Published in final edited form as:
Nat Chem. 2021 May ; 13(5): 480–487. doi:10.1038/s41557-021-00646-w.

Stereoelectronic Effects in Stabilizing Protein–*N*-Glycan Interactions Revealed by Experiment and Machine Learning

Maziar S. Ardejani[†], **Louis Noodleman**[‡], **Evan T. Powers**[†], **Jeffery W. Kelly**[†,§]

[†] Department of Chemistry, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, United States

[‡] Department of Integrative Structural and Computational Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, United States

[§] The Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, United States

Abstract

The energetics of protein-carbohydrate interactions, central to many life processes, cannot yet be predictably manipulated. This is mostly due to an incomplete quantitative understanding of the enthalpic and entropic basis of these interactions in aqueous solution. Here, we show that stereoelectronic effects contribute significantly to stabilizing protein–*N*-glycan interactions in the context of a cooperatively folding protein. Double-mutant cycle analyses of the folding data from 52 electronically-varied *N*-glycoproteins demonstrate an enthalpy-entropy compensation depending on the electronics of the interacting side-chains. Linear and non-linear models obtained using quantum mechanical calculations and machine learning explain up to 79 and 97 % of the experimental interaction energy variability as inferred from the R^2 value of the respective models. Notably, protein-carbohydrate interaction energies strongly correlate with the molecular orbital energy gaps of the interacting substructures. This suggests that stereoelectronic effects must be given a greater weight than previously thought for accurately modelling the short-range dispersive van der Waals interactions between the *N*-glycan and the protein.

Cooperative folding is an emergent property that links the *N*-glycosylation status of an Asn amino acid building block to the systems-level properties of a *N*-glycoprotein.[1, 2] The attachment site(s) of a glycan, its composition, and the amino acid sequence of the protein

Corresponding author name(s): Jeffery W. Kelly.
Author Contributions
M.S.A., L.N., E.T.P and J.W.K. conceived and designed the experiments. M.S.A. carried out the experiments and performed the data analysis. M.S.A., L.N., E.T.P and J.W.K. co-wrote the paper.

Competing Interests Statement
The authors declare no competing interests.

Data availability
All data generated or analyzed during this study are included in this published article (and its supplementary information files).

Code availability
All of the analysis done in this study are carried out using previously published codes and there is no custom code generated during the current study.

all have co-evolved to regulate protein function and maintain optimum thermodynamic and kinetic stability.[3-7] Reaching the same goal by design has been the pursuit of many protein chemists and engineers.[8-10] However, identifying glycosylation sites that can stabilize proteins is difficult, thus success is almost always based on trial-and-error.[11-13] The dearth of engineering guidelines for the systematic design of stabilizing *N*-glycosylation sites arises from our inadequate understanding of the physics underlying carbohydrate-glycoprotein interactions. For instance, the error associated with the current force fields used for *N*-glycoprotein modelling exceeds the measured stabilization energy.[14] This deficit merits the scrutiny of the thermodynamic and electronic origins of protein–*N*-glycan interactions.

Stabilizing carbohydrate–protein-side-chain interactions involve quantum mechanical (QM) effects that are often influenced by the molecular context. An example of such a stability-enhancing interaction occurs in the framework of the $\mathrm{Phe_{i-2}}$-$\mathrm{Xxx_{i-1}}$-$\mathrm{Asn_{i}}$-$\mathrm{Gly_{i+1}}$-$\mathrm{Thr_{i+2}}$ sequence, known as an Enhanced Aromatic Sequon (EAS). Glycosylation of Asn at position “i” of the EAS stabilizes reverse turns in most proteins, e.g., in WW domains, through a face-to-face interaction between the i-2 side-chain aromatic ring and the sugar attached to the Asn side-chain (Fig. 1).[15, 16]

In this report, we use experiment and theory (Fig. 1) to unravel the fundamental physical determinants of weak stabilizing protein–*N*-glycan interactions. We employ cooperatively folding WW domains and a double-mutation cycle approach (Fig. 1) to measure protein–*N*-glycan interaction free energies.[17, 18] This approach requires thoughtful diversification of the chemical identity of the EAS interacting moieties (X and Y in the red box in Fig. 1). Having 52 ΔΔG values arising from these perturbations enables extraction of information about the electronic origins of the EAS interaction. This diversified dataset of interacting residues not only minimizes the effect of potential experimental errors, but also increases the probability that the electronic feature(s) leading to stabilization are more significantly represented and hence are more accessible to statistical identification. This data is fed to machine learning algorithms to identify explanatory models that quantitatively relate the electronic structure of the variable interacting fragments to the experimentally observed free energies (the red trajectory in the Fig. 1). We find that stabilizing protein–*N*-glycan interactions can be explained through optimization of three factors: *i*) electrostatic complementarity, *ii*) non-polar surface burial, and *iii*) molecular orbital interactions between the *N*-linked carbohydrate and the proximal interacting amino acid side-chain, indicating that stereoelectronic effects must be given greater importance than previously considered.

Results and Discussion:

Harnessing peptide and carbohydrate chemistry to expand the repertoire of electronically-varied glycoprotein folding thermodynamics

The mutational perturbations were carefully chosen here to sample a small continuum of chemical space. We mutated the $\mathrm{Phe_{i-2}}$ in the EAS to a series of aromatic and non-aromatic amino acids (Fig. 2), while maintaining the glycan constant as *N*-acetyl-D-glucosamine (GlcNAc). These natural and isosteric unnatural amino acids have varied electronic properties (Fig. 2) as evidenced by the density functional theory (DFT)-calculated electrostatic surface potentials (ESPs). We also synthesized and characterized an analogous series of Pin WW variants with galactose (Gal) as the monosaccharide at the “i” *N*-glycosylation EAS site (Fig. 2). Galactose was chosen because it has the highest chance of utilizing electrostatic interactions amongst commonly available sugars (Supplementary Fig. 1).[19] Therefore, we hypothesized that if there is a significant electrostatic component to the protein–*N*-glycan interaction, variation of the electronic properties of the amino acid side-chains that interact with galactose should enable its quantification. A matching series of non-glycosylated Pin WW domain variants was also made to serve as the baseline for the double-mutant cycle analyses in the GlcNAc and galactose series.

The glycoprotein variants and their non-glycosylated counterparts were prepared using solid phase peptide synthesis by way of an Fmoc-strategy (See SI, Supplementary Figs. 2-5). The Fmoc-protected L-Asparagine(beta-D-galactopyranose-tetraacetate)-OH was prepared by modifying the published reaction scheme (See SI, Supplementary Figs. 2-4). The differential thermal unfolding thermodynamics of the *N*-glycosylated WW domains and their non-glycosylated counterparts were experimentally measured as described previously (See SI).[16] Circular dichroism (CD) and two-dimensional NMR spectroscopy (Supplementary Figs. 6-9) verified that mutating the glycan at the “i” EAS position and/or the side-chain at the “i-2” position does not change the structure of WW domain detectably.

