Cryptic SARS-CoV2-spike-with-sugar interactions revealed by 'universal' saturation transfer analysis

Charles J. Buchanan[2]ψ, Ben Gaunt[1]ψ, Peter J. Harrison[3,12], Audrey Le Bas[3], Aziz Khan[2], Andrew M. Giltrap[1,2], Philip N. Ward[3], Maud Dumoux[1], Sergio Daga[5], Nicola Picchiotti[8,9], Margherita Baldassarri[5], Elisa Benetti[7], Chiara Fallerini[5], Francesca Fava[5,6], Annarita Giliberti[5], Panagiotis I. Koukos[11], Abirami Lakshminarayanan[2], Xiaochao Xue[2,4], Georgios Papadakis[2], Lachlan P. Deimel[4], Virgínia Casablancas-Antràs[2], Timothy D.W. Claridge[2], Alexandre M.J.J Bonvin[11], Quentin J. Sattentau[4], Simone Furini[7], Marco Gori[8,10], Jiangdong Huo[1,3], Raymond J. Owens[1,3], Alessandra Renieri[5,6], GEN-COVID Multicenter Study, James H. Naismith*[1,3], Andrew Baldwin*[2], Benjamin G. Davis*[1,2]

[1] The Rosalind Franklin Institute, Harwell Science & Innovation Campus, OX11 0FA, UK

[2] Department of Chemistry, University of Oxford, Oxford, OX1 3TA

[3] Division of Structural Biology, University of Oxford, The Wellcome Centre for Human Genetics, Headington, Oxford, OX3 7BN, UK

[4] Sir William Dunn School of Pathology, South Parks Road, Oxford, UK

[5] Medical Genetics, University of Siena, Siena, Italy

[6] Azienda Ospedaliera Universitaria Senese, Siena, Italy

[7] Department of Medical Biotechnologies, University of Siena, Siena, Italy

[8] Department of Information Engineering and Mathematics, University of Siena, Italy

[9] Department of Mathematics, University of Pavia, Pavia, Italy

[10] Université Côte d'Azur, Inria, CNRS, I3S, Maasai

[11] Bijvoet Centre for Biomolecular Research, Faculty of Science, Utrecht University, Netherlands

[12] Diamond Light Source, Harwell Science & Innovation Campus, Oxfordshire, UK

ψ These authors contributed equally.

* Correspinding authors: naismith@strubi.ox.ac.uk, andrew.baldwin@chem.ox.ac.uk, ben.davis@chem.ox.ac.uk

Abstract
~400 words

Host-expressed proteins on both host-cell and pathogen surfaces are widely exploited by pathogens, mediating cell entry (and exit) and influencing disease progression and transmission. This is highlighted by the diverse modes of coronavirus entry into cells and their consequent differing pathogenicity that is of direct relevance to the current SARS-CoV-2 pandemic. Host-expressed viral surface proteins bear post-translational modifications such as glycosylation that are essential for function but can confound or limit certain current biophysical methods used for dissecting key interactions. Several human coronaviruses attach to host cell-surface *N*-linked glycans that include forms of sialic acid. There remains, however, conflicting evidence as to if or how SARS-associated coronaviruses might use such a mechanism. Here, we show that novel protein NMR methods allow a complete and comprehensive analysis of the magnetization transfer caused by interactions between even heavily modified proteins and relevant ligands to generate quantitative binding data in a general manner. Our method couples direct, objective resonance-identification via a deconvolution algorithm with quantitative analysis using Bloch-McConnell equations to obtain interaction parameters (e.g. K_D, k_{Ex}), which together enable structural modelling. By using an automated and openly available workflow, this method can be readily applied in a range of systems. This complete treatment of so-called 'saturation transfer' between protein and ligand now enables a general analysis of solution-phase ligand-protein binding beyond previously perceived limits of exchange rates, concentration or system – this allows 'universal' saturation transfer analysis (uSTA). uSTA proves critical in mapping direct interaction between natural sialoside sugar ligands and SARS-CoV-2-spike glycoprotein by quantitating ligand signal in spectral regions otherwise occluded by resonances from mobile spike-protein glycans (that also include sialosides). Using uSTA, 'end on'-binding by SARS-CoV-2-spike protein to sialoside glycan is revealed, which contrasts with an observed 'extended surface'-binding for previously validated heparin sugar ligands. Quantitative use of uSTA-derived restraints pinpoints likely binding modes to an intrinsically disordered region of the *N*-terminal domain of SARS-CoV-2-spike trimer. Consistent with this, glycan binding is minimally perturbed by antibodies that neutralize via binding the ACE2-binding domain (RBD) but strongly disrupted in the B1.1.7 and B1.351 variants-of-concern that possess hotspot mutations around the identified site. An analysis of beneficial genetic variances in cohorts of patients from early 2020 suggests a possible model in which A-lineage-SARS-CoV-2 may have exploited a specific sialylated-polylactosamine motif found on tetraantennary human *N*-linked-glycoproteins in deeper lung. Since cell-surface glycans are widely relevant to biology and pathology, uSTA can now provide a ready, quantitative method for widespread analysis of complex, host-derived and post-translationally modified proteins with putative ligands relevant to disease even in previously confounding complex systems.

Introduction

~800 words

Both SARS-CoV-1[1-3] and SARS-CoV-2[4,5] gain entry to host cells through the use of receptor-binding domains (RBDs) of their respective spike proteins that bind human cell-surface protein ACE2. Although other coronaviruses that infect humans, including MERS,[6] are thought to use cell-surface sugar sialosides[6-10] as initial attachment points prior to binding entry receptor proteins, the corresponding interaction (despite predictions[10,11]) has remained unclear for either of the SARS-causing viruses to date. In preliminary disclosures, contradictory format-dependent observations have been made[12,13] as to whether complex sialosides are or are not bound.[12,14] Instead, glycosaminoglycans on proteoglycans, such as heparin, have been suggested as a primary cooperative glycan attachment point.[14,15] Those studies suggesting sugar binding have all, to date, implicated possible binding sites in or close to the RBD. Nonetheless, the region of the spike protein that binds sialosides in other coronaviruses, the N-terminal domain (NTD), also appears in SARS-CoV-2-spike to display a putative glycan binding groove in currently determined structures (**Supplementary Figure S1**).[4,16,17] The unresolved role of host cell surface sialosides in this pathogen has been highlighted as an important open question.[15] Clear and more detailed elucidation of sugar-binding by SARS-CoV-2-spike is therefore needed.

The classical 'saturation transfer difference' (STD) NMR experiment[18] is a widely-used method to reveal interactions between putative small molecule ligands and proteins[19] without the need for labelling/modification of the ligand or protein, or the need to attach either ligand or protein to a surface or sensor. In the STD experiment the nuclear spin systems inside a given protein are moved out of equilibrium using a 'saturation' pulse. When a ligand binds to the protein, the two cross-relax, which results in magnetization being passed from the ligand to the protein (**Supplementary Figure S2,3**). In a dynamic binding equilibrium, the two then dissociate and the cycle repeats during the saturation pulse. Although 'saturation' is not itself a measurable phenomenon, after a pre-set duration, the remaining reduced signal on the ligand is recorded, and the difference in signal for each resonance (the 'saturation transfer difference') from a reference spectrum indicates that ligand and protein have made contact. This magnetization transfer has often been colloquially described in terms of modes of 'saturation transfer' and, indeed, sometimes given an interpreted converse direction of 'saturation transfer' from protein-to-ligand[20] although, in practice, signal on the ligand is in fact reduced by this process.

Complex, highly-modified protein systems prove difficult to analyze by current STD methods for several reasons. Firstly, mammalian proteins, or those derived from expression in mammalian hosts such as SARS-CoV-2-spike, often bear large and highly mobile glycans. Critically, in the case of glycoproteins that may themselves bind glycans this leads to contributions to the spectra that overlap with putative ligand resonances thereby obscuring signal. Secondly, the NMR spectra of glycan ligands are themselves complex, comprising many overlapped resonances and their multiplets. This can limit the accurate determination of intensities needed in STD. Finally,

it has long been perceived that STD is only possible under certain kinetic regimes and/or under certain limiting assumptions about the ligand-to-protein binding equilibrium. This has meant that certain regimes and systems have been considered inaccessible to STD.

Here, by devising a new, complete formulation of magnetization / 'saturation' transfer (**Supplementary Figures S5, S6, S8**), we demonstrate that the range of regimes for which the experiment can be considered reliable is substantially larger than generally assumed (**Supplementary Figure S4**). By implementing a rigorous theoretical description of the experiment, coupled with a computational approach based on a Bayesian deconvolution algorithm to accurately extract data, it proved possible to reliably and quantitatively determine precise on and off rates and K_Ds describing the interaction.

This 'universal' saturation transfer analysis (uSTA) then only requires a series of specific simplified magnetization transfer spectra to be acquired on a protein-only sample, and then with a range of ligand concentrations (**Supplementary Figures S5, S6, S8**). The signal intensity from individual resonances in these spectra are obtained automatically. These are then converted into on and off rates (and K_Ds) via complete theoretical treatment. These can also provide per resonance transfer efficiencies that, when used as constraints to high-level (e.g. HADDOCK) computational modelling environments provide exact structural models. In this way, uSTA analysis provides an automated pipeline from raw NMR free induction decay (FID) signals all the way to protein•ligand structures in a freely available form for the non-expert.

uSTA proves effective in a range of model ligand-protein systems and, by evaluating multiple glycan ligands, identifies a glycan-binding pocket in the NTD of A-lineage[21] SARS-CoV-2 spike as the most potent, and distinct from other SARS-CoV-2-spike interactions. Notably, this binding is strongly ablated in B1.1.7 and B1.351 lineage[21] variants of SARS-CoV-2 spike providing now an additional, potential molecular explanation for widely different infectivities. These data also prove to be consistent with observations made through correlative genetic analyses with disease severity in SARS-CoV-2-positive patients infected in early 2020 with A-lineage[21] virus, suggesting a model by which cell-surface glycans and the modulation of binding by spike-to-sialosides plays a role in infection, disease progression and disease evolution.

Results
~ 4500 words

Analysis of the Current Limits to 'STD'

As noted, the classical 'STD' experiment cannot be reliably applied to mammalian-derived proteins to quantitatively survey ligand binding (**Supplementary Figures S6,S8**). Our theoretical analyses (**Supplementary Figure S4**), however, suggested that many common assumptions or limits that are thought to govern the applicability of the current STD experiment might in fact be circumvented and we set out to devise a complete treatment that might accomplish this (**Supplementary Figures S5,S6,S8**).

First, and perhaps most notably, appreciable transfer of signal can in fact occur for a wider range of timescales and strength of protein•ligand interaction than previously recognised (**Supplementary Figures S2,S4**). The sensitivity of such experiments is strictly independent of chemical exchange, in that the chemical shift difference between the free and bound ligand conformation is not in itself relevant to the mechanism of 'saturation'/magnetization transfer. Instead, such modes are governed by the number of ligands that come into contact with protein and the cross relaxation between the two – this does not restrict the saturation transfer to the 'fast' exchange regime, as is commonly argued.[20]

Second, in seeking to precisely quantitate 'STD', our analysis also reveals that it is vital to systematically vary both the protein and the ligand concentration in order to robustly separate the exchange parameters from concentration-independent relaxation processes (**Supplementary Figure S4**).[20,22] In this way, forward and backward rates (and K_D) can be consistently obtained via experiments at multiple concentrations coupled with global, complete analyses of cross-relaxation.

Third, as in the case of heavily glycosylated human-derived proteins covered by highly mobile sugars[23] such as SARS-CoV-2 spike, the problems may be greatly exacerbated. The conformational flexibility of these groups is such that their R_2 relaxation rates become extremely short[23-25] leading to greatly confounded difference spectra. In conventional STD experiments, using for example the 'group epitope' method,[26] relaxation filters are added at the end of an experiment where protein signal is hoped to evenly decay away, ideally leaving only signal from ligands. However, in a protein that contains both mobile modifications as well as mobile disordered regions, such as SARS-CoV-2-spike, this approach is no longer viable since protein signal remains even after aggressive use of such relaxation filters (**Supplementary Figures S5,S6,S8**). Moreover, relaxation filters inevitably result in reduced overall sensitivity and effects of intra-ligand cross-relaxation via the nuclear Overhauser effect (NOE) and ROE inhibit the elucidation of atom-specific data. In principle, relaxation filters however would not be necessary at all with reliable extraction of ligand-only signal. We considered that baseline subtraction of residual signal transfer using reference to samples containing only protein could prove possible if methods for precise resonance identification could be developed.

Finally, NMR spectra from biological ligands such as sugars can be extremely complex, containing a large number of multiplets across diverse chemical shifts, each of which can be highly overlapped. In order to accurately quantify such a spectrum for use in saturation transfer methods, it is necessary to accurately determine the degree of magnetization (and its change) for each and all observed proton resonances. This too would be addressed by precise resonance identification (**Supplementary Figures S5,S6,S8**).

Design of uSTA based on a Comprehensive Treatment of nOe

In order to completely and quantitatively model signal intensity via uSTA we first considered full aspects of the appropriate spin physics (**Supplementary Figure S2**). In brief, initially, the protein and ligand resonances are at equilibrium. Saturation transfer experiments[18] then move spin systems inside the protein that are within the bandwith of the 'pulse' out of equilibrium (via a 'saturation pulse', **Supplementary Figure S2**). These resonances then cross-relax with adjacent spins in the protein (spin-diffusion) which results in wide-ranging sets of spin systems within the protein that are effectively out of equilibrium. During this period, previously free ligand binds the protein and the two molecules cross relax via NOE (magnetization is passed from ligand that is in equilibrium to protein that is not). Modelling the degree to which signal has passed between the ligand and protein is complex; transfer depends on the on and off rates (and hence the K_D), and the intrinsic cross-relaxation rates, which depend on the tumbling time of the complex (**Supplementary Figure S4**). A model for the complex signal transfer suitable for fitting to data has therefore not, to our knowledge, been addressed to date. However, these factors would, in principle, be straightforward to accommodate in a theoretical analysis of the transfer, using fully comprehensive numerical approaches. In this approach, all relevant interactions could be fully and hence quantitatively described by modified Bloch-McConnell equations. Indeed, Bloch-McConnell approaches prove successful in other protein NMR methods involving dynamical processes such as CEST[27,28] or DEST.[29] Although it is more parametrically complex to analyze such magnetization transfer data using this formalism, all that would be required, in principle, is calculation[30] of the ratio of signal loss before and after the saturation pulse for each and every identified resonance as well as the total concentration of protein and ligand (**Supplementary Figure S5, S6, S8**). In this way, the theoretical description of the experiment would allow calculation of the expected transfer of signal with varying K_{ex} and K_D values (see **Methods**).

Second, as noted above, central to this process is precise resonance identification and accurate extraction of signal intensity. This was made possible by the design of a Bayesian computational method[31] to detect ligand and protein resonances in the raw spectral data, even when these are highly complex and/or obscured by competing signals (**Supplementary Figure S5, S6, S8**). By this method, spectra could be automatically reduced to a series of constituent peaks and intensities (**Figure 1A,B**).[32,33] To accomplish this, it was necessary to accurately sum the intensity from each multiplet that the resonances are split into via scalar coupling. On a

fundamental mathematical level these resonances in the NMR spectrum can be described as a convolution between a peak shape and a set of weighted delta functions (a 'delta matrix', see **Figure 1A,B** and **Methods**). The uSTA algorithm was therefore also designed to exploit this relationship to allow automatic detection of resonances and iteratively return a set of unique delta functions that describe the peak position and intensities; this process is executed simultaneously for both spectra acquired (pulse on and pulse off) in order to exactly and precisely determine not only peak positions, but the change in intensity due to magnetization transfer. Pleasingly, the algorithm successfully returned reconvolved spectra that were close to identical to the original spectra (**Figures 1, 2, 3**).

The consequence was not only resonance identification ('peak-picking') but also combined multiplicity-determination and overall signal quantification (the latter being critical to precise uSTA, see below) (**Figure 1B**). The results of this algorithm were also demonstrated/tested in the analysis of overlapped 1D spectra (**Figures 1B** – see also overlaps in all subsequent analyses), where the algorithm successfully returned the locations and intensities of the 3 source functions used to simulate the overlapped spectrum shown. In this way, and in an iterative manner, a 'near perfect' peak shape was supplied to give an accurate representation of the spectrum (**Figure 1A**). Moreover, when sufficient peaks are combined, then this process also allows the determination of multiplicities for a given more complex resonance through the combination of multiple peaks that may be similarly re-convolved and compared (**Figure 1B**). The success of the method was immediately evident from the 'raw' (unperturbed/reference) data; raw data and algorithmically-derived spectra were essentially identical in all cases (see all **Figures** for such overlaps). The peak positions identified by the algorithm also mapped exactly to the expected positions of resonances as established using conventional, manual NMR assignment (see **Supporting Methods**).

Importantly, in the highlighted context of precise resonance identification this enabled signal intensity determination in the background of other (possibly confounding) resonances (**Supplementary Figure S5, S6, S8**). Moreover, those background resonances themselves could be similarly analyzed. In this way the two contributing components found in any saturation transfer experiment (i.e. protein (P) and ligand (L)) were therefore determined with dramatically improved precision (**Figure 1C**). With prior knowledge of the peak positions following analysis of the 1D NMR spectrum, only these positions could then be considered as possibilities for providing signal in the difference spectrum. This approach thereby provided precision and enabled a detailed quantitative analysis via uSTA by revealing signals that would be otherwise 'hidden'.

