RESEARCH ARTICLE SUMMARY

CORONAVIRUS

Pathogen-sugar interactions revealed by universal saturation transfer analysis

Charles J. Buchanan†, Ben Gaunt†, Peter J. Harrison, Yun Yang, Jiwei Liu, Aziz Khan, Andrew M. Giltrap, Audrey Le Bas, Philip N. Ward, Kapil Gupta, Maud Dumoux, Tiong Kit Tan, Lisa Schimaski, Sergio Daga, Nicola Picchiotti, Margherita Baldassarri, Elisa Benetti, Chiara Fallerini, Francesca Fava, Annarita Giliberti, Panagiotis I. Koukos, Matthew J. Davy, Abirami Lakshminarayanan, Xiaochao Xue, Georgios Papadakis, Lachlan P. Deimel, Virgínia Casablancas-Antràs, Timothy D. W. Claridge, Alexandre M. J. J. Bonvin, Quentin J. Sattentau, Simone Furini, Marco Gori, Jiandong Huo, Raymond J. Owens, Christiane Schaffitzel, Imre Berger, Alessandra Renieri, GEN-COVID Multicenter Study, James H. Naismith*, Andrew J. Baldwin*, Benjamin G. Davis*

INTRODUCTION: The surface proteins found on both pathogens and host cells mediate cell entry (and exit) and influence disease progression and transmission. Both types of proteins can bear host-generated posttranslational modifications, such as glycosylation, that are essential for function but can confound current biophysical methods used for dissecting key interactions. Several human viruses (including non-SARS coronaviruses) attach to host cell surface N-linked glycans that include forms of sialic acid (sialosides). There remains, however, conflicting evidence as to whether or how SARS-associated coronaviruses might use such a mechanism. In the absence of an appropriate biochemical assay, the ability to analyze the binding of such glycans to heavily modified proteins and resolve this issue is limited.

RATIONALE: We developed and demonstrated a quantitative extension of "saturation transfer" protein nuclear magnetic resonance (NMR) methods to a complete mathematical model of the magnetization transfer caused by interactions between protein and ligand. The designed method couples object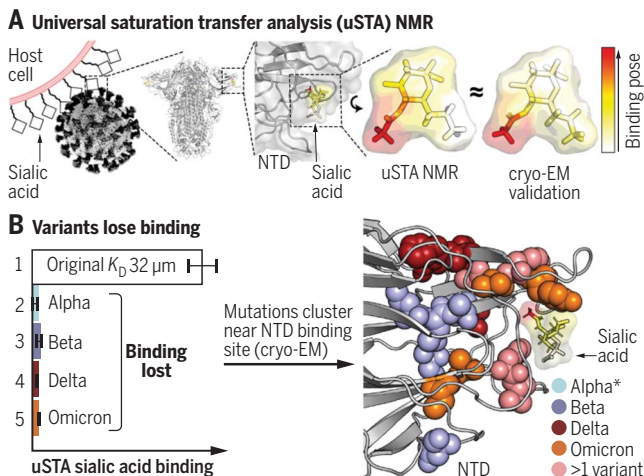ive resonance identification and intensity measurement in NMR spectra (via a deconvolution algorithm) with Bloch-McConnell analysis of magnetization transfer (as judged by this resonance signal intensity) to enable a structural, kinetic, and thermodynamic analysis of ligand binding. Such quantification is beyond previously perceived limits of exchange rates, concentration, or system and therefore represents a potentially universal saturation transfer analysis (uSTA) method.

RESULTS: In an automated workflow, uSTA can be applied to a range of even heavily modified protein systems in a general manner to obtain quantitative binding interaction parameters (K_D, k_{Ex}). uSTA proved critical in mapping direct interactions between sialoside sugar ligands and relevant virus surface attachment glycoproteins, including multiple variants of both severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spike protein and influenza H1N1 hemagglutinin protein. It was successful in quantitating ligand NMR signals in spectral regions otherwise occluded by resonances from mobile protein glycans. In early-pandemic (December 2019) B-origin-lineage SARS-CoV-2 spike trimer, a clear "end-on" binding mode of sialoside sugars to spike was revealed by uSTA.