Thermodynamic basis of the stabilizing effect of *N*-glycosylation

The resulting temperature-induced unfolding profiles (Fig. 3a) were used to quantify and parse the thermodynamics of individual protein–*N*-glycan interactions (Supplementary Figs. 10-36). The influence of a given protein–*N*-glycan interaction on the stability of the WW domain was quantified by $\Delta\Delta G_{glyc} = \Delta G_{fold,glyc} - \Delta G_{fold,nonglyc}$ where $\Delta G_{fold,glyc}$ and $\Delta G_{fold,nonglyc}$ are the folding free energies of the *N*-glycosylated and non-glycosylated Pin WW variants, respectively.[20, 16]. Varying the amino acid side-chain at the “i-2” and *N*-glycan at the “i” positions in the EAS modulated the extent to which *N*-GlcNAcylation and *N*-galactosylation of the sequon stabilized the glycosylated WW domain (Fig. 3b). The GlcNAc attached to the Asn side-chain N at the “i” position tends to more strongly enhance the stability of the WW domain than galactose. This difference was even more pronounced in the case of non-aromatic sequons, i.e., when the “i-2” position of the EAS is mutated to a residue with aliphatic side chain (Fig. 3b). Furthermore, a normal distribution of $\Delta\Delta G_{glyc}$ values is obtained (analysis not shown) as a result of a thoughtful selection of mutational perturbations.

To further scrutinize the thermodynamic origins of the stabilizing effect of *N*-GlcNAcylation and *N*-galactosylation, we examined the enthalpic (H) and entropic (S) contributions to the free energy of glycosylation. The quantities of $\Delta\Delta H_{glyc}$ and $\Delta\Delta S_{glyc}$ were independently obtained using two different mathematical analyses of the CD temperature-induced unfolding profiles (Supplementary Figs. 10-36). We observed a negative enthalpic signature for most of the variants (Fig. 4 and Supplementary Fig. 37). However, for some of the stabilized variants, e.g., Gal-PheCN and GlcNAc-PheNO2, we detected a high, positive enthalpy of glycosylation. Remarkably, this unfavorable enthalpy change, is compensated for by an increase in entropy, netting a favorable change in the free-energy of glycosylation (Fig. 4 and Supplementary Fig. 37). This apparent Enthalpy-Entropy Compensation (EEC) is similar to the general EEC observed in carbohydrate-protein interactions[21] and in other molecular recognition systems.[22-24] The EEC observed here cannot be just a mathematical artifact for two reasons. First, the $\Delta\Delta H_{glyc}$ and $\Delta\Delta S_{glyc}$ estimates are physically relevant as they show a significant correlation with $\Delta\Delta G_{glyc}$ (R^2= 0.64 and 0.53, respectively, Supplementary Fig. 38).[23] Second, the measurement temperature, T=333.15 K, falls outside the T_c-2σ (340 K) and T_c+2σ (371 K) range at 95% confidence intervals.[25] The compensating temperature, T_c=356 K, is the slope of the linear fit to ΔΔH versus ΔΔS and σ is the estimated standard error of the fit (Supplementary Fig. 39).

The stabilizing effect of *N*-glycosylation does not entirely derive from enthalpic contributions. Some of the glycovariants with an electron-deficient benzene ring at the “i-2” position (e.g., PheNO2 and PheF3) fail to establish a negative $\Delta\Delta H_{glyc}$ (Supplementary Fig. 37). This effect is more pronounced in the galactose series. These variants are however stabilized through a compensating positive $\Delta\Delta S_{glyc}$ (Supplementary Fig. 37). This means ΔS of folding for these glycosylated variants is larger than that of their non-glycosylated counterparts. A widely accepted explanation for EEC argues that a stronger electronic interaction among interacting moieties (yielding a more negative ΔH) leads to reduced degrees of freedom (DoF) in either or both components of the interaction. This reduced DoF, which lowers the overall conformational entropy, compensates for the enthalpy decrease.[26] A complementary explanation of EEC postulates that a significant part of EEC arises from solvent-mediated effects. During the folding of the glycoprotein, water molecules initially solvating the unfolded state undergo femtosecond rearrangements and make enthalpic and entropic contributions to the folding. The number of waters hydrating and the strength of their interactions with the system can also change during the folding; the contribution of these rearrangements to the enthalpy and entropy of folding will also be mostly compensatory.[27]

One way to unravel the electronic origins of the observed entropic component of the glycosylation-mediated stabilization would be to use physics-based modeling of all variants studied here. However, despite advances in the physics of water dynamics, these contributions are still difficult to estimate.[28, 29]

Quantum mechanical description of the variable protein–*N*-glycan interaction subsystems

Glycans are thought to interact with proteins through a number of solvent-driven and electronic mechanisms originating from the structure of the interacting fragments. Correlation analysis using empirical factors such as the water-octanol transfer free energies (Supplementary Fig. 40) and Hammett constants (Supplementary Fig. 41) of interacting fragments show these factors can only explain a small fraction of variability in glycosylation-mediated stabilization. Structural and energetic analyses led to the hypothesis that van der Waals forces arising from stacking of a monosaccharide onto the “i-2” aromatic ring would be the dominant contributor to the stabilizing effect of a *N*-GlcNAcylated EAS (Fig. 2, top left).[20] Although often considered as minor contributors[30], van der Waals interactions because of their multiplicity could be crucial for obtaining qualitatively reliable and quantitatively accurate descriptions of protein–*N*-glycan interactions.[31–33]

The WW domain is among the smallest cooperatively folding proteins, yet it is too large for *ab initio* molecular dynamics calculations. However, we have designed the experiments herein in such a way that the only variable in our double-mutant cycle analyses is the minimal variation in two fixed positions, i.e., the “i” and “i-2” side chains in the EAS. The fold of the WW domain, as reflected by the very similar NMR spectra amongst the variants (Figs. S6-S9), reliably enforces a face-to-face interaction geometry. The minimal mutational perturbations do not alter the microenvironment of the interaction site and hence the effect of the interaction site on the nature of the interaction is not accessible to statistical analysis. This precise control of mutational perturbations enabled us to hypothesize that the observed variability of the experimental $\Delta\Delta G_{glyc}$ is likely to be a mathematical function of the electronic structure of the variable fragments. To identify this function, we aimed to use machine learning to find a model that would quantitatively relate the electronic structure of the variable interacting fragments at the “i” and “i-2” positions to the experimentally observed $\Delta\Delta G_{glyc}$. The first step in achieving this aim was to translate the interaction system into machine-understandable information. The prior knowledge of electronic substituent effects being important in the interaction subsystem dictated this chemoinformatic translation to be done at the quantum mechanical (QM) level. Therefore, many QM parameters were calculated to achieve a comprehensive electronic description of the variable interacting subsystems in the context of the WW domain system (Supplementary Fig. 42).

The electronic structures of variable subsystem fragments were computed in isolation at the B3LYP/6-31G(d,p) level of theory as implemented in Gaussian 09[34] (see the supplementary information for details). The DFT-obtained electronic structures were used to calculate three series of electronic descriptors of the interacting subsystem. The first series are “mechanistically interpretable” descriptors that are calculated using QM polarizability and dipole moments of isolated fragments and are meant to roughly represent dispersion, dipole-dipole and dipole-induced dipole interactions.[35] The second series of the fragment descriptors were calculated through the quantitative analysis of ESPs at the van der Waals surface of each isolated fragment.[36, 37] The descriptors of each pair of the interacting fragments in every subsystem were summed or multiplied to obtain the “subsystem descriptor” for each subsystem.[38]

The third series of electronic descriptors were calculated using molecular orbital (MO) energies since weak interactions measured here may have some influence from the interfragmentary forces in the region of small MO overlap. These descriptors were obtained through the subtraction of the energy levels of every combination of highest occupied MOs (HOMOs) and lowest unoccupied MOs (LUMOs) of each isolated fragment in the interaction subsystem. Inclusion of the frontier MO energy gaps in the descriptor set was motivated by the recent reports that CH-**π** interactions involve strong interactions among MOs of the CH and **π** system.[39, 40] The probability of MOs being involved in interfragmentary dispersion and orbital interactions depends on the energy gap between the interacting MOs, among other factors. Accurate descriptions of dispersion interactions have been successfully calculated using models incorporating both occupied and unoccupied orbitals.[41, 42] Therefore, we hypothesized that if dispersion and MO interactions are significant contributors to the protein–*N*-glycan interaction, there should be a high correlation between some of the MO energy gaps and the experimental $\Delta\Delta G_{glyc}$.