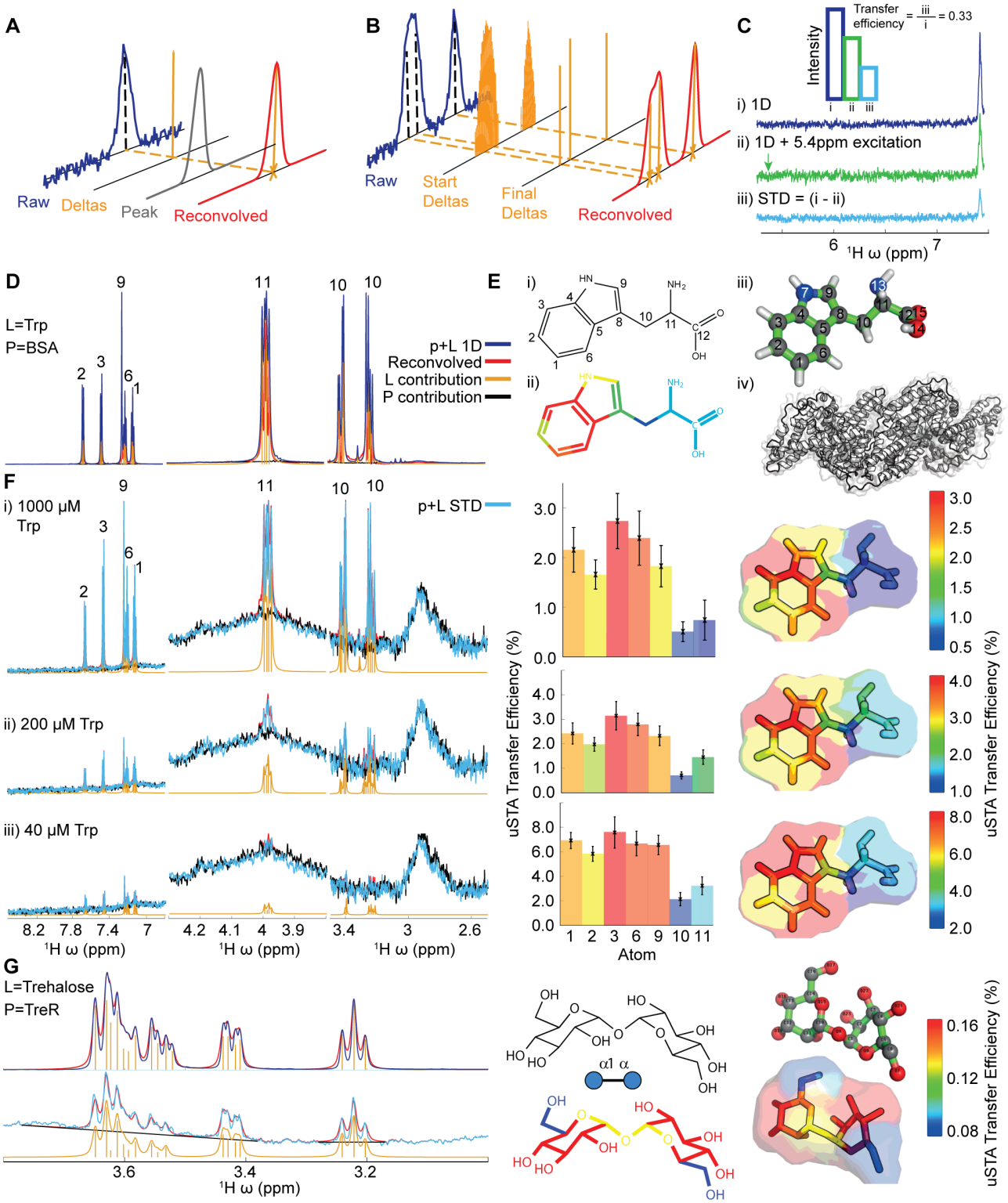


Figure 1. **Development of the uSTA Method.**
A),**B**): Schematic of process for automatically quantifying the signal intensities in NMR spectra using deconvolution. The analysis determined the number of peaks that can give rise to the signal, and return a simulated spectrum by convolving these with a peak shape function. Precise peak positions and their intensities are returned, which can be attributed to single protons observed in a spectrum.
C): The STD NMR experiment compares two 1D NMR spectra, where the second involves a specific saturation pulse that aims to hit the protein but miss the ligand. The uSTA method requires these two spectra to be analysed in pairs, one that contains the raw signal, and the second that is the difference between the two. We define the 'transfer efficiency' as the fractional signal that has passed from the ligand to the protein. This measurement allows us to perform a detailed and

quantitative analysis that enables us to obtain structural, kinetic and thermodynamic information about the interaction between small molecules and proteins.
D)**,E), F)**: Application of uSTA to study the interaction between bovine serum albumin (BSA) and L-Tryptophan (Trp). In **D** the 1D ^{1}H-NMR spectrum of the mixture at 200 µM Trp and 5 µM BSA (= p+L, **blue**) is dominated by ligand but yet the ligand (L) and protein (P) can still be deconvolved using universal deconvolution, using a reference obtained from a sample containing protein only. This reveals contributions from individual multiplets originating from the ligand (**yellow**) and the protein only baseline (**black**) allowing precise recapitulation of the sum (**red**). In **E** application of universal deconvolution to STD spectra with varying concentrations of tryptophan allows uSTA analysis using ligand resonances identified in **D**. This, in turn allows signal intensity in the STD spectrum (p + L STD, **light blue**) to be determined with high precision. While signal-to-noise in the STD increases considerably with increasing ligand concentration, the measured atom-specific transfer efficiencies as determined by uSTA (C) are remarkably similar (**D, middle bar charts** and **right transfer efficiency 'maps'**), showing that the primary contact between protein and ligand occurs on the C4a-C7a bridge of the indole aromatic ring.
G): Application of the same uSTA workflow allows precise determination of even weakly binding sugar ligand trehalose (Glc-α1,1α-Glc) to *E. coli* trehalose repressor TreR. Again, the uSTA allows determination of transfer efficiencies with atom-specific precision.

Validation of uSTA's Precision in Model Systems

To fully establish a working system and test the full potential of the putative uSTA method (**Figure 1C**) we chose next to examine an archetypal, yet challenging, ligand-to-protein interaction. This was implemented in an automated manner through software that governed the uSTA workflow and that therefore reduced artefacts arising from subjective, manual analyses (**Supplementary Figure S6**). The binding of L-tryptophan (Trp) to bovine serum albumin (BSA, **Figure 1D**) is a long-standing benchmark[34] in STD analyses due to the plasticity of this interaction based on supposed hydrophobicity but also due to a lack of corresponding fully-determined, unambiguous crystal structures. This is therefore a system that standard NMR/STD methods are perceived[20,35] to have already delineated well. We therefore set out to see if we could improve upon this benchmark through even more precise analyses via uSTA (**Figure 1D-F**).

As for a standard STD experiment, 1D ^{1}H-NMR spectra were determined for both ligand and protein. In addition, mixed spectra (p + L) containing both protein and excess ligand were determined with and without irradiation at frequencies corresponding to prominent resonance within the protein. Using the newly designed uSTA protocol, each of these spectra were processed automatically (see **Supplementary Figure S6** and **Methods**). The deconvolved spectrum for ligand determined in the presence of protein was matched with high accuracy by uSTA (**Figure 1D**). Moreover, we were able to achieve highly consistent 'heat maps', comprising atom-specific transfer efficiencies over a range of ligand concentrations that describe the pose of the ligand in the presence of the protein (**Figure 1F** see also **Supplementary Figures S7**,**10**) – these were determined not only at a single concentration but a variety of concentrations, even as low as 40 μM where the ability of uSTA to extract accurate signal proved critical (**Figure 1F, left**). These maps used observed proton data mapped onto heteroatoms to enable visual comparison of binding modes. Maps were strikingly consistent across a range of concentrations suggesting that a single, consistent pose was adopted by Trp when bound by BSA. The determined 'maps' of this pose suggested much stronger interaction of protein with the heteroaromatic indole side-chain of Trp, consistent with prior observations and known crystal structures with other hydrophobic ligands.[36] They also revealed quantitative subtleties to this interaction, suggesting that a stronger interaction / 'grip' is felt by the centre of the indole ring (across bond C4a–C7a) than at its periphery, an observation that to our knowledge had not been previously possible using standard methods.

Next, having demonstrated high precision of the method in a benchmark system and greater capability enabled by automated workflow (**Supplementary Figure S6,S8**) that gives rise to consequent quantitative analysis, we chose to analyze a more testing system based instead upon sugar-protein interaction. The trehalose repressor system has been analyzed previously by STD methods.[37] The sugar ligand trehalose binds only weakly in this system and gives rise to a more challenging ligand-to-protein interaction analysis. Nonetheless, the uSTA workflow again successfully and rapidly determined atom-specific transfer efficiencies with high precision and resolution (**Figure 1G**). It also revealed hotspots of binding around OH-3 and OH-4 in a subtle,

atom-precise manner with graduation to reduced binding around both pyranoside rings with only minimal/less binding of the primary hydroxyl (**Figure 1G**). This p+L interaction also proved consistent with prior X-ray crystal structural analysis,[38] although it should be noted that very low ligand occupancy in these prior analyses means that uSTA once again revealed greater detail than could be confidently gained previously.

Determination of K_D in a direct manner using uSTA.

In this way, via uSTA, variation of concentration also allowed determination of binding constants directly from spectra. The transfer of magnetization between ligand and protein here was quantitatively analyzed using modified Bloch-McConnell equations,[39] (see **Methods**) that account for intrinsic relaxation, cross-relaxation and protein-ligand binding to directly provide measurements of equilibrium binding K_D and associated kinetics (k_{ex}). In the Trp•BSA system this readily revealed a K_D = 32 ± 12 μM and k_{on} = 6300 ± 2300 $M^{-1}s^{-1}$, k_{off} = 0.20 ± 0.08 s^{-1}, consistent with a prior determination of K_D = 40 μM by isothermal calorimetry.[35] It should be noted that this direct method proved only possible due to the ability of the uSTA method to deconvolute a true signal with sufficient precision, even at the lower concentrations used despite lower signal (**Figure 1F**); uSTA through its precision vitally enabled atom-mapping and quantitation for ligand binding over previous methods. Critically, these values were fully consistent with all observed NMR data and independently obtained measures of K_D.

uSTA allows observation of Cryptic Sugar-Binding with SARS-CoV-2 Spike

As noted above, the binding of glycans as a pre-attachment mode during COVID-19 infection has been predicted, although not yet clearly delineated. Our analysis of the 1D protein ^{1}H-NMR spectrum of the purified prefusion-stabilised ectodomain construct[4] of intact trimeric SARS-CoV-2-spike protein (see **Supplementary Figure S7** for preparation details) revealed extensive glycosylation on the protein with sufficient mobility to generate a strong ^{1}H-NMR resonance in the region 3.4-4.0 ppm (**Figure 2B(i)**). Whilst lacking detail, these resonances displayed chemical shifts consistent with the described mixed patterns of oligomannose, hybrid and complex *N*-glycosylation found on SARS-CoV-2-spike after human expression.[40] As such, these mobile 'internal' glycans on SARS-CoV-2-spike contain glycan residues – i.e. sialosides – that are also potential competing ligands for any putative pre-attachment interactions. Therefore, their presence in the protein NMR analysis presented clear confounding issues for typical STD analyses. As such, SARS-CoV-2-spike represented a stringent and important test of the uSTA method.

We evaluated a representative panel of both natural and site-specifically-modified unnatural sialosides (**Figure 2A** and **Figure 3**). Strikingly, whilst use of standard methods provided an ambiguous assessment (**Supplementary Figure S11**), use of uSTA immediately revealed binders

and non-binders (**Figure 3** and **Supplementary Figures S8**,**S10,S11** and **Supplementary Table S7**).

First, the simplest sialioside, *N*-acetyl-neuraminic acid (**1**), was tested as a mixture of its mutarotating anomers (**1α ⇔ 1β**). When analyzed using uSTA, these revealed (**Figure 4** and **Supplementary Table S7** and **Supplementary Figures S11**) low affinity but clear interactions by **1α** as a ligand, mediated primarily by NHAc-5, but no measurable interactions by **1β**. It should be noted that the detection of selective α-anomer interaction, despite the great dominance of the β-anomer in solution, provided yet another further demonstration of the precision of the uSTA method (**Figure 4** and **5A** and **Supplementary Figure S11**). Together these data not only unambiguously confirmed interaction of SARS-CoV-2-spike with the smallest natural sialoside sugar but also revealed striking configurational selectivity (**α** over **β**) by the protein. This **α-**selectivity is notably consistent with the near-exclusive occurrence of **α-**sialosides on host cell surfaces as their α- but not β-linked conjugates (**Figure 4**).

Next, having confirmed simple, selective monosaccharide α-sialoside binding we explored extended, α-sialoside ligands that might give further insight into natural endogenous human sugar ligands that could be bound by SARS-CoV-2, as well as unnatural variants that could potentially modulate and even interrupt such binding. First, oligosaccharides of α-sialic acid were explored. Sialosides are often found appended to galactosyl (Gal/GalNAc) residues in either α2,3- (**2**) linked or α2,6- linked form (**3**); both were tested (**Figure 3A(i,ii)**). Both exhibited consistent, more extended binding surfaces with greater potential for protein interaction. Qualitatively this suggested a stronger binding affinity (see below for quantitative analysis). Common features of both binding modes were observed: as for monosaccharide **1**, NHAc-5 acetamide group at C5 of the terminal sialic acid residues (Sia) proved to be a key binding hotspot for both **2** and **3**. Differences between **2** and **3** were also observed: the α2,6-trisaccharide (**3**) system displayed a more extended binding face yet with less intense binding hot-spots (**Figure 3A**, and corresponding lower overall interaction, see below) engaging additionally the side-chain glycerol moiety (C7-C9) of the terminal Sia acid as the OH-4 C4 hydroxyl of the galactosyl (Gal) residue. The interaction with α2,3-trisaccharide (**2**) was tighter (see below) and more specific to NHAc-5 of the Sia.

The extended interaction displayed by the C7-C9 Sia side-chain in some of these natural ligands caused us to consider whether modification of this side-chain might engage further and/or alternative binding modes. To this end, we also constructed modified unnatural sialosides to explore this possibility (**4**-**6**, **Figure 2A**). In the simplest variant we replaced the OH-9 hydroxyl group of sialic acid Sia to generate deoxygenated azide **4**; this too interacted in an anomer-specific manner with only the α-variant displaying interaction, despite again the dominance of the β-anomer in solution, as for **1**. The atom-specific map of **4** (**Figure 4C** and **Supplementary Table S7**) suggested that some interaction was lost through the removal of OH-9 (consistent also with binding analysis – see below). Notably, however, when C9 was instead extended by a biphenyl-

carboxamide (BPC) moiety to generate **5**, a strong *additional* interaction was generated with this introduced hydrophobic portion (**Figure 2B,C,D** and **Figure 3B,C**). It is particularly noteworthy that uSTA allowed exquisite detail to be determined in this unusual hybrid ligand (**Figure 2B,C,D**) that could not have been determined using standard methods, as these would have been dominated simply by the most potent hydrophobic interaction (with ligand sugar protons lost within glycoprotein sugar responses). In this way, we could demonstrate that such modified ligands bind through *all* portions of their surfaces but a distinct difference of interaction is observed in hybrid ligands: a greater contact was seen with the hydrophobic moiety than with the carbohydrate moiety. Such is the precision of uSTA, graded binding even within these two portions of the ligand was also determined (see atom-specific scoring, **Figure 2C,D**). For example in **5**, greatest binding for the hydrophobic moiety was seen at the tip with lowering, graduated binding towards the amide junction at C9. In the sugar portion, despite a quite different contact type, individual gradation was also observed: the C7-C9 sidechain bound most strongly but a clear significant contribution from the NHAc-5 group can also be discerned (**Figure 2C,D**). Finally, a variant **6** that is a hybrid with both hydrophobic BPC-moiety and extended unnatural trisaccharide was also synthetically generated (**Supplementary Figure S12**) to test the relative dominance of the two most potent moieties identified by uSTA. Interestingly, significantly extended binding interfaces were seen in both, suggesting a generated synergy in binding. In this way, uSTA rapidly allowed the mapping and iterative design of natural and unnatural sugar ligands for SARS-CoV-2-spike.

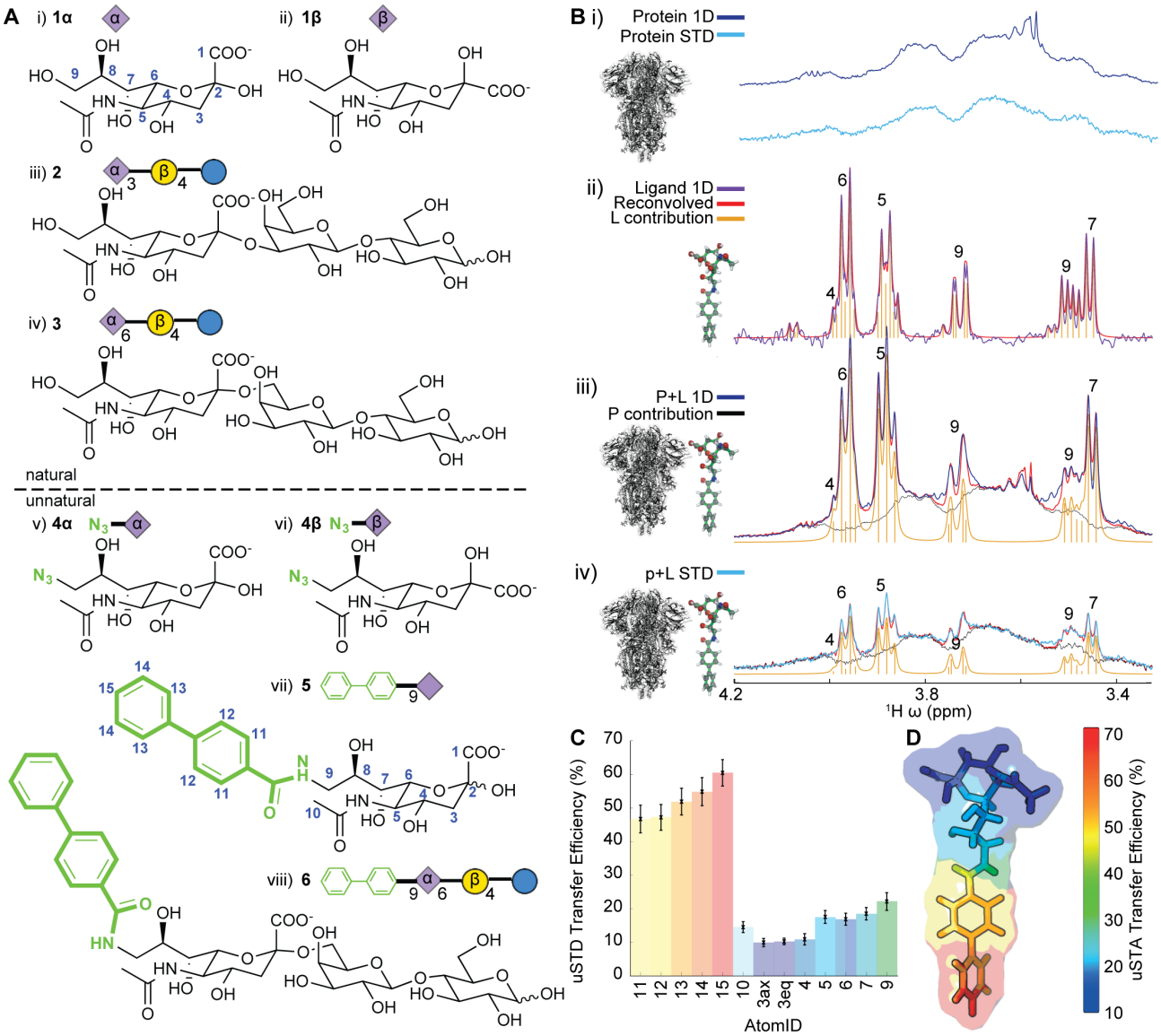


Figure 2. **uSTA Reveals Interaction of Sialosides with SARS-CoV-2-spike Protein**
A): A panel of natural, unnatural and hybrid variant sialoside sugars **1**-**6** was used to probe interaction between sialic acid moieties and spike. Unnatural variations (**4-6**, **green**) allowed mapping of C7-C9 side-chain interactions in the sialoside whilst use of extended sugars probed differing cell-surface glycan structures.
B): Application of the uSTA workflow (**Supplementary Figure S6**) to SARS-CoV-2-spike protein (shown in detail for **5,** see also **Figure 3**) **i)** The 1D ^{1}H-NMR of SARS-CoV-2-spike protein shows considerable signal in the glycan-associated region despite protein size, indicative of mobile internal glycans in spike protein. This effectively masks traditional analyses, as without careful subtraction of the protein's contributions to the spectrum (**Supplementary Figure S8**), the ligand cannot be effectively studied. The uSTA process of: **ii)** ligand peak assignment and deconvolution → **iii)** p + L peak assignment and deconvolution → **iv)** application to p + L STD yields precise atom-specific transfer efficiencies (**Supplementary Figure S6**). Note how in **ii)** individual multiplet components, have been assigned (**yellow**); the back-calculated deconvolved spectrum (**red**) is an extremely close match for the raw data (**purple**). In **iii)** the spectrum is a complex superposition of the ligand spectrum (and protein only yet uSTA again accurately deconvolves the spectrum revealing the contribution of protein only (**black**) and the ligand peaks (**yellow**). Using these data, uSTA analysis of the STD spectrum in **iv)** pinpoints ligand peaks and signal intensities.
C),**D**): Using these intensities, atom-specific transfer efficiencies can be determined with high precision and reveal in hybrid **5** the details of both the unnatural BPC moiety and the natural sialic-acid moiety. Although the aromatic BPC dominates the interaction for the unnatural ligand **5**, the

subtleties of the associated sugar contribution in this ligand can nonetheless be determined (**Supplementary Figure S5,S6**).

