteostasis machinery (*4*, *5*). Intrinsically disordered proteins have fewer and more hydrophilic APRs as they lack the necessary hydrophobicity to make a globular fold (*6*). Nevertheless, their APRs can be highly susceptible to aggregation since they are constantly exposed, which may explain the prominent role of specific intrinsically disordered proteins, such as tau and α-synuclein, in aggregation-related disorders. Insoluble aggregates, from either globular or intrinsically disordered proteins, lead to the loss of function of the affected proteins and are often toxic to cells. This toxicity is strongly associated with a wide range of human diseases and aging, including Alzheimer's and Parkinson's diseases (*7*, *8*).

The evolutionary persistence of APRs is a result of their necessity for protein stability, as the forces that drive aggregation, i.e., hydrophobicity and β sheet propensity, are also crucial for the folding of globular proteins (*9*). Nevertheless, throughout evolution, the potency of APRs has been minimized by the presence of residues that suppress aggregation propensity, known as aggregation gatekeepers (*10*). Specifically, charged amino acids (Arg, Lys, Asp, and Glu) and proline (Pro) are enriched in the regions immediately flanking APRs, as they kinetically and thermodynamically disfavor aggregation (*2*, *11*–*14*). The introduction of charges generates repulsion forces that strongly reduce aggregation propensity, while Pro is incompatible with the β strand conformations associated with protein aggregation. Because of their antiaggregation properties, gatekeepers are essential to maintain the overall fitness of cells, as they affect protein synthesis and degradation rates and can even act as molecular signals to recruit chaperones to non-native states (*15*, *16*). Aggregation gatekeepers are evolutionarily conserved despite destabilizing the native structure, showing that these residues constitute a functional class specifically devoted to proteostasis (*17*). Accordingly, mutations that remove gatekeeper residues are more often associated with human diseases than neutral polymorphisms (*18*).

Many proteins are chemically modified during or shortly after translation to assist protein folding and increase the stability of the native structure. Given this intimate connection with protein folding, it is perhaps expected that protein co- and posttranslational modifications (herein referred to as PTMs) are gradually becoming associated with protein aggregation events. Several studies have shown that PTMs can directly—or indirectly—increase or decrease the aggregation potency of proteins associated with common aggregation diseases (*19*–*21*). For example, phosphorylation interferes directly with amyloid-β fibrillary structure maturation (*22*), whereas in tau molecules, it reduces microtubule binding affinity, thus increasing the concentration of soluble tau and resulting in later-stage aggregation (*23*). In recent years, the reversible O–*N*-acetylglucosamine (GlcNAc) modification has been shown to directly inhibit protein aggregation in many neurodegenerative diseases and indirectly promote cytoprotection against a wide range of cellular stresses (*24*, *25*). Nevertheless, it is unclear whether other PTM types constitute a general mechanism of aggregation prevention across proteomes.

The most abundant category of PTMs involves the enzymatic addition of functional groups to amino acid side chains, increasing their size and chemical complexity. Many PTM types have chemical properties reminiscent of gatekeeper residues as they often add bulk chains—likely incompatible with β-aggregation—and charges—potentially causing charge repulsion. Negatively charged residues (Asp and Glu) have historically been used to mimic the phosphorylated state of proteins, as phosphorylation adds a negative charge to

[1] Switch Laboratory, VIB Center for Brain and Disease Research, 3000 Leuven, Belgium. [2] Switch Laboratory, Department of Cellular and Molecular Medicine, KU Leuven, 3000 Leuven, Belgium. [3] Laboratory for Molecular Diagnosis, Center for Human Genetics, KU Leuven, 3000 Leuven, Belgium.
*Corresponding author. Email: joost.schymkowitz@kuleuven.be (J.S.); frederic.rousseau@kuleuven.be (F.R.)

the amino acid side chain (*26*). Furthermore, positively charged residues (Arg and Lys) are susceptible to many types of PTMs, such as acetylation or methylation. For these reasons, we hypothesize that PTMs could have been selected throughout evolution at the flanks of APRs as an intrinsic factor to protect proteins against aggregation, thus expanding the current repertoire of aggregation gatekeepers. In this work, we scanned the entire human proteome with a widely used protein aggregation prediction algorithm, TANGO, to analyze the frequency of the most abundant PTM types in APRs and their surrounding residues. Our findings show that N-glycosylation is enriched, is conserved, and commonly replaces unmodified gatekeeper residues at these positions. Using biophysical assays on N-glycosylated and unmodified aggregation-prone peptides, we show that this modification mitigates aggregation in vitro through steric hindrance. Analysis of the structural properties of proteins with APRs flanked by N-glycosylation indicates a preferential association with topologically complex domains that have a high aggregation propensity. Last, reanalysis of proteomics data that measures changes in protein solubility after treatment of mouse Neuro2a cells with an N-glycosylation inhibitor shows the aggregation of specific proteins.

RESULTS

While most PTM types are disfavored around APRs, N-glycosylation is enriched

Unmodified aggregation gatekeepers (Arg, Lys, Asp, Glu, and Pro) are enriched in the positions immediately surrounding APRs. At least one of these amino acids is found within the three neighboring residues—on either side—in more than 90% of all APRs identified by TANGO (*1*, *2*), a widely used protein aggregation predictor. Therefore, to investigate the potential role of the most common types of PTMs as aggregation gatekeepers, we calculated their relative enrichment in and around human APRs compared to the proteome average (Fig. 1A). First, human proteins were scanned with TANGO, which identified 84,537 APRs (TANGO score of >10 and length of 5 to 15 residues). The three residues preceding and succeeding APRs were labeled as gatekeeping regions (GRs) and all other residues as distal regions (DRs). GRs were further labeled as GR1, GR2, or GR3 and as N terminus (N-ter) or C terminus (C-ter) based on their relative position to the APRs. Next, PTM sites that have been experimentally identified in human proteins were collected from dbPTM (*27*) and the O-GlcNAcAtlas (*28*) and were mapped to the dataset. Only PTM types with enough observations to ensure accurate statistics were kept (at least 1200 sites), which resulted in 17 PTM types across 571,759 unique sites (table S1).

Our findings show that PTMs, in general, are highly depleted in APRs and GRs (Fig. 1, B and C), which means that most PTM types occur more frequently in residues that are located far away from APRs. This is expected as APRs are normally partially or completely buried in the folded structure, while PTM sites must be solvent accessible to be recognized by their modifying enzyme (fig. S1) (*29*, *30*). Another protein property that has been strongly associated with the occurrence of PTMs is structure disorder (*30*). However, APRs and their GRs are predominantly found in structured domains, which could explain why PTM types that are often observed in intrinsically disordered regions, such as phosphorylation or O-glycosylation, are disfavored (fig. S2, A and B). Therefore, we repeated the analysis only with residues that are solvent accessible and hence more readily modified and/or with residues that are structurally ordered. Nevertheless, similar enrichment patterns were observed, suggesting that most PTM types are not intended to protect APRs against aggregation (figs. S2C, S3, and S4). This is also the case for O-GlcNAcylation, despite reports showing that this modification markedly slows down the aggregation of specific proteins involved in neurodegeneration, including tau and α-synuclein (*24*). One reason for this is that O-GlcNAcylation protects against aggregation in a limited set of proteins, so it may not be a general antiaggregation mechanism across proteins. In addition, most proteins in which O-GlcNAcylation has been found to protect against aggregation are intrinsically disordered. Therefore, their aggregation is mainly driven by hydrophilic regions with high β sheet propensity (*10*), which are more difficult to identify by computational aggregation predictors (*6*).

In contrast to all other PTM types analyzed, N-glycosylation is substantially enriched in APRs and GRs, especially at the N-terminal side (Fig. 1C). Moreover, restricting the analysis only to exposed and/or structurally ordered residues further increased this enrichment (figs. S2C, S3, and S4).

N-glycosites flanking APRs are evolutionarily selected

N-glycosylation is one of the most common protein modifications in eukaryotic cells. It occurs in nearly all proteins that enter the secretory pathway (SP) and has essential roles in protein folding and quality control (*31*, *32*). The attachment of an N-glycan to an asparagine residue requires the recognition of a canonical sequence motif or sequon (Asn-X-Thr/Ser, where X ≠ Pro). This reaction is catalyzed by an oligosaccharyltransferase (OST) on the luminal side of the endoplasmic reticulum (ER), often cotranslationally.

Since TANGO is a sequence-based predictor, we assessed whether the enrichment detected above was an artifact stemming from the Asn-X-Thr/Ser sequon being polar—and hence likely to be recognized as a gatekeeper when it flanks an APR—instead of a biological signal from the N-glycan. To check this, we compared the relative enrichment of sequons in proteins that have been experimentally determined to undergo N-glycosylation (SP glycosylated) to sequons that are either not glycosylated (SP nonglycosylated) or cannot be glycosylated because of their subcellular location (non-SP). An enrichment was only observed in APRs and GRs for glycosylated sequons (Fig. 2A). This is highlighted in transmembrane proteins from the SP, as only those sequons in domains predicted to be in the extracellular or the lumenal side, and therefore can get glycosylated, showed an enrichment in these regions (fig. S5, A and B). Moreover, the enrichment was not present in sequons of artificial protein sequences that were randomly generated keeping the same amino acid composition of SP proteins (SP randomized) or non-SP proteins (non–SP randomized), further indicating that it does not arise from sequence bias (Fig. 2A). Since the comparison between DRs, which include the majority of residues in a protein, to GRs and APRs can exaggerate the statistical significance of the enrichments, we repeated the analysis by focusing only on DRs that are 4 to 10 amino acids away from the APRs. Despite this adjustment, glycosylated sequons were still enriched N-terminally of APRs, and, overall, the enrichment profiles of the different protein groups remained largely unchanged (fig. S5C). Last, we observed similar enrichment patterns when using a different aggregation predictor [CamSol (*33*); fig. S5D]. Together, these results indicate that the enrichment of glycosylated sequons observed in APRs and GRs neither arises from a bias due to