Taken together, the three descriptor sets summarized above comprise 357 QM descriptors of protein–*N*-glycan interaction subsystems. This machine-learnable QM data together with the experimental $\Delta\Delta G_{glyc}$ dataset were fed into machine learning algorithms to find the most relevant system descriptors on which to base a theoretical model that is both accurate and precise (Supplementary Fig. 42). "System identification" using such a model should uncover hidden effects that control the complex interaction networks, and simple rules to explain them.[43]

Using machine learning to identify the electronic origins of the stabilizing protein–*N*-glycan interactions

We used machine learning algorithms to build explanatory models that would quantitatively relate the electronic structure of the "i" and "i-2" variable interacting fragments in the EAS to the experimentally observed $\Delta\Delta G_{glyc}$. One of the advantages of such fragment-based quantitative structure-energy relationships is that they do not require a prior knowledge of the context-dependent complexity of local interactions. These complexities will be implicitly identified and included in the model during its construction. This is achieved as the process of machine learning selects the most relevant descriptors and optimizes their coefficients to build the final model.

The number of descriptors here (p=357) is much larger than that of the experimental $\Delta\Delta G_{glyc}$ measurements (n=52). To tackle this "large p small n" problem and to defy the "curse of dimensionality", we used a dimension reduction methodology to select the most relevant descriptors. We used a hybrid method based on principal component analysis (PCA), PCA-ranking, that combines subjective and objective descriptor selection criteria.[44] The objective selection is based on the relationship between the principal components (PCs) of the descriptor space and the dependent variable, i.e., $\Delta\Delta G_{glyc}$. In the subjective selection, descriptors are chosen solely considering the relationship between PCs themselves. In the hybrid method, descriptors are selected based on both high variances of the PCs and high correlations of PCs with $\Delta\Delta G_{glyc}$ (Supplementary Fig. 43 and 44). The descriptor subset selected using PCA-ranking, where p < n, is then fed to the "Least absolute shrinkage and selection operator" (LASSO) algorithm to learn a linear explanatory model.

The linear model, built using the three quantum chemical descriptors chosen by LASSO, explains up to 79% of the experimentally observed $\Delta\Delta G_{glyc}$ variance (R^2, Fig. 5a). A LASSO model of randomly generated (non-chemical) features performs poorly (Supplementary Fig. 45). The quantum chemical LASSO model has the following form:

$$\Delta\Delta G_{glyc} = c_1 q_1 + c_2 q_2 + c_3 q_3 + c_4$$

where $c_1 = -0.086$ kcal.mol^{-1}.Å^{-2} and $q_1 = \mathrm{PSA}_{sug} + \mathrm{NSA}_{int}$ (the sum of the surface area of sugar where the ESP is positive, i.e., blue-to-teal regions of ESP in Fig. 2 and the surface area of the protein side-chain where the ESP is negative, i.e., yellow-to-red regions of ESP in Fig. 2); $c_2 = -0.071$ kcal.mol^{-1}.Å^{-2} and q_2 = the sum of the quantum mechanical non-polar surface area (qNPSA)[45] of both fragments; $c_3 = -0.2$ and $q_3 = \mathrm{HOMO\text{-}5}_{sug} - \mathrm{LUMO{+}1}_{int}$ (the energy gap between the 6th highest occupied molecular orbital on the sugar and the 2nd lowest unoccupied molecular orbital on the protein interactor side-chain in kcal.mol^{-1}); and c_4 is −0.2 kcal.mol^{-1}.

Statistical interactions between these descriptors can unravel important interdependencies between the underlying physical effects they represent. For example, we have previously shown that the strength of hydrogen bonding interactions in proteins depends on their microenvironment polarity.[18] Such statistical interactions between the descriptors can be captured by a nonlinear model. Therefore, we hypothesized whether a nonlinear model using the selected descriptors would have a greater explanatory power over the linear model. To test this hypothesis, we used a random forest (RF) algorithm to explore the explanatory potential of such nonlinear models. RF algorithms use bootstrap aggregation, where localized models are built through random and independent sampling of uniformly distributed subsets of data. These local models are then combined to generate a final regressor.[46] A nonlinear model, built using a RF algorithm and the three QM descriptors selected by PCA-ranking–LASSO, explains up to 97 % of $\Delta\Delta G_{glyc}$ variance (R^2, Fig. 5b) and 75 % of the unseen (Out-Of-Bag, OOB score) variance. The OOB score is the mean prediction errors on each and every random set of samples using a prediction model that is built without those samples.[46] The RF model built using randomly generated descriptors does not exhibit an explanatory power comparable to that of the RF model based on the QM descriptors (Supplementary Fig. 45). The performance of the RF model implies potential nonlinear relationships between stereoelectronic factors and protein–*N*-glycan interactions. Therefore, a linear model may not be general and that the strength and type of electronic contributions to protein–*N*-glycan interactions may be context dependent.

The appearance of the quantum mechanical non-polar surface area (qNPSA)[45] in the model suggests that a portion of the stabilizing protein–*N*-glycan interaction is mediated through the non-polar surface of the interacting fragments. An example of such effects could be the hydrophobic burial of the non-polar surfaces of the interacting fragments in water. Selection of PSA_{sug}+NSA_{int} by the machine-learned model can be interpreted based on the physical importance of an electrostatic complementarity between the electron-rich regions of the “i-2” protein side-chain and electron-poor regions of the “i” carbohydrate in the EAS. Here, the negative and positive surface area are the integration of van der Waals surface areas where the ESP is negative, i.e., electron-rich, and positive, i.e., electron-poor (yellow-to-red and teal-to-blue regions of the ESP in Fig. 2), respectively.

Notably, out of the three parameters, the most statistically important parameter for both the linear and the nonlinear models is a particular MO energy gap (Fig. 5c). The explanatory power of one MO energy gap in Fig. 5c is likely indicative of the importance of multiple interfragmentary HOMO-LUMO interactions contributing to $\Delta\Delta G_{glyc}$. Compatible with this hypothesis, several MO energy gaps show a high correlation with the experimental $\Delta\Delta G_{glyc}$ (Fig. 6a, Supplementary Fig. 46 and 47). The extent of correlation and the abundance of highly correlating HOMO-LUMO energy gaps are somewhat dependent on the basis sets used for the DFT calculation (Fig. 6a, Supplementary Fig. 46 and 47). Notably, HOMO-LUMO energy gaps calculated using randomly generated MO energies do not show a significant correlation with the experimental $\Delta\Delta G_{glyc}$ (Supplementary Fig. 46 and 47). This hyperdependence of $\Delta\Delta G_{glyc}$ on the HOMO-LUMO but not HOMO-HOMO or LUMO-

LUMO energy gaps suggests the importance of HOMO-LUMO interactions in stabilizing *N*-glycosylation in the context of the EAS. Several molecular orbital energy gaps having the capacity to be a significant part of the theoretical model is also consistent with recent reports that the CH-π interaction involves strong interactions among multiple molecular orbitals of the interacting fragments.[39, 40] This type of multivalent frontier molecular orbital interaction is compatible with the notion that CH-π interactions in the WW domain system may involve overlap between multiple CH antibonding orbitals (σ^*_{CH}) of the sugar and bonding (π) orbitals of the aromatic interactor moieties that are overrepresented in the dataset.[47] Effects similar to this have been previously observed in interaction between methyl CHs and the π system of carbonyl groups in proteins.[40]