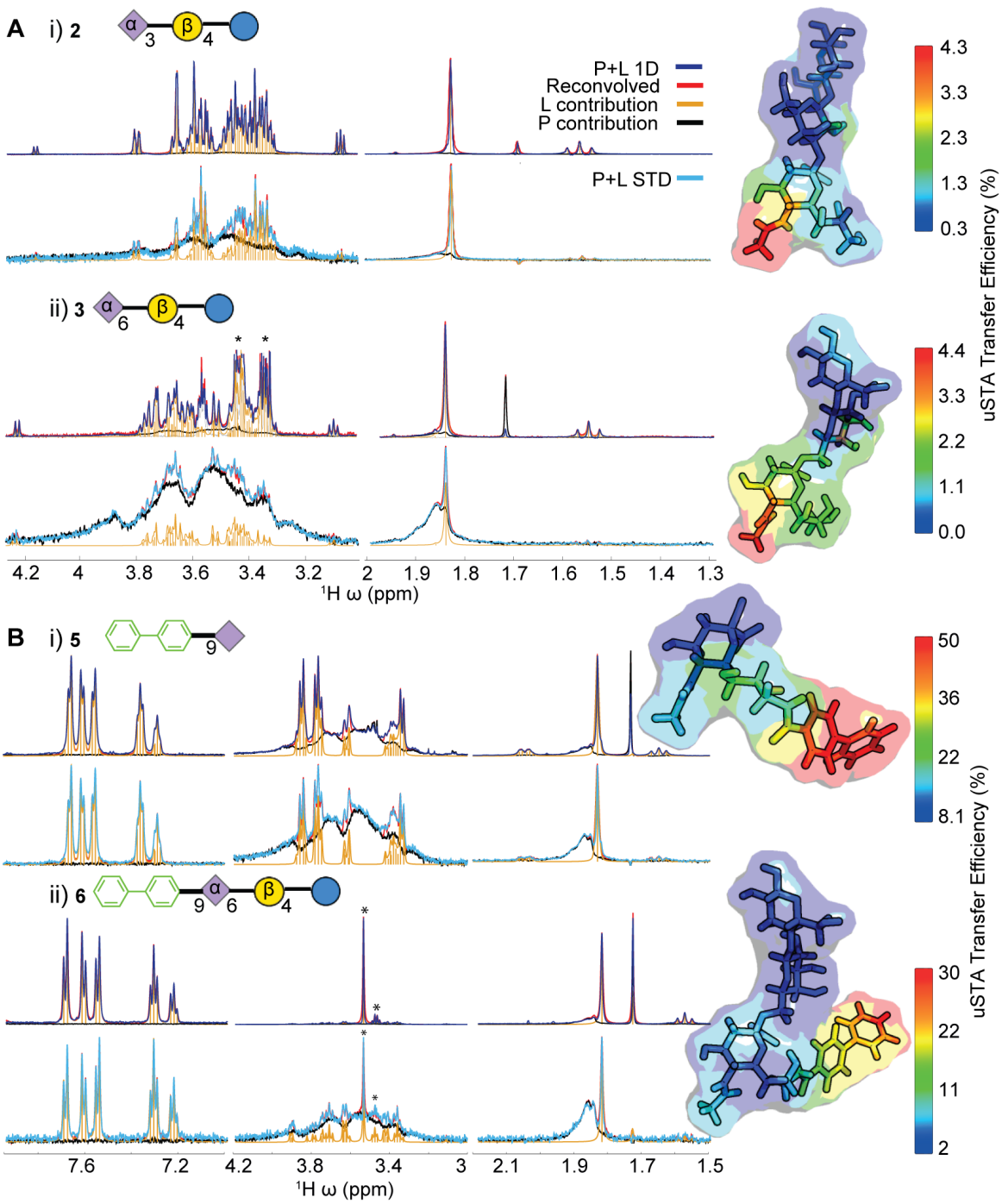


Figure 3. Observation of Sialoside Binding to SARS-CoV-2-spike Protein via uSTA. Application of the uSTA workflow (**Supplementary Figure S6**) reveals atom-specific binding modes to spike protein for both natural e.g. sialoside trisaccharides Siaα2,3Galβ1,4Glc (**A**) and Siaα2,6Galβ1,4Glc (**B**) and unnatural hybrid variants 9-BPC-Sia (**C**) and 9-BPC-Siaα2,6Galβ1,4Glc (**D**). uSTA deconvolution of the 1D spectrum of the p + L mixture (**green**), together with the STD spectrum (**blue**) allows atom- specific transfer efficiencies to be calculated, and rendered onto the molecules as 'maps' (**Supplementary Figure S6,S8**). Notably, even when protein preparations contain protein reagents such as glycerol (*), this could be effectively included in the analysis and subtracted.

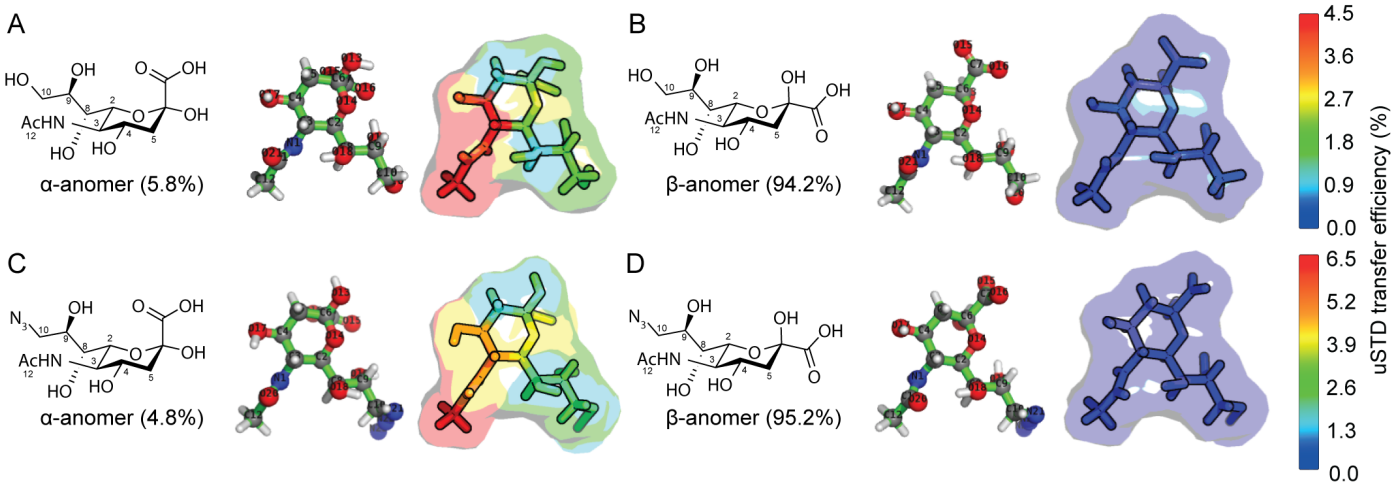


Figure 4. Simultaneous observation of binding even within dominated equilibria using uSTA In the spectra of sialic acid (**1**) and azido-sialic acid (**4**), both α and β anomeric forms could be readily identified, with the overall population being dominated by the β form (94 and 95%, respectively). Even despite this strong population difference, application of the uSTA method following assignment of resonances from the two forms allowed determination of binding surfaces simultaneously. Spike shows strong binding preference for the α anomers, as revealed by surfaces (**A**, **C**) even though its population is minor overall; both surfaces were highly similar for these two simpler monosaccharides but closely resemble those of extended trisaccharides **2** and **3** (Figure **3**). While the β form is dominant in terms of population and overall contribution to the 1D NMR spectra, its ability to bind spike (**B**, **D**), and hence its proportional contribution to the STD spectrum, was found to be significantly lower than the α form.

Determination of binding potency of sialosides to spike protein.

Through the use of uSTA and complementary methods we also sought to assess the relative binding affinity of spike trimer to our ligands. First, the uSTA method allowed a rapid and ready comparison of the relative binding efficiencies of all of the sialosides at pinpointed ligand sites (**Figure 5B**). These revealed that when judged by the level of engagement of the tip-sugar sialoside NHAc-5 group, the natural α2,3-trisaccharide **2** and the hybrid, unnatural monosaccharide **5** were the most potent 'end-on' binders. They also confirmed the conclusions drawn from the qualitative analysis (see above): that loss of the OH-9 (to form azide in **4**) was detrimental to this binding, whereas extension through the BPC moiety proved enhancing (in both **6** > **3** and **5** >> **1**). Interestingly, the hybrid BPC-trisaccharide **6** displayed apparently lower relative binding affinity than the simple hybrid monosaccharide **5**.

Next, using variable concentrations of the most potent natural ligands, α2,3-trisaccharide **2** [6μM spike, **2** at 12μM, 60μM, 1mM and 2mM excitation at 5.3 ppm] and variable concentrations of spike protein we used the uSTA method to directly determine an affinity constant (**Figure 5C**), K_D = 32 ± 12 μM with k_{on} = 630 ± 230 $M^{-1}s^{-1}$, k_{off} = 0.08 s^{-1}. We also probed binding in a different mode by measuring the affinity of spike trimer to the α2,3-trisaccharide **2** when displayed in an unorganized form on a modified surface (see **Supplementary Figure S13**) using surface plasmon resonance analysis; the latter generated a corresponding K_D = 23 ± 6 μM (k_{on} = 1030 $M^{-1}s^{-1}$). These similar determined values for either single soluble glycan binding to spike protein or when displayed on a surface suggests combined avidity for attachment through multivalency[41,42] might require proper organization on a cell-surface to prove physiologically relevant in infection (see also **Discussion**).

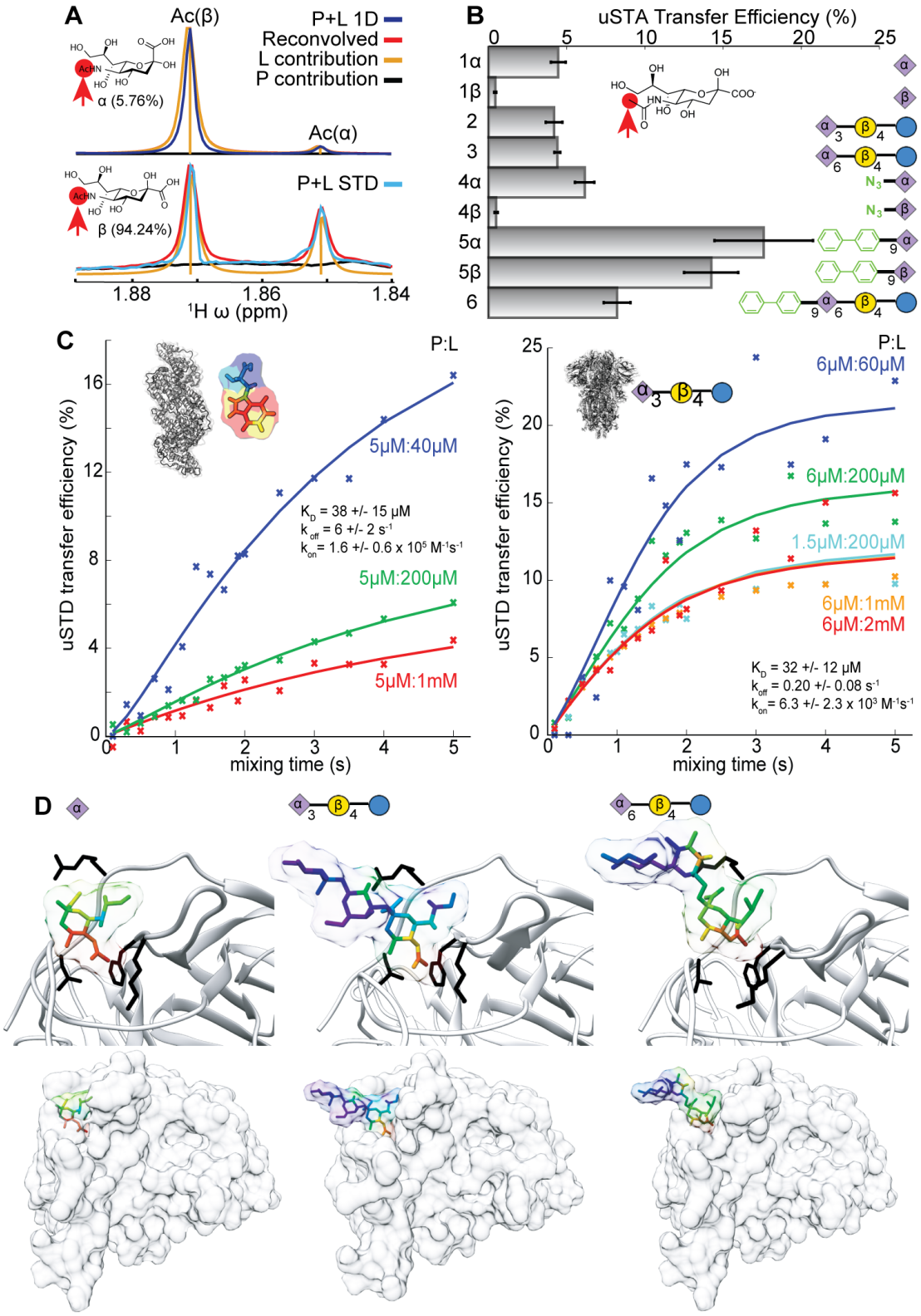


Figure 5. uSTA Reveals Selectivity Modes for SARS-CoV-2-spike Protein Sugar-Binding. **A**): uSTA reveals strong alpha over beta selectivity in sialosides for 1 and 4. Ligand **1** exists in a mutarotating mixture of configurational isomers. Despite the strong concentration dominance of **1**β > **1**α (**top**) in this equilibrium, uSTA shows the strong selective binding of **1**α > **1**β (**bottom**). See also **Figure 4**.

B): Normalized uSTA transfer efficiencies of the NAc methyl protons can be determined for each ligand studied here. We see not only the same selectivity for Siaα > Siaβ for **1** and **4**, but also the enhancing effect of unnatural modification BPC (**6** > **3** and **5** >> **1**). The specific uSTA value for **5**

and **6** for the biphenylcarboxamide (BPC) moiety ring is approximately 10x that of the NAc, and we see no preference for α / β forms, raising the possibility that the BPC may be interacting with the protein via a different site (see also **Discussion**). Errors were determined through a boot-strapping procedure where mixing times were sampled with replacement, allowing for the construction of histogram of values in the various parameters that robustly reflect their fitting errors.
2mM overall concentration of Siaalpha is 100 μM. The observed transfer efficiency of Sia α-anomer at this concentration is similar to that observed for 2,3-SiaLac at 100 μM suggesting similar K_Ds. By contrast, the Sia β-anomer is 1.9mM, and at this concentration its transfer efficiency is extremely small.
C): Quantitative analysis of the STD build up curves using a modified set of Bloch-McConnell equations that account for binding and cross relaxation allow us to determine thermodynamic and kinetic parameters that describe the interaction, K_D, k_{on}, and k_{off}. The values obtain are indicated, where the errors come from a boot-strapping procedure. BSA•Trp system (**left**); SARS-CoV-2•**2** system (**right**).
D): Coupling uSTA with an integrative modelling approach such as HADDOCK (High Ambiguity Driven DOCKing)[43,44] allows generation and, by quantitative scoring against the experimental uSTA data, selection of models that provide atomistic insights into the binding of sugars to the SARS-CoV-2-spike protein. as shown here by superposition of uSTA 'map' onto cluster. uSTA mapping the interaction between SARS-CoV- 2-spike (based on PDB ID PDB 7c2l, [17]) with ligands **1a**, **2** and **3** identifies the NHAc-5 methyl group of the tip sialic acid residue making the strongest interaction with the protein. By filtering HADDOCK models against this information, we obtain structural models that that describe the interaction between ligand and protein (**Supplementary Figure S14**). Most strikingly, we see the same pattern of interactions between protein and sialic acid moiety in each case, where the NAc methyl pocket is described by F79, L77, L18 and R78, supported by S255. Additionally, the OH-2 of the galactosyl residue and the COO^--1 carboxylate of the sialic acid both interact with R21, which is only possible for the α form, not the β form, thus explaining the selectivity. Although sequence and structural homology are low (**Supplementary Figure S1**), MERS-spike protein also appears to possess a corresponding NHAc-binding pocket characterized by an aromatic (F39)–hydrogen-bonding (D36)–hydrophobic (I132) triad.[9]

uSTA Allows delineation of spike sugar-binding site.

The uSTA method also allowed use of atom-specific data describing binding to be combined with detailed modelling algorithms. uSTA analyses of ligands **1**-**4** identified a series of atoms as consistently providing the highest transfer efficiencies, most notably the acetamide methyl group of the sialic acid residue in all. We sought to combine our measurements with the information-driven *HADDOCK* (High Ambiguity Driven *DOCKing*) method as an example a highly flexible, information-driven docking approach[43,44] capable of creating detailed atomistic models for binding interactions. Having identified likely sets of interacting residues using sequence alignment (**Supplementary Figure S1**) and structural homology of SARS-CoV-2-spike to other coronavirus spikes (PDB ID 6q06, ref. [9]), models were generated with HADDOCK for **1α** (the alpha anomer of sialic acid), and the two natural trisaccharides **2** , **3**. The resulting ensembles of models were quantitatively scored against the uSTA data (**Supplementary Figure S14**). In each case, a single narrow cluster of poses emerged (**Figure 5D**) where the acetamide NHAc methyl group of the sialic acid moiety is held within a pocket delineated by a SARS-CoV-2-spike sidechain triad of F79, T259 and L18 that appear to mediate aromatic, carbonyl-hydrogen-bonding and hydrophobic interactions, respectively. Additionally, the OH-2 of the galactosyl residue and the COO^--1 carboxylate of the sialic acid are predicted to both interact with R21 providing a structural rationale for why the α anomer binds but the β anomer does not. Notably, C1 carboxylate is a known, key recognition point in certain Sia-specific lectins.[45] Interestingly, although the sequence and structural homology to MERS coronavirus spike protein in this region is seemingly low, the MERS-spike protein possesses a corresponding NHAc-binding pocket characterized by an aromatic (F39), hydrogen-bonding (D36) and hydrophobic (I132) triad.[9] uSTA analysis therefore unlocked relevant distance restraints that revealed consistent ligand poses in a set of model complexes that now pinpoints a glycan binding pocket in the NTD. Notably, this site corresponds to the sites of reported hotspot mutations in SARS-CoV-2 variants (see below).

uSTA allows comparison of SARS-CoV-2 glycan attachment mechanism and evolution.