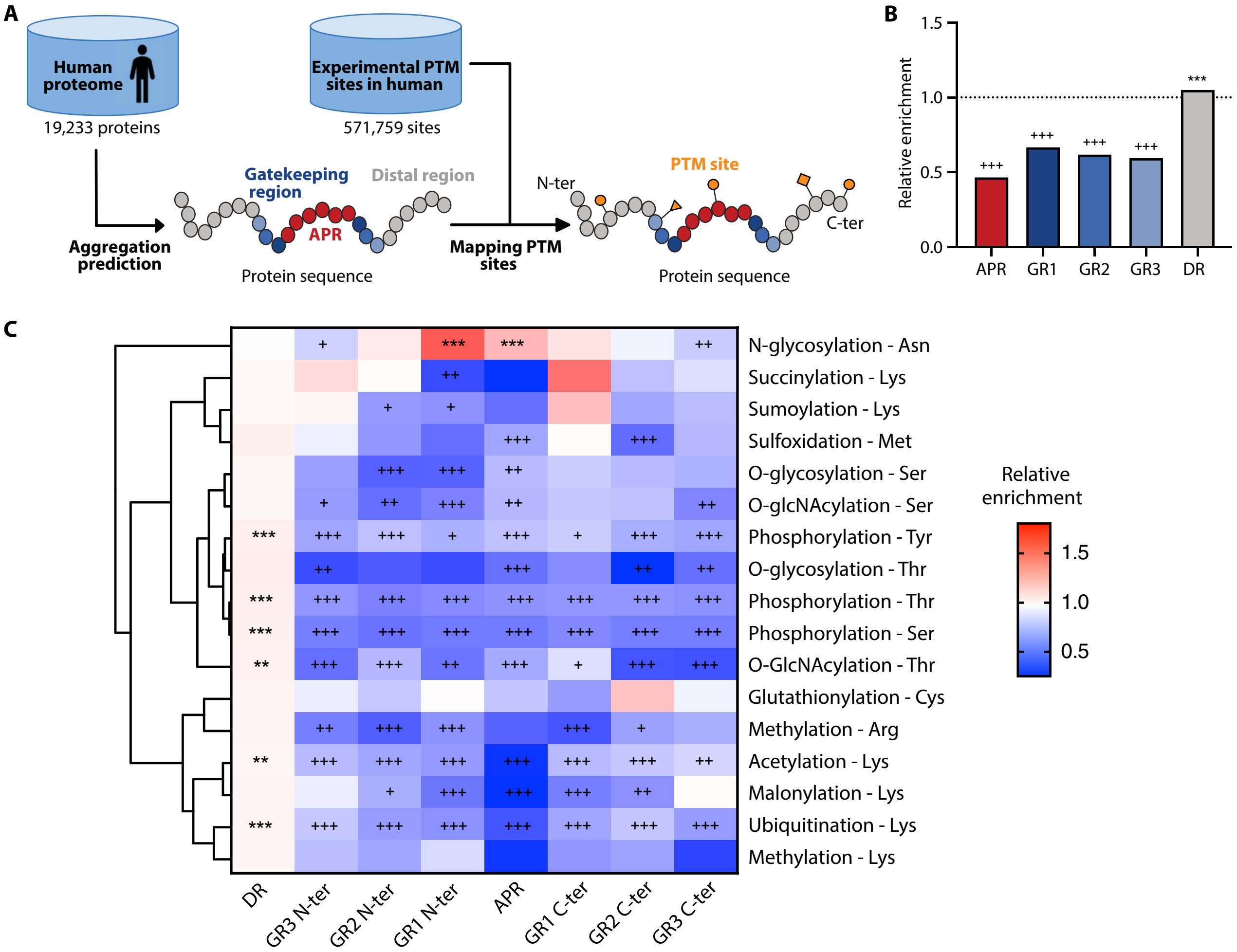


Fig. 1. Relative enrichment of different PTM types in APRs and GRs. (**A**) Flow chart illustrating the preparation of the dataset. (**B**) Bar plot showing the relative enrichment (odds ratio) of all PTM sites in APRs, GRs, and DRs relative to the background (all protein regions). (**C**) Heatmap showing the relative enrichment for each of the 17 types of PTMs. Columns indicate the different protein regions, while rows show the PTM types. Rows are clustered on the basis of Pearson correlation as a distance measure. The number of observations for each PTM type can be found in table S1. Statistical significance was determined by Fisher's exact test with false discovery rate correction [(B) and (C)]. Crosses and asterisks indicate that a region has a significantly lower (higher) frequency than the background. + and *$P \leq 0.05$, ++ and **$P \leq 0.01$, and +++ and ***$P \leq 0.001$.

the sequon composition nor the choice of the aggregation predictor and, instead, is a direct result of N-glycosylation. Calculating the ratio between the relative enrichments of glycosylated sequons against the relative enrichments of nonglycosylated sequons showed that there are three regions under positive selective pressure to be glycosylated, which we named enriched positions (EPs): GR2 N-ter, GR1 N-ter, and APR (Fig. 2B). There are 1155 N-glycosites in EPs distributed in 858 unique proteins (14% of all SP proteins; Fig. 2C). A list containing all human N-glycosites at EPs is provided in table S2.

N-glycosylation efficiency is highly influenced by the primary sequence context of glycosylation acceptor sites (*34*, *35*). Therefore, the specific sequence composition of APRs, GR2 N-ter, and GR1 N-ter could favor glycosylation efficiency. That is, the strong selection observed at EPs might arise from the OST binding preferentially to them. To assess this, we predicted the glycosylation efficiency of human N-glycosites using a model developed by Huang *et al.* (*36*). Briefly, the authors used site-directed saturation mutagenesis to determine which residues improved or suppressed N-glycosylation efficiency. On the basis of their model, N-glycosites in EPs are predicted to have a slightly lower N-glycosylation efficiency compared to other N-glycosites (Fig. 2D). This suggests that sequence composition is not driving the overabundance of glycosylated sites in these regions and, thus, hints at an actual shared functional role of N-glycosylation on these sites. To corroborate this, we looked at the conservation of human sequons in a dataset of 100 mammalian species from the UCSC genome browser (*37*), as high conservation is commonly associated with an essential biological function. Specifically, we analyzed whether a canonical N-glycosylation motif was present in the different mammalian species at the exact positions that are aligned to human sequons. As suspected, N-glycosites in EPs have higher conservation compared to all other N-glycosites, as well as to nonglycosylated sequons (Fig. 2E).

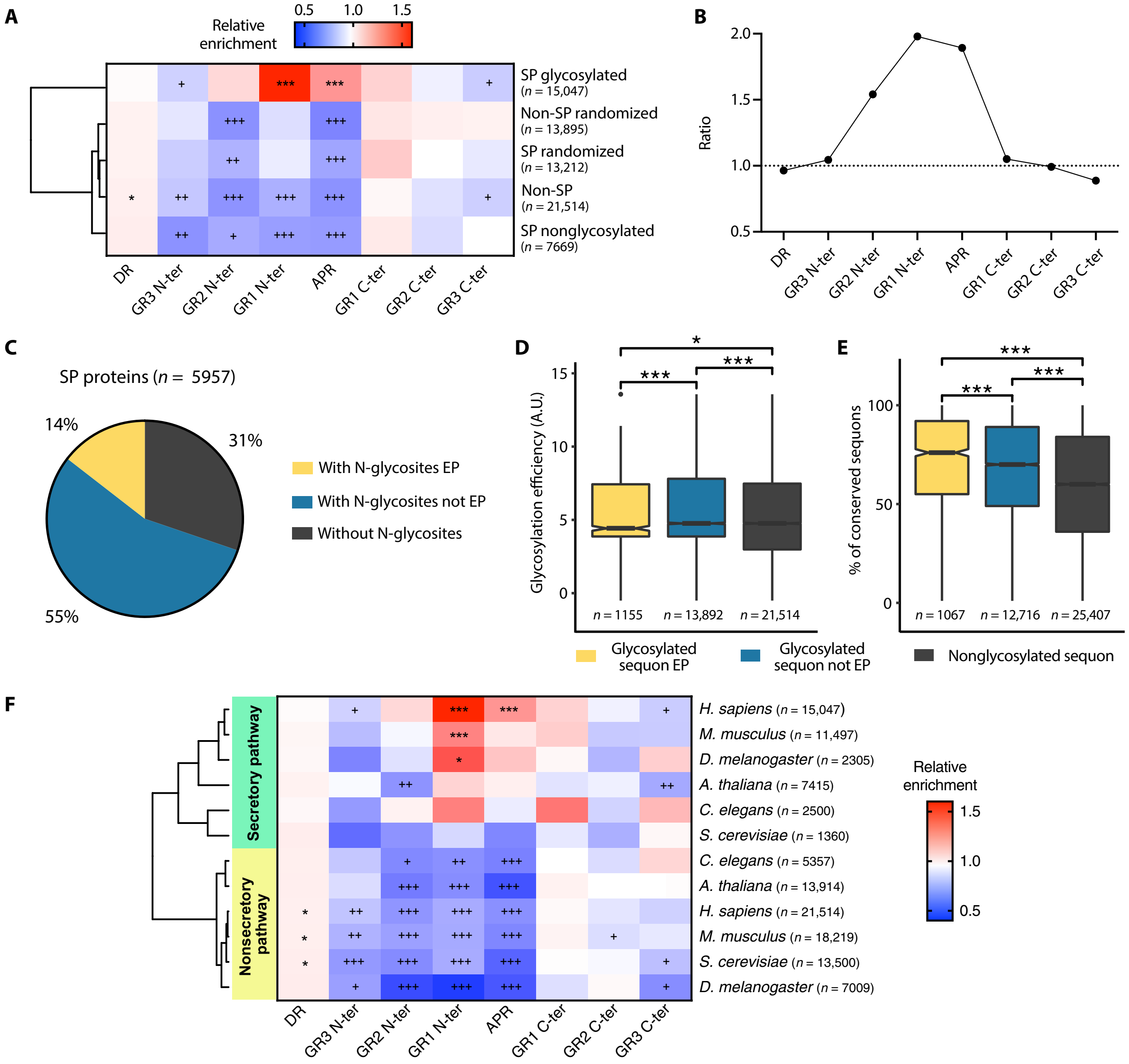


Fig. 2. Functional assessment of N-glycosylation in APRs and GRs. (**A**) Heatmap showing the relative enrichment of finding a sequon in each region (columns) for different groups of proteins (rows). Rows are clustered on the basis of Pearson correlation as a distance measure. The number of sequons observed in each group is indicated. (**B**) Ratio between the relative enrichments of glycosylated sequons versus nonglycosylated sequons. (**C**) Fraction of known SP proteins with at least one N-glycosite in EPs (yellow), with N-glycosites that are not in EPs (blue) or without N-glycosites (back). (**D**) Box plot showing the glycosylation efficiency of glycosylated sequons in EPs (yellow), rest of glycosylated sequons (blue), and nonglycosylated sequons (black) of human proteins. A.U., arbitrary units. (**E**) Box plot showing the conservation of human sequons in a set of 100 mammalian species for the same categories as (D). (**F**) Heatmap showing the relative enrichment of finding a sequon in each region (columns) for SP and non-SP proteins in five different eukaryotic species. For all species, the enrichment profiles of SP proteins are clustered together, while the same is true for non-SP proteins. Rows are clustered on the basis of Pearson correlation as a distance measure. The number of sequons observed in each group is indicated. Statistical significance was determined by Fisher's exact test with false discovery rate correction [(A) and (F)] or by unpaired Wilcoxon test with Bonferroni correction for multiple comparisons [(D) and (E)]. Crosses and asterisks in the heatmaps indicate that a region has a significantly lower (higher) frequency than the background. + and *$P \leq 0.05$, ++ and **$P \leq 0.01$, and +++ and ***$P \leq 0.001$.

The N-glycosylation pathway in the ER is very conserved across all eukaryotes (38, 39). Therefore, we next investigated whether a similar enrichment pattern was present in other eukaryote model organisms. Given that the number of experimentally verified N-glycosites in nonhuman species is very low, we assumed all sequons in SP proteins to be glycosylated. Notably, a similar enrichment pattern was found for sequons in SP proteins of other animals (*Mus musculus*, *Drosophila melanogaster*, and *Caenorhabditis elegans*) and plants (*Arabidopsis thaliana*), clustering together with the human SP enrichment profile (Fig. 2F). Similarly, in these species, sequons of proteins that cannot get glycosylated (non-SP) were not enriched at EPs. For yeast (*Saccharomyces cerevisiae*), although its SP enrichment profile clustered together with the rest of SP profiles, no enrichment was observed at these positions.

All of the above underlines a high selective pressure for N-glycosites in EPs to be preserved in evolution, pointing to a similar functional role for N-glycans in these sites in higher eukaryotes. Since protein aggregation is generally detrimental for cells, we hypothesized that N-glycosylation is selected in these positions to protect against aggregation. That is, this modification could be a previously unrecognized class of aggregation gatekeeper.

N-glycosites flanking APRs behave as and replace aggregation gatekeeper residues

The presence and number of unmodified gatekeeping residues (Arg, Lys, Asp, Glu, and Pro) flanking an APR correlate strongly with its aggregation propensity (*40*). To investigate whether N-glycosites flanking APRs act as aggregation gatekeepers, we analyzed the aggregation propensity (TANGO score) of APRs containing glycosylated and nonglycosylated sequons at EPs. We found that APRs flanked by N-terminally glycosylated sequons at GR1 N-ter and GR2 N-ter have significantly higher aggregation propensities than those flanked by nonglycosylated sequons in the same positions (Fig. 3A). However, despite having a higher aggregation propensity on average, these APRs are flanked by significantly fewer unmodified gatekeeping residues (Fig. 3B). Moreover, while for nonglycosylated sequons the number of unmodified gatekeepers increases with APR strength, for glycosylated sequons the number remains low and constant across different APR strength bins (fig. S6A). Since unmodified gatekeeping residues are crucial to avoid aggregation, especially for very strong APRs, these data suggest that N-glycans are replacing them in these positions, thus potentially taking their function as aggregation breakers. In contrast, glycosylated sequons in GRs that are not under selective pressure (GR3 N-ter, GR1 C-ter, GR2 C-ter, and GR3 C-ter) did not show a significant difference in aggregation propensity or in the number of flanking unmodified gatekeeping residues compared to nonglycosylated sequons (fig. S6, B and C). Glycosylated sequons in APRs did not show a difference in any of these analyses either (Fig. 3, A and B), despite being under positive selective pressure. A possible explanation is that APRs comprise a much larger region (5 to 15 amino acids), which adds noise to the analysis.