This mode contrasted with "extended-surface side"-binding for heparin sugar ligands. uSTA-derived restraints used in structural modeling suggested sialoside-glycan binding sites in a β sheet–rich region of spike N-terminal domain (NTD), distant from the receptor-binding domain (RBD) that binds ACE2 co-receptor and that has been identified as the site for other sugar interactions. Consistent with this NTD site being a previously unknown sialoside sugar-binding pocket, uSTA-sialoside binding was minimally perturbed by antibodies that neutralize the ACE2-binding RBD domain. Strikingly, uSTA also shows that this sialoside binding is disrupted in spike from multiple variants of concern (B1.1.7/alpha, B1.351/beta, B.1.617.2/delta, and B.1.1.529/omicron) that emerged later in the pandemic (September 2020 onward). Notably, these variants possess multiple hotspot mutations in the NTD. End-on sialoside binding in a B-origin-lineage spike-NTD pocket was pinpointed by cryo-EM to a previously unknown site that is created from residues that are notably mutated or are in regions where mutations occur in variants of concern (e.g., His[69], Val[70], and Tyr[145] in alpha and omicron). An analysis of beneficial genetic variances correlated with disease severity in cohorts of patients from early 2020 suggests a model in which this site in the NTD of B-origin-lineage SARS-CoV-2 (but not in later variants) may have exploited a specific sialylated polylactosamine motif found on tetra-antennary human N-linked glycoproteins, known to be present in deeper human lung.

CONCLUSION: Together, these results confirm a distinctive sugar-binding mode mediated by the unusual NTD of B-origin-lineage SARS-CoV-2 spike protein that is lost in later variants. This may implicate modulation of binding by SARS-CoV-2 virus to human cell surface sugars as a determinant of virulence and/or zoonosis. More generally, because cell surface glycans are widely relevant to biology and pathology, the uSTA method can now provide ready, quantitative, widespread analysis of complex, host-derived, and posttranslationally modified proteins in their binding to putative ligands, which may be relevant to disease, even in previously confounding complex systems. ∎

uSTA identifies a sugar-binding site and pose in the unusual NTD of early-pandemic SARS-CoV-2 spike. (**A**) This site and pose, which were confirmed by cryo–electron microscopy, are heavily mutated in variants of concern. (**B**) Analyses reveal a loss of sialoside-spike binding, rationalized by clustering of mutations around the binding site. *All alpha-variant mutations are in >1 variant.

A Universal saturation transfer analysis (uSTA) NMR

Host cell / Sialic acid / NTD / Sialic acid / uSTA NMR / cryo-EM validation / Binding pose

B Variants lose binding

1 Original K_D 32 μm
2 Alpha
3 Beta — Binding lost
4 Delta
5 Omicron

uSTA sialic acid binding

Mutations cluster near NTD binding site (cryo-EM)

Sialic acid

Alpha* / Beta / Delta / Omicron / >1 variant

NTD

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abm3125

CORONAVIRUS

Pathogen-sugar interactions revealed by universal saturation transfer analysis

Charles J. Buchanan[1,2,3]†, Ben Gaunt[1]†, Peter J. Harrison[4,5], Yun Yang[1,4], Jiwei Liu[1], Aziz Khan[1,2], Andrew M. Giltrap[1,2], Audrey Le Bas[1,4], Philip N. Ward[4], Kapil Gupta[6], Maud Dumoux[1], Tiong Kit Tan[7], Lisa Schimaski[7], Sergio Daga[8,9], Nicola Picchiotti[10,11], Margherita Baldassarri[8,9], Elisa Benetti[9], Chiara Fallerini[8,9], Francesca Fava[8,9,12], Annarita Giliberti[8,9], Panagiotis I. Koukos[13], Matthew J. Davy[1], Abirami Lakshminarayanan[1,2], Xiaochao Xue[2,14], Georgios Papadakis[2], Lachlan P. Deimel[14], Virgínia Casablancas-Antràs[2,3], Timothy D. W. Claridge[2], Alexandre M. J. J. Bonvin[13], Quentin J. Sattentau[14], Simone Furini[9], Marco Gori[10,15], Jiandong Huo[1,4], Raymond J. Owens[1,4], Christiane Schaffitzel[13], Imre Berger[13], Alessandra Renieri[8,9,12], GEN-COVID Multicenter Study, James H. Naismith[1,4]*, Andrew J. Baldwin[1,2,3]*, Benjamin G. Davis[1,2,16]*

