Stereoelectronic effects in stabilizing protein-\(N\)-glycan interactions revealed by experiment and machine learning

Maziar S. Ardejani\textsuperscript{1,1}, Louis Noodleman\textsuperscript{2,2}, Evan T. Powers\textsuperscript{1} and Jeffery W. Kelly\textsuperscript{1,3,4}

The energetics of protein-carbohydrate interactions, central to many life processes, cannot yet be manipulated predictably. 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 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 nonlinear 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, the 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 an \(N\)-glycoprotein.\textsuperscript{14} The attachment site(s) of a glycan, its composition and the amino-acid sequence of the protein have all co-evolved to regulate protein function and maintain optimum thermodynamic and kinetic stability.\textsuperscript{3–7} Reaching the same goal by design has been the pursuit of many protein chemists and engineers.\textsuperscript{8–10} However, identifying glycosylation and maintaining optimum thermodynamic and kinetic stability\textsuperscript{3–7}.

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 Phe\textsubscript{i−2}–Xxx\textsubscript{i−1}–Asn\textsubscript{i}–Gly\textsubscript{i+1}–Thr\textsubscript{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, for example, 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).\textsuperscript{15,16}

In this Article, 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.\textsuperscript{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 \(\Delta\Delta G\) values arising from these perturbations enables extraction of information about the electronic origins of the EAS interaction.

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 Phe\textsubscript{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 electrostatic surface potentials (ESPs) calculated by density functional theory (DFT). 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 among the commonly available sugars (Supplementary Fig. 1).\textsuperscript{19} 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 (Supplementary Information and Supplementary Figs. 2–5). The Fmoc-protected L-asparagine (β-D-galactopyranose-tetraacetate)-OH was prepared by modifying the published reaction scheme (Supplementary Information and Supplementary Figs. 2–4)14. The differential thermal unfolding thermodynamics of the N-glycosylated WW domains and their non-glycosylated counterparts were experimentally measured as described previously (Supplementary Information)15. Circular dichroism (CD) and two-dimensional (2D) 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 the WW domain in a detectable manner.

**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 ΔΔGglyc = ΔGfold,glyc − ΔGfold,complex, where ΔGfold,glyc and ΔGfold,complex are the folding free energies of the N-glycosylated and non-glycosylated Pin WW variants, respectively16,21. Varying the amino-acid side chain at the ‘i−2’ position and N-glycan at the ‘i’ position 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, that is, when the ‘i−2’ position of the EAS is mutated to a residue with an aliphatic side chain (Fig. 3b). Furthermore, a normal distribution of ΔΔGglyc 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 ΔΔHglyc and ΔΔSglyc 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, for example, Gal-PheCN and GlcNAc-PheNO2, we detected a high positive enthalpy of glycosylation. Remarkably, this unfavourable enthalpy change is compensated by an increase in entropy, netting a favourable 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 interactions22 and in other molecular recognition systems23–25. The EEC observed here cannot be just a mathematical artefact for two reasons. First, the ΔΔHglyc and ΔΔSglyc estimates are physically relevant as they show a relatively high correlation with ΔΔGglyc (R²=0.64 and 0.53, respectively; Supplementary Fig. 38)24. Second, the measurement temperature, Tc=333.15 K, falls outside the Tc−2σ (340 K) and Tc+2σ (371 K) range at 95% confidence intervals26. The compensating temperature, Tc=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 (for example, PhenoN, and PheF) fail to exhibit a negative ΔΔHglyc (Supplementary Fig. 37). This effect is more pronounced in the galactose series. These variants are, however, stabilized through a compensating positive ΔΔSglyc (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 reduced DoF, leading degrees of freedom (DoF) in either or both components of the interaction. This reduced DoF,
Fig. 2 | An expanded repertoire of electronically varied N-glycosylated proteins constructed using chemical incorporation of natural and unnatural amino acids. Top left: our solution structure of a five-residue EAS in loop1 of the Pin WW domain shows a face-to-face stacking interaction between the Phe aromatic side chain at the 'i − 2' interactor position and the α-face of GlcNAc at the 'i' position (PDB 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 chains at the 'i − 2' position of the EAS in our dataset.
which lowers the overall conformational entropy, compensates for the enthalpy decrease. A complementary explanation of EEC postulates that a part of EEC arises from solvent-mediated effects. However, despite advances in the physics of water dynamics, these contributions are still difficult to estimate. The postulates that a part of EEC arises from solvent-mediated effects.