To gain more insight into the role of N-glycans as gatekeepers of aggregation, we looked at the conservation of human glycosylated and nonglycosylated sequons throughout mammalian evolution. In particular, we focused on sequons at GR1 N-ter since this position showed the highest enrichment and strongest selective pressure when it is glycosylated (Fig. 2, A and B). Each sequon at GR1 N-ter was mapped to the multiz100way dataset (*37*), a dataset containing multiple sequence alignments of 100 mammalian species to the human genome. We then calculated the average number of unmodified gatekeepers that are found aligned to the three residues upstream and downstream of human APRs for orthologs in which the sequon is present and orthologs for which it is absent. An example of this can be seen for the N-glycosite at position 439 of the basal cell adhesion molecule protein (BCAM; Fig. 3C). In agreement with our previous analyses, we found that when the sequon was conserved in other species, it was usually flanked by only a small number of unmodified gatekeepers (fewer than one on average; Fig. 3C). However, the same regions had a higher number of unmodified gatekeepers in BCAM orthologs for which the sequon was not conserved (more than two on average; Fig. 3C). Interestingly, this was a general observation for most N-glycosites at GR1 N-ter, even for those that were next to very strong APRs (Fig. 3D). On the other hand, since nonglycosylated sequons are already protected by a high number of unmodified gatekeepers, particularly in the case of strong APRs, their absence in a species did not lead to a significant increase in these residues around the APRs (Fig. 3D).

A similar observation was obtained when analyzing protein paralogs, particularly the serpin superfamily of protease inhibitors. In humans, most serpins are classified into two clades: the extracellular "clade A" and the intracellular "clade B" (*41*, *42*). We found that many extracellular serpins have a glycosylated sequon flanking a very strong APR that is conserved in both clades (Fig. 3, E and F). However, in intracellular serpins, this APR is flanked instead by one or more unmodified gatekeeping residues, evidencing again an analogous function for N-glycans and unmodified gatekeepers (Fig. 3F).

N-glycosylation efficiently inhibits peptide aggregation in vitro by steric hindrance

The bioinformatics analysis presented above hints at a protective role of N-glycosylation against the aggregation of its cognate APRs. To experimentally assess this, we measured the aggregation kinetics and solubility of peptides with and without an N-glycan attached (Fig. 4A). Short aggregating peptides were used instead of full proteins to mimic exposed APRs and to ensure the interpretability of our findings.

After an N-glycan precursor ($Glc_3Man_9GlcNAc_2$) is transferred to a protein, it is processed in the ER by removal of the glucose residues as part of the quality control process (*31*). Then, the protein moves to the Golgi apparatus, where the carbohydrate is further processed into an extensive array of mature and complex N-glycoforms (fig. S7) (*43*). This raises the question whether there is a particular glycoform that confers protection against aggregation or, instead, if it is an intrinsic effect of all glycoforms. The genomes of higher eukaryotes encode two STT3 proteins (STT3A and STT3B), which are the catalytic subunits of two distinct OST complexes (*38*). The STT3A complex is associated with the protein translocation channel and glycosylates the majority of sites as they emerge into the ER lumen, while specific N-glycosites that are skipped by the STT3A complex are modified posttranslationally by the STT3B complex. That is, the addition of most N-glycans takes place, while a protein is being translated and, therefore, before it folds (*44*). During this time, an APR is exposed and at risk of engaging in non-native interactions, such as aggregation. Therefore, we reasoned that this is the most vulnerable time point in a protein lifespan—when it is most in need of antiaggregation mechanisms—and decided to use the $Man_9GlcNAc_2$ (Man_9) glycoform for our analyses, since it is the minimal carbohydrate structure that can be found attached to nascent polypeptides during their folding and before they leave the ER (fig. S7).

We analyzed 10 human APRs with a flanking N-glycosite (table S3). To investigate whether any structural constraints explain why the enrichment in our previous analysis was only observed in the N-terminal flank, we chose five APRs that were modified in the N-terminal site and five in the C-terminal site. Man_9 variants for each APR were compared to their unmodified versions. In addition, GlcNAc versions of each peptide were made to determine whether a much shorter N-glycan form can inhibit aggregation. All peptides in a set were dissolved to a concentration in which the unmodified

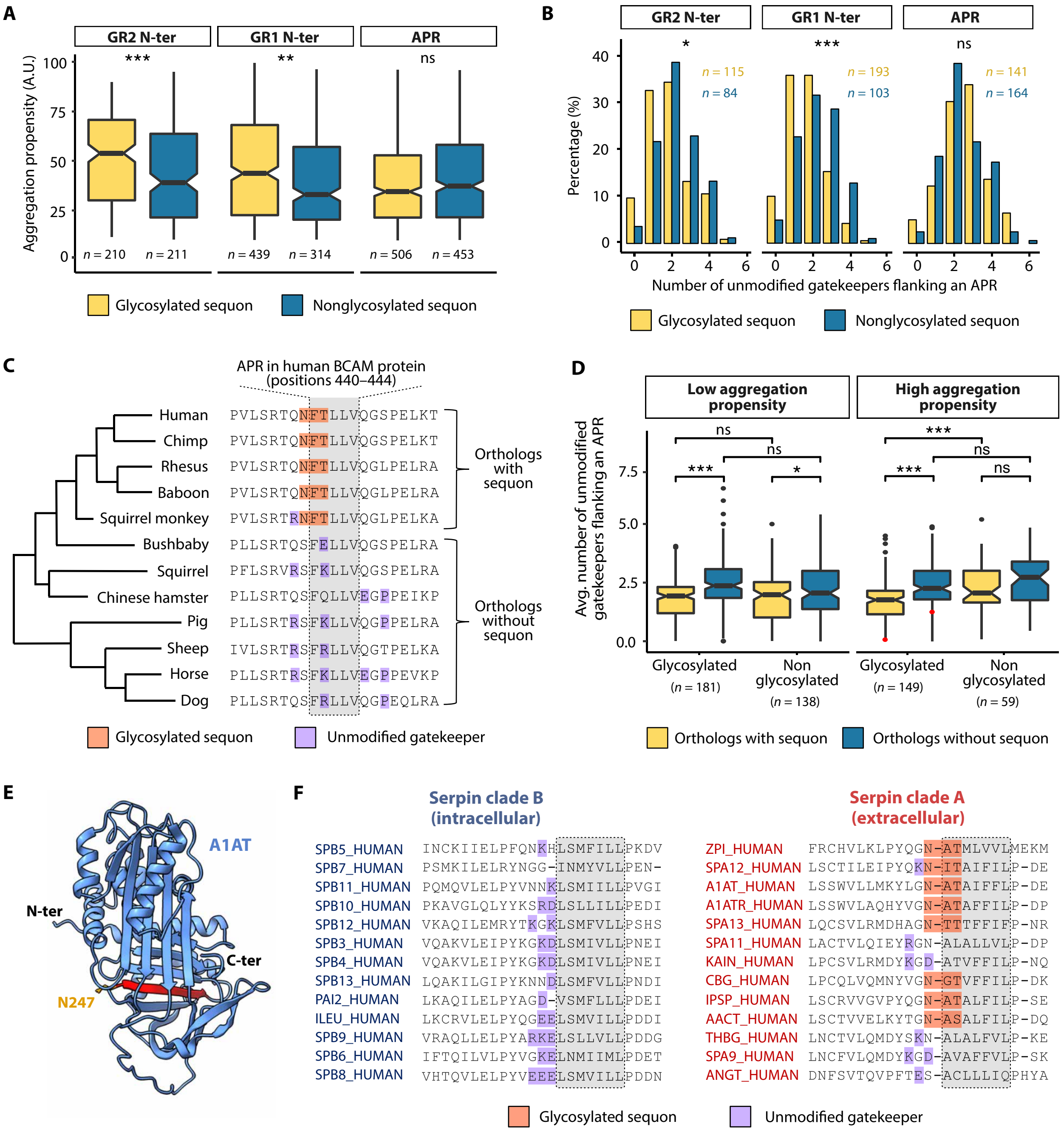
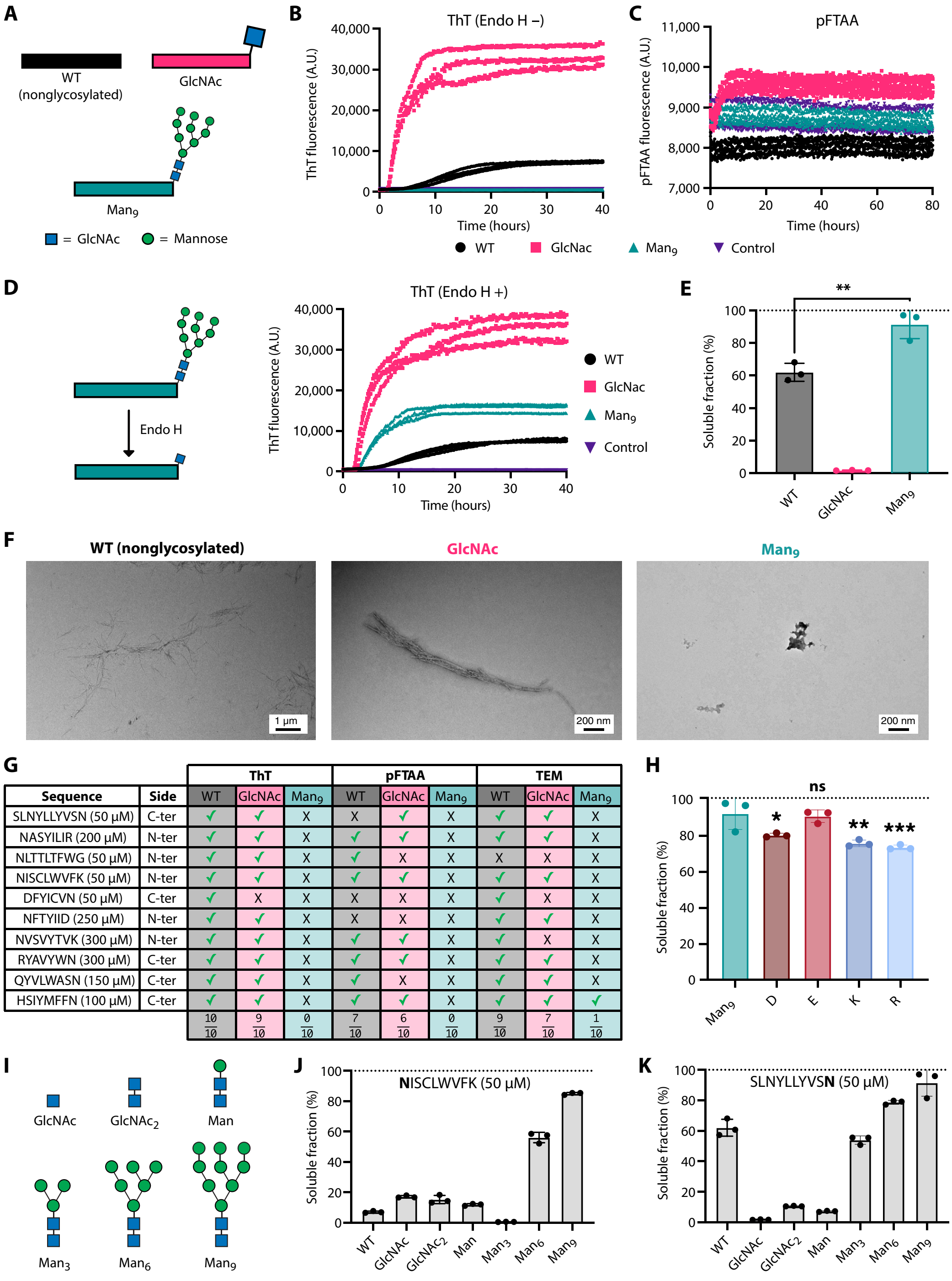


Fig. 3. N-glycosites at EPs behave as aggregation gatekeepers. (**A**) Box plot showing the aggregation propensity (TANGO scores) of APRs that have glycosylated or nonglycosylated sequons for each EP. (**B**) Distribution of the number of unmodified gatekeeper residues flanking (three positions upstream and downstream) strong APRs that have glycosylated or nonglycosylated sequons for each EP. Strong APRs (TANGO score ≥ 50) were used to ensure that a high evolutionary pressure is acting on the APRs to mitigate their aggregation. (**C**) Subset of a multiple sequence alignment for the glycosylated site at position 439 of the BCAM protein. Glycosylated sequons and unmodified gatekeepers that are flanking or within the aligned APR are indicated. The positions of the human APR in the alignment are colored in gray. (**D**) Box plot showing the average number of unmodified gatekeepers flanking or within aligned APRs for all glycosylated and nonglycosylated human sequons at GR1 N-ter when these are conserved (orthologs with sequon) or not (orthologs without sequon) in 100 mammalian species. APRs are divided into two categories: weak if the TANGO score is <50 or strong if the TANGO score is ≥50. Red dots indicate the values for the BCAM site shown in (C). (**E**) Example of a serpin structure (alpha-1 antitrypsin; A1AT) obtained from AlphaFold. A1AT has an N-glycosite (orange) at the N-terminal flank of a very conserved APR (red). (**F**) Multiple sequence alignment showing the conserved APR (in gray) for intracellular and extracellular serpins. N-glycosylated sequons or unmodified gatekeepers, three residues upstream of the APR, are highlighted. Statistical significance was determined by unpaired Wilcoxon test with Bonferroni correction for multiple comparisons [(A) to (C)]. The number of glycosylated or nonglycosylated sequons in each region is indicated. *$P \leq 0.05$, **$P \leq 0.01$, and ***$P \leq 0.001$. ns, not significant.