To investigate how interactions among molecular orbitals could be related to the stabilizing effect of *N*-glycosylation, we analyzed the electronic structure of the Asn-GlcNAc–Phe complex. The natural bond orbital (NBO) calculations of this complex reveal the possibility of donor-acceptor type interactions between “bonding” and “antibonding” MOs of Asn-GlcNAc and Phe moieties (Fig. 6b and Supplementary Table 1).[48] Among a number of stabilizing NBO interactions are two sets of $\pi_{C=C}{\rightarrow}\sigma^*_{C\text{-}H}$ and $\pi_{C=O}{\rightarrow}\sigma^*_{C\beta\text{-}H}$ interactions which have second order interaction energies up to 0.47 kcal/mol (Table S1). Moreover, a complex set of weak nonbonding interactions appear in the reduced density gradient (peaks denoted with red triangles in Supplementary Fig. 48) when GlcNAc and Phe side-chains are positioned close to each other in their native-like configuration. Formation of these distinct interfragmentary electron density regions requires establishment of new MOs when GlcNAc and Phe side-chains interact. The gas-phase energy decomposition analysis (EDA) also shows that dispersion and orbital interactions are a significant part of the attractive interaction between Phe or Ala and GlcNAcylated or galactosylated side-chains (Figs 6C-E and Supplementary Fig. 49). The NBO analysis was done using the B3LYP/6–31G(d,p) level of theory with implicit solvation effects and employing empirical dispersion corrections as implemented in Gaussian 16.[49] The EDA was done using BLYP/TZ2P as implemented in ADF[50] (see the Supplementary Information for details).

Conclusion

Examples of rational engineering and systematic design of molecular systems involving carbohydrates remain extremely limited. For example, the EAS is the only portable structural module available for conferring glycosylation-mediated stabilization on a protein. Even the applicability of the EAS is limited to proteins that contain certain kinds of β-turns. This bottleneck mainly arises from a lack of a fundamental understanding of how *N*-glycosylation can stabilize proteins. The goal of the study presented here was to deepen our understanding of the physical basis of stabilizing protein–*N*-glycan interactions down to the electronic level. To approach this, we built a dataset comprising the differential folding free energy information for 52 molecularly matched pairs of glycosylated and non-glycosylated proteins, each hosting an electronically unique protein side-chain–*N*-glycan combination. Thus, the structurally subtle but electronically effectual variation at only two positions differentiates these glycoproteins. Through thermodynamic analysis of these proteins, we have shown that protein–*N*-glycan interactions are rather complex, being dependent on wide-ranging entropy-enthalpy compensation effects. The possibility of these compensatory effects being dependent on the conformational and solvent-driven entropy underscores the importance of considering protein and solvent dynamics for developing systematic methods for molecular glycoengineering. At the electronic level, with the help of DFT calculations and machine learning, we discovered that stabilizing protein–*N*-glycan interactions mainly result from optimization of three factors, namely, electrostatic complementarity, non-polar surface burial, and multiple molecular orbital interactions between the *N*-linked carbohydrate and the interacting amino acid side-chain.

These observations imply that the short-range dispersive interactions between carbohydrates and proteins should follow the energetic and geometric rules of frontier molecular orbital interactions. Molecular orbital energies, both of occupied and virtually vacant, are physical as they have been subject to microscopic observation.[51] Stereoelectronic effects involving CH-**π** electron delocalization in conformationally constrained protein–*N*-glycan interfaces seem to involve an ensemble of HOMO-LUMO overlaps (Figs. 6a and 6b and Supplementary Table 1). These findings are suggestive of multivalent weak $\boldsymbol{\pi}_{C=C}\rightarrow\boldsymbol{\sigma}^*_{C\text{-}H}$ and/or $\boldsymbol{\pi}_{C=O}\rightarrow\boldsymbol{\sigma}^*_{C\beta\text{-}H}$ frontier molecular orbital interactions between carbohydrate and protein side-chains not unlike the intramolecular $n\rightarrow\boldsymbol{\pi}^*$ interactions observed in proteins.[52, 40] The orbital-orbital interactions are highly orientation dependent, which could be one of the reasons why the design of stabilizing protein–*N*-glycan interactions has been methodologically challenging. Our results infer that for accurate modelling of protein-carbohydrate interactions, stereoelectronic effects must be given greater weight than previously thought. The structure-energy relationships tabulated here serve as much-needed guidelines for the improvement of molecular force fields used in the simulation and design of systems involving protein-carbohydrate interactions. Improving these force fields will enable many industrial and therapeutic applications that rely on these interactions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was funded by Grants GM51105 (JWK) and GM100934 (LN) from the National Institutes of Health. The authors thank Gerard J. Kroon and Jane Dyson for the help with the protein NMR experiments; David E. Mortenson for the help with chemical synthesis experiments; Kendall N. Houk, Marcus Jäger and David E. Mortenson for helpful discussions; Jean-Christophe Ducom and Wenge Han for the help with the computational infrastructure; and Colleen Fearns and KyungA Lee for the critical reading of the manuscript. Correspondence and requests for materials should be addressed to J.W.K.

References

1. Hebert DN, Lamriben L, Powers ET, Kelly JW The intrinsic and extrinsic effects of N-linked glycans on glycoproteostasis. Nat. Chem. Biol. 10**,** 902–910 (2014). [PubMed: 25325701] **,**
2. Varki A Biological roles of glycans. Glycobiology 27**,** 3–49 (2016). [PubMed: 27558841] **,**
3. Banks DD The effect of glycosylation on the folding kinetics of erythropoietin. J. Mol. Biol. 412**,** 536–550 (2011). [PubMed: 21839094] **,**
4. Lynch CJ, Lane DA N-linked glycan stabilization of the VWF A2 domain. Blood 127**,** 1711 (2016). [PubMed: 26773038] **,**
5. Ressler VT, Raines RT Consequences of the endogenous N-glycosylation of human ribonuclease 1. Biochemistry 58**,** 987–996 (2019). [PubMed: 30633504] **,**

6. Tan NY, et al. Sequence-based protein stabilization in the absence of glycosylation. Nat. Commun. 5, 3099 (2014). [PubMed: 24434425] ,

7. Yuzwa SA, et al. Increasing o-glcnac slows neurodegeneration and stabilizes tau against aggregation. Nat. Chem. Biol. 8, 393 (2012). [PubMed: 22366723] ,

8. Chang MM, et al. Small-molecule control of antibody N-glycosylation in engineered mammalian cells. Nat. Chem. Biol. 15, 730–736 (2019). [PubMed: 31110306] ,

9. Elliott S, et al. Enhancement of therapeutic protein in vivo activities through glycoengineering. Nat. Biotechnol. 21, 414–421 (2003). [PubMed: 12612588] ,

10. Laughrey ZR, Kiehna SE, Riemen AJ, Waters ML Carbohydrate−**π** interactions: What are they worth? J. Am. Chem. Soc. 130, 14625–14633 (2008). [PubMed: 18844354] ,

11. Chaffey PK, et al. Structural insight into the stabilizing effect of O-glycosylation. Biochemistry 56, 2897–2906 (2017). [PubMed: 28494147] ,