QM 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 energy (Supplementary Fig. 40) and Hammett constants (Supplementary Fig. 41) of interacting fragments show that 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’ aromatic ring would be the dominant contributor to the stabilizing effect of a N-GlcNAcyalted EAS (Fig. 2, top left). Although often considered as minor contributors, van der Waals interactions, because of their multiplicity, could be crucial for obtaining qualitatively reliable and quantitatively accurate descriptions of protein–N-glycan interactions.

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, that is, the ‘i’ and ‘i − 2’ side chains in the EAS. The fold of the WW domain, as reflected by the very similar NMR spectra among the variants (Supplementary Figs. 6–9), reliably enforces a face-to-face interaction geometry. The minimal mutational perturbations do not alter the microenvironment of the interaction site, so 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 ΔΔGϕ屋 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 ΔΔGϕ屋. 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 that this chemoinformatic translation be done at the 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 [1] (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 comprises '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 [2]. The second series of fragment descriptors were calculated through the quantitative analysis of ESPs at the van der Waals surface of each isolated fragment [3]. The descriptors of each pair of interacting fragments in every subsystem were summed or multiplied to obtain the 'subsystem descriptor' for each subsystem [4].

The third series of electronic descriptors were calculated using MO energies, because 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 systems [5]. 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 calculated using models incorporating both occupied and unoccupied orbitals [6,7]. Therefore, we hypothesized that if dispersion and MO interactions are major contributors to the protein–N-glycan interaction, there should be a high correlation between some of the MO energy gaps and the experimental ΔΔΔΔGglyc.

Taken together, the three descriptor sets summarized above comprise 357 QM descriptors of protein–N-glycan interaction subsystems. These machine-learnable QM data, together with the experimental ΔΔΔΔGglyc 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, as well as simple rules to explain them [8].

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 ΔΔΔΔGglyc. One of the advantages of such fragment-based quantitative structure–energy relationships is that they do not require 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 (n = 357) is much larger than that of the experimental ΔΔΔΔGglyc 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—which combines subjective and objective descriptor selection criteria [9]. The objective selection is based on the relationship between the principal components (PCs) of the descriptor space and the dependent variable, that is, ΔΔΔΔGglyc. 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 ΔΔΔΔGglyc (Supplementary Figs. 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 ΔΔΔΔGglyc variance (R², 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:

\[
\Delta \Delta \Delta \Delta G_{\text{glyc}} = c_1 q_1 + c_2 q_2 + c_3 q_3 + c_4
\]
models. RF algorithms use bootstrap aggregation, where localized algorithm to explore the explanatory potential of such nonlinear linear model. To test this hypothesis, we used a random forest (RF) selected descriptors would have a greater explanatory power over the Therefore, we hypothesized whether a nonlinear model using the importance interdependencies between the underlying physical between the descriptors can be captured by a nonlinear model. 

where \( c_1 = -0.086 \text{ kcal mol}^{-1} \text{Å}^{-2} \) and \( q_4 = \text{PSA}_{\text{ag}} + \text{NSA}_{\text{ag}} \) (the sum of the surface area of sugar where the ESP is positive, that is, blue-to-teal regions of ESP in Fig. 2 and the surface area of the protein side chain where the ESP is negative, that is, yellow-to-red regions of ESP in Fig. 2); \( c_1 = -0.071 \text{ kcal mol}^{-1} \text{Å}^{-2} \) and \( q_4 \) is the sum of the quantum mechanical non-polar surface area (qNPSA)\(^{46}\) of both fragments; \( c_1 = -0.2 \) and \( q_1 = \text{HOMO-5}_{\text{ag}} - \text{LUMO} + 1_{\text{ag}} \) (the energy gap between the sixth highest occupied molecular orbital on the sugar and the second lowest unoccupied molecular orbital on the protein interactor side chain in kcal mol\(^{-1}\)); and \( c_1 \) is \(-0.2 \text{ kcal mol}^{-1}\).