Several suggested SARS-CoV-2 glycan attachment mechanisms have to date focused on the role of spike RBD.[13,15] First, we compared the relative potency of the sialoside binding identified here to previously identified[15] binding of heparin motifs (**Figure 6A,B**). Heparin tetrasaccharides **7** and **8** (**Figure 6A**) were selected to allow a near residue-for-residue comparison with our identified trisaccharide (and extended trisaccharide) motifs; **7** and **8** differ from each other only at a single glycan residue (residue 2) site. Unlike the 'end-on' binding of sialosides, uSTA revealed an extended non-localized binding interface (**Figure 6B**).

Next, we examined the possible evolution of sialoside binding over emerging lineages of SARS-CoV-2 that have been observed to date.[21] Two notable 'variants of concern' B1.1.7 (so-called Kent) and B1.351 (so-called South African) have seen some of the highest levels of enhanced transmission and/or infectivity in current phases of the pandemic. Strikingly, structural

analysis revealed that key mutational hotspots in these variants sit around or in the glycan binding pocket identified by uSTA (**Figure 6C**). When corresponding B1.1.7 and B.1.351 spike-protein variants were probed by uSTA, these displayed dramatically ablated binding towards sialoside **2** as compared to A-lineage spike (**Figure 6D** and **Supplementary Figure S15**).

Finally, to explore the possible role of sialoside binding in relation to ACE2 binding we also used uSTA to test the effect upon binding in the presence of a known potent neutralising antibody of ACE2•spike binding, C5 (**Figure 6D,E** and **Supplementary Figure S16**).[46] Notably, assessment of binding to sialoside **2** in the presence and absence of antibody at a concentration sufficient to saturate the RBD led to only slight reduction in binding. Uniformly modulated atomic transfer efficiencies were consistent with a maintained glycan binding pocket with undisrupted topology and mode of binding. This suggested that sialoside binding is a mechanism that operates in addition to and potentially cooperatively with ACE2 binding in SARS-CoV-2. It also confirmed that the primary sialoside glycan binding site SARS-CoV-2-spike is distinct from the RBD.

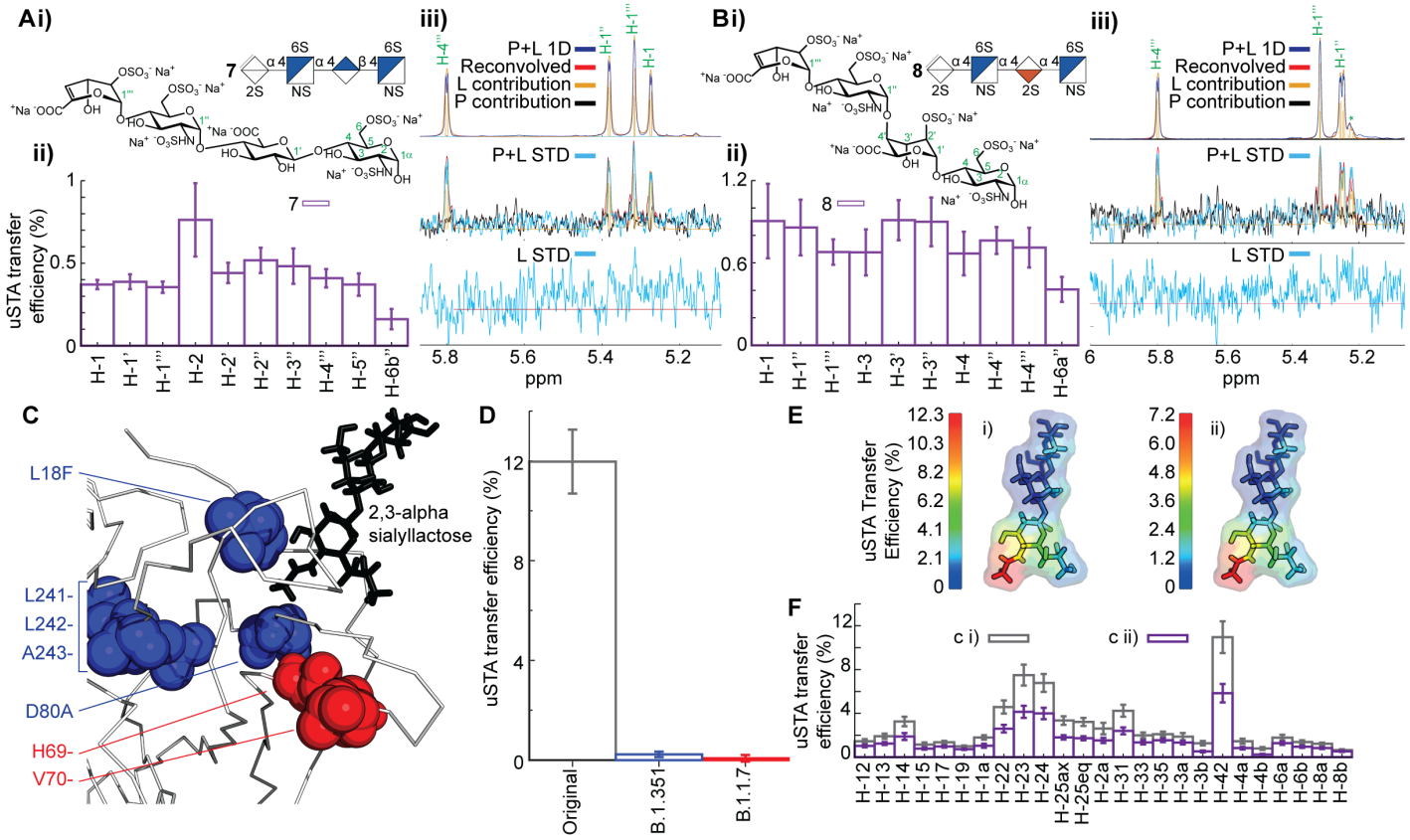


Figure 6. Comparison of SARS-CoV-2 glycan attachment mechanism and evolution.
A)i, **B)i** The binding of two heparin-derived tetrasaccharides (**7**, **8**) to A-lineage SARS-CoV-2 spike protein is demonstrated through a protein and ligand STD response and a corresponding lack of ligand only STD response **A)iii**, **B)iii**. Atom specific bindings are shown **A)ii**, **B)ii**. The assignments are shown in green, with a * indicating overlapped, ambiguous assignments.
C): Mutations to the glycan binding pocket identified by uSTA are found in the B.1.351 (blue) and B.1.1.7 (red) lineage variants of SARS-CoV-2-spike.
D): These ablated binding of the 2,3-alpha sialyllactoside **2** as measured by the transfer efficiency of the NAc protons.
E,F): When uSTA was used in the presence of the potent neutralizing nanobody C5 to probe A-lineage SARS-CoV-2 spike (spike (**E i**), nanobody-plus-spike (**E ii**)) essentially similar binding pattern was observed with uniformly modulated the atomic transfer efficiencies (**F**).

Identification of sialoside trisaccharide as a putative ligand in A-lineage SARS-CoV-2 correlates with genetic variation in clinical samples.

The identification of a putative binding site glycan in sialoside trisaccharide **2** as a functional ligand and hence potential attachment point for SARS-CoV-2 highlights a sugar motif (and related structures) found commonly on cell surfaces (attached both as glycolipid and glycoprotein glyco-conjugates). This, in turn, suggested that there may be impact of glycosylation function in humans upon infection by SARS-CoV-2 and hence in the presentation and pathology of COVID19 disease. Strikingly, analysis following whole exome sequencing of an early 2020 cohort of 533 COVID-19-positive patients (see **Supplementary Table S1**) identified two glycan-associated genes within the top five most-influential genes upon disease severity. Specifically, recursive feature elimination applied to a Least Absolute Shrinkage and Selection Operator (LASSO)-based[47] logistic regression model identified: *LGALS3BP* (4th out of >18000 analyzed genes) and *B3GNT8* (5th out of >18000) (**Figure 7A** and **Supplementary Figure S17**). Rare variants in these two genes were beneficially associated with less severe disease outcome (**Figure 6B,C**, see also **Supplementary Tables S2-6** for specific *B3GNT8* and *LGALS3BP* genetic variants, *B3GNT8* chi-squared five categories, *B3GNT8* chi-squared 2x2, *LGALS3BP* chi-squared five categories, *LGALS3BP* chi-squared 2x2, respectively).

LGALS3BP encodes for a secreted protein Galectin-3-binding-protein (Gal-3-BP, also known as Mac-2-BP) that is an apparent partner and blocker of a specific member (Gal-3) of the galectin class of carbohydrate-binding proteins.[48] Galectins are soluble, typically secreted and implicated in a wide-range of cellular functions.[49] Notably Gal-3 binds the so-called poly-*N*-acetyl-lactosamine [polyLacNAc or $(Gal\text{-}GlcNAc)_n$] chain-extension variants found in tetraantennary *N*-linked glycoproteins (**Figure 7D**), including those displaying sialyl-Gal-GlcNAc motifs, close analogues of our discovered ligands for SARS-CoV-2 spike **2** and **3**.[50,51] Notably, rare variants in *LGALS3BP* were present in in 9/114 (~8%) of a/pauci-symptomatic subjects or mildly-affected patients compared to 8/419 (< 2%) of the remaining 419 patients who required more intensive care: oxygen support, CPAP/BiPAP or intubation; none of the 69 most seriously affected patients (intubated) carried variants of *LGALS3BP* (**Figure 7B**). Intact *LGALS3BP*-gene-product Gal-3-BP therefore appears correlated with more severe COVID19 outcome.

The other implicated gene, *B3GNT8* encodes for a protein glycosyltransferase β-1,3-*N*-acetyl-glucosaminyltransferase-8 (GlcNAcT8 or β3GnT8) that is responsible for the anchor-point initiation of poly-*N*-acetyl-lactosamine (polyLacNAc) synthesis from tetraantennary *N*-linked glycoproteins (**Figure 7D**).[52] Again, rare variants in *B3GNT8* were present in in 11/114 (~10%) of a/pauci-symptomatic subjects or mildly-affected patients compared to 10/419 (~2%) of the remaining 419 patients who required more intensive care (**Figure 7C**).

Finally, to test the possible roles of sialosides in mediating viral entry suggested by this model (**Figure 7D**), we tested the consequences of reducing cell-surface sialosides. Consistent with this model, when ACE-2-expressing mammalian (MDCK) cells were treated with a broad-

activity sialidase (*Arthrobacter ureafaciens* neuraminidase, see **Methods**), the consequent uptake of pseudovirus displaying SARS-CoV-2-lineage-A-spike and capable of expressing luciferase reporter was significantly reduced (**Figure 7E**).

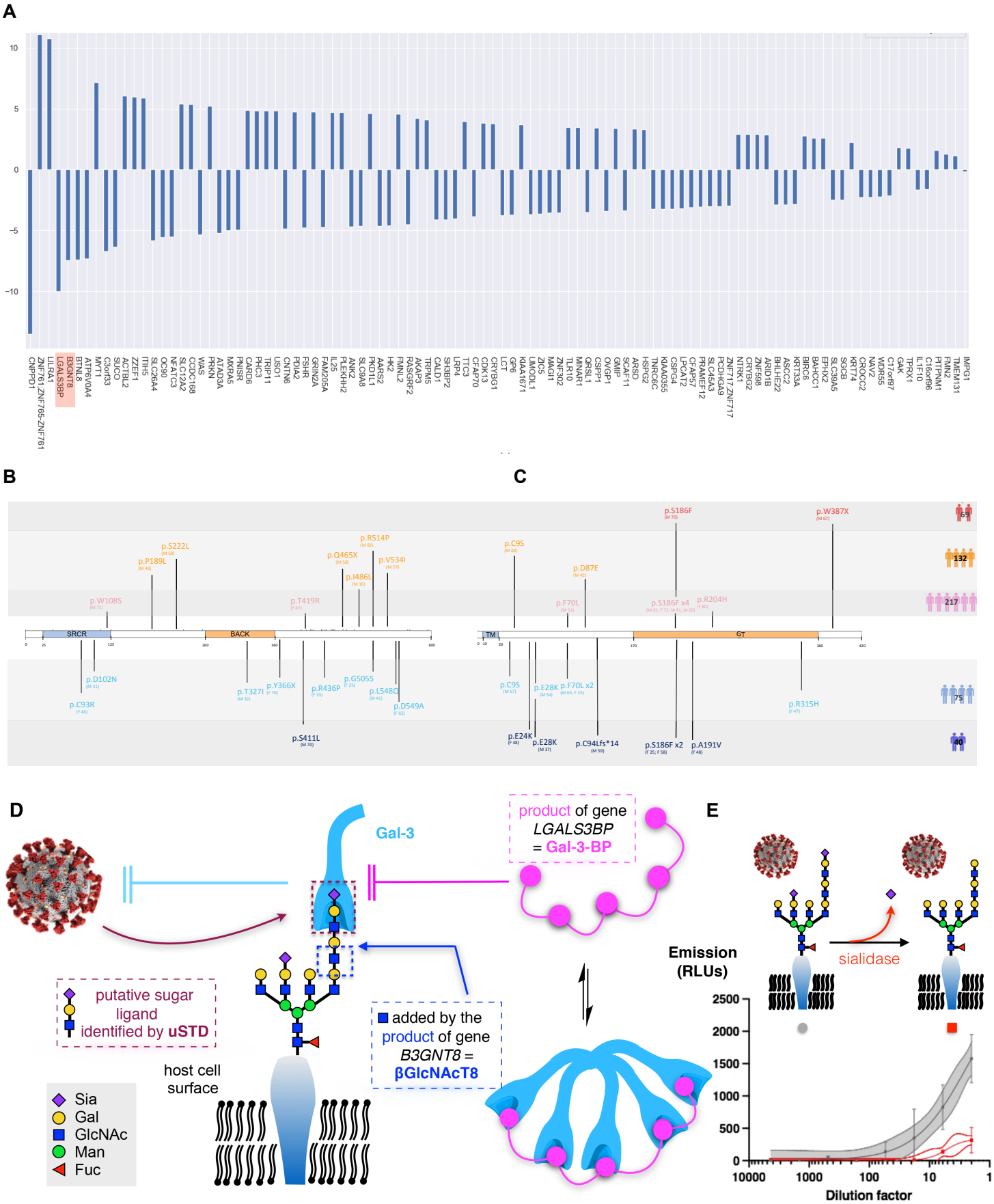


Figure 7. Analyses of Early 2020 SARS-CoV-2-positive Patients Reveals Glycan-Associated Genes Suggesting a Model of Glycan Interaction Consistent with uSTA Observations of Sialoside Binding in Lineage-A SARS-CoV-2.

A): Histogram of the LASSO-based Logistic Regression weightings after Recursive Feature Elimination analysis of 533 SARS-CoV-2-positive patients. Positive weights score susceptible response of gene variance to COVID19 disease, whereas negative weights confer protective action through variance. Variation in glycan-associated genes *B3GNT8* and *LGALS3BP* score 2nd and 3rd out of all (>18000) genes as the most protective, respectively (**highlighted red**).
B),**C**): Distribution of rare variants in *B3GNT8* and *LGALS3BP*. **B**) Rare beneficial mutations distributed along the Gal-3-BP protein product of *LGALS3BP*, divided into the Scavenger Receptor Cysteine-Rich (SRCR)-domain (**light blue**) and the BACK domain (**light orange**). **C)** Rare beneficial mutations distribution along the βGlcNAcT8 protein product of *B3GNT8* divided into the predicted transmembrane (TM) domain (**light blue**) and glycosyltransferase catalytic (GT) domain (**light orange**) which catalyzes the transfer of polyLacNAc-initiating GlcNAc onto tetraantennary *N*-linked glycoproteins (see also **D**). The different colours of the mutation bands (top to bottom) refer to the severity grading of the COVID19-postive patients who carried that specific mutation (**red**: Hospitalized intubated; **orange**: Hospitalized CPAP/BiPAP; **pink**: Hospitalized Oxygen Support; **light blue**: Hospitalized w/o Oxygen Support; **blue**: Not hospitalized a/paucisymptomatic).
D): A proposed coherent model consistent with observation of implicated *B3GNT8* and *LGALS3BP* genes and observed Sia-Gal-Glc trisaccharide as ligand by uSTA. Strikingly, although independently identified, *B3GNT8* and *LGALS3BP* produce gene products βGlcNAcT8 and Gal-3-BP, respectively, that manipulate and/or engage with processes associated with a common polyLacNAc-extended chain motif found on tetraantennary *N*-linked glycoproteins. A model emerges in which any associated loss-of-function from variance leads either to loss of polyLacNAc-extended chain (due to loss of initiation by βGlcNAcT8) or enhanced sequestration of by Gal-3 polyLacNAc-extended chain (which is antagonized by Gal-3-BP). Both, would potentially lead to reduced access of virus spike to uSTA-identified motifs. **E**): ACE-2-expressing MDCK cells were pre-treated with sialidase before a 4 h pulse with SARS-CoV-2 pseudovirus. After 48 h, pseudoviral entry was evaluated via luciferase reporter assay. Data are represented as the fitted logarithmic curve ± the 95% confidence intervals (n = 3).

Discussion

~800 words

Prior analyses of sugar interaction with coronavirus spike proteins have generated varied observations and conclusions as to possible mechanisms. These have largely employed more traditional, surface-display biophysical methods for analyzing interaction (e.g. 'glycochip'/glycan-array binding analyses). We have devised a protein saturation transfer difference method (uSTA) for assessing ligand•protein binding applied here to SARS-CoV-2 spike protein (**Supplementary Figures S5,S6**). Such mammal-derived pathogen proteins are typically a challenging problem in NMR analysis, complicated by the presence and signal arising from mobile, protein-linked glycans, However, uSTA has effectively and robustly identified intermolecular sugar spike interactions. Our workflow does not require the use of relaxation filters and is automated by freely distributed software. With these data we have derived structural models for three biologically relevant ligands that reveal the sialic acid to be in a consistent binding pose in each.