12. Chen MM, et al. Perturbing the folding energy landscape of the bacterial immunity protein im7 by site-specific N-linked glycosylation. Proc. Natl. Acad. Sci. 107, 22528–22533 (2010). [PubMed: 21148421] ,

13. Gavrilov Y, Shental-Bechor D, Greenblatt HM, Levy Y Glycosylation may reduce protein thermodynamic stability by inducing a conformational distortion. J. Phys. Chem. Lett. 6, 3572–3577 (2015). [PubMed: 26722726] ,

14. Woods RJ Predicting the structures of glycans, glycoproteins, and their complexes. Chem. Rev. 118, 8005–8024 (2018). [PubMed: 30091597] ,

15. Culyba EK, et al. Protein native-state stabilization by placing aromatic side chains in *N*-glycosylated reverse turns. Science 331, 571–575 (2011). [PubMed: 21292975] ,

16. Price JL, Powers DL, Powers ET, Kelly JW Glycosylation of the enhanced aromatic sequon is similarly stabilizing in three distinct reverse turn contexts. Proc. Natl. Acad. Sci. 108, 14127–14132 (2011). [PubMed: 21825145] ,

17. Ardejani MS, Powers ET, Kelly JW Using cooperatively folded peptides to measure interaction energies and conformational propensities. Acc. Chem. Res. 50, 1875–1882 (2017). [PubMed: 28723063] ,

18. Gao JM, Bosco DA, Powers ET, Kelly JW Localized thermodynamic coupling between hydrogen bonding and microenvironment polarity substantially stabilizes proteins. Nat. Struct. Mol. Biol. 16, 684–U681 (2009). [PubMed: 19525973] ,

19. Hudson KL, et al. Carbohydrate–aromatic interactions in proteins. J. Am. Chem. Soc. 137, 15152–15160 (2015). [PubMed: 26561965] ,

20. Chen W, et al. Structural and energetic basis of carbohydrate–aromatic packing interactions in proteins. J. Am. Chem. Soc. 135, 9877–9884 (2013). [PubMed: 23742246] ,

21. García-Hernández E, et al. Structural energetics of protein–carbohydrate interactions: Insights derived from the study of lysozyme binding to its natural saccharide inhibitors. Protein Sci. 12, 135–142 (2003). [PubMed: 12493836] ,

22. Fox JM, et al. The molecular origin of enthalpy/entropy compensation in biomolecular recognition. Annu. Rev. Biophys. 47, 223–250 (2018). [PubMed: 29505727] ,

23. Krug RR, Hunter WG, Grieger RA Statistical interpretation of enthalpy–entropy compensation. Nature 261, 566–567 (1976).,

24. Qian H, Hopfield JJ Entropy-enthalpy compensation: Perturbation and relaxation in thermodynamic systems. J. Chem. Phys. 105, 9292–9298 (1996).,

25. Sharp K Entropy—enthalpy compensation: Fact or artifact? Protein Sci. 10, 661–667 (2001). [PubMed: 11344335] ,

26. Bigman LS, Levy Y Entropy-enthalpy compensation in conjugated proteins. Chem. Phys. 514, 95–105 (2018).,

27. Grunwald E, Steel C Solvent reorganization and thermodynamic enthalpy-entropy compensation. J. Am. Chem. Soc. 117, 5687–5692 (1995).,

28. Hassan SA Implicit treatment of solvent dispersion forces in protein simulations. J. Comput. Chem. 35, 1621–1629 (2014). [PubMed: 24919463] ,

29. Zhong D, Pal SK, Zewail AH Biological water: A critique. Chem. Phys. Lett. 503, 1–11 (2011).,

30. Yang L, Adam C, Nichol GS, Cockroft SL How much do van der Waals dispersion forces contribute to molecular recognition in solution? Nat. Chem. 5, 1006 (2013). [PubMed: 24256863] ,

31. Grimme S Density functional theory with london dispersion corrections. Wiley Interdiscip. Rev. Comput. Mol. Sci. 1, 211–228 (2011).,

32. Hermann J, DiStasio RA, Tkatchenko A First-principles models for van der Waals interactions in molecules and materials: Concepts, theory, and applications. Chem. Rev. 117, 4714–4758 (2017). [PubMed: 28272886] ,

33. Wagner C, et al. Non-additivity of molecule-surface van der Waals potentials from force measurements. Nat. Commun. 5, 5568 (2014). [PubMed: 25424490] ,

34. Frisch MJ, et al. Gaussian 09 revision a.2. 2009.

35. Pang S-K Quantum-chemically-calculated mechanistically interpretable molecular descriptors for drug-action mechanism study – a case study of anthracycline anticancer antibiotics. RSC Adv. 6, 74426–74435 (2016).,

36. Lu T, Chen F Quantitative analysis of molecular surface based on improved marching tetrahedra algorithm. J. Mol. Graph. Model. 38, 314–323 (2012). [PubMed: 23085170] ,

37. Murray JS, et al. Statistically-based interaction indices derived from molecular surface electrostatic potentials: A general interaction properties function (GIPF). J. Mol. Struct: THEOCHEM 307, 55–64 (1994).,

38. Pham T-L, et al. Learning structure-property relationship in crystalline materials: A study of lanthanide–transition metal alloys. J. Chem. Phys. 148, 204106 (2018). [PubMed: 29865801] ,

39. Li J, Zhang R-Q Strong orbital interaction in a weak CH-π hydrogen bonding system. Sci. Rep. 6, 22304 (2016). [PubMed: 26927609] ,

40. Perras FA, et al. Observation of CH⋯π interactions between methyl and carbonyl groups in proteins. Angew. Chem. Int. Ed. 56, 7564–7567 (2017).,

41. Iwata S Dispersion energy evaluated by using locally projected occupied and excited molecular orbitals for molecular interaction. J. Chem. Phys. 135, 094101 (2011). [PubMed: 21913747] ,

42. Kapuy E, Kozmutza C Calculation of the dispersion interaction energy by using localized molecular orbitals. J. Chem. Phys. 94, 5565–5573 (1991).,

43. Ardejani MS, Orner BP Obey the peptide assembly rules. Science 340, 561–562 (2013). [PubMed: 23641105] ,

44. Jalali-Heravi M, Shahbazikhah P, Zekavat B, Ardejani MS Principal component analysis-ranking as a variable selection method for the simulation of 13C nuclear magnetic resonance spectra of xanthones using artificial neural networks. QSAR Comb. Sci. 26, 764–772 (2007).,

45. Schaftenaar G, de Vlieg J Quantum mechanical polar surface area. J. Comput. Aided Mol. Des. 26, 311–318 (2012). [PubMed: 22391921] ,

46. Breiman L Random forests. Mach. Learn. 45, 5–32 (2001).,

47. Plevin MJ, Bryce DL, Boisbouvier J Direct detection of CH/π interactions in proteins. Nat. Chem. 2, 466–471 (2010). [PubMed: 20489715] ,

48. Glendening ED, Landis CR, Weinhold F Natural bond orbital methods. Wiley Interdiscip. Rev. Comput. Mol. Sci. 2, 1–42 (2012).,

49. Frisch MJ, et al. Gaussian 16 rev. C.01. Wallingford, CT; 2016.

50. Baerends EJ, et al. ADF2017, SCM. Theoretical Chemistry, Vrije Universiteit, Amsterdam, The Netherlands.

51. Patera LL, Queck F, Scheuerer P, Repp J Mapping orbital changes upon electron transfer with tunnelling microscopy on insulators. Nature 566, 245–248 (2019). [PubMed: 30760911] ,