Many pathogens exploit host cell-surface glycans. However, precise analyses of glycan ligands binding with heavily modified pathogen proteins can be confounded by overlapping sugar signals and/or compounded with known experimental constraints. Universal saturation transfer analysis (uSTA) builds on existing nuclear magnetic resonance spectroscopy to provide an automated workflow for quantitating protein-ligand interactions. uSTA reveals that early-pandemic, B-origin-lineage severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spike trimer binds sialoside sugars in an "end-on" manner. uSTA-guided modeling and a high-resolution cryo–electron microscopy structure implicate the spike N-terminal domain (NTD) and confirm end-on binding. This finding rationalizes the effect of NTD mutations that abolish sugar binding in SARS-CoV-2 variants of concern. Together with genetic variance analyses in early pandemic patient cohorts, this binding implicates a sialylated polylactosamine motif found on tetraantennary N-linked glycoproteins deep in the human lung as potentially relevant to virulence and/or zoonosis.

Sialosides are present in glycans that are anchored to human cells, and they mediate binding that is central to cell-cell communication in human physiology and that is at the heart of many host-pathogen interactions. One of the most well-known examples is that of influenza virus, which binds to sialosides with its hemagglutinin (HA or H) protein and cleaves off sialic acid from the infected cell with its neuraminidase (NA or N) protein; HxNx variants of influenza with different HA or NA protein types have a profound effect on zoonosis and pathogenicity (1).

The Middle East respiratory syndrome [MERS (2)] virus, which is related to severe acute respiratory syndrome coronavirus 1 and 2 (SARS-CoV-1 and -2), has been shown to exploit cell-surface sugar sialosides (2–6) as part of an attachment strategy. Both SARS-CoV-1 (7–9) and SARS-CoV-2 (10, 11) are known to 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, but whether these viruses engage sialosides as part of the infection cycle has, despite predictions (6, 12), remained unclear. Preliminary reports as to whether complex sialosides are or are not bound are contradictory and format-dependent (13–15). Glycosaminoglycans on proteoglycans such as heparin have been identified as a primary cooperative glycan attachment point (16, 17). Studies reporting sugar binding have so far implicated binding sites in or close to the RBD of the spike protein. Surprisingly, the N-terminal domain (NTD, fig. S1), which has a putative glycan binding fold (10, 18, 19) and binds sialosides in other non-SARS coronaviruses (including MERS), has been less explored. The NTD has no confirmed function

in SARS-CoV-2, and yet neutralization by antibodies against this domain suggests a potentially important function in viral replication. The unresolved role of host cell-surface sialosides for this pathogen has been noted as an important open question (17). The hypothesized roles for sugar interactions (20) in both virulence (21) and zoonosis (1) indicate that there is an urgent need for precise, quantitative, and robust methods for analysis.

In principle, magnetization transfer in protein nuclear magnetic resonance (NMR) spectroscopy could meet this need, as it can measure ligand binding in its native state without the need for additional labeling or modification of either ligand or protein (e.g., attachment to surface or sensor) (figs. S2 and S3; see text S1 for more details). Saturation transfer difference (STD) (22), which has been widely used to gauge qualitative ligand-protein interactions (23), detects the transfer of magnetization while they are bound via "cross-relaxation."

In reality, complex, highly modified protein systems have proven difficult to analyze in a quantitative manner with current methods for several reasons. First, mammalian proteins (or those derived by pathogens from expression in infected mammalian hosts) often bear large, highly mobile glycans. Critically, in the case of glycoproteins such as SARS-CoV-2 spike that may themselves bind glycans, this leads to contributions to protein NMR spectra that may overlap with putative glycan ligand resonances, thereby obscuring needed signal. Second, the NMR spectra of glycan ligands are themselves complex, comprising many overlapped resonances as multiplets and limiting the accurate determination of signal intensities. Finally, STD is commonly described as limited to specific kinetic regimes and/or ligand-to-protein binding equilibrium positions (24). As a result, many regimes and systems have been considered inaccessible to STD.