		ThT			pFTAA			TEM		
Sequence	Side	WT	GlcNAc	Man$_9$	WT	GlcNAc	Man$_9$	WT	GlcNAc	Man$_9$
SLNYLLYVSN (50 μM)	C-ter	✓	✓	X	X	✓	X	✓	✓	X
NASYILIR (200 μM)	N-ter	✓	✓	X	✓	✓	X	✓	✓	X
NLTTLTFWG (50 μM)	N-ter	✓	✓	X	✓	X	X	X	X	X
NISCLWVFK (50 μM)	N-ter	✓	✓	X	✓	✓	X	✓	✓	X
DFYICVN (50 μM)	C-ter	✓	X	X	X	X	X	✓	X	X
NFTYIID (250 μM)	N-ter	✓	✓	X	X	X	X	✓	✓	X
NVSVYTVK (300 μM)	N-ter	✓	✓	X	✓	✓	X	✓	X	X
RYAVYWN (300 μM)	C-ter	✓	✓	X	✓	✓	X	✓	✓	X
QYVLWASN (150 μM)	C-ter	✓	✓	X	✓	X	X	✓	✓	X
HSIYMFFN (100 μM)	C-ter	✓	✓	X	✓	✓	X	✓	✓	✓
		$\frac{10}{10}$	$\frac{9}{10}$	$\frac{0}{10}$	$\frac{7}{10}$	$\frac{6}{10}$	$\frac{0}{10}$	$\frac{9}{10}$	$\frac{7}{10}$	$\frac{1}{10}$

Fig. 4. In vitro analysis of N-glycosylated peptides. (**A**) Schematic representation of the peptide variants and experimental design. An aggregation core is flanked by either a nonglycosylated Asn (WT), GlcNAc, or Man$_9$. (**B** and **C**) ThT binding (B) and pFTAA binding (C) kinetics of the SLNYLLYVSN peptide set ($n = 3$). Vehicle control fluorescence is shown in purple. (**D**) ThT binding after incubation with 1 μl (500 U) of Endo H enzyme, which cleaves the bond between two GlcNAc subunits directly proximal to the asparagine residue of the glycopeptide ($n = 3$). Vehicle control fluorescence is shown in purple. (**E**) Percentage of the concentration of peptide in the soluble fraction after ultracentrifugation for the SLNYLLYVSN peptide set ($n = 3$). Unpaired *t* test was used to assess significance. (**F**) TEM images for the SLNYLLYVSN peptide set after q days of incubation. (**G**) Combined results for all peptide sets. Peptides were classified on whether they showed or did not show kinetics based on ThT and pFTAA assays (marked with ticks or crosses, respectively) and whether they formed or did not form fibrillar aggregates detectable by TEM imaging (marked with ticks or crosses, respectively). (**H**) Percentage of soluble fraction for the charged residue variants [Asp (D), Glu (E), Lys (K), and Arg (R); $n = 3$]. Man$_9$ values were reused from (E). Unpaired *t* test was used to assess significance against Man9. (**I**) Schematic representation of the structures of the different glycoforms analyzed. (**J** and **K**) Percentage of soluble fraction after ultracentrifugation for the nonglycosylated and glycoforms versions of NISCLWVFK (J) and SLNYLLYVSN (K) peptide sets ($n = 3$). Nonglycosylated and Man$_9$ peptides values were reused from fig. S10 and from (E). *N* represents biological replicates. Bars represent means and error bars SD. *$P \leq 0.05$, **$P \leq 0.01$, and ***$P \leq 0.001$.

variant displayed dye-binding aggregation kinetics with Thioflavin-T (ThT). The results for the peptide set derived from SLNYLLYVSN are shown in Fig. 4, B to H. ThT-binding experiments revealed that aggregates were formed by the nonglycosylated and GlcNAc peptides, while for Man_9, no fluorescent signal was observed (Fig. 4B). Since ThT is positively charged, we also used an alternative amyloid dye, the negatively charged pentameric formyl thiophene acetic acid (pFTAA), to overcome cases of dye-specific failure based on charge repulsion. Nevertheless, no pFTAA fluorescence signal was observed for Man_9 (Fig. 4C). Incubating the Man_9 peptide with Endo H, an enzyme that catalyzes the conversion of Man_9 into GlcNAc, resulted in a strong ThT fluorescent signal (Fig. 4D), suggesting that the Man_9 glycoform was inhibiting the aggregation of the peptide. However, since Man_9 is a huge molecule, its size could hinder the binding of the fluorescent dyes to a potential aggregated structure. To dismiss this possibility, we used an orthogonal assay that measures the concentration of soluble peptide left once the aggregation reaction has reached an equilibrium. Briefly, peptides were incubated for a week and then subjected to ultracentrifugation. End-point solubility measurements of this peptide set showed that Man_9 substantially improved APR solubility compared to nonglycosylated and GlcNAc peptides (Fig. 4E). We reached similar conclusions by transmission electron microscopy (TEM) imaging where no aggregated species were observed for the Man_9 peptide, while both nonglycosylated and GlcNAc peptides formed fibrillar aggregated structures (Fig. 4F). Together, these results indicate that Man_9 strongly inhibits the formation of aggregates. The combined results of the 10 APRs analyzed confirmed the generality of these findings (Fig. 4G and figs. S8 to S16). Unexpectedly, while the computational analysis showed selection only for N-glycosites at the N-terminal flanks of APRs, the in vitro experiments revealed that N-glycosylation can inhibit aggregation in both flanks. This indicates that the preference for N-terminal flanks observed computationally does not arise from any APR-intrinsic structural constraint and, therefore, other biological factors may be responsible (see Discussion).

To investigate the differences between N-glycans with unmodified gatekeeper residues, we made peptides in which the modified Asn residue was replaced by each of the four charged residues (Asp, Glu, Arg, and Lys) since these are known to strongly oppose aggregation. For the SLNYLLYVSN peptide set, Man_9 was more soluble than all peptide versions with charged residues, apart from Glu (Fig. 4H). Furthermore, in each APR set, Man_9 was as good or better at improving the solubility of peptides compared to their charged counterparts (figs. S8 to S16). This enhanced solubility could partially explain why N-glycosylation is selected over unmodified gatekeeping residues in some proteins.

While Man_9 showed complete or strong inhibition of aggregation in all peptide sets, GlcNAc's capability of inhibiting aggregation was substantially lower. Moreover, in some peptide sets, GlcNAc actually enhanced aggregation (Fig. 4 and fig. S16). Previous studies have proposed that the large size and hydrophilicity of glycans prevent the aggregation of protein pharmaceutical products through steric hindrance (45, 46). Therefore, we hypothesized that the difference in size between the two glycoforms might be responsible for the degree of inhibition observed. To assess this, we measured, in two of the peptide sets, the solubility of four additional glycoforms: $GlcNAc_2$, $ManGlcNAc_2$ (Man), $Man_3GlcNAc_2$ (Man_3), and $Man_6GlcNAc_2$ (Man_6) (Fig. 4I). GlcNAc, $GlcNAc_2$, and Man caused a minor and similar increase in solubility for the NISCLWVFK peptide compared to its unmodified version (Fig. 4J) and were actually found to be more insoluble for the SLNYLLYVSN peptide (Fig. 4K). A possible explanation could be the presence of glycoform-specific interactions, leading to stacking between the hydrophobic faces of sugars or between aromatic residues and sugars of different peptides (47). Conversely, Man_6 and Man_9 caused a substantial and size-dependent increase in solubility in both peptide sets (Fig. 4, J and K), supporting that steric hindrance may be the mechanism behind aggregation inhibition. These results provide direct evidence that different glycoforms confer distinct levels of protection against aggregation. Moreover, the more potent inhibitory effect of Man_9 on aggregation supports the idea that N-glycan–mediated protection against aggregation occurs before protein folding, since N-glycans are trimmed in the Golgi once proteins leave the ER.

N-glycosylation protects against aggregation in hard-to-fold proteins

Of all APRs in proteins that follow the SP, only around 7% are flanked by N-glycans at EPs (Fig. 5A). Why do some APRs, or the proteins bearing those APRs, require the extra level of protection granted by N-glycosylation? To answer this, we built a random forest classifier that predicts which APRs are protected by N-glycans using different features related to structural topology and aggregation, both at the APR and protein domain levels (see Methods). We decided to use features of individual protein domains instead of features from full proteins, as domains are independent evolutionary units that often fold independently from each other (48). Domains were extracted using CATH-Gene3D (49, 50). Since the number of protected and unprotected APRs is quite different and random forests are known to be sensitive to class imbalance, we trained two different models with opposite resampling techniques. The relative contact order of domains bearing the APRs was the most important feature in both models (Fig. 5B and fig. S17A). The relative contact order is a widely used metric to describe the complexity of a polypeptide fold as it measures the average sequence separation between contacting residues in a protein structure, which has been shown to correlate with folding times (51). When comparing domains with at least one APR, those with an N-glycosite at EPs have a significantly higher relative contact order (fig. S17B). Moreover, while high contact order domains without protected APRs generally have lower aggregation propensities, the ones with N-glycosites at EPs usually contain much stronger APRs (Fig. 5C). Thus, N-glycosylation protects APRs of complex domains with overall high aggregation propensities. As expected, other parameters determined to be important by both models were the solvent accessibility of APRs and the number of unmodified gatekeeping residues flanking them (Fig. 5B and fig. S17, C and D). N-glycosylation constrains part of the APR to be solvent accessible to avoid steric clashes, while from our previous analyses, we know that N-glycosylation replaces unmodified gatekeeping residues at EPs. The oxidizing environment of the ER allows for the formation of disulfide bridges, which help stabilize the native fold of SP proteins. Nevertheless, the number of disulfide bridges in a domain had low importance in the prediction (Fig. 5B).

The high relative contact order observed in domains bearing protected APRs could be indicative of an enrichment for a specific fold topology, as most folds have lower contact orders than domains with protected APRs (Fig. 5D). To investigate this, we looked at the relative enrichment of protected APRs in different CATH domain