52. Bartlett GJ, Choudhary A, Raines RT, Woolfson DN n→π* interactions in proteins. Nat. Chem. Biol. 6, 615–620 (2010). [PubMed: 20622857] ,

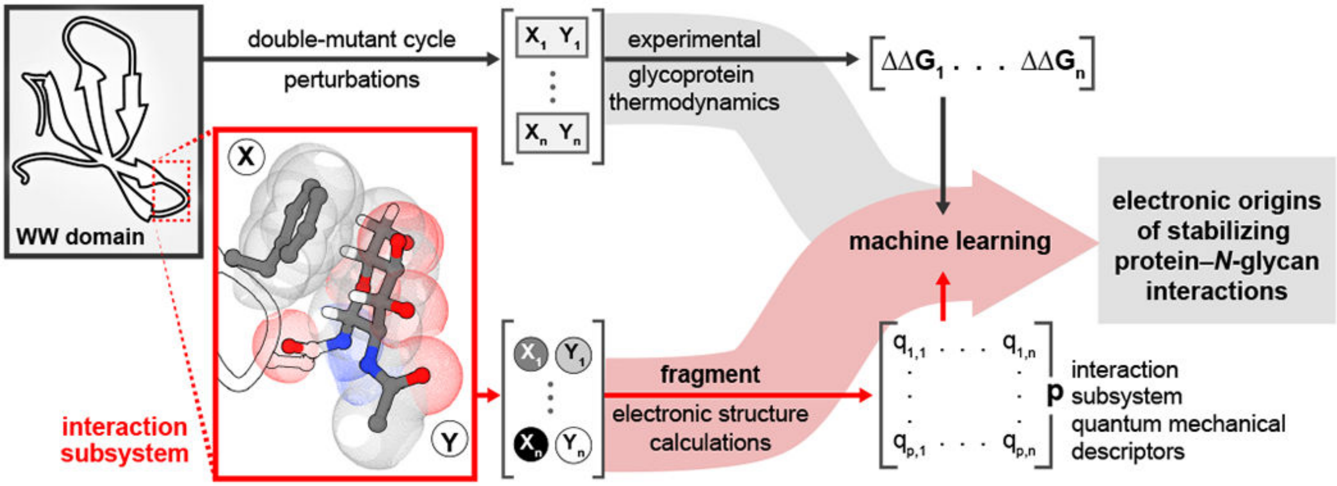

Figure 1 |.

A strategy combining experiment (gray trajectory) and theory (red trajectory) enables probing the thermodynamic and electronic origins of protein-carbohydrate interactions. WW domain (gray box) hosts the interaction subsystem (dotted red box) of the enhanced aromatic sequon (EAS). Perturbation of the EAS interaction via chemical incorporation of two guest residues (X and Y) and double-mutant cycle analysis provide a series of experimental ΔΔG of glycosylation values (n=52). The fragment-based ab initio calculations generate a large set (p=357) of QM descriptors of the interaction subsystem. Statistical and machine-learned models that relate the experimental ΔΔG to QM properties lead to the discovery of the electronic origins of protein–*N*-glycan interactions.

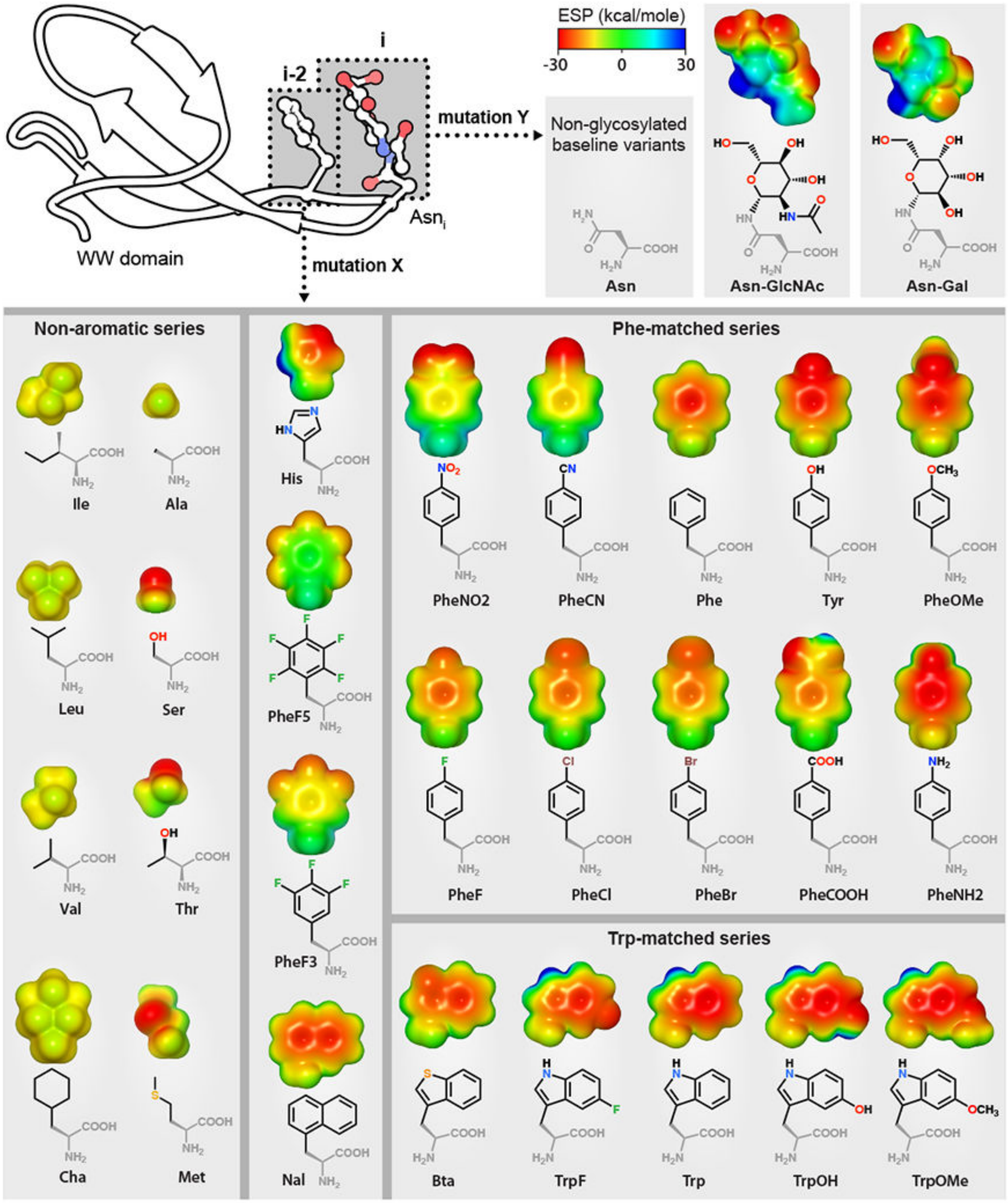

Figure 2 |.

An expanded repertoire of electronically-varied N-glycosylated proteins is constructed using chemical incorporation of natural and unnatural amino acids. Top left: Our solution structure of a 5-residue EAS in loop1 of the Pin WW domain shows a face-to-face stacking interaction between Phe aromatic side-chain at the “i-2” interactor position and the α-face of GlcNAc at the “i” position (PDB ID 2M9F). Top right: Structures and electrostatic surface potentials (ESPs) of the monosaccharide attached to the Asn side-chain amide N in the EAS.