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.\(^{45}\) 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.\(^{45}\) A nonlinear model, built using a RF algorithm and the three QM descriptors selected by PCA ranking–LASSO, explains up to 97% of \( \Delta \Delta G_{\text{glyc}} \) 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.\(^{45}\) 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 qNPSA\(^{46}\) 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$_{\alpha}$ and PSA$_{\beta}$ 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 (that is, electron-rich) and positive (that is, electron-poor) (yellow-to-red and teal-to-blue regions of the ESP in Fig. 2), respectively.

Notably, 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 probably indicative of the importance of multiple interfragmentary HOMO–LUMO interactions contributing to $\Delta \Delta G_{\text{glyc}}$. Compatible with this hypothesis, several MO energy gaps show a high correlation with the experimental $\Delta \Delta G_{\text{glyc}}$ (Fig. 6a and Supplementary Figs. 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 and Supplementary Figs. 46 and 47). Notably, HOMO–LUMO energy gaps calculated using randomly generated MO energy gaps do not show a substantial correlation with the experimental $\Delta \Delta G_{\text{glyc}}$ (Supplementary Figs. 46 and 47). This hyperdependence of $\Delta \Delta G_{\text{glyc}}$ 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 MO energy gaps having the capacity to be an essential part of the theoretical model is also consistent with recent reports that the CH–π interaction involves strong interactions among multiple MOs of the interacting fragments. This type of multivalent frontier MO interaction is compatible with the notion that CH–π interactions in the WW domain system may involve overlap between multiple CH antibonding orbitals (σ$_{\text{C-H}}$) of the sugar and bonding (π) orbitals of the aromatic interactor moieties that are overrepresented in the dataset. Effects similar to this have been observed previously in interactions between methyl CHs and the π system of carbonyl groups in proteins.

To investigate how interactions among MOs could be related to the stabilizing effect of N-glycosylation, we analysed 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 the Asn-GlcNAc and Phe moieties (Fig. 6b and Supplementary Table 1). Among a number of stabilizing NBO interactions are two sets of σ$_{\text{C=C}}$ → σ$_{\text{C-H}}$ and σ$_{\text{C=O}}$ → σ$_{\text{C=H}}$ interactions, which have second-order interaction energies up to 0.47 kcal mol$^{-1}$ (Supplementary Table 1). Moreover, a complex set of weak non-bonding 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 orientation. 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 an important part of the attractive interaction between Phe or Ala and GlcNAcylated or galactosylated side chains (Fig. 6c–e and Supplementary Fig. 49).

The NBO analysis was done using the B3LYP/6-31G(d,p) level of theory with implicit solvent effects and employing empirical dispersion corrections as implemented in Gaussian 16. The EDA was done using BLYP/TZ2P as implemented in ADF (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 EEC 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 have discovered that stabilizing protein–N-glycan interactions mainly result from optimization of three factors, namely, electrostatic complementarity, non-polar surface burial, and multiple MO 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 MO interactions. MO energies, both occupied and virtually vacant, are physical, as they have been subject to microscopic observation. Stereoelectronic effects involving CH–π electron delocalization in conformationally constrained protein–N-glycan interfaces seem to involve an ensemble of HOMO–LUMO overlaps (Fig. 6a,b and Supplementary Table 1). These findings are suggestive of multivalent weak π→σ$_{\text{C-H}}$ and/or π→σ$_{\text{C=H}}$ frontier MO interactions between carbohydrate and protein side chains, not unlike the intramolecular n→π* interactions observed in proteins. 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 indicate 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.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41557-021-00646-w.