Intriguingly we find a clear link between our data and genetic analyses in patients that have been correlated with severity of disease. This suggests not only that cell-surface glycans and the modulation of binding to spike-to-sialosides may have played a role in infection and disease progression but also now identifies two glycan-associated genes. Genetic variations of *LGALS3BP* (which produces galectin-blocker Gal-3-BP) and *B3GNT8* (which produces glycosyltransferase β3GnT8) that reduce corresponding gene products (Gal-3-BP and β3GnT8, respectively) prove beneficial. Strikingly, despite their independent identification here, both Gal-3-BP and β3GnT8 interact around a common glycan motif: the polyLacNAc chain-extension variants found in tetraantennary *N*-linked glycoproteins (**Figure 7D**). The simplest explanation is thus that modulation of such glycoproteins have played a role in infection and disease progression. Consistent with the sialoside ligands we identify here, these glycoproteins contain Sia-Gal-GlcNAc motifs within *N*-linked-polyLacNAc-chains. Notably, these motifs have recently been identified in the deeper human lung.[53]

These data lead us to suggest that A-lineage SARS-CoV-2 virus may have exploited glycan-mediated attachment to host cells (**Figure 7D**) using *N*-linked-polyLacNAc-chains as a foothold. Reduction of Gal-3-BP function would allow its target, the lectin Gal-3, to bind more effectively to *N*-linked-polyLacNAc-chains, thereby competing with SARS-CoV-2 virus. Similarly, loss of β3GnT8 function would ablate the production of foothold *N*-linked-polyLacNAc-chains directly denying the virus a foothold. We cannot exclude other possible mechanisms including, for example, the role of *N*-linked-polyLacNAc-chains in T-cell regulation.[54] Notably, this analysis of the influence of genetic variation upon susceptibility to virus was confined to 'first wave' patients infected with A-lineage SARS-CoV-2. Our discovery here also that in B-lineage virus such binding to certain sialosides may be ablated further highlights the dynamic role that sugar binding may play in virus evolution and may be linked, as was previously suggested for H5N1 influenza A virus, to changing sugar preference during zoonotic transitions.[55]

Analyses of the glycosylation of spike protein itself suggests at least 66 *N*-linked glycosylation sites on Asn are occupied in the complete spike trimer[40] with some *O*-glycosylation. This glycosylation, at least as determined from mammalian cell expression, appears to be of a variety of types including oligomannose, hybrid and complex type – the latter two displaying tip sialic acid residues. Comparison of the glycosylation states of viral surfaces reveals[56] that SARS-CoV-2 is more lightly glycosylated than other viruses suggesting, at least in part, that an evolved role of these glycans is not to shield the virus for immune evasion, unlike other viruses such as heavily glycosylated HIV-1. Recently, advanced simulations have suggested a role of *N*-glycans in altering the conformational dynamics of the RBD with functional consequence.[57] Moreover, the glycans in the N-terminal region of spike protein are located on highly mobile loops, and it is likely that the sugars are also themselves dynamic and flexible with potential for various additional modes of interaction with not only receptors but potentially even with spike protein itself intra-molecularly. This raises the intriguing possibility that in SARS-CoV-2 such a mechanism could modulate inter-molecular glycan attachment affinity through intra-molecular use of its own 'internal glycans' (also sialosides) as competitors. If so, any such interaction may have therefore further contributed to the cryptic nature of sialoside binding by SARS-CoV-2. Notably, titration of trisaccharide **2** into spike trimer had no measurable effect upon thermal denaturation, suggesting that any binding does not grossly affect protein stability or structure (**Supplementary Figure S18**).

More generally, uSTA now provides a comprehensive treatment of saturation transfer that is a ready method for not only identifying ligands and their binding parameters but that can also be linked to identifying binding sites even in relevant heavily-modified proteins. It could also potentially be used to similarly restrain structure calculation refinements where atomic resolution in the vicinity of a ligand is not possible. Overall, therefore, the uSTA method provides a new pipeline to provide, with minimal user input, multi-faceted analysis of ligand engagement even for previously challenging protein systems, such as those exploited by pathogens.

Author Contributions

BG, CJB, TDWC and AJB ran NMR experiments. Under the supervision of AJB, CJB developed the underlying deconvolution algorithm, UnidecNMR. AJB developed the modified Bloch-McConnell K_D analysis method. CJB performed the processing, uSTA applications and surrounding analysis. CJB and AK assigned key uSTA spectra. PH, ALB purified spike protein and prepared samples for NMR analysis; PH and ALB conducted thermal denaturation assays. AK synthesized 9-BPC-Neu5Ac, expressed and purified NmCSS and Pd2,6ST, assigned proton peaks in ^{1}H NMR spectra of sialosides, designed TOCSY NMR experiments and performed the SPR experiments. AMG and AK designed and synthesized 9-BPC-Sia2,6Lac, AMG generated reagents for SPR studies. AK, AMG and AJB analysed SPR data. JH cloned spike-BAP; PW and MD expressed spike protein. AL aided design of protein-ligand NMR experiments, prepared samples and contributed to data analysis. AR is coordinating the GEN-COVID Consortium and supervised the genotype-phenotype correlation of glycosylation-associated genes. AG performed WES experiments. EB performed WES alignment and joint call which was supervised by SF. NP performed logistic regression analysis which was supervised by MG. MB and FF collected and supervised the collection of clinical data. CF and SD perform specific analysis on *B3GNT8* and *LGALS3BP* genes. LPD and QJS assisted with experimental design and conducted viral pseudotype assays. AMJJB and PIK performed the structural modelling of the SARS-CoV-2-spike•sugar interactions. AJB and AMJJB scored these against the NMR data to obtain structural models for the complex. VCA, CJB, AJB and BGD produced figures. BGD, AJB, TDWC and JHN supervised and designed experiments, analyzed data and wrote the manuscript. All authors read and commented on the manuscript.

Acknowledgments

Upgrades of the 600 MHz and 950 MHz spectrometers were funded by the Wellcome Trust (Grant Ref: 095872/Z/10/Z) and the Engineering and Physical Sciences Research Council (Grant Ref: EP/R029849/1), respectively, and by the University of Oxford Institutional Strategic Support Fund, the John Fell Fund, and the Edward Penley Abraham Cephalosporin Fund. We thank the Wellcome Trust for support (20289/Z16/Z). The Rosalind Franklin Institute is supported by the EPSRC.

AMJJB and PIK acknowledge financial support from the European Union Horizon 2020 projects BioExcel (823830) and EOSC-Hub (777536) and the IMI-CARE project (101005077)

We would like to thank Professor Xi Chen at the University of California at Davis for providing the plasmid encoding for Pd2,6ST; Mikhail Kutuzov and Omer Dushek for useful discussions and assistance with SPR experiments; Dr Andrew Quigley and Nadisha Gamage (Diamond Light Source) for help with thermal stability assays; Professor Christina Redfield for assistance with high field Bruker instrumentation and Dr Lindsay Baker for insights into the structural models.

Data Availability Statement

Raw spectral data for Figures 1, 2, 3, and 4A, the data used to plot Figures 4B, C and relevant uSTA code have been deposited on the University of Oxford ORAdata server at https://ora.ox.ac.uk/objects/uuid:121699de-68d5-4b6f-b77e-cee85895da6c All associated spectral data is also shown in **Supplementary Table S7**. HADDOCK datasets and results for this work can be found at https://github.com/haddocking/SARS-COV2-NTD-SIA-modelling and https://doi.org/10.5281/zenodo.4271288. All associated clinical data have been supplied in **Supplementary Tables S1-S6**, conducted under trial (NCT04549831, www.clinicaltrial.org).

Methods

Protein expression and purification.

The gene encoding amino acids 1-1208 of the SARS-CoV-2 spike glycoprotein ectodomain [with mutations of RRAR > GSAS at residues 682-685 (the furin cleavage site) and KV > PP at residues 986-987, as well as inclusion of a T4 fibritin trimerisation domain], using the construct previously described as template,[58] was cloned into the pOPINTTGneo-BAP vector using the forward primer (5'- GTCCAAGTTTATACTGAATTCCTCAAGCAGGCCACCATGTTCGTGTTCCTGGTGCTG -3') and the reverse primer (5'- GTCATTCAGCAAGCTTAAAAAGGTAGAAAGTAATAC -3'), resulting in an aviTag plus 6His in the 3' terminus of the construct.

Expi293 cells (Thermofisher Scientific) were used to express the Spike-Bap protein. The cells were cultured in expi293 expression media (Thermofisher Scientific) and were transfected using PEI MAX 40kDa (Polyscience) if cells were >95% viable and had reached a density of between 1.5 - 2 x 10^6 cells per ml. Following transfection, cells were cultured at 37 ˚C and 5% CO_2 at 120rpm for 17h. Enhancers (6mM Valproic Acid, 6.5mM Sodium Propionate, 50mM Glucose - Sigma) were then added and protein was expressed at 30 ˚C for 5 days before purification.

The media in which the protein spike-bap was secreted was supplemented with 1x PBS buffer at pH 7.4 (1:1 v/v) and 5mM $NiSO_4$. The pH was adjusted with NaOH to pH 7.4 and filtered using a 0.8um filter. The mixture was stirred at 150 rpm for 2 hours at room temperature. The Spike protein was purified on an Akta Express system (GE Healthcare) using a 5mL His trap FF GE Healthcare column in PBS, 20mM imidazole pH 7.4 and eluted in PBS, 300mM imidazole pH 7.4. The protein was then injected on a High load superdex 200 16/600 gel filtration column (GE Healthcare) in deuterated PBS buffer pH 7.4. The eluted protein was concentrated using an Amicon Ultra-4 100kDa concentrator at 2000 RPM, 16°C, pre-washed multiple times with deuterated PBS) to a concentration of roughly 1 mg/mL.

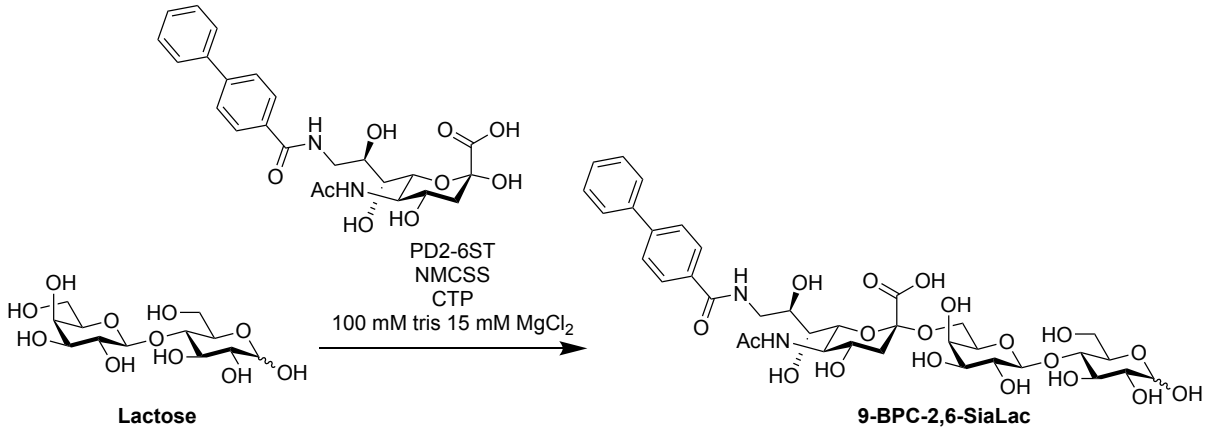
Errors

The errors in the transfer efficiencies were estimated using a bootstrapping procedure. Specifically sample STD spectra were assembled through taking random combinations with replacement of mixing times, and the analysis to obtain the transfer efficiency was performed on each. This process was repeated 100 times to enable evaluation of the mean and standard deviation transfer efficiency for each residue. Mean values correspond well with the value from the original analysis, and so we take the standard deviation as our estimate in uncertainty.

Reagent sources.

6'-Sialyllactose sodium salt and 3'-Sialyllactose sodium salt were purchased from Carbosynth and used directly: 6'-Sialyllactose sodium salt- CAS-157574-76-0, 35890-39-2, 3'-Sialyllactose sodium salt - CAS-128596-80-5,35890-38-1. Bovine serum Albumin (BSA) and L-tryptophan were purchased from Sigma Aldrich. Heparin sodium salt, from porcine intestinal mucosa, IU≥100/mg was purchased from Alfa Aesar. All other chemicals were purchased from commercial suppliers (Alfa Aesar, Acros, Sigma Aldrich, Merck, Carbosynth, Fisher, Fluorochem, VWR) and used as supplied, unless otherwise stated.

*Synthesis of 9-BPC-6'-Sialyllactose (**6**).*

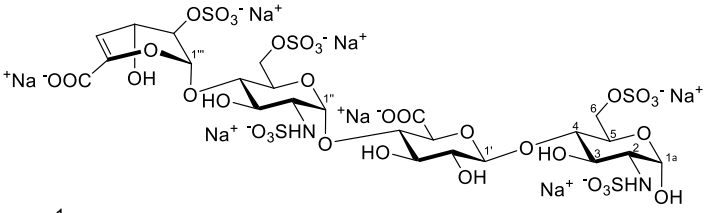


Lactose (30 mg, 0.087 mmol, 1 eq), BPC-Neu5Ac (48 mg, 0.0974 mmol, 1.12 eq, prepared using a procedure adapted from ref. [59]) and CTP disodium salt (92 mg, 1.74 mmol, 2 eq) were dissolved in buffer 100 mM Tris, 15 mM $MgCl_2$ pH 8.5, 8.39 mL- to give achieve a final concentration of 10 mM with respect to the lactose acceptor. Pd-26ST enzyme (261 mL, 1.2 wt%, 1.38 mg/mL) and NmCSS enzyme (107.4 mL, 1 wt %, 2.73 mg/mL) were added to the reaction and the mixture shaken at 37 °C. After 22 h the reaction mixture was concentrated *in vacuo* and then purified via column chromatography (EtOAc:IPA:H_2O 3:2:1→2.5:2:1 and then EtOAc:IPA:H_2O: 4:2:1). Fractions containing the desired compound were concentrated, redissolved in water and lyophilised to yield the product as a white powder (17.7 mg, 0.0218 mmol, 25%) – Note: the reaction conversion was very high, however the yield is reflective of difficult isolation of the product. **LRMS:** m/z (ES-) 811 $[M-H]^-$; **HRMS**: m/z (ESI-): calc. for $C_{36}H_{47}O_{19}N_2$ $[M-H]^-$ 811.2779, found 811.2773; **^{1}H NMR** (400 MHz, D_2O) δ 7.90 (2H, d, *J* = 8.5 Hz, ArH), 7.82 (2H, d, *J* = 8.3 Hz, ArH), 7.80 – 7.71 (2H, m), 7.61 – 7.53 (2H, m), 7.53 – 7.45 (1H, m), 5.16 (0.4H, d, *J* = 3.8 Hz, H-1a (alpha)), 4.63 (1H, d, *J* = 8.0 Hz, H-1a (beta)), 4.41 (1H, d, *J* = 7.8 Hz, H-1b), 4.09 (1H, ddd, *J* = 8.8, 7.2, 3.2 Hz, H-8c), 4.04 – 3.48 (m, 16H), 3.35 – 3.26 (1H, m, H-2a (beta)), 2.73 (1H, dd, J = 12.4, 4.7 Hz, H-3c_{eq}), 2.02 (3H, s, $NCOCH_3$), 1.77 (1H, t, *J* = 12.2 Hz, H-3c_{ax}).

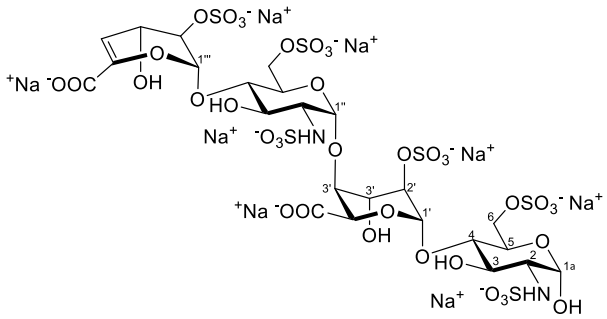
*Synthesis of Heparin Tetrasaacharides (**7, 8**).*

Heparin sodium salt (UFH, 100mg) was dissolved in the enzyme preparations in 100mM Tris Acetate, 5Mm $Ca(OAc)_2$, pH 6.8 to achieve a final concentration of 30 mg/mL UFH and 1.5mg/mL Hep-1-SUMO. The reaction was incubated in a water bath at 28°C and terminated after 36hrs by heating at 95 ˚C for 5 min. Upon centrifugation (2000g, 2min), the supernatant was filtered using a 0.2μm filter (Regenerated Cellulose, Sartorius) and lyophilised to yield a mixture of heparin oligosaccharides as a foam. These were then fractionated based on their degree of polymerisation on a Superdex Peptide 10/300 column (GE-Healthcare), following a procedure adapted from the procedure adapted from ref. [60,61]. Briefly, the column was equilibrated before each run with 2CV of 0.5M $(NH_4)_2CO_3$ and the elution was monitored at 232nm. Sample application (100uL injection per run, ~10mg in sugar content) was followed by an isocratic elution with 0.5M $(NH_4)_2CO_3$ at a flow rate of 0.6mL/min. The fraction collector was set to collect 1mL fractions until well-resolved peaks were detected; at that point, fractionation was manually controlled and tetrasaccharide-containing fractions from sequential chromatographic runs were pooled together and subjected to serial lyophilisation rounds. Approximately 15mg of heparin tetrasaccharides were recovered upon removal of the volatile salt; these were further analysed with strong anion exchange HPLC on a SAX Propac PA1 column (9x250mm, Thermo Scientific) using H_2O, pH 3.5 as solvent A and 2M NaCl (HPLC grade) pH 3.5 as solvent B. The column was equilibrated with solvent A for 30min prior to sample application (500uL injection per run, 10mg/mL) and the target oligosaccharides were eluted following a linear gradient from 30% to 70% solvent B over 180min, at a flow rate of 1mL/min. Fractions corresponding to the same retention times were pooled, neutralized with saturated $NaHCO_3$ (HPLC grade) and lyophilized. Desalting was performed on an AKTA Purifier system (GE healthcare) by connecting three 5mL HiTrap Desalting columns (GE Healthcare) in a

row. Elution with water was performed at a flow rate of 6mL/min and monitored at 232nm; UV absorbing fractions were pooled and lyophilized to yield the pure target tetrasaccharides.