Bottom: Structures and ESPs of the aromatic and non-aromatic amino acid side-chain at the “i-2” position of the EAS in our data set.

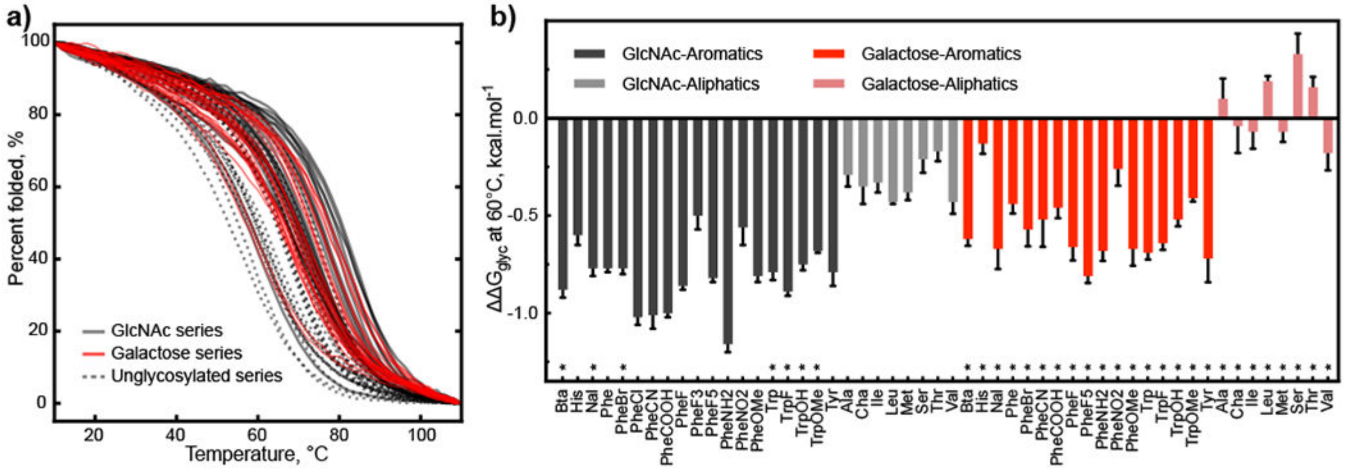

Figure 3 |.

The effect of N-glycosylation on the temperature-induced unfolding profile depends on the chemical identity of the sugar and amino acid at the interaction site. a) The unfolding profiles were obtained using variable temperature circular dichroism (CD) spectropolarimetry. b) Quantitative analysis of the protein unfolding profiles reveals that installation of GlcNAc or galactose at position “i” can stabilize the protein structure to different extents based on the “i-2” aromatic or aliphatic side-chain. ΔΔGglyc values were obtained by fitting the temperature-induced CD unfolding profiles (Supplementary Figs. 10-36) to the van’t Hoff or Taylor’s series expansion equations. Experimental measurements made in the current study are marked by asterisks; whereas the rest of values are obtained by reanalyzing the data from WW domains previously studied[20, 16]. The error bars represent standard error of mean (SEM).

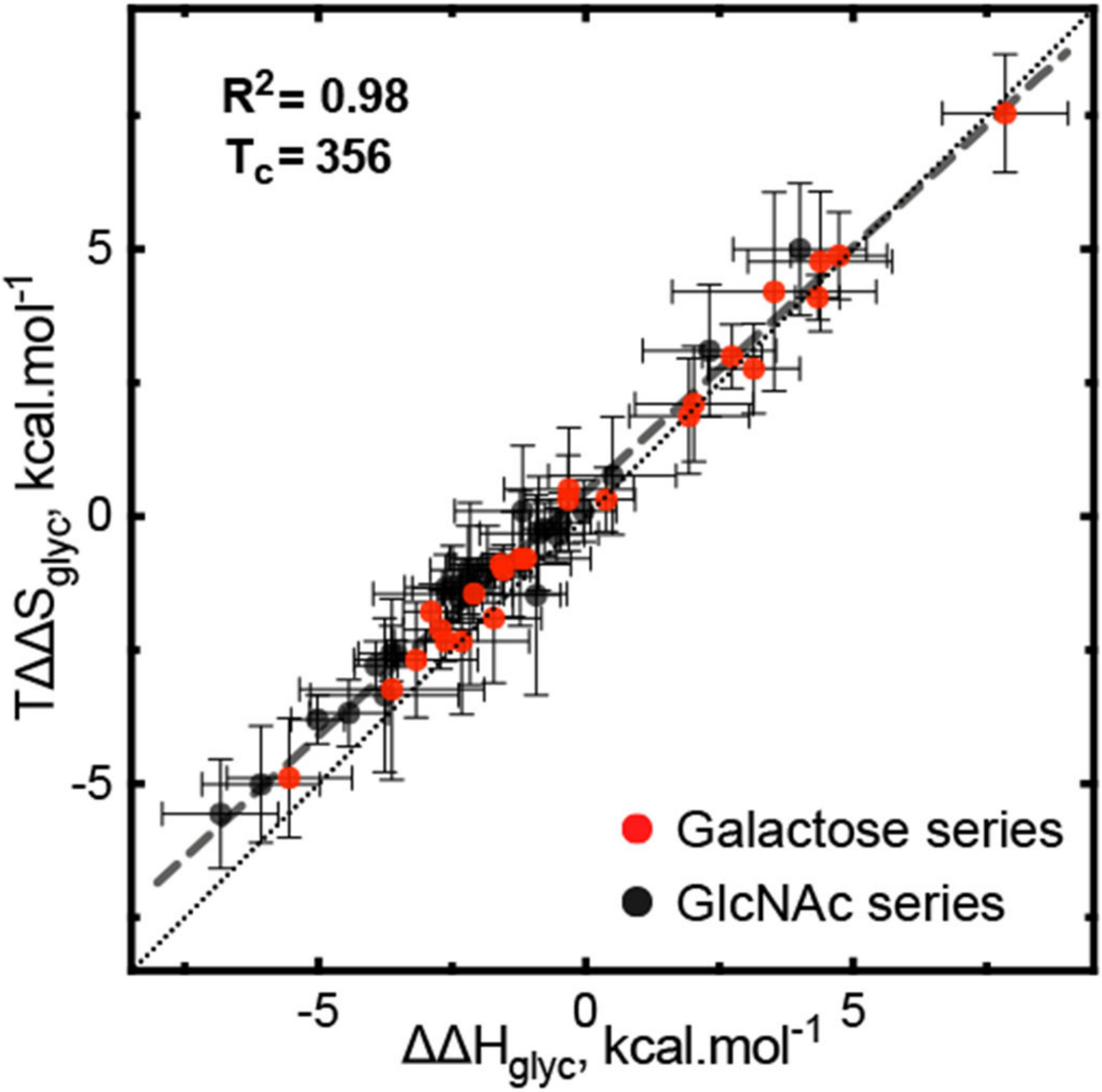

Figure 4 |.

The thermodynamic origin of $\Delta\Delta G_{glyc}$ differs among the electronically-varied WW glycovariants. The thermal unfolding of the *N*-glycosylated WW variants exhibits a diverging entropy-enthalpy compensation. The entropy is expressed as TΔΔS at T=333.16K. The error bars represent standard error of mean (SEM).