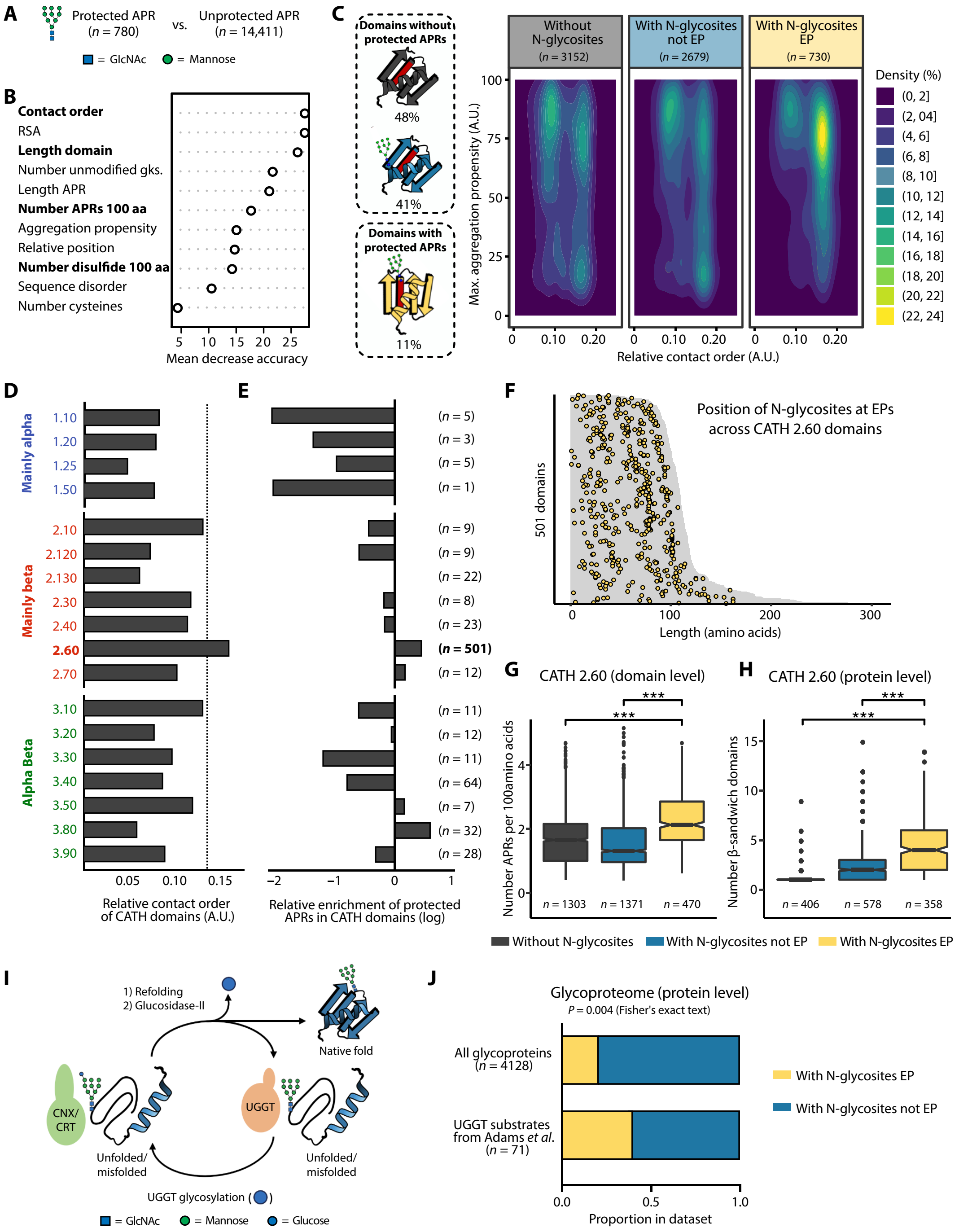


Fig. 5. N-glycosylation protects against aggregation in hard-to-fold proteins. (**A**) A random forest classifier was built to predict which APRs in CATH domains are protected (with an N-glycosite at an EP) or unprotected (all others). (**B**) Variable importance plot for the predictive model built using random undersampling. Higher values indicate that a variable is more important for the model. Domain-specific variables are highlighted in bold, while APR-specific variables are not highlighted. aa, amino acid. (**C**) Left: Schematic representation and percentage of domains classified in three categories: with N-glycosites in EPs (yellow), with N-glycosites not in EPs (blue) and without N-glycosites (black). APRs are colored in red. Right: A two-dimensional density plot showing the relative contact order and the maximum aggregation propensity for domains in each category. (**D**) Average relative contact order of domains in each CATH architecture. The dotted line indicates the average value for domains containing an N-glycosite at an EP. (**E**) Relative enrichment of finding a protected APR in each CATH architecture. The number of protected APRs present in each architecture is indicated. (**F**) Heatmap showing the position of N-glycosites at EPs in β-sandwich domains with at least one protected APR. Domains are sorted by length and colored in gray. (**G**) Box plot showing the number of APRs per 100 amino acids in β-sandwich domains. The number of domains in each category is indicated. (**H**) Box plot showing the number of β-sandwich domains in proteins with at least one of these domains. The number of proteins in each category is indicated. (**I**) Quality control system of glycoproteins. (**J**) Fraction of UGGT substrates that have at least an N-glycosite in an EP compared to the same fraction in all glycoproteins. Statistical significance was determined by unpaired Wilcoxon test with Bonferroni correction for multiple comparisons (G and H). ***$P \leq 0.001$.

architectures. As background, we used all SP protein domains with at least one APR. There was a depletion of protected APRs in domain architectures of the class "Mainly alpha," while domain architectures with more β sheet content were more abundant (Fig. 5E). In particular, the "CATH 2.60" domain architecture, also known as β-sandwich, was highly enriched and included the majority of N-glycosites at EPs, which are distributed throughout the entire fold (Fig. 5, E and F). β-sandwich domains are characterized by two opposing antiparallel β sheets and span a large number of fold superfamilies, including the immunoglobulin-like fold, and it has been linked to many neurodegenerative diseases associated with the formation of protein aggregates (*52*, *53*). Moreover, β-sandwich domains are frequently organized in linear arrays within multidomain proteins, which have a higher risk of forming domain-swapped misfolded species (*54*). A deeper analysis of β-sandwich domains showed that those with N-glycosites at EPs have stronger and higher number of APRs than the rest of β-sandwich domains, including other domains that are also N-glycosylated (Fig. 5G and fig. S17E). Furthermore, β-sandwich domains containing APRs protected by N-glycans are found in larger multidomain proteins, with, on average, five β-sandwich domains per protein (Fig. 5H and fig. S17F).

N-glycosylation plays a crucial role in glycoprotein quality control (Fig. 5I), as it acts as the attachment site for the ER soluble and membrane-bound lectin chaperones calreticulin and calnexin (*31*). These chaperones have been shown to direct protein folding, reduce aggregation, retain misfolded or immature proteins within the ER and target aberrant proteins for degradation (*55*). Upon release from the lectin chaperones, correctly folded proteins are transported to the Golgi apparatus. However, nascent chains that are not properly folded can be recognized by the protein folding sensor uridine 5′-diphosphate–glucose:glycoprotein glucosyltransferase (UGGT) and then directed for rebinding to the lectin chaperones (*56*). That is, UGGT substrates are prone to misfold and require multiple rounds of chaperone binding. Recently, Adams *et al.* (*57*) identified 71 bona fide human UGGT substrates using quantitative proteomics in human embryonic kidney–293 cells. Proteins containing N-glycosites in EPs are significantly enriched in UGGT substrates when compared to other glycoproteins (Fig. 5J).

Our findings show that the protection of APRs through N-glycans is linked to biophysical properties that challenge protein folding, such as structural complexity, a higher number of APRs, and higher aggregation propensities. Moreover, this protection is enriched in UGGT substrates, which require multiple rounds of chaperone association to reach their native conformations. Therefore, it appears that these sites are strongly correlated with folding challenges, consistent with the idea that N-glycans mitigate aggregation before folding. In addition, the fact that most of domains that require this antiaggregation mechanism have the same topology suggests that their folding landscapes, populated by similar folding intermediates (*58*), might have coevolved together with N-glycosylation to avoid aggregation.

Absence of N-glycosylation Neuro2a cells specifically increases protein aggregation

If N-glycosylation is a common mechanism against protein aggregation in eukaryotes, then its inhibition should affect protein solubility across proteomes. Complete inhibition of N-glycosylation can be accomplished with the small molecule tunicamycin, which blocks the transfer of GlcNAc-phosphate to phosphorylated dolichol (Fig. 6A). In animal and plant cells, inhibition of N-glycosylation with tunicamycin leads to misfolding and aggregation inside the ER (*59*–*61*), triggering the unfolded protein response.

To investigate which particular glycoproteins aggregate in the absence of N-glycosylation, we reanalyzed a proteomics dataset from Sui *et al.* (*62*). Briefly, in this study, they measured the changes in proteome solubility in the mouse Neuro2a cell line after treatment with five different stresses, including tunicamycin. Our analysis found that after treatment with tunicamycin, around 20% of the proteins identified by tandem mass spectrometry (MS/MS) with an N-glycosite at an EP are more insoluble (Fig. 6B and table S3). In the majority of these aggregated proteins, the N-glycosite is located within a β-sandwich domain (table S4). In contrast, just 10% of all other N-glycoproteins are more insoluble. This suggests that the absence of N-glycosylation at EPs is more detrimental to protein solubility compared to sites that are far from APRs. However, because of the small number of proteins identified by MS/MS, the difference between these two groups was not statistically significant ($P = 0.31$ by Fisher's exact test). Expectedly, an even smaller percentage of nonglycosylated ER proteins are more insoluble after tunicamycin treatment. When looking at proteins that are more soluble after treatment with tunicamycin, no enrichment was found for proteins with an N-glycosite at an EP (Fig. 6B). The same analysis was performed by looking at the solubility changes under the other four stresses. However, none of these affected the solubility of proteins identified with an N-glycosite in an EP (Fig. 6, C to F). Together, these results suggest that inhibiting N-glycosylation leads to a decrease in protein solubility, especially in proteins where N-glycans act as aggregation gatekeepers (Fig. 6G).

DISCUSSION

Our work demonstrates that N-glycans are enriched, are highly conserved, and commonly replace unmodified gatekeeper residues in sequence segments with an intrinsic capacity to aggregate, here referred to as APRs, in nearly a thousand human proteins. In addition, we show that N-glycans suppress the aggregation of APRs in vitro and that their inhibition in mouse Neuro2a cells leads to a specific aggregation of newly made proteins. Together, these findings suggest that, among its many molecular functions, N-glycosylation constitutes a functional mechanism directly dedicated to the control of protein aggregation in eukaryotes.

Many studies have shown that N-glycosylation prevents the aggregation of glycoproteins in cells through diverse indirect molecular mechanisms. For example, N-glycans can affect the folding process by restricting the conformational entropy of the unfolded protein and stabilizing specific secondary structural elements, preventing the formation of folding intermediates prone to aggregate (*55*, *63*). Moreover, the association of glycoproteins with ER lectin chaperones increases folding efficiency while decreasing aggregation propensity (*55*). Direct inhibition of aggregation by N-glycans has also been described, particularly in recombinant therapeutic proteins (*64*). For the production of therapeutic antibodies, such as bevacizumab, N-glycosylation sites have been engineered near APRs to mitigate aggregation (*65*). However, the conditions in which biotherapeutics are produced are far from those found in cells, as often these proteins are manufactured and stored at very high concentrations for extended periods of time. Instead, our work points to a widely conserved cellular strategy in which N-glycans