Compound **7**. White solid. ^{1}H NMR (700 MHz, D_2O) δ 5.91 (dd, $J_{4''',3'''}$ 4.7 Hz, $J_{4''',2'''}$ 1.4 Hz, 1H, H-4'''), 5.50 (d, $J_{1'',2''}$ 3.8 Hz, 1H, H-1''), 5.43 (d, $J_{1''',2'''}$ 2.6 Hz, 1H, H-1'''), 5.39 (d, $J_{1,2}$ 3.5 Hz, 1H, H-1), 4.55 (m, 1H, H-2'''), 4.52 (d, $J_{1',2'}$ 8.0 Hz, 1H, H-1'), 4.29 – 4.26 (m, 3H, H-6a'', 6a, 6b), 4.25 (dd, $J_{3''',4'''}$ 4.8 Hz, $J_{3''',2'''}$ 2.1 Hz, 1H, H-3'''), 4.13 (dd, $J_{6b'',6a''}$ 11.3 Hz, $J_{6b'',5''}$ 2.0 Hz, 1H, H-6b''), 4.08 (ddd, $J_{5,4}$ 9.7 Hz, $J_{5,6a}$ 4.2 Hz, $J_{5,6b}$ 2.7 Hz, 1H, H-5), 3.91 (ddd, $J_{5'',4''}$ 10.1 Hz, $J_{5'',6a''}$ 4.2 Hz, $J_{5'',6b''}$ 2.2 Hz, 1H, H-5''), 3.79 – 3.70 (m, 4H, H-3', H-4', H-5', H-4''), 3.68 – 3.59 (m, 2H, H-4, H-3), 3.56 (dd, $J_{3'',2''}$ 10.6 Hz, $J_{3'',4''}$ 8.8 Hz, 1H, H-3''), 3.31 (dd, $J_{2',3'}$ 9.5 Hz, $J_{2',1'}$ 7.9 Hz, 1H, H-2'), 3.23 (dd, $J_{2'',3''}$ 10.6 Hz, $J_{2'',1''}$ 3.8 Hz, 1H, H-2''), 3.20 (dd, $J_{2\alpha,3\alpha}$ 10.2 Hz, $J_{2\alpha,1\alpha}$ 3.6 Hz, 1H, H-2), 2.97 (dd, $J_{2\beta,3\beta}$ 10.0 Hz, $J_{2\beta,1\beta}$ 8.4 Hz, H-2β). ^{13}C NMR (176 MHz, D_2O) δ 169.3 (C-6''' - COOH), 160.2 (C-6' – COOH), 144.8 (C-5'''), 106.0 (C-4'''), 102.1 (C-1'), 97.7 (C-1''), 97.1 (C-1'''), 91.0 (C-1), 78.8 (C-4), 78.0 (C-4''), 77.5 (C-4'), 76.4, 75.9 (C-3',5'), 74.5 (C-2'''), 72.9 (C-2'), 69.53 (C-3''), 69.3 (C-3), 68.7 (C-5''), 68.1 (C-5), 66.7 (C-6), 66.1 (C-6''), 62.8 (C-3'''), 57.6 (C-2), 57.5 (C-2''). Data consistent with previous reports[4]. LRMS m/z (ESI-): Found 535.5 $[M-2H]^{2-}$; HRMS: m/z (ESI-) calc. for $C_{24}H_{36}N_2O_{35}S_5^{2-}$ $[M-2H]^{2-}$ 535.9857, found 535.9858.



Compound **8**. White solid. ^{1}H NMR (700 MHz, D_2O) δ 5.91 (dd, $J_{4''',3'''}$ 4.8 Hz, $J_{4''',2'''}$ 1.4 Hz, 1H, H-4'''), 5.43 (d, $J_{1''',2'''}$ 1.9 Hz, 1H, H-1'''), 5.38 (dd, J 3.6, 1.7 Hz, 2H, H-1, H-1''), 5.13 (d, $J_{1',2'}$ 3.4 Hz, 1H, H-1'), 4.67 (d, $J_{5',4'}$ 3.0 Hz, 1H, H-5'), 4.62 (d, $J_{1\beta,2\beta}$ 8.3 Hz, 0.1H, H-1β), 4.55 (m, 1H, H-2'''), 4.30 (dd, $J_{6a,6b}$ 11.4 Hz, $J_{6a,5}$ 4.9 Hz, 1H, H-6a), 4.27 (dd, $J_{6b'',6a''}$ 10.7 Hz, $J_{6b'',5''}$ 1.6 Hz, 1H, H-6b''), 4.25 – 4.21 (m, 3H, H-6b, H-2', H-3'''), 4.18 (dd, $J_{6a'',6b''}$ 11.3 Hz, $J_{6a'',5''}$ 2.0 Hz, 1H, H-6a''), 4.12 (dd, $J_{3',2'}$ 6.7 Hz, $J_{3',4'}$ 3.9 Hz, 1H, H-3'), 4.07 – 4.02 (m, 2H, H-4', H-5), 3.97 (m, 1H, H-5''), 3.75 (at, J 9.5 Hz, 1H, H-4''), 3.67 (at, J 9.4 Hz, 1H, H-4), 3.62 (dd, $J_{3,2}$ 10.4 Hz, $J_{3,4}$ 8.9 Hz, 1H, H-3), 3.55 (dd, $J_{3'',2''}$ 10.6 Hz, $J_{3'',4''}$ 8.8 Hz, 1H, H-3''), 3.22 (dd, $J_{2'',3''}$ 10.6 Hz, $J_{2'',1''}$ 3.6 Hz, 1H, H-2''), 3.18 (dd, $J_{2\alpha,3}$ 10.3 Hz, $J_{2\alpha,1\alpha}$ 3.6 Hz, 1H, H-2α), 2.97 (dd, $J_{2\beta,3}$ 10.1 Hz, $J_{2\beta,1\beta}$ 8.4 Hz, H-2β, 0.1H). ^{13}C NMR (176 MHz, D_2O) δ 174.4 (C-6' – COOH), 169.2 (C-6''' – COOH), 144.8 (C-5'''), 105.8 (C-4'''), 99.4 (C-1'), 97.2 (C-1'''), 96.5 (C1''), 91.0 (C-1), 78.1 (C-4''), 76.9 (C-4), 76.6 (C-2'), 76.3 (C-4'), 74.4 (C-2'''), 69.7 (C-5'), 69.6 (C-3'), 69.5 (C-3''), 69.4 (C-3), 68.8 (C-5''), 68.4 (C-5), 67.0 (C-6), 66.2 (C-6''), 62.7 (C-3'''), 57.9 (C-2), 57.6 (C-2''). Data consistent with previous reports[4]. LRMS m/z (ESI-): Found 575.5 $[M-2H]^{2-}$; HRMS: m/z (ESI-) calc. for $C_{24}H_{36}N_2O_{38}S_6^{2-}$ $[M-2H]^{2-}$ 575.9641, found 575.9651.

Protein NMR Experiments.

All NMR experiments in **NMR Methods Table 1** were conducted at 15 °C on a Bruker AVANCE NEO 600 MHz spectrometer with CPRHe-QR-1H/19F/13C/15N-5mm-Z helium-cooled cryo-probe. Samples were stored in a Bruker SampleJet sample loader while not in magnet, at 4 °C.

1D 1H NMR spectra with w5 water suppression were acquired using the Bruker pulse sequence zggpw5, using the smooth square Bruker shape SMSQ.10.100 for the pulsed field gradients. The spectrum was centered on the water peak, and the receiver gain adjusted. Typical acquisition parameters were sweep width of 9615.39 Hz, 16 scans per transient (NS), with 4 dummy scans, 32768 complex points (TD) and a recycle delay (d1) of 1 s for a total acquisition time of 54 s. Reference 1D spectra of protein only were acquired similarly with 16384 scans per transient with a total acquisition time of 12.5 hrs.

An STD experiment with excitation sculpted water suppression was developed from the Bruker pulse sequence stddiffesgp.2. The saturation was achieved using a concatenated series of 50ms Gaussian shaped pulses to achieve the desired total saturation time (d20). The shape of the pulses was specified by the Bruker shape file Gaus.1.1000, where the pulse is divided into 1000 steps and the standard deviation for the Gaussian shape is 165 steps. The field of the pulse was set to 200Hz, which was calculated internally through scaling the power of the high power 90° pulse.

The total relaxation delay was set to 5s, during which the saturation pulse was applied. The experiment was acquired in an interleaved fashion, with each individual excitation frequency being repeated 8 times (L4) until the total desired number of scans was achieved. Again, the spectrum was centred on the water peak, and the receiver gain optimised. Following recording of the fid, and prior to the recycle delay, a pair of water selective pulses are applied to destroy any unwanted magnetisation. For all gradients (excitation sculpting and spoil), the duration was 3 ms using the smooth-square shape SMSQ10.100. Typical acquisition parameters were sweep width of 9615.39 Hz with typically 128 scans per transient (NS=16 * L4=8), 32768 complex points in the direct dimension and 16 dummy scans were executed prior to data acquisition.

In a typical experiment, two excitation frequencies were required, one exciting protein, and one exciting far from the protein (+20,000Hz, +33ppm from the carrier). A range of mixing times were acquired to allow us to carefully quantify the buildup curve to obtain K_D values. A typical set of values used was 0.1s, 0.3s, 0.5s, 0.7s, 0.9s, 1.1s, 1.3s, 1.5s, 1.7s, 1.9s, 2.0s, 2.5s, 3.0s, 3.5s, 4.0s and 5.0s.

Off and on-resonance spectra were acquired for 16 saturation times, giving a total acquisition time of 8.7 hrs

The experiment was acquired as a pseudo 3D experiment, with each spectrum being acquired at a chosen set of excitation frequencies and mixing times. Relaxation delays were set to 10 s for BSA + tryptophan STDs, and were 5 s otherwise.

NMR Methods Table 1:

Sample:	Ligand conc / µM	Protein conc / µM	On-resonance frequency / ppm	Total acquisition time	Used in
BSA + Tryp	**1000**	**5**			
STD			5.44	13.4hrs	Fig 1 D & F i)
STD 10 ms Gaussian duration			5.44	6.7hrs	
STD 25 ms Gaussian duration			5.44	6.7hrs	
STD 50 ms Gaussian duration			5.44	6.7hrs	
BSA + Tryp	**200**	**5**			
STD			5.44	13.4hrs	Fig 1 F ii)
BSA + Tryp	**40**	**5**			
STD			5.44	13.4hrs	Fig 1 F iii)
BSA	**None**	**5**			
1D				12.5hrs	
STD			5.44	13.4hrs	Fig 1D, F i), F ii) & F iii)
SPIKE	**None**	**1.5**			
1D				12.5hrs	

STD			5.44	8.7hrs	Fig 2 B i), B iii), B iv) & C Fig 3 A ii), B i) & B ii)
SPIKE + Neu5Ac 1	**200**	**1.5**			
STD			5.44	8.7hrs	
SPIKE + 9-azido-Neu5Ac 4	**200**	**1.5**			
STD			5.44	8.7hrs	
SPIKE + BPC-Neu5Ac 5	**200**	**1.5**			
STD			5.44	8.7hrs	Fig 2 B iii), B iv) & C Fig 3 B i)
SPIKE + 6-SiaLac 3	**200**	**1.5**			
STD			5.44	2.5hrs	Fig 3 A ii)
STD			5.44	9.83hrs	Fig 3 A ii)
SPIKE + 3-SiaLac 2	**200**	**1.5**			
STD			5.44	2.5hrs	
STD			5.44	9.83hrs	
STD			5.44	8.7hrs	
SPIKE + BPC-6-SiaLac 6	**200**	**1.5**			
STD			5.44	8.7hrs	Fig 3 B ii)
SPIKE	**None**	**6**			
STD			5.44	8.7hrs	S8
STD			8.00	7.8hrs	Fig 3 A i)
STD var freq 1			List1	29.3hrs	
SPIKE + 3-SiaLac 2	**2000**	**6**			
STD			5.44	8.7hrs	Fig 5 C
STD var freq 2			List2	14.1hrs	
SPIKE + 3-SiaLac 2	**1000**	**6**			
STD			5.44	8.7hrs	Fig 5 C S6 & S8
STD			8.00	7.8hrs	Fig 3 A i) S6 & S8
STD var freq 3			List3	14.7hrs	
SPIKE + 3-SiaLac 2	**60**	**6**			
STD			5.44	8.7hrs	Fig 5 C
SPIKE + 3-SiaLac 2	**12**	**6**			
STD			5.44	8.7hrs	Fig 5 C
Neu5Ac 1	**1000**	**None**			
1D				1 minute	
3-SiaLac 2	**1000**	**None**			
1D				1 minute	
STD			5.44	8.7hrs	S6 & S8
STD			8.00	7.8hrs	Fig 3 A i) S6 & S8

STD var freq 2			List2	14.1hrs	
STD var freq 3			List3	14.7hrs	
6-SiaLac 3	**1000**	**None**			
1D				1 minute	
STD			5.44	8.7hrs	Fig 3 A ii)
STD			8.00	7.8hrs	
BPC-6-SiaLac 6	**500**	**None**			
1D				18 minutes	
9-azido-Neu5Ac 4	**1000**	**None**			
1D				1 minute	
BPC-Neu5Ac 5	**1000**	**None**			
1D				1 minute	Fig 2 B ii)
B.1.1.7-SPIKE	**None**	**6**			
STD			8.00	7.8 h	
B.1.1.7-SPIKE + 3-SiaLac 2	**2000**	**6**			
STD			8.00	7.8 h	

For STD 10-50 ms Gaussian experiments, the saturation times used were every other time from the default STD: 0.1 s, 0.5 s, 0.9 s, 1.3 s, 1.7 s, 2 s, 3 s, 4 s.

List1: For STD var freq 1, the on-resonance frequencies in Hz relative to an offset of 2820.61 Hz are: 337.89, 422.36, 524.93, 736.11, 914.10, 1276.12, 1336.46, 1380.21, 1458.64, 1556.69, 1693.96, 2494.93, 2597.50, 2790.58, 2930.86, 3362.27, 3663.95, 3986.75, 4099.88, 4326.15, 4484.53, 4703.25, 4896.33, 5824.01, 6006.53 and 6208.65. The saturation times used were 2 s, 3 s, 4 s and 5 s.

List2: For STD var freq 2, the on-resonance frequencies in Hz relative to an offset of 2820.61 Hz are: -2399.99, -1979.99, -1530.00, -1050.01, -330.021, 338.096, 1679.95, 1829.94, 1979.94, 2129.94, 2279.94 and 2579.93. The saturation times used were 0.1 s, 0.5 s, 2 s and 5 s.

List3: For STD var freq 3, the on-resonance frequencies in Hz relative to an offset of 2820.61 Hz are: -2579.98, -2459.99, -2339.99, -2039.99, -1488.00, -1120.03, -345.02, 311.97, 1079.96, 1379.95, 1679.95, 1979.94, 2279.94 and 2579.93. The saturation times used were 0.1 s, 0.3 s, 0.5 s and 0.9 s.

List4: For STD var freq 4, the on-resonance frequencies in Hz relative to an offset of 2820.61 Hz are: -2461.55, -1973.77, -1518.69, -1270.72, -693.69, -274.80, 280.08, 808.02, 1047.05, 2055.57, 2630.61 and 2979.65. The saturation times used were 0.1 s, 0.5 s, 0.9 s, 2 s, 3 s and 4 s.

Spectra were also acquired on a 600 MHz spectrometer with Bruker Avance III HD console and 5mm TCI CryoProbe, running TopSpin 3.2.6, recorded in **NMR Methods Table 2**, and a 950 MHz spectrometer with Bruker Avance III HD console and 5mm TCI CryoProbe, running TopSpin 3.6.1, recorded in **NMR Methods Table 3**. The 950 used a SampleJet sample changer. Samples were stored at 15 °C. The parameters used for the STD experiments were the same as above, with the following varying by instrument:

On the 600, typical acquisition parameters were sweep width of 9615.39 Hz with typically 128 scans per transient (NS=16 * L4=8), 32768 complex points in the direct dimension and 2 dummy scans, executed prior to data acquisition.

On the 950, typical acquisition parameters were sweep width of 15243.90 Hz with typically 128 scans per transient (NS=16 * L4=8), 32768 complex points in the direct dimension and 2 dummy scans, executed prior to data acquisition.

NMR Methods Table 2:

Sample:	Ligand conc. / µM	Protein conc. / µM	On-resonance frequency / ppm	Total acquisition time	Used in
Trp STD var freq 4	**1000**	**None**	 List 4	 19.2hrs	 S7
BSA STD var freq 4	**None**	**5**	 List 4	 19.2hrs	 S7
BSA STD	**None**	**10**	 5.44	 7.9hrs	
BSA + Trp STD var freq 4	**1000**	**5**	 List 4	 19.2hrs	 S7
BSA + Trp STD	**25**	**10**	 5.44	 11.9hrs	
SPIKE + Neu5Ac 1 STD	**2000**	**3**	 8.00	 7.9hrs	
SPIKE STD	**None**	**3**	 -1.00	 7.9hrs	
Heparin B STD	**500**	**None**	 -1.00	 7.9hrs	
Heparin C STD	**500**	**None**	 -1.00	 7.9hrs	
SPIKE + Heparin B STD	**500**	**3**	 -1.00	 7.9hrs	
SPIKE + Heparin C STD	**500**	**3**	 -1.00	 7.9hrs	
SPIKE B.1.351 STD STD	**None**	**6**	 -1.00 5.44	 7.9hrs 7.9hrs	
3-SiaLac 2 STD	**1000**	**None**	 -1.00	 7.9hrs	
SPIKE B.1.351 + 3-SiaLac 2 STD	**60**	**6**	 -1.00	 7.9hrs	
SPIKE B.1.351 + 3-SiaLac 2 STD	**200**	**6**	 -1.00	 7.9hrs	
SPIKE B.1.351 + 3-SiaLac 2 STD	**1000**	**6**	 -1.00	 7.9hrs	

SPIKE B.1.351 + 3-SiaLac 2	**2000**	**6**		
STD			-1.00	7.9hrs

NMR Methods Table 2.1:

Sample:	Ligand conc. / µM	Protein conc. / µM	Nanobody conc. / µM	On-resonance frequency / ppm	Total acquisition time	Used in
SPIKE	**None**	**6**	**None**			
STD				8.00	7.9hrs	
3-SiaLac 2	**2000**	**None**	**None**			
STD				8.00	7.9hrs	
C5 Nanobody	**None**	**None**	**6**			
STD				8.00	7.9hrs	
SPIKE + 3-SiaLac 2	**2000**	**None**	**6**			
STD				8.00	7.9hrs	
SPIKE + C5 Nanobody	**None**	**6**	**6**			
STD				8.00	7.9hrs	
SPIKE + 3-SiaLac 2 + C5 Nanobody	**2000**	**6**	**6**			
STD				8.00	7.9hrs	
3-SiaLac 2 + C5 Nanobody	**2000**	**None**	**6**			
STD				8.00	7.9hrs	

NMR Methods Table 3:

Sample:	Ligand conc / µM	Protein conc / µM	On-resonance frequency / ppm	Total acquisition time	Used in
SPIKE	**None**	**6**			
STD			8.00	7.2hrs	Fig 4 A-D S9 A & B
SPIKE + Neu5Ac	**2000**	**3**			
STD			5.44	7.2hrs	Fig 4 A & B S9 A
SPIKE + 9-Azido-Neu5Ac	**2000**	**3**			
STD			5.44	7.2hrs	Fig 4 C & D S9 B

Neu5Ac STD	**2000**	**None**	5.44	7.2hrs	Fig 4 A & B S9 A
Neu5Ac STD	**5000**	**None**	5.44	7.2hrs	
9-Azido-Neu5Ac STD	**2000**	**None**	5.44	7.2hrs	Fig 4 C & D S9 B

uSTA Data Analysis.

NMR spectra with a range of excitation frequencies and mixing times were acquired on ligand only, protein only, and mixed protein/ligand samples (**Supplementary Figure S6**).