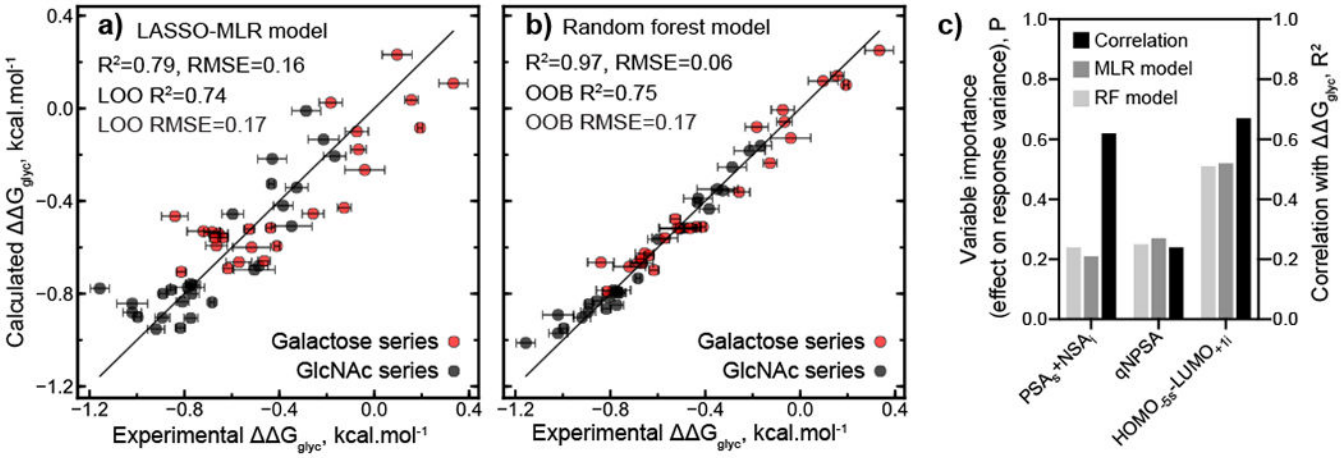

Figure 5 |.

Machine learning extracts the stereoelectronic factors that explain the stabilizing protein–*N*-glycan interactions. a) A linear model selected by the LASSO algorithm explains 79 % of variability of the experimentally measured $\Delta\Delta G_{glyc}$ and 74 % of the unseen (leave-one-out, LOO) variance. b) A random forest (RF) nonlinear model explains up to 97 percent of $\Delta\Delta G_{glyc}$ variance and 75 percent of the unseen (out-of-bag, OOB) variance. c) Relative importance of quantum-chemical descriptors does not vary among the linear and nonlinear models.

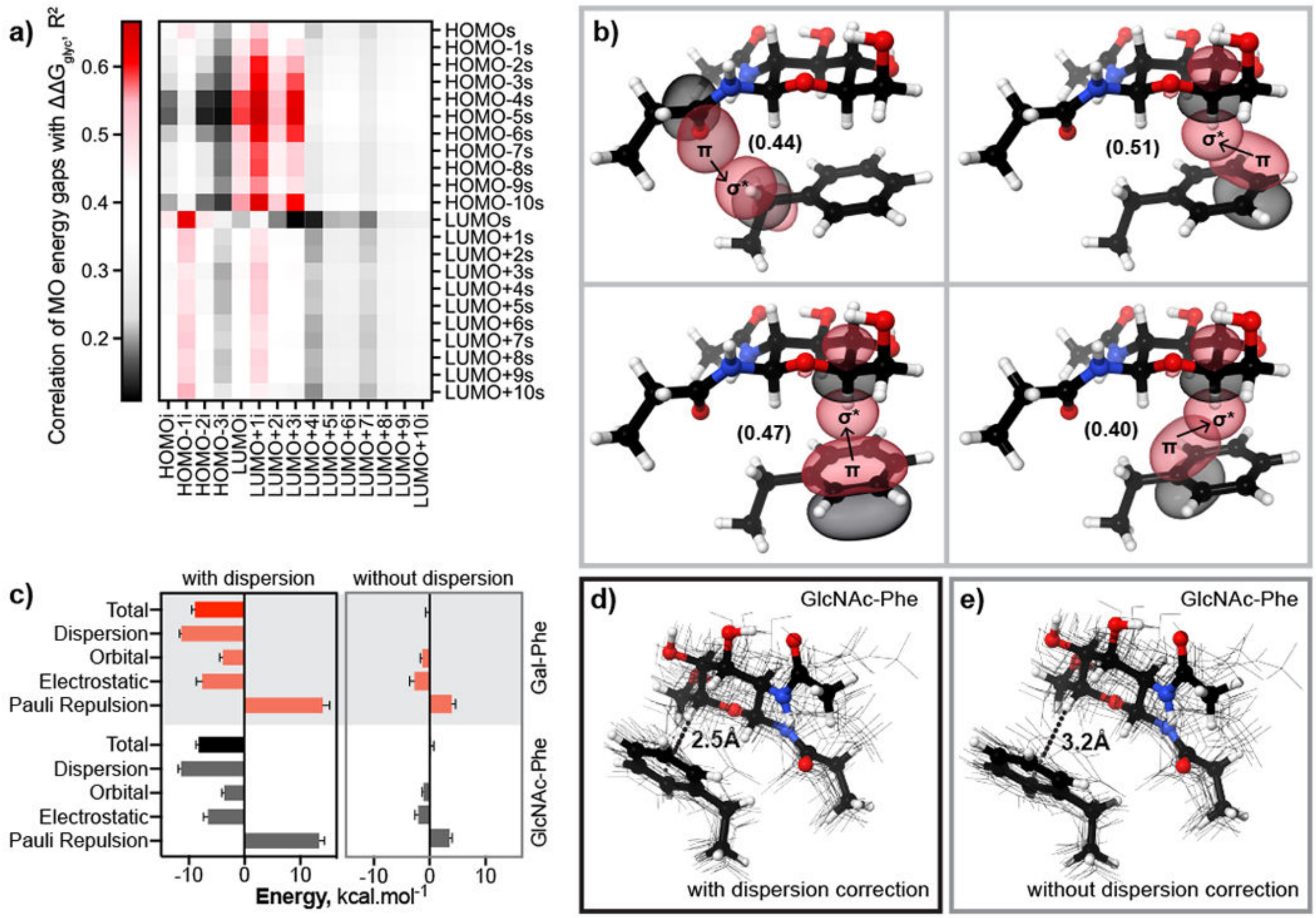

Figure 6 |.

Interactions between molecular orbitals contribute to the stabilizing effect of *N*-glycosylation. a) High correlations are observed between certain HOMO-LUMO, but not LUMO-LUMO or HOMO-HOMO energy gaps, and $\Delta\Delta G_{glyc}$. MO energy gaps were calculated at the B3LYP/6-31G(d,p) level of theory. b) NBO computed orbital interactions corresponding to $\pi_{C=O} \rightarrow \sigma^*_{C\beta\text{-}H}$ or $\pi_{C=C} \rightarrow \sigma^*_{C\text{-}H}$. Second order interaction energies are given in parentheses in kcal.mol^{-1}. c, d and e) Energy Decomposition Analysis (EDA) of Phe-GlcNAc and Phe-galactose interactions. Twenty NMR-obtained structures of GlcNAc-Phe (d and e) and Gal-Phe (Supplementary Fig. 49) whose geometries were optimized through minimization of QM energy and used for EDA. The lowest energy structures are shown in stick representation and rest of the quantum-mechanically optimized ensemble is shown in wire representation. The distances shown in c and e are between the center of sugar's axial hydrogen and the center of the aromatic ring. Each EDA and its corresponding geometry optimization were done with and without dispersion corrections shown in the left and right panels of c, respectively. The error bars represent SEM.