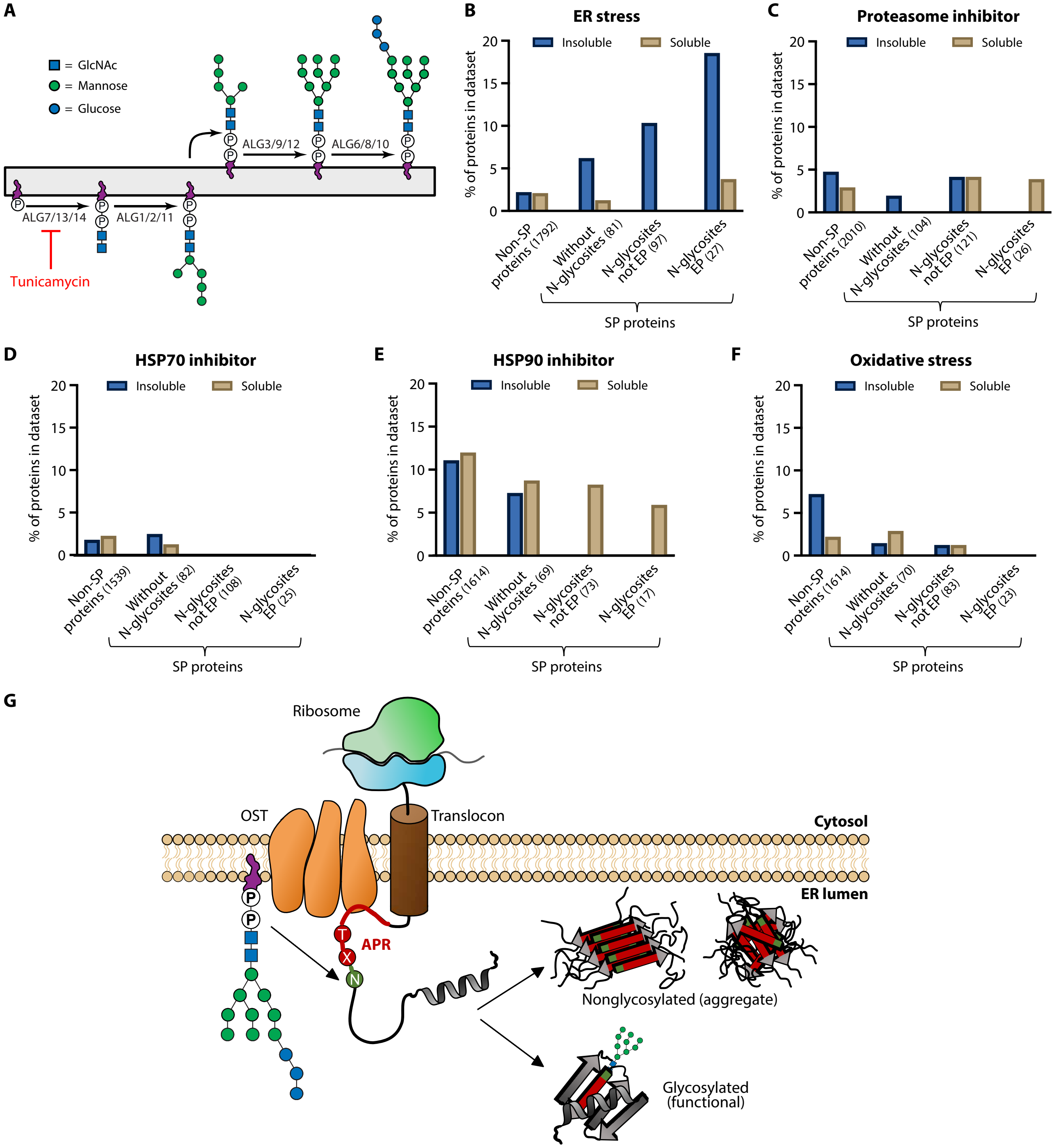


Fig. 6. Absence of N-glycosylation in Neuro2a cells specifically increases protein aggregation. (**A**) A simplified overview of the process of glycan precursor synthesis. Tunicamycin (in red) completely blocks the enzymatic activity of UDP-N-acetylglucosamine—dolichyl-phosphate N-acetylglucosaminephosphotransferase (encoded by *ALG7*), eliminating the production of all glycan precursors and the complete inhibition of N-glycosylation. (**B** to **F**) Percentage of proteins that are enriched in the insoluble or soluble fraction in each protein group after ER stress (B), treatment with a proteasome inhibitor (C), treatment with an HSP70 inhibitor (D), treatment with an HSP90 inhibitor (E), or oxidative stress (F) relative to the number of proteins identified by MS in each protein group (background). The total number of proteins (background) in each group and stress is indicated. (**G**) During translocation, the OST can glycosylate proteins before these are folded. When N-glycans are attached at the flanks of an APR, they shield this region from aggregation, leading to a glycosylated functional protein. However, the absence of N-glycosylation, specifically at the flanks of an APR, can lead to misfolding and aggregation of the affected proteins. The ultrastructure of these aggregates (amorphous or fibrillar) is not yet known.

directly hinder the formation of aggregates during folding under physiological conditions.

An unexpected result from our computational analysis is that only N-glycans located in the N-terminal flanks of APRs are enriched and under selection and share similar features to unmodified gatekeeper residues (Figs. 2 and 3). However, placing N-glycans on either side of APRs in vitro strongly suppresses their aggregation (Fig. 4). Since most N-glycans are cotranslationally attached to proteins by STT3A, it appears possible that the preferential addition of this modification to the N-terminal flanks is coupled with translation. It has been proposed that the initiation of aggregation may occur within polysomes, where identical unfolded nascent chains reach high local concentrations (*10*, *66*). Under this framework, N-glycosylating the N-terminal side of an APR will immediately shield it from potential cotranslational non-native interactions, including aggregation, as this side is translated before the rest of the APR sequence. Consistent with this hypothesis, previously identified human STT3B-dependent N-glycosites (*67*), which can only be attached posttranslationally, are significantly depleted at EPs (fig. S18). Moreover, overexpression of STT3B only partially rescues STT3A-deficient cells, despite STT3B acting downstream of STT3A, which enables it to glycosylate sites missed by STT3A (*68*). Eukaryotic species lacking the STT3A ortholog, such as *S. cerevisiae*, can only perform N-glycosylation posttranslationally (*69*). Unlike the other eukaryotic species analyzed, N-glycosites at EPs were found to be depleted in yeast proteins (Fig. 2F). Despite all this circumstantial evidence, future studies are required to determine whether N-glycosylation is specifically suppressing aggregation during translation.

One question remains: Why is N-glycosylation the only modification found to broadly act as an aggregation gatekeeper? Although we do not rule out that other PTM types not investigated here may act as gatekeepers, the answer probably again lies in the cotranslational nature of this modification. First, PTMs require acceptor sites to be accessible to the modifying enzyme, precluding regions that are buried or structurally too rigid when the protein is folded, such as APRs and their GRs. The placement of N-glycosylation in bacteria, which takes place posttranslationally, is restricted only to flexible segments (*70*). Therefore, coupling N-glycosylation with folding increases the number of sites that can be modified. Second, during folding, APRs are exposed and at risk of aggregation. Consequently, protein folding exerts a dual selection pressure on the glycosylation process (*38*). On the one hand, sites that destabilize the native structure are under negative selection (*71*), while sites that optimize folding, in this case, by reducing aggregation, are under positive selection and are likely to become essential (Fig. 2E). An additional consequence of this shift in the temporal sequence of maturation events has been the coevolution of N-glycans with the ER chaperone machinery, leading to a very specific quality control system for secretory and membrane glycoproteins (*38*). Recently, a similar coadaptation process was described between chaperone specificity and protein composition to explain the preference of heat shock 70 kDa proteins (Hsp70s) for positively charged residues in bacteria (*16*).

We found a higher proportion of aggregated proteins with N-glycans acting as gatekeeper residues compared to other glycoproteins after treatment of mouse Neuro2a cells with tunicamycin (Fig. 6B). Tunicamycin treatment has been extensively used as a model to mimic type I congenital disorders of glycosylation (CDG-I) (*72*, *73*). These are rare groups of metabolic diseases that affect specific sugar transferases and enzymes involved in the synthesis and transfer of N-glycans, thus leading to the improper N-glycosylation of proteins, which causes various symptoms potentially affecting multiple organs (*74*, *75*). It has been reported that several CDG-I can lead to ER stress and activate the unfolded protein response due to misfolded hypoglycosylated proteins unable to leave the ER (*76*). On the basis of our findings, we hypothesize that the formation of protein aggregates resulting from a loss of N-glycans may provide an additional molecular cause of ER stress in CDGs and may contribute to the pathomechanism of these disorders. Future efforts should be made to determine whether there is a direct relationship between these genetic disorders and protein aggregation.

METHODS

Human proteome dataset

The human proteome was obtained from UniProtKB/Swiss-Prot database (reference proteome UP000005640; release 2022_02). The dataset contains 19,379 proteins, after excluding sequences with nonstandard amino acids (e.g., selenocysteine), sequences with <25 amino acids and those with >10,000 residues and after filtering at 90% sequence identity using the CD-hit algorithm (*77*). Signal peptides and transmembrane domains were identified using deepTMHMM (*78*) and removed from the analyses to avoid biases. In addition, deepTMHMM provides information about the overall topology of the protein. Experimentally annotated protein PTM sites were obtained from dbPTM (*27*) and from the GlcNAcAtlas (*28*) and were mapped to the proteome. Only those PTM types with more than 1200 sites were retained.

Information about protein subcellular location was extracted from UniProt. Proteins known to reside in the ER, Golgi apparatus, cell membrane, lysosomes, or extracellular space were labeled as part of the SP. On the other hand, proteins known to reside in the cytoplasm, nucleus, or mitochondria were labeled as part of the non-SP. Proteins labeled both as SP and non-SP were excluded from further analyses unless specified otherwise.

Structural information was added to the dataset for each protein using the structures from the AlphaFold database (*79*, *80*). Absolute solvent accessibility values were calculated with DSSP based on these structures (*81*, *82*). Then, the relative solvent accessibility (RSA) values were calculated by dividing the absolute solvent accessibility values by residue-specific maximal accessibility values, as extracted from Tien *et al.* (*83*). Residues with RSA values of <0.2 were labeled as buried. Disordered regions were identified using the predicted local-distance difference test (pLDDT) score provided in the AlphaFold models, as regions with low confidence scores have been shown to overlap largely with intrinsically disorder regions (*84*). Residues with pLDDT scores of <50 were labeled as structurally disordered.

Protein aggregation prediction

APRs were predicted computationally using TANGO (*1*) under physiological conditions (pH at 7.5, temperature at 298 K, protein concentration at 1 mM, and ionic strength at 0.15 M). In this study, APRs are defined as segments between 5 and 15 amino acids in length, each with an aggregation score of at least 10. GRs are defined as the three residues immediately downstream and upstream of APRs. All other residues are defined as DRs. GRs were further labeled as GR1, GR2, or GR3 and as N-ter or C-ter based on their position to the APR.

APRs were also identified with CamSol (*33*). CamSol calculates an intrinsic solubility profile where regions with a score higher than 1 are highly soluble, while scores smaller than −1 are poorly soluble (aggregation-prone). CamSol APRs are defined as segments between 5 and 15 amino acids in length, each with a solubility score smaller than −1. GRs and DRs are defined in the same way as above.

Identification of sequons

All human proteins were scanned for N-glycosylation sequons (Asn-X-Thr/Ser, where X ≠ Pro). Sequons known to be experimentally glycosylated based on dbPTM annotations were labeled as "SP glycosylated," regardless of their protein subcellular location. This can include some mislabeled "non-SP" proteins. Sequons in proteins from the SP without dbPTM annotations were labeled as "SP nonglycosylated." Sequons in proteins that do not follow the SP and thus cannot be glycosylated were labeled as "non-SP."

Relative enrichment calculation

For all PTM types and sequons, the frequency in each region was calculated by taking all experimentally defined PTM sites in APRs, GRs, and DRs versus all sites that could receive a PTM in each region

$$\text{Frequency} = \frac{\text{Number of PTM sites in a region for a specific PTM type}}{\text{Number of residues that could be modified in that region}}$$

For example, for serine phosphorylation

$$\text{Frequency} = \frac{\text{Number of phosphorylated serines in region}}{\text{Number of serines in that region}}$$

The relative enrichment (odds ratio) was obtained by dividing the frequency in each region by the frequency in the entire proteome (background). To avoid biases, only proteins that contain PTM sites are used as background.

Generation of randomized protein sequences

Two datasets with 5000 randomized protein sequences were built with a custom R script using the amino acid sequence composition and the average residue length as proteins from the SP ("SP randomized") and proteins that do not follow the SP ("non–SP randomized").

Eukaryotic proteome dataset

The proteomes of five other eukaryotic species were analyzed in the same way as the human proteome and include representatives from the animal [*M. musculus* (UP000000589), *D. melanogaster* (UP000000803), and *C. elegans* (UP000001940)], plant [*A. thaliana* (UP000006548)], and fungal [*S. cerevisiae* (UP000002311)] kingdom.

The relative enrichments of glycosylated and nonglycosylated sequons were determined for each eukaryotic species. However, since experimentally identified N-glycosites for these organisms are scarce, all sequons in SP proteins were considered glycosylated [unless topological annotations by deepTMHMM (*78*) predicted the site to be facing the cytoplasm, where glycosylation does not occur].

Sequon conservation analysis

The multiz100way (*37*) is a dataset containing multiple sequence alignments of 100 mammalian species to the human genome (hg38). Human N-glycosylation sequons (N-X-T/S, where X ≠ P) were mapped to this dataset to calculate their conservation. A human sequon is considered conserved in a species if in that species there is an N-glycosylation canonical motif aligned to it.

Peptide set design

To construct a set of aggregating peptides with N-glycans at the flanks, APRs were selected from the human proteome containing an N-glycosylation site at the N-terminal or C-terminal flank. To facilitate accurate concentration determination of peptides through absorbance measurements at 280 nm, only APRs containing Trp and/or Tyr were considered. Twenty APRs were synthesized and screened for ThT-binding kinetics, from which a final set of 10 APRs was selected on the basis of their kinetic profile. Five of these sequences had the N-glycosylation site in the N-terminal, while the other five were in the C-terminal. Seven variants for each peptide sequence were produced: nonmodified (WT), GlcNAc, Man_9, and each of the charged residues (D, E, K, and R). For two specific peptide sets, four more variants were produced: GlcNAc_2, Man, Man_3, and Man_6.

Peptide aggregation kinetics

All peptides, except the GlcNAc_2, Man, Man_3, and Man_9 variants, were synthesized in-house using an Intavis Multipep RSi solid-phase peptide synthesis robot. The complex glycoform peptide variants were ordered from Chemitope Glycopeptide. Stocks were then diluted to the appropriate peptide concentration in phosphate-buffered saline with a final concentration of 5% dimethyl sulfoxide. The concentration of each peptide set was selected on the basis of a screening of ThT-binding kinetics to allow for a lag phase shorter than 72 hours and longer than 2 hours. TCEP [tris(2-carboxyethyl) phosphine, 1 mM] was included in solutions of peptides containing cysteine or methionine residues to disrupt disulfide bond formation. For ThT- and pFTAA-binding kinetics, 10 μM ThT or 1 μM pFTAA was added to the peptide samples. Dye binding was measured over time through excitation at 440 nm and emission at 480 and 520 nm, for ThT and pFTAA, respectively, in a FLUOstar OMEGA.