To analyse an STD dataset, two projections were created by summing over all 1D spectra, and summing over all corresponding STD spectra. These two projections provide exceptionally high signal to noise, suitable for detailed analysis and reliable peak detection. The UnidecNMR algorithm was first executed on the raw spectra, to identify peak positions and intensities. Having identified possible peak positions, the algorithm then analyses the STD spectra but only allowing resonances in places already identified in the 1D spectrum. Both analyses are conducted using the protein only baselines for accurate effective subtraction of the protein baseline without the need to use relaxation filters (**Supplementary Figure S8**).

The results for the ligand only spectra were first analysed. In each case, excellent agreement with the known assignments was obtained, providing us with confidence in the algorithm. The mixed protein/ligand spectrum was then analysed, which returns very similar results to ligand only case. Contributions from the protein, although small however, are typically evident in the spectra justifying the explicit inclusion of the protein only baseline during the analysis. When analysing the mixture, we included the protein only background as a peak shape whose contribution to the spectrum can be freely adjusted. In this way, the spectra of protein/ligand mixtures could be accurately and quickly deconvolved, with the identified ligand resonances occurring in precisely the positions expected from the ligand only spectra. The results from the previous steps were then used to analyse the STD spectra. As these have much lower signal to noise, we fixed the ligand peak positions to be only those previously identified. Otherwise the protocol performed as described previously, where we used a protein only STD data to provide a baseline.

These analyses allow us to define a ‘transfer efficiency’, which is simply the ratio of the signal from a given multiplet in the STD spectrum, to the total expected in the raw 1D experiment. To obtain ‘per atom’ transfer efficiencies, signal from the various pre-assumed components on the multiplets from each resonance were first summed before calculating the ratio. In the software, this is achieved by manually annotating the initial peak list using information obtained from independent assignment experiments (see **Supplementary Figures S19,S20**).

Over the course of the project, it became clear that subtracting the transfer efficiencies obtained from a ligand only sample was an essential part of the method (**Supplementary Figure S9,S10**). Depending on the precise relationship between the chemical shift of excitation, the location of the ligand peaks, and the excitation profile of the Gaussian train, we observed small apparent STD transfer in the ligand only sample that cannot be attributed to ligand binding, arising from a small residual excitation of ligand protons, followed by internal cross relaxation. It is likely that this excitation occurs at least in part via resonances of the ligand that are exchange broadened, such as OH protons, which are not directly observed in the spectrum. When exciting far from the protein, zero ligand excitation is observed, as we would expect, but when exciting close to the methyls, or in the aromatic region, residual ligand excitation could be detected in ligand only samples (**Supplementary Figure S9,S10**). Without the ligand only correction, the uSTA surface may appear to be highly dependent on choice of excitation frequency. However, with the ligand correction, the relative uSTA profiles become invariant with excitation frequency. In

general therefore, we advise acquiring these routinely, and so the uSTA analysis assumes the presence of this data (**Supplementary Figure S6,S8**). The invariance of relative transfer efficiency with excitation frequency suggest that the internal evolution of magnetization within the protein during saturation (likely on the ms timescale) is much faster than the effective cross relaxation rate between protein and ligand (occurring on the seconds timescale).

Having identified the relevant resonances of interest, and performed both a protein, and residual ligand subtraction, the spectra were then re-analysed without first summing over the different mixing times, in order to develop the quantitative atom-specific build-up curves. These were quantitatively analyzed as described below to obtain K_D and k_{off} rates. The values we obtain performing this analysis on BSA/Trp closely match those measured by ITC, and the values we measure for ligand **2** and SPIKE are in good agreement with those measured by SPR as described in the text.

The coverage of protons over the ligands studied here was variable, as for example, there are no protons on carboxyl groups. To enable a complete surface to be rendered, the transfer efficiencies for each proton were calculated as described above, and the value is then transferred to the adjacent heteroatom. For heteroatoms not connected to an observed proton, a $1/r^6$ weighted average score was calculated. This approach allows us to define a unique surface. Caution should be exercised when quantitatively interpreting such surfaces where there are no reliable measurements of the heteroatom.

In practice, raw unformatted fids are submitted to the uSTA pipeline, and the various steps are performed largely automatically, where a user needs to manually adjust processing settings such as phasing and choosing which regions to focus on, iteratively adjust the peak shape to get a good match between the final reconvolved spectrum and the raw data, and input manual atomic assignments for each observed multiplet. The uSTA analysis pipeline then provides a user with a report that shows the results of the various stages of analysis, and uses pymol to render the surfaces. The final transfer efficiencies delivered by the program can be combined with a folder containing a series of HADDOCK models to provide final structural models (**Figure 5**).

Quantitative analysis via uSTA.

In principle, a complete description of the saturation transfer experiment can be achieved via the Bloch-McConnell equations. If we can setup a density matrix describing all the spins in the system, their interactions and their rates of chemical change in an evolution matrix R, then we can follow the system with time according to:

$$\rho(t) = \rho(0)e^{-Rt}$$

The challenge comes from the number of spins that needs to be included, and the need to accurately describe all the interactions between them, which will need to also include how these are modulated by molecular motions in order to get an accurate description of the relaxation processes. This is illustrated by the CORCEMA method,[62] that takes a static structure of a protein/ligand complex and estimates STD transfers. The calculations performed to arrive at cross relaxation rates assumes the complex is rigid, which is a poor approximation for a protein, and because of the large number of spins involved, the calculation is sufficiently intensive such that this calculation cannot be routinely used to fitted to experimental data.

It would be very desirable to extract quantitative structural parameters, as well as chemical properties such as interaction strengths and association/dissociation rates directly from STD data. In what follows we develop a simple quantitative model for the STD experiment to achieve this goal. We will treat the system as comprising just two spins, one to represent the ligand and one to represent the protein, and we allow the two spins to exist either in isolation, or in a bound state. We can safely neglect scalar coupling and so we only need to allow the x,y and z basis operators for each spin, together with an identity operator to ensure the system returns to thermal equilibrium at long times. As such, our evolution matrix R will be square matrices with 13 x 13 elements.

For the spin part, our model requires us to consider the chemical shift of the ligand in the free and bound states, and the chemical shifts of the protein in the free and bound states. In practice however, it is sufficient to set the free protein state on resonance with the pulse, and

ensure that the free ligand chemical shift is far removed from the pulse as we would expect to have in the experiment.

The longitudinal and transverse relaxation rates are calculated for the free and bound states using a simple model assuming in each state there are two coupled spins separated by a distance R. In addition, cross relaxation between ligand and protein is allowed only when the two are bound. The relaxation rates are characterised by an effective distance, and an effective correlation time.

$$R_1 = \frac{1}{4}K\left(J(0)+3J(\omega)+6J(2\omega)\right)$$
$$R_2 = \frac{1}{4}K\left(\frac{5}{2}J(0)+\frac{9}{2}J(\omega)+3J(2\omega)\right)$$
$$\sigma = \frac{1}{4}K\left(6J(2\omega)-J(0)\right)$$

Which are each parameterized in terms of an interaction constant $K = \left(\frac{\mu_0 \hbar \gamma_H^2}{4\pi r^3}\right)^2$ and a spectral density function $J(\omega) = \frac{2}{5}\frac{\tau}{1+\omega^2\tau^2}$. The longitudinal and transverse relaxation rates R_1 and R_2 describe auto-relaxation of diagonal z, and xy elements respectively. The cross-relaxation rates σ describes cross relaxation and couples z elements between the ligand and protein in the bound state. We ensure that the system returns to equilibrium at long times by adding elements of the form $R_1 M_0$ or σM_0 linking the identify element, and the z matrix elements. Overall, the relaxation part of the model is parametised by two correlation times, one for the ligand, and one for the protein/complex, and three distances, one for the ligand auto relaxation rates, one for protein auto relaxation rates, and one for the protein/ligand separation.

Finally, the chemical kinetics govern the rates at which the spins can interconvert. We will take a simple model where $PL \leftrightarrows P + L$, whose dissociation constant is given by:

$$K_D = k_{on} / k_{off} = [P][L]/[PL]$$

The free protein concentration can be determined from knowledge of the K_D, and the total ligand and protein concentrations:

$$[P] = \frac{1}{2}\left(P_{Tot} - L_{Tot} - K_D + \sqrt{\left(L_{Tot} + K_D - P_{Tot}\right)^2 + 4P_{Tot}K_D}\right)$$

From which the bound protein concentration, and the free and bound ligand concentrations can be easily calculated.

The density matrix is initialised with the free and bound protein/ligand concentrations assigned to the relevant z operators. It was found to be important to additionally include a factor that accounts for the increased proton density within the protein. The saturation pulse is then applied either as a concatenated series of Gaussian pulses whose duration and peak power in Hz needs to be specified, exactly matching the pulse shapes and durations used in the experiment (see NMR methods above).

Build up curves and transfer efficiencies can be easily simulated using this model and compared to data, and the various parameters can be optimized to fit to the data. In total, the model is characterized by nine parameters: K_D, k_{off}, the correlation times of the ligand and the protein, the three distances described above and the proton density within the protein. There is substantial correlation between the effects of the various parameters. By obtaining data at various

protein and ligand concentrations however, it is possible to break this degeneracy and obtain well described values as described in the text.

In practical terms, the initial rate of the buildup curve is predominantly affected by the cross-relaxation rate and the off rate, and the final height of the buildup curve is mostly influenced by the proton density in the protein and K_D. Software to perform this analysis has been directly incorporated into the uSTA software.

Parameters Fitted by the Model

Overall the model is parameterised by a set of values that characterise the intrinsic and cross relaxation. From tG and rIS(ligand) we estimate R_1 and R_2 of the ligand, from tE, rIS(protein) we obtain R_1 and R_2 of the protein, and from tE and rIS(complex) we calculate the cross relaxation rate. These values are combined with a factor that accounts for the larger number of spins present in the protein, 'fac', and the on and off rates, to complete a set of 8 parameters that specify our model. The distances should be considered 'effective' values that parasitize the relaxation rates though in principle it should be possible to obtain physical insights form their interpretation. The concentration independent relaxation rates can be separated from the exchange rates by comparing the curves as a ruction of ligand and protein concentration. By treating the system as comprising of two spins we are effectively assuming that the cross relaxation within the protein is very efficient. In the STD experiment, saturation pulses are applied for several second, which is sufficient for near-saturating spin diffusion within a protein.

Thermostability assays.

Thermal stability assays were performed using a NanoTemper Prometheus NT.48 (Membrane Protein Laboratory, Diamond Light Source). To 11 uL of 2 uM Spike (deuterated PBS), 2 uL of trisaccharide 2 (deuterated PBS) was titrated to give final concentrations of 0.1, 0.2, 0.4, 0.8, 1.6 and 2.0 mM. Samples were then loaded into capillaries and heated from 15 to 95 ˚C. Analysis was performed using PR.ThermControl v2.3.1 software.

SPR binding measurement assays.

All experiments were performed on a Biacore T200 instrument. For the immobilisation of SiaLac onto sensor chip a flow rate 10 μL/min was used in a buffer solution of HBP-EP (0.01 M HEPES pH 7.4, 0.15 M NaCl, 3 mM EDTA, 0.005% v/v Surfactant P20). A CM5 sensor chip (carboxymethylated dextran) was equilibrated with HBS-EP buffer at 20°C. The chip was activated by injecting a mixture of *N*-hydroxysuccinimide (50 mM) and EDC-HCl (200 mM) for 10 min followed by a 2 min wash step with buffer. Ethylenediamine (1 M in PBS) was then injected for 7 min followed by a 2 min wash step followed by ethanolamine-HCl (1 M, pH 8.5) for 10 min and then a further 1 min wash step. Finally, SiaLac-IME (5.6 mM in PBS) was injected over 10 min and a final 2 min wash step was performed (see **Supplementary Figure S13**)

For analysis of spike binding a flow rate 10 μL/min was used at 16 °C. Serial dilutions of spike (0.19, 0.50, 1.36, and 3.68 μM) were injected for 30 s association and 150 s dissociation starting with the lowest concentration. Buffer only runs were carried out before injection of spike and after the first two dilutions. BSA (3.03 μM in PBS) was used as a negative control and a mouse serum in a 100-fold dilution was used as a positive control.

Analysis of SPR Data.

To analyse the SPR data, we assume an equilibrium of the form $PL \rightleftarrows P + L$ characterised by a dissociation constant $K_D = \frac{k_{on}}{k_{off}} = \frac{[PL]}{[P][L]}$. To follow the kinetics of binding and dissociation, we

assume the SPR response Γ is proportional to the bound complex $\Gamma = \kappa[PL]$, which leads to the following kinetic equation:

$$\frac{d\Gamma}{dt} + \frac{k_{off}}{\kappa}\Gamma - \frac{k_{on}}{\kappa}[P][L] = 0$$

This can be solved when restrained by the total number of binding sites, $L_{tot} = [L] + [PL]$. Under conditions of constant flow, we assume the free protein concentration is constant, which leads to the following:

$$\Gamma_{on} = \frac{\kappa k_{on} L_{tot}[P]}{k_{off} + k_{on}[P]}\left(1 - e^{-(k_{off} + k_{on}[P])t}\right)$$

And similarly, for dissociation where we take the concentration of free protein to be zero:

$$\Gamma_{off} = \kappa R e^{-k_{off} t}$$

The recovery of the chip was not complete after each protein concentration and wash step, as has been observed for shear-induced lectin-ligand binding with glycans immobilised onto a chip surface.[63] Nevertheless, the data were well explained by a global analysis where the on and off rates were held to be identical for each replicate, but the value of k was allowed to vary slightly between runs, and an additional constant was introduced to Γ_{off} to account for incomplete recovery of the SPR signal following standard approaches. Concentration of spike was insufficient to get the plateau region of the binding, and so the specific time values taken for the on rate affect the fitted values.

Modelling of the N-terminal Domain of SARS-CoV-2 with Glycans.

We modelled the structure of the N-terminal domain (NTD) on Protein Data Bank (PDB) entry 7c2l [17] since it provided significantly better coverage of the area of interest when compared to the majority of the templates available on the PDB as of July 15th 2020. The models were created with Modeller [64], using the '`automodel`' protocol without refining the '`loop`'. We generated 10 models and ranked them by their DOPE score [65], selecting the top 5 for ensemble docking.

Docking of 3'-sialyllactose to SARS-CoV-2 NTD

We docked 3'-sialyllactose to NTD with version 2.4 of the HADDOCK webserver. [43,44] The binding site on NTD was defined by comparison with PDB entry 6q06 [9], a complex of MERS-CoV spike protein and 2,3-sialyl-N-acetyl-lactosamine. The binding site could not directly be mapped because of conformational differences between the NTDs of MERS-CoV and SARS-CoV-2, but by inspection a region with similar properties (aromatics, methyl groups and positively charged residues) could be identified. We defined in HADDOCK the sialic acid as 'active' and residues 18, 19, 20, 21, 22, 68, 76, 77, 78, 79, 244, 254, 255, 256, 258 and 259 of NTD as 'passive', meaning the sialic acid needs to make contact with at least one of the NTD residues but there is no penalty if it doesn't contact all of them, thus allowing the compound to freely explore the binding pocket. Since only one restraint was used we disabled the random removal of restraints. Following our small molecule docking recommended settings [66] we skipped the 'hot' parts of the semi-flexible simulated annealing protocol ('`initiosteps`' and '`cool1_steps`' set to 0) and also lowered the starting temperature of the last two substages to 500 and 300K, respectively ('`tadinit2_t`' and '`tadinit3_t`' to 500 and 300, respectively). Clustering was performed based on 'RMSD' with a distance cut-off of 2Å and the scoring function was modified to:

$$\text{HADDOCKscore} = 1.0 * \mathrm{E_{vdW}} + 0.1 * \mathrm{E_{elec}} + 1.0 * \mathrm{E_{desol}} + 0.1 * \mathrm{E_{AIR}}$$

All other settings were kept to their default values. Finally, the atom-specific transfer efficiencies determined by uSTA were used to filter cluster candidates.

Genetic analysis of clinical samples.

Variant calling: Reads were mapped to the hg19 reference genome by the Burrow-Wheeler aligner BWA. Variants calling was performed according to the GATK4 best practice guidelines. Namely, duplicates were first removed by *MarkDuplicates*, and base qualities were recalibrated using *BaseRecalibration* and *ApplyBQSR*. *HaplotypeCaller* was used to calculate Genomic VCF files for each sample, which were then used for multi-sample calling by *GenomicDBImport* and *GenotypeGVCF*. In order to improve the specificity-sensitivity balance, variants quality scores were calculated by *VariantRecalibrator* and *ApplyVQSR*, and only variants with estimated truth sensitivity above 99.9% were retained. Variants were annotated by ANNOVAR.

Rare variant selection: Missense, splicing and loss of function variants with a frequency lower than 0.01 according to ExAC_NFE (Non Finnish European ExAC Database) were considered for further analyses. A score of 0 was assigned to each samples where the gene is not mutated and the score of 1 was assigned when at least one variant is present on the gene.

Genes prioritization by Logistic regression.

Discriminating genes in COVID-19 disease were interpreted in a framework of feature selection analysis using a customized feature selection approach based on the recursive feature elimination algorithm applied to the LASSO (Least Absolute Shrinkage and Selection Operator) logistic regression model. Specifically, for a set of n samples $\{x_i, y_i\}$ $(i = 1, \dots, n)$ each of which consists of p input features $x_{i,k} \in \chi_i$ $k = 1, \dots, p$ and one output variable $y_i \in Y$, these features assumed the meaning of genes, whereas the samples were the patients involved in the study. The space $\chi = \chi_1 \times \chi_2 \dots \times \chi_p$ was denoted "input space", whereas the "hypothesis space" was the space of all the possible functions $f: \chi \rightarrow Y$ mapping the inputs to the output. Given the number of features (p) is substantially higher than the number of samples (n), LASSO regularization[47] has the effect of shrinking the estimated coefficients to zero, providing a feature selection method for sparse solutions within the classification tasks. Feature selection methods based on such regularization structures (Embedded methods) were most applicable to our scope because they were computationally tractable and strictly connected with the classification task of the ML algorithm.

As the baseline algorithm for the embedded method, we adopted the Logistic Regression (LR) model that is a state-of-the-art ML algorithm for binary classification tasks with probabilistic interpretation. It models the log-odds of the posterior success probability of a binary variable as the linear combination of the input:

$$\log \frac{Pr(Y = 1|X = \boldsymbol{x})}{1 - Pr(Y = 1|X = \boldsymbol{x})} = \beta_0 + \sum_{k=1}^{p} \beta_k \, x_k,$$

where $\boldsymbol{x}$ is the input vector, β_k are the coefficients of the regression and X and Y are the random variables representing the input and the output respectively. The loss function to be minimized is given by the binary cross-entropy loss

$$-\sum_{i=1}^{n} [\mathrm{y_i} \ \log \hat{y}_i \ \ (1 - \mathrm{y_i}) \ \log(1 - \hat{y}_i)],$$

where $\hat{y} = Pr(Y = 1|X = \boldsymbol{x})$ is the predicted target variable and y is the true label. As already introduced, in order to enforce both the sparsity and the interpretability of the results, the model is trained with the additional LASSO regularization term:

$$\lambda \sum_{k=1}^{p} |\beta_k|.$$

In this way, the absolute value of the surviving weights of the LR algorithm was interpreted as the feature importances of the subset of most relevant genes for the task. Since a feature ranking criterion can become sub-optimal when the subset of removed features is large [67], we applied Recursive Feature Elimination (RFE) methodology. For each step of the procedure, we fitted the model and removed the features with smallest ranking criteria in a recursive manner until a certain number of features was reached.