Received: 26 November 2019; Accepted: 27 January 2021; Published online: 15 March 2021

References

1. Hebert, D. N., Lamriben, L., Powers, E. T. & Kelly, J. W. The intrinsic and extrinsic effects of N-linked glycans on glycoproteostasis. Nat. Chem. Biol. 10, 902–910 (2014).
2. Varli, A. Biological roles of glycans. Glycobiology 27, 3–49 (2016).
3. Banks, D. D. The effect of glycosylation on the folding kinetics of erythropoietin. J. Mol. Biol. 412, 536–550 (2011).
28. Grunwald, E. & Steel, C. Solvent relaxation and thermodynamic entropy–enthalpy compensation. J. Am. Chem. Soc. 105, 95–105 (2013).

29. Hussain, S. A. Implicit treatment of solvent dispersion forces in protein simulations. J. Comput. Chem. 35, 1621–1629 (2014).

30. Zhong, D., Pal, S. K. & Zewail, A. H. Biological water: a critique. Chem. Phys. Lett. 503, 1–11 (2011).

31. Yang, L., Adam, C., Nichol, G. S. & Cockroft, S. L. How much do van der Waals dispersion forces contribute to molecular recognition in solution? Nat. Chem. 5, 1006–1010 (2013).

32. Grimmie, S. Density functional theory with London dispersion corrections. WIREs Comput. Mol. Sci. 1, 211–228 (2011).

33. Hermann, J., DiStasio, R. A. & Tkatchenko, A. First-principles models for van der Waals interactions in molecules and materials: concepts, theory and applications. Chem. Rev. 117, 4714–4758 (2017).

34. Wagner, C. et al. Non-additivity of molecule-surface force of van der Waals potentials from force measurements. Nat. Commun. 5, 5568 (2014).

35. Frisch, M. J. et al. Gaussian 09, Revision A.2 (Gaussian, 2009).

36. 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).

37. Lu, T. & Chen, F. Quantitative analysis of molecular surface based on improved matching tetrahedra algorithm. J. Mol. Graph. Model. 38, 314–323 (2012).

38. Murray, J. S. 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).

39. 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).

40. Li, J. & Zhang, R.-Q. Strong orbital interaction in a weak CH⋯π hydrogen bonding system. Sci. Rep. 6, 22304 (2016).

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

42. Iwata, S. Dispersion energy evaluated by using locally projected occupied and excited molecular orbitals for molecular interaction. J. Chem. Phys. 135, 094101 (2011).

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

44. Ardejani, M. S. & Orner, B. P. Obey the peptide assembly rules. WIREs Comput. Mol. Sci. 3, 311–318 (2012).

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

46. Plevin, M. J., Bryce, D. L. & Boisbourdier, J. Direct detection of CH⋯π interactions in proteins. Nat. Chem. 2, 466–471 (2010).

47. Ghendinger, E. D., Landis, C. R. & Weinhold, F. Natural bond orbital methods. WIREs Comput. Mol. Sci. 2, 1–42 (2012).

48. Frisch, M. J. et al. Gaussian 16, Revision C.01 (Gaussian, 2016).

49. Schäfer, G. A. & de Vlueg. J. Quantum mechanical polar surface area. J. Comput. Aided Mol. Des. 26, 311–318 (2012).

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

51. Hussain, S. A. Implicit treatment of solvent dispersion forces in protein simulations. J. Comput. Chem. 35, 1621–1629 (2014).
Data availability
All data generated or analysed during this study are included in this paper and its Supplementary Information.

Code availability
All of the analysis done in this study was carried out using previously published codes. No custom code was generated during the current study.

Acknowledgements
This work was funded by grants GM51105 (J.W.K.) and GM100934 (L.N.) from the National Institutes of Health. The authors thank G. J. Kroon and J. Dyson for their help with the protein NMR experiments, D. E. Mortenson for help with the chemical synthesis experiments, K. N. Houk, M. Jäger and D. E. Mortenson for helpful discussions, J.-C. Ducom and W. Han for help with the computational infrastructure and C. Fears and K. Lee for critical reading of the manuscript.

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. Correspondence and requests for materials should be addressed to J.W.K.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41557-021-00646-w.
Correspondence and requests for materials should be addressed to J.W.K.
Peer review information Nature Chemistry thanks Christopher Bauer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at www.nature.com/reprints.