For Endo H treatment, endoglycosidase H (500 U, 1 μl; New England Biolabs, catalog no. P0702) was added to each of the SLNYLLYVSN peptide samples. Aggregation kinetics were measured over time as above.

Endpoint solubility

For endpoint solubility concentrations, peptide preparations were left at room temperature for a week at an initial concentration equal to the one used in for aggregation kinetics. Peptides were subsequently subjected to ultracentrifugation at 100,000*g* for 1 hour at 4°C. Supernatant concentrations were measured using reversed-phase high-performance liquid chromatography (RP-HPLC). Concentrations were measured with RP-HPLC instead of using absorbance measurements at 280 nm since it is more accurate for low concentrations.

TEM imaging

Peptide solutions were incubated for a week at room temperature at the same concentrations of previous experiments. Suspensions (5 μl) of each peptide solution were added on 400-mesh carbon-coated copper grids, which were negatively stained using uranyl acetate. Grids were examined with a JEM-1400 120-kV transmission electron microscope.

Machine learning

All human APRs were classified as protected (with an N-glycosylation site at an EP) and unprotected (all others).To predict which APRs are protected by N-glycans, a random forest classifier (randomForest R package, number of trees = 500, mtry = 3) was trained using several features of the APRs (RSA, length, number of unmodified gatekeepers, relative position within the domain, aggregation propensity, sequence disorder, and number of cysteines) and of the protein domains bearing such APRs (contact order, length, number of APRs per 100 amino acids, and number of disulfide bonds per 100 amino acids). Domain boundaries were extracted using CATH-Gene3D (*49*, *50*). Domains smaller than 25 residues, larger than 500 residues or with transmembrane regions were filtered out. The contact order for each domain was calculated as defined by Plaxco *et al.* (*51*). The number of disulfide bonds was extracted from UniProt. Since the number of observations in each class is substantially different, random undersampling and random oversampling (ROSE R package) were used to avoid biases. In random undersampling, observations from the majority class are randomly removed until a balanced class distribution is achieved. On the other hand, in random oversampling using ROSE, observations from the minority class are synthetically generated using a smoothed bootstrap technique to match the number of instances in the majority class. Feature importance was evaluated with the mean decrease accuracy plot, which indicates how much accuracy the model loses when excluding each variable.

Reanalysis of Neuro2a cells proteomics

Proteins that are significantly altered in solubility after treatment with ER stress (tunicamycin), oxidative stress (arsenite), proteasome inhibitor (MG132), HSP90 inhibitor (novobiocin), or HSP70 (Ver-155008) inhibitor were downloaded from Sui *et al.* (*62*). The number of proteins that are more soluble or insoluble after treatment with each stress in one of the following groups was determined: non-SP proteins (cytoplasmic proteins), SP without N-glycosites, SP with N-glycosites not in EPs, and SP with N-glycosites in EPs. Then, a percentage was calculated by dividing it by the total number of proteins identified (MS/MS proteome) after each stress in the different groups.

Statistics

GraphPad prism or R software were used to perform the different statistical tests. The tests used in each analysis are specified in the corresponding figure. P values are represented as $*P \leq 0.05$, $**P \leq 0.01$, and $***P \leq 0.001$.

Visualizations

Visualizations were performed with GraphPad prism or custom R scripts using the packages ggplot2 (*85*) and ComplexHeatmap (*86*). ChimeraX was used to visualize protein structures (*87*).

Supplementary Materials

This PDF file includes:

Figs. S1 to S18
Tables S1, S3 and S4
Legend for table S2

Other Supplementary Material for this manuscript includes the following:

Table S2

REFERENCES AND NOTES

1. A. M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, L. Serrano, Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* **22**, 1302–1306 (2004).
2. F. Rousseau, L. Serrano, J. W. H. Schymkowitz, How evolutionary pressure against protein aggregation shaped chaperone specificity. *J. Mol. Biol.* **355**, 1037–1047 (2006).
3. R. Prabakaran, D. Goel, S. Kumar, M. M. Gromiha, Aggregation prone regions in human proteome: Insights from large-scale data analyses. *Proteins* **85**, 1099–1118 (2017).
4. J. Tyedmers, A. Mogk, B. Bukau, Cellular strategies for controlling protein aggregation. *Nat. Rev. Mol. Cell Biol.* **11**, 777–788 (2010).
5. H. Saibil, Chaperone machines for protein folding, unfolding and disaggregation. *Nat. Rev. Mol. Cell Biol.* **14**, 630–642 (2013).
6. J. Santos, I. Pallarès, V. Iglesias, S. Ventura, Cryptic amyloidogenic regions in intrinsically disordered proteins: Function and disease association. *Comput. Struct. Biotechnol. J.* **19**, 4192–4206 (2021).
7. F. Chiti, C. M. Dobson, Protein misfolding, amyloid formation, and human disease: A summary of progress over the last decade. *Annu. Rev. Biochem.* **86**, 27–68 (2017).
8. M. G. Iadanza, M. P. Jackson, E. W. Hewitt, N. A. Ranson, S. E. Radford, A new era for understanding amyloid structures and disease. *Nat. Rev. Mol. Cell Biol.* **19**, 755–773 (2018).
9. T. Langenberg, R. Gallardo, R. van der Kant, N. Louros, E. Michiels, R. Duran-Romana, B. Houben, R. Cassio, H. Wilkinson, T. Garcia, C. Ulens, J. Van Durme, F. Rousseau, J. Schymkowitz, Thermodynamic and evolutionary coupling between the native and amyloid state of globular proteins. *Cell Rep.* **31**, 107512 (2020).
10. B. Houben, F. Rousseau, J. Schymkowitz, Protein structure and aggregation: A marriage of necessity ruled by aggregation gatekeepers. *Trends Biochem. Sci.* **47**, 194–205 (2022).
11. E. Monsellier, M. Ramazzotti, N. Taddei, F. Chiti, Aggregation propensity of the human proteome. *PLoS Comput. Biol.* **4**, e1000199 (2008).
12. A. K. Buell, G. G. Tartaglia, N. R. Birkett, C. A. Waudby, M. Vendruscolo, X. Salvatella, M. E. Welland, C. M. Dobson, T. P. J. Knowles, Position-dependent electrostatic protection against protein aggregation. *Chembiochem* **10**, 1309–1312 (2009).
13. B. N. Markiewicz, R. Oyola, D. Du, F. Gai, Aggregation gatekeeper and controlled assembly of Trpzip β-hairpins. *Biochemistry* **53**, 1146–1154 (2014).
14. R. Sant'Anna, C. Braga, N. Varejão, K. M. Pimenta, R. Graña-Montes, A. Alves, J. Cortines, Y. Cordeiro, S. Ventura, D. Foguel, The importance of a gatekeeper residue on the aggregation of transthyretin: Implications for transthyretin-related amyloidoses. *J. Biol. Chem.* **289**, 28324–28337 (2014).
15. J. Beerten, W. Jonckheere, S. Rudyak, J. Xu, H. Wilkinson, F. De Smet, J. Schymkowitz, F. Rousseau, Aggregation gatekeepers modulate protein homeostasis of aggregating sequences and affect bacterial fitness. *Protein Eng. Des. Sel.* **25**, 357–366 (2012).
16. B. Houben, E. Michiels, M. Ramakers, K. Konstantoulea, N. Louros, J. Verniers, R. van der Kant, M. De Vleeschouwer, N. Chicoria, T. Vanpoucke, R. Gallardo, J. Schymkowitz, F. Rousseau, Autonomous aggregation suppression by acidic residues explains why chaperones favour basic residues. *EMBO J.* **39**, e102864 (2020).
17. G. De Baets, J. Van Durme, F. Rousseau, J. Schymkowitz, A genome-wide sequence-structure analysis suggests aggregation gatekeepers constitute an evolutionary constrained functional class. *J. Mol. Biol.* **426**, 2405–2412 (2014).
18. G. De Baets, L. Van Doorn, F. Rousseau, J. Schymkowitz, Increased aggregation is more frequently associated to human disease-associated mutations than to neutral polymorphisms. *PLoS Comput. Biol.* **11**, e1004374 (2015).
19. L. N. Schaffert, W. G. Carter, Do post-translational modifications influence protein aggregation in neurodegenerative diseases: A systematic review. *Brain Sci.* **10**, 232 (2020).
20. C. Alquezar, S. Arya, A. W. Kao, Tau post-translational modifications: Dynamic transformers of tau function, degradation, and aggregation. *Front. Neurol.* **11**, 595532 (2021).
21. P. J. Barrett, J. Timothy Greenamyre, Post-translational modification of α-synuclein in Parkinson's disease. *Brain Res.* **1628**, 247–253 (2015).
22. N. Rezaei-Ghaleh, S. Kumar, J. Walter, M. Zweckstetter, Phosphorylation interferes with maturation of amyloid-β fibrillar structure in the N terminus. *J. Biol. Chem.* **291**, 16059–16067 (2016).
23. C.-X. Gong, F. Liu, I. Grundke-Iqbal, K. Iqbal, Post-translational modifications of tau protein in Alzheimer's disease. *J. Neural Transm.* **112**, 813–838 (2005).
24. P. Ryan, M. Xu, A. K. Davey, J. J. Danon, G. D. Mellick, M. Kassiou, S. Rudrawar, *O*-GlcNAc modification protects against protein misfolding and aggregation in neurodegenerative disease. *ACS Chem. Nerosci.* **10**, 2209–2221 (2019).
25. M. R. Martinez, T. B. Dias, P. S. Natov, N. E. Zachara, Stress-induced *O*-GlcNAcylation: An adaptive process of injured cells. *Biochem. Soc. Trans.* **45**, 237–249 (2017).
26. S. M. Pearlman, Z. Serber, J. E. Ferrell Jr., A mechanism for the evolution of phosphorylation sites. *Cell* **147**, 934–946 (2011).
27. Z. Li, S. Li, M. Luo, J.-H. Jhong, W. Li, L. Yao, Y. Pang, Z. Wang, R. Wang, R. Ma, J. Yu, Y. Huang, X. Zhu, Q. Cheng, H. Feng, J. Zhang, C. Wang, J. B.-K. Hsu, W.-C. Chang, F.-X. Wei, H.-D. Huang, T.-Y. Lee, dbPTM in 2022: An updated database for exploring regulatory networks and functional associations of protein post-translational modifications. *Nucleic Acids Res.* **50**, D471–D479 (2022).