The fundamental hyper-parameter of LR is the strength of the LASSO term tuned with a grid search procedure on the accuracy of the 10-fold cross-validation. The k-fold cross-validation provided the partition of the dataset into k batches, then exploited k^{-1} batches for the training and the remaining test batch as a test, by repeating this procedure k times. In the grid search method, a cross validation procedure was carried out for each value of the regularization hyperparameter in the range [10^{-4}, ..., 10^{6}]. Specifically, the optimal regularization parameter is chosen by selecting the most parsimonious parameter whose cross-validation average accuracy falls in the range of the best one along with its standard deviation. During the fitting procedure, the class unbalancing was tackled by penalizing the mis-classification of minority class with a multiplicative factor inversely proportional to the class frequencies. For the RFE, the number of excluded features at each steps of the algorithm as well as the final number of features was fixed at 100. All data pre-processing and the RFE procedure was coded in Python; the LR model was used, as included, in the scikit-learn module with the liblinear coordinate descent optimization algorithm.

Pseudoviral Cell-entry Asssay.

A spike-expressing lentivirus in HEK 293T cells was generated using a two-plasmid system. The first plasmid encoded an NL4.3ΔEnvΔVifΔVpr backbone and firefly luciferase, and the second encoded SARS-CoV-2 Spike (Wuhan strain). Spike expression on pseudovirus was confirmed using an ELISA-based assay, capturing with an ACE-2 IgA Fc fusion protein (Absolute Antibody). Pseudovirus was detected via CR3022, anti-Human IgG-HRP (Jackson ImmunoResearch Europe) and 1-Step Ultra TMB ELISA substrate (Thermo Fisher).

$1x10^4$ ACE-2-expressing MDCK cells in PBS were seeded in a flat-bottom 96-well plate and treated with 0.3mU *Arthrobacter ureafaciens* neuraminidase (Merck) or PBS (mock) for 30 minutes (37°C, 5% CO_2). After repeated washing, cells were resuspended in serially diluted pseudovirus in 10% FCS DMEM (Thermo Fisher) for 4 h. After removing excess pseudovirus, cells were incubated in 1% FCS DMEM for 48 h. Cells were washed and subsequently lysed using Reporter Lysis Buffer (Promega) and freeze-thaw cycling. Lysate was incubated with Bright-Glo (Promega) according to the manufacturer's instructions and luminescence was acquired using a SpectraMax M5 (Molecular Devices). Data were analysed using Prism 9 (version 9.1.0 for Macintosh).

References

1 Li, W. *et al.* Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* **426**, 450-454, doi:10.1038/nature02145 (2003).

2 Kuba, K. *et al.* A crucial role of angiotensin converting enzyme 2 (ACE2) in SARS coronavirus–induced lung injury. *Nature Medicine* **11**, 875-879, doi:10.1038/nm1267 (2005).

3 Li, F., Li, W., Farzan, M. & Harrison, S. C. Structure of SARS Coronavirus Spike Receptor-Binding Domain Complexed with Receptor. *Science* **309**, 1864, doi:10.1126/science.1116480 (2005).

4 Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260-1263, doi:papers3://publication/doi/10.1126/science.abb2507 (2020).

5 Yan, R. *et al.* Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* **367**, 1444, doi:10.1126/science.abb2762 (2020).

6 Li, W. *et al.* Identification of sialic acid-binding function for the Middle East respiratory syndrome coronavirus spike glycoprotein. *Proceedings of the National Academy of Sciences* **114**, E8508, doi:10.1073/pnas.1712592114 (2017).

7 Schwegmann-Weßels, C. & Herrler, G. Sialic acids as receptor determinants for coronaviruses. *Glycoconjugate Journal* **23**, 51-58, doi:10.1007/s10719-006-5437-9 (2006).

8 Hulswit, R. J. G. *et al.* Human coronaviruses OC43 and HKU1 bind to 9-O-acetylated sialic acids via a conserved receptor-binding site in spike protein domain A. *Proceedings of the National Academy of Sciences* **116**, 2681, doi:10.1073/pnas.1809667116 (2019).

9 Park, Y.-J. *et al.* Structures of MERS-CoV spike glycoprotein in complex with sialoside attachment receptors. *Nat Struct Mol Biol* **26**, 1151-1157, doi:papers3://publication/doi/10.1038/s41594-019-0334-7 (2019).

10 Qing, E., Hantak, M., Perlman, S. & Gallagher, T. Distinct Roles for Sialoside and Protein Receptors in Coronavirus Infection. *mBio* **11**, e02764-02719, doi:10.1128/mBio.02764-19 (2020).

11 Morniroli, D., Giannì, M. L., Consales, A., Pietrasanta, C. & Mosca, F. Human Sialome and Coronavirus Disease-2019 (COVID-19) Pandemic: An Understated Correlation? *Frontiers in Immunology* **11**, doi:10.3389/fimmu.2020.01480 (2020).

12 Alexander N., B. *et al.* The SARS-COV-2 Spike Protein Binds Sialic Acids, and Enables Rapid Detection in a Lateral Flow Point of Care Diagnostic Device. *ChemRxiv*, doi:10.26434/chemrxiv.12465680.v1 (2020).

13 Nguyen, L. *et al.* Sialic acid-Dependent Binding and Viral Entry of SARS-CoV-2. *bioRxiv*, 2021.2003.2008.434228, doi:10.1101/2021.03.08.434228 (2021).

14 Hao, W. *et al.* Binding of the SARS-CoV-2 Spike Protein to Glycans. *bioRxiv*, 2020.2005.2017.100537, doi:10.1101/2020.05.17.100537 (2020).

15 Clausen, T. M. *et al.* SARS-CoV-2 Infection Depends on Cellular Heparan Sulfate and ACE2. *Cell* **183**, 1043-1057.e1015, doi:10.1016/j.cell.2020.09.033 (2020).

16 Behloul, N., Baha, S., Shi, R. & Meng, J. Role of the GTNGTKR motif in the N-terminal receptor-binding domain of the SARS-CoV-2 spike protein. *Virus Research* **286**, 198058, doi:10.1016/j.virusres.2020.198058 (2020).

17 Chi, X. *et al.* A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science*, eabc6952, doi:10.1126/science.abc6952 (2020).

18 Mayer, M. & Meyer, B. Characterization of Ligand Binding by Saturation Transfer Difference NMR Spectroscopy. *Angewandte Chemie International Edition* **38**, 1784-1788, doi:10.1002/(SICI)1521-3773(19990614)38:12<1784::AID-ANIE1784>3.0.CO;2-Q (1999).

19 Wagstaff, J. L., Taylor, S. L. & Howard, M. J. Recent developments and applications of saturation transfer difference nuclear magnetic resonance (STD NMR) spectroscopy. *Molecular BioSystems* **9**, 571-577, doi:10.1039/C2MB25395J (2013).

20 Angulo, J., Enríquez-Navas, P. M. & Nieto, P. M. Ligand–Receptor Binding Affinities from Saturation Transfer Difference (STD) NMR Spectroscopy: The Binding Isotherm of STD

Initial Growth Rates. *Chemistry – A European Journal* **16**, 7803-7812, doi:10.1002/chem.200903528 (2010).
21 Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology* **5**, 1403-1407, doi:10.1038/s41564-020-0770-5 (2020).
22 Huang, R., Bonnichon, A., Claridge, T. D. W. & Leung, I. K. H. Protein-ligand binding affinity determination by the waterLOGSY method: An optimised approach considering ligand rebinding. *Scientific Reports* **7**, 43727, doi:10.1038/srep43727 (2017).
23 Baldwin, A. J. *et al.* Cytochrome Display on Amyloid Fibrils. *Journal of the American Chemical Society* **128**, 2162-2163, doi:10.1021/ja0565673 (2006).
24 Carver, J. A., Aquilina, J. A., Truscott, R. J. W. & Ralston, G. B. Identification by 1H NMR spectroscopy of flexible C-terminal extensions in bovine lens α-crystallin. *FEBS Letters* **311**, 143-149, doi:10.1016/0014-5793(92)81386-Z (1992).
25 Hsu, S.-T. D. *et al.* Structure and dynamics of a ribosome-bound nascent chain by NMR spectroscopy. *Proceedings of the National Academy of Sciences* **104**, 16516, doi:10.1073/pnas.0704664104 (2007).
26 Mayer, M. & Meyer, B. Group Epitope Mapping by Saturation Transfer Difference NMR To Identify Segments of a Ligand in Direct Contact with a Protein Receptor. *Journal of the American Chemical Society* **123**, 6108-6117, doi:10.1021/ja0100120 (2001).
27 Vallurupalli, P., Bouvignies, G. & Kay, L. E. Studying “Invisible” Excited Protein States in Slow Exchange with a Major State Conformation. *Journal of the American Chemical Society* **134**, 8148-8161, doi:10.1021/ja3001419 (2012).
28 Xie, T., Saleh, T., Rossi, P. & Kalodimos, C. G. Conformational states dynamically populated by a kinase determine its function. *Science* **370**, eabc2754, doi:10.1126/science.abc2754 (2020).
29 Fawzi, N. L., Ying, J., Torchia, D. A. & Clore, G. M. Probing exchange kinetics and atomic resolution dynamics in high-molecular-weight complexes using dark-state exchange saturation transfer NMR spectroscopy. *Nature Protocols* **7**, 1523-1533, doi:10.1038/nprot.2012.077 (2012).
30 Allard, P., Helgstrand, M. & Härd, T. The Complete Homogeneous Master Equation for a Heteronuclear Two-Spin System in the Basis of Cartesian Product Operators. *Journal of Magnetic Resonance* **134**, 7-16, doi:10.1006/jmre.1998.1509 (1998).
31 Marty, M. T. *et al.* Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to Polydisperse Ensembles. *Analytical Chemistry* **87**, 4370-4376, doi:10.1021/acs.analchem.5b00140 (2015).
32 Richardson, W. H. Bayesian-Based Iterative Method of Image Restoration*. *J. Opt. Soc. Am.* **62**, 55-59, doi:10.1364/JOSA.62.000055 (1972).
33 Lucy, L. B. An iterative technique for the rectification of observed distributions. *Astronomical Journal* **79**, 745 (1974).
34 McMenamy, R. H. & Oncley, J. L. The Specific Binding of L-Tryptophan to Serum Albumin. *Journal of Biological Chemistry* **233**, 1436-1447 (1958).
35 Fielding, L., Rutherford, S. & Fletcher, D. Determination of protein–ligand binding affinity by NMR: observations from serum albumin model systems. *Magnetic Resonance in Chemistry* **43**, 463-470, doi:10.1002/mrc.1574 (2005).
36 Bujacz, A., Zielinski, K. & Sekula, B. Structural studies of bovine, equine, and leporine serum albumin complexes with naproxen. *Proteins: Structure, Function, and Bioinformatics* **82**, 2199-2208, doi:10.1002/prot.24583 (2014).
37 Pérez-Victoria, I. *et al.* Saturation transfer difference NMR reveals functionally essential kinetic differences for a sugar-binding repressor protein. *Chemical Communications*, 5862-5864, doi:10.1039/B913489A (2009).
38 Hars, U., Horlacher, R., Boos, W., Welte, W. & Diederichs, K. Crystal structure of the effector-binding domain of the trehalose-repressor of Escherichia coli, a member of the LacI family, in its complexes with inducer trehalose-6-phosphate and noninducer trehalose. *Protein Science* **7**, 2511-2521, doi:10.1002/pro.5560071204 (1998).

39 McConnell, H. M. Reaction Rates by Nuclear Magnetic Resonance. *J. Chem. Phys.* **28**, 430, doi: 10.1063/1.1744152 (1958).
40 Watanabe, Y., Allen, J. D., Wrapp, D., McLellan, J. S. & Crispin, M. Site-specific glycan analysis of the SARS-CoV-2 spike. *Science*, doi:papers3://publication/doi/10.1126/science.abb9983 (2020).
41 Mammen, M., Choi, S.-K. & Whitesides, G. M. Polyvalent Interactions in Biological Systems: Implications for Design and Use of Multivalent Ligands and Inhibitors. *Angewandte Chemie International Edition* **37**, 2754-2794, doi:10.1002/(SICI)1521-3773(19981102)37:20<2754::AID-ANIE2754>3.0.CO;2-3 (1998).
42 Lee, R. T. & Lee, Y. C. Affinity enhancement by multivalent lectin–carbohydrate interaction. *Glycoconjugate Journal* **17**, 543-551, doi:10.1023/A:1011070425430 (2000).
43 Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. *Journal of the American Chemical Society* **125**, 1731-1737, doi:10.1021/ja026939x (2003).
44 van Zundert, G. C. P. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology* **428**, 720-725, doi:10.1016/j.jmb.2015.09.014 (2016).
45 Imberty, A. *et al.* An Unusual Carbohydrate Binding Site Revealed by the Structures of Two Maackia amurensis Lectins Complexed with Sialic Acid-containing Oligosaccharides. *Journal of Biological Chemistry* **275**, 17541-17548 (2000).
46 Huo, J. *et al.* Neutralizing nanobodies bind SARS-CoV-2 spike RBD and block interaction with ACE2. *Nature Structural & Molecular Biology* **27**, 846-854, doi:10.1038/s41594-020-0469-6 (2020).
47 Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267-288, doi:10.1111/j.2517-6161.1996.tb02080.x (1996).
48 Koths, K., Taylor, E., Halenbeck, R., Casipit, C. & Wang, A. Cloning and characterization of a human Mac-2-binding protein, a new member of the superfamily defined by the macrophage scavenger receptor cysteine-rich domain. *Journal of Biological Chemistry* **268**, 14245-14249 (1993).
49 Johannes, L., Jacob, R. & Leffler, H. Galectins at a glance. *Journal of Cell Science* **131**, jcs208884, doi:10.1242/jcs.208884 (2018).
50 Stowell, S. R. *et al.* Galectin-1, -2, and -3 Exhibit Differential Recognition of Sialylated Glycans and Blood Group Antigens. *Journal of Biological Chemistry* **283**, 10109-10123, doi:10.1074/jbc.M709545200 (2008).
51 Kamili, N. A. *et al.* Key regulators of galectin–glycan interactions. *PROTEOMICS* **16**, 3111-3125, doi:10.1002/pmic.201600116 (2016).
52 Togayachi, A. *et al.* beta 3GNT2 (B3GNT2), A MAJOR POLYLACTOSAMINE SYNTHASE: ANALYSIS OF B3GNT2-DEFICIENT MICE. *Methods in Enzymology, Vol 502: Protein Engineering for Therapeutics, Pt A* **479**, 185-204, doi:papers3://publication/doi/10.1016/S0076-6879(10)79011-X (2010).
53 Jia, N. *et al.* The Human Lung Glycome Reveals Novel Glycan Ligands for Influenza A Virus. *Scientific Reports* **10**, doi:papers3://publication/doi/10.1038/s41598-020-62074-z (2020).
54 Toscano, M. A. *et al.* Differential glycosylation of T H 1, T H 2 and T H -17 effector cells selectively regulates susceptibility to cell death. *Nat Immunol* **8**, 825-834, doi:papers3://publication/doi/10.1038/ni1482 (2007).
55 Yamada, S. *et al.* Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type receptors. *Nature* **444**, 378-382, doi:10.1038/nature05264 (2006).
56 Watanabe, Y. *et al.* Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. *Nature Communications* **11**, 2688, doi:10.1038/s41467-020-16567-0 (2020).
57 Casalino, L. *et al.* Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Central Science*, doi:10.1021/acscentsci.0c01056 (2020).
58 Huo, J. *et al.* Neutralization of SARS-CoV-2 by Destruction of the Prefusion Spike. *Cell Host & Microbe* **28**, 445-454.e446, doi:10.1016/j.chom.2020.06.010 (2020).

59 Collins, B. E. *et al.* High-Affinity Ligand Probes of CD22 Overcome the Threshold Set by cis Ligands to Allow for Binding, Endocytosis, and Killing of B Cells. *The Journal of Immunology* **177**, 2994, doi:10.4049/jimmunol.177.5.2994 (2006).
60 Chuang, W.-L., McAllister, H. & Rabenstein, D. L. Chromatographic methods for product-profile analysis and isolation of oligosaccharides produced by heparinase-catalyzed depolymerization of heparin. *Journal of Chromatography A* **932**, 65-74, doi:10.1016/S0021-9673(01)01241-9 (2001).
61 Powell, A. K., Ahmed, Y. A., Yates, E. A. & Turnbull, J. E. Generating heparan sulfate saccharide libraries for glycomics applications. *Nature Protocols* **5**, 821-833, doi:10.1038/nprot.2010.17 (2010).
62 Jayalakshmi, V. & Krishna, N. R. Complete Relaxation and Conformational Exchange Matrix (CORCEMA) Analysis of Intermolecular Saturation Transfer Effects in Reversibly Forming Ligand–Receptor Complexes. *Journal of Magnetic Resonance* **155**, 106-118, doi:10.1006/jmre.2001.2499 (2002).
63 Nakamura, K. *et al.* Immobilized glycosylated Fmoc-amino acid for SPR: comparative studies of lectin-binding to linear or biantennary diLacNAc structures. *Carbohydrate Research* **382**, 77-85, doi:10.1016/j.carres.2013.10.003 (2013).
64 Sali, A. Comparative protein modeling by satisfaction of spatial restraints. *Molecular Medicine Today* **1**, 270-277, doi:10.1016/S1357-4310(95)91170-7 (1995).
65 Shen, M.-y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Science* **15**, 2507-2524, doi:10.1110/ps.062416606 (2006).
66 Koukos, P. I., Xue, L. C. & Bonvin, A. M. J. J. Protein–ligand pose and affinity prediction: Lessons from D3R Grand Challenge 3. *Journal of Computer-Aided Molecular Design* **33**, 83-91, doi:10.1007/s10822-018-0148-4 (2019).
67 Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* **46**, 389-422, doi:10.1023/A:1012487302797 (2002).
68 Lepre, C. A., Moore, J. M. & Peng, J. W. Theory and Applications of NMR-Based Screening in Pharmaceutical Research. *Chemical Reviews* **104**, 3641-3676, doi:10.1021/cr030409h (2004).
69 Viegas, A., Manso, J., Nobrega, F. L. & Cabrita, E. J. Saturation-Transfer Difference (STD) NMR: A Simple and Fast Method for Ligand Screening and Characterization of Protein Binding. *Journal of Chemical Education* **88**, 990-994, doi:10.1021/ed101169t (2011).
70 Abragam, A. *The Principles of Nuclear Magnetism*. (Clarendon Press, 1961).
71 Taylor, J. R. & Taylor, S. L. L. J. R. *Introduction To Error Analysis: The Study of Uncertainties in Physical Measurements*. (University Science Books, 1997).