28. J. Ma, Y. Li, C. Hou, C. Wu, O-GlcNAcAtlas: A database of experimentally identified O-GlcNAc sites and proteins. *Glycobiology* **31**, 719–723 (2021).
29. C. N. I. Pang, A. Hayen, M. R. Wilkins, Surface accessibility of protein post-translational modifications. *J. Proteome Res.* **6**, 1833–1845 (2007).
30. I. Bludau, S. Willems, W.-F. Zeng, M. T. Strauss, F. M. Hansen, M. C. Tanzer, O. Karayel, B. A. Schulman, M. Mann, The structural context of posttranslational modifications at a proteome-wide scale. *PLoS Biol.* **20**, e3001636 (2022).
31. A. Helenius, M. Aebi, Roles of N-linked glycans in the endoplasmic reticulum. *Annu. Rev. Biochem.* **73**, 1019–1049 (2004).
32. A. Varki, Biological roles of glycans. *Glycobiology* **27**, 3–49 (2017).
33. P. Sormanni, F. A. Aprile, M. Vendruscolo, The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.* **427**, 478–490 (2015).
34. H. L. H. Malaby, W. R. Kobertz, The middle X residue influences cotranslational N-glycosylation consensus site skipping. *Biochemistry* **53**, 4884–4893 (2014).
35. M. Igura, D. Kohda, Quantitative assessment of the preferences for the amino acid residues flanking archaeal N-linked glycosylation sites. *Glycobiology* **21**, 575–583 (2011).
36. Y.-W. Huang, H.-I. Yang, Y.-T. Wu, T.-L. Hsu, T.-W. Lin, J. W. Kelly, C.-H. Wong, Residues comprising the enhanced aromatic sequon influence protein N-glycosylation efficiency. *J. Am. Chem. Soc.* **139**, 12947–12955 (2017).
37. M. Blanchette, W. J. Kent, C. Riemer, L. Elnitski, A. F. A. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, D. Haussler, W. Miller, Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).
38. M. Aebi, N-linked protein glycosylation in the ER. *Biochim. Biophys. Acta* **1833**, 2430–2437 (2013).
39. J. Lombard, The multiple evolutionary origins of the eukaryotic N-glycosylation pathway. *Biol. Direct* **11**, 36 (2016).
40. J. Reumers, S. Maurer-Stroh, J. Schymkowitz, F. D. Rousseau, Protein sequences encode safeguards against aggregation. *Hum. Mutat.* **30**, 431–437 (2009).
41. R. H. P. Law, Q. Zhang, S. McGowan, A. M. Buckle, G. A. Silverman, W. Wong, C. J. Rosado, C. G. Langendorf, R. N. Pike, P. I. Bird, J. C. Whisstock, An overview of the serpin superfamily. *Genome Biol.* **7**, 216 (2006).
42. M. A. Spence, M. D. Mortimer, A. M. Buckle, B. Q. Minh, C. J. Jackson, A comprehensive phylogenetic analysis of the serpin superfamily. *Mol. Biol. Evol.* **38**, 2915–2929 (2021).
43. P. Stanley, Golgi glycosylation. *Cold Spring Harb. Perspect. Biol.* **3**, a005199 (2011).
44. N. A. Cherepanova, S. V. Venev, J. D. Leszyk, S. A. Shaffer, R. Gilmore, Quantitative glycoproteomics reveals new classes of STT3A-and STT3B-dependent N-glycosylation sites. *J. Cell Biol.* **218**, 2782–2796 (2019).
45. H. Nakamura, M. Kiyoshi, M. Anraku, N. Hashii, N. Oda-Ueda, T. Ueda, T. Ohkuri, Glycosylation decreases aggregation and immunogenicity of adalimumab Fab secreted from *Pichia pastoris*. *J. Biochem.* **169**, 435–443 (2021).
46. R. J. Solá, K. Griebenow, Effects of glycosylation on the stability of protein pharmaceuticals. *J. Pharm. Sci.* **98**, 1223–1245 (2009).
47. P. E. Mason, A. Lerbret, M. L. Saboungi, G. W. Neilson, C. E. Dempsey, J. W. Brady, Glucose interactions with a model peptide. *Proteins* **79**, 2224–2232 (2011).
48. J.-H. Han, S. Batey, A. A. Nickson, S. A. Teichmann, J. Clarke, The folding and evolution of multidomain proteins. *Nat. Rev. Mol. Cell Biol.* **8**, 319–330 (2007).
49. I. Sillitoe, N. Bordin, N. Dawson, V. P. Waman, P. Ashford, H. M. Scholes, C. S. Pang, L. Woodridge, C. Rauer, N. Sen, M. Abbasian, S. Le Cornu, S. D. Lam, K. Berka, I. H. Varekova, R. Svobodova, J. Lees, C. A. Orengo, CATH: Increased structural coverage of functional space. *Nucleic Acids Res.* **49**, D266–D273 (2021).
50. T. E. Lewis, I. Sillitoe, N. Dawson, S. D. Lam, T. Clarke, D. Lee, C. Orengo, J. Lees, Gene3D: Extensive prediction of globular domains in proteins. *Nucleic Acids Res.* **46**, D435–D439 (2018).
51. K. W. Plaxco, K. T. Simons, D. Baker, Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994 (1998).
52. G. Merlini, R. L. Comenzo, D. C. Seldin, A. Wechalekar, M. A. Gertz, Immunoglobulin light chain amyloidosis. *Expert Rev. Hematol.* **7**, 143–156 (2014).
53. L. I. Grad, S. M. Fernando, N. R. Cashman, From molecule to molecule and cell to cell: Prion-like mechanisms in amyotrophic lateral sclerosis. *Neurobiol. Dis.* **77**, 257–265 (2015).
54. A. Borgia, K. R. Kemplen, M. B. Borgia, A. Soranno, S. Shammas, B. Wunderlich, D. Nettels, R. B. Best, J. Clarke, B. Schuler, Transient misfolding dominates multidomain protein folding. *Nat. Commun.* **6**, 8861 (2015).
55. D. N. Hebert, L. Lamriben, E. T. Powers, J. W. Kelly, The intrinsic and extrinsic effects of N-linked glycans on glycoproteostasis. *Nat. Chem. Biol.* **10**, 902–910 (2014).
56. M. Sousa, A. J. Parodi, The molecular basis for the recognition of misfolded glycoproteins by the UDP-Glc: Glycoprotein glucosyltransferase. *EMBO J.* **14**, 4196–4203 (1995).
57. B. M. Adams, N. P. Canniff, K. P. Guay, I. S. B. Larsen, D. N. Hebert, Quantitative glycoproteomics reveals cellular substrate selectivity of the ER protein quality control sensors UGGT1 and UGGT2. *eLife* **9**, e63997 (2020).
58. P. Neudecker, P. Robustelli, A. Cavalli, P. Walsh, P. Lundstrom, A. Zarrine-Afsar, S. Sharpe, M. Vendruscolo, L. E. Kay, Structure of an intermediate state in protein folding and aggregation. *Science* **336**, 362–366 (2012).
59. J. S. Tkacz, J. O. Lampen, Tunicamycin inhibition of polyisoprenyl N-acetylglucosaminyl pyrophosphate formation in calf-liver microsomes. *Biochem. Biophys. Res. Commun.* **65**, 248–257 (1975).
60. T. Marquardt, A. Helenius, Misfolding and aggregation of newly synthesized proteins in the endoplasmic reticulum. *J. Cell Biol.* **117**, 505–513 (1992).
61. F. Sparvoli, F. Faoro, M. G. Daminati, A. Ceriotti, R. Bollini, Misfolding and aggregation of vacuolar glycoproteins in plant cells. *Plant J.* **24**, 825–836 (2000).
62. X. Sui, D. E. V. Pires, A. R. Ormsby, D. Cox, S. Nie, G. Vecchi, M. Vendruscolo, D. B. Ascher, G. E. Reid, D. M. Hatters, Widespread remodeling of proteome solubility in response to different protein homeostasis stresses. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 2422–2431 (2020).
63. J. J. Caramelo, A. J. Parodi, A sweet code for glycoprotein folding. *FEBS Lett.* **589**, 3379–3387 (2015).
64. Q. Zhou, H. Qiu, The mechanistic impact of N-glycosylation on stability, pharmacokinetics, and immunogenicity of therapeutic proteins. *J. Pharm. Sci.* **108**, 1366–1377 (2019).
65. F. Courtois, N. J. Agrawal, T. M. Lauer, B. L. Trout, Rational design of therapeutic mAbs against aggregation through protein engineering and incorporation of glycosylation motifs applied to bevacizumab. *MAbs* **8**, 99–112 (2016).
66. F. Brandt, S. A. Etchells, J. O. Ortiz, A. H. Elcock, F. U. Hartl, W. Baumeister, The native 3D organization of bacterial polysomes. *Cell* **136**, 261–271 (2009).
67. S. Shrimal, S. F. Trueman, R. Gilmore, Extreme C-terminal sites are posttranslocationally glycosylated by the STT3B isoform of the OST. *J. Cell Biol.* **201**, 81–95 (2013).
68. S. Shrimal, B. G. Ng, M.-E. Losfeld, R. Gilmore, H. H. Freeze, Mutations in STT3A and STT3B cause two congenital disorders of glycosylation. *Hum. Mol. Genet.* **22**, 4638–4645 (2013).
69. S. Shrimal, N. A. Cherepanova, E. C. Mandon, S. V. Venev, R. Gilmore, Asparagine-linked glycosylation is not directly coupled to protein translocation across the endoplasmic reticulum in *Saccharomyces cerevisiae*. *Mol. Biol. Cell* **30**, 2626–2638 (2019).
70. C. Lizak, S. Gerber, S. Numao, M. Aebi, K. P. Locher, X-ray structure of a bacterial oligosaccharyltransferase. *Nature* **474**, 350–355 (2011).
71. M. L. Medus, G. E. Gomez, L. F. Zacchi, P. M. Couto, C. A. Labriola, M. S. Labanda, R. C. Bielsa, E. M. Clérico, B. L. Schulz, J. J. Caramelo, N-glycosylation triggers a dual selection pressure in eukaryotic secretory proteins. *Sci. Rep.* **7**, 8788 (2017).
72. M. Rita Lecca, U. Wagner, A. Patrignani, E. G. Berger, T. Hennet, Genome-wide analysis of the unfolded protein response in fibroblasts from congenital disorders of glycosylation type-I patients. *FASEB J.* **19**, 1–21 (2005).
73. P. de Haas, M. I. de Jonge, H. J. P. M. Koenen, B. Joosten, M. C. H. Janssen, L. de Boer, W. J. A. J. Hendriks, D. J. Lefeber, A. Cambi, Evaluation of cell models to study monocyte functions in PMM2 congenital disorders of glycosylation. *Front. Immunol.* **13**, 869031 (2022).
74. M. P. Wilson, G. Matthijs, The evolving genetic landscape of congenital disorders of glycosylation. *Biochim. Biophys. Acta Gen. Subj.* **1865**, 129976 (2021).
75. P. Yuste-Checa, A. I. Vega, C. Martín-Higueras, C. Medrano, A. Gámez, L. R. Desviat, M. Ugarte, C. Pérez-Cerdá, B. Pérez, DPAGT1-CDG: Functional analysis of disease-causing pathogenic mutations and role of endoplasmic reticulum stress. *PLOS ONE* **12**, e0179456 (2017).
76. L. Sun, Y. Zhao, K. Zhou, H. H. Freeze, Y.-W. Zhang, H. Xu, Insufficient ER-stress response causes selective mouse cerebellar granule cell degeneration resembling that seen in congenital disorders of glycosylation. *Mol. Brain* **6**, 52 (2013).
77. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
78. J. Hallgren, K. D. Tsirigos, M. D. Pedersen, J. J. A. Armenteros, P. Marcatili, H. Nielsen, A. Krogh, O. Winther, DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. bioRxiv 2022.04.08.487609 [Preprint] (2022). https://doi.org/10.1101/2022.04.08.487609.
79. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
80. M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Zidek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, AlphaFold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
81. W. Kabsch, C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
82. R. P. Joosten, T. A. te Beek, E. Krieger, M. L. Hekkelman, R. W. Hooft, R. Schneider, C. Sander, G. Vriend, A series of PDB related databases for everyday needs. *Nucleic Acids Res.* **39**, D411–D419 (2011).

83. M. Z. Tien, A. G. Meyer, D. K. Sydykova, S. J. Spielman, C. O. Wilke, Maximum allowed solvent accessibilites of residues in proteins. *PLOS ONE* **8**, e80635 (2013).
84. K. M. Ruff, R. V. Pappu, AlphaFold and implications for intrinsically disordered proteins. *J. Mol. Biol.* **433**, 167208 (2021).
85. H. Wickham, in *ggplot2* (Springer, 2016), pp. 189–201.
86. Z. Gu, R. Eils, M. Schlesner, Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
87. T. D. Goddard, C. C. Huang, E. C. Meng, E. F. Pettersen, G. S. Couch, J. H. Morris, T. E. Ferrin, UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).

Acknowledgments: We acknowledge the Electron Microscopy Platform of the VIB–KU Leuven Center for Brain and Disease Research for support and assistance in this work. **Funding:** This work was supported by the Flanders Institute for Biotechnology (VIB, grant no. C0401 to F.R. and J.S.); the Fund for Scientific Research Flanders (FWO, project grant G053420N to J.S.; and Postdoctoral Fellowships 12P0919N and 12P0922N to N.L., 12S3722N to B.H., and 1289023N to M.P.W.); the Jaeken-Theunissen CDG Fund (to G.M. and M.P.W.); and a CELSA fund grant (CELSA/21/027 to G.M. and M.P.W.). **Author contributions:** R.D.-R., B.H., N.L., M.D.V., F.R., and J.S. conceptualized this study. R.D.-R., F.R., J.S., N.L., B.H., M.P.W., and M.D.V. designed the experiments and in silico analyses. R.D.-R. performed all the in vitro experimental work, as well as all in silico analyses. M.D.V. performed peptide synthesis. F.R., J.S., N.L., and G.M. supervised this study. F.R., J.S., and N.L. acquired funding. R.D.-R., F.R., and J.S. wrote the manuscript. All authors proofread and corrected the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions are present in the paper and/or the Supplementary Materials. The codes of all custom R scripts used for data analysis are found publicly on Zenodo (https://doi.org/10.5281/zenodo.10429472).

Submitted 13 September 2023
Accepted 28 December 2023
Published 31 January 2024
10.1126/sciadv.adk8173