Using a rigorous theoretical description, coupled with a computational approach based on a Bayesian deconvolution algorithm to objectively and accurately extract signal from all observed resonances, we have undertaken an optimized reformulation of the magnetization/ "saturation" transfer protocol (figs. S4, S5, S6, and S8). This approach reliably and quantitatively determines precise binding rates (k_{on}, k_{off}, k_{ex}), constants (K_D), and interaction "maps" across a wide range of regimes (fig. S4), including systems previously thought to be intractable.

Design of uSTA based on a comprehensive treatment of ligand-protein magnetization transfer

While using existing STD methodology to study the interaction between the SARS-CoV-2 spike protein and sialosides, we noted several challenges that resulted in the development of

[1]Rosalind Franklin Institute, Harwell Science and Innovation Campus, Oxford OX11 0FA, UK. [2]Department of Chemistry, University of Oxford, Oxford OX1 3TA, UK. [3]Kavli Institute of Nanoscience Discovery, University of Oxford, Oxford OX1 3QU, UK. [4]Division of Structural Biology, Wellcome Centre for Human Genetics, University of Oxford, Headington, Oxford OX3 7BN, UK. [5]Diamond Light Source, Harwell Science and Innovation Campus, Oxfordshire, UK. [6]Max Planck Bristol Centre for Minimal Biology, University of Bristol, Bristol, UK. [7]MRC Human Immunology Unit, Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK. [8]Medical Genetics, University of Siena, Siena, Italy. [9]Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, Siena, Italy. [10]Department of Information Engineering and Mathematics, University of Siena, Siena, Italy. [11]Department of Mathematics, University of Pavia, Pavia, Italy. [12]Genetica Medica, Azienda Ospedaliera Universitaria Senese, Siena, Italy. [13]Bijvoet Centre for Biomolecular Research, Faculty of Science, Utrecht University, Utrecht, Netherlands. [14]Sir William Dunn School of Pathology, Oxford, UK. [15]Maasai, I3S CNRS, Université Côte d'Azur, Nice, France. [16]Department of Pharmacology, University of Oxford, Oxford OX1 3QT, UK.
*Corresponding author. Email: naismith@strubi.ox.ac.uk (J.H.N.); andrew.baldwin@chem.ox.ac.uk (A.J.B.); ben.davis@rfi.ac.uk (B.G.D.)
†These authors contributed equally to this work.

uSTA (see text S3 for more details). Our theoretical analyses (fig. S4) suggested that many common assumptions or limits that are thought to govern the applicability of magnetization transfer might in fact be circumvented, and we set out to devise a complete treatment that might accomplish this (figs. S5, S6, and S8). This resulted in five specific methodological changes that resulted in a more sensitive, accurate, quantitative, and general method for studying the interactions between biomolecules and ligands (summarized in figs. S5 and S8 and discussed in detail in text S4).

1) We noted the discrepancy between K_Ds determined using existing STD methods and those obtained using other biophysical methods ([25]). We performed a theoretical analysis using the Bloch-McConnell equations, a rigorous formulation for studying the evolution of magnetization in exchanging systems that has been widely used to analyze chemical exchange saturation transfer ([26], [27]), dark-state exchange saturation transfer ([28]), and Carr-Purcell-Meiboom-Gill ([29]) NMR data to describe protein motion. This analysis not only allowed us to explain this discrepancy, but also enabled fitting of data to give accurate k_{on}, k_{off}, and K_D values for protein-ligand interactions that were in excellent accord with alternative measurements (Fig. 1G and fig. S13); we also found that the range of k_{on} and k_{off} in which the experiment is applicable is far wider than previously recognized (fig. S4).

2) In mammalian proteins, contributions from glycans on the surface of the protein could not be removed from the spectrum by means of relaxation filters used in epitope mapping ([30]) without compromising the sensitivity of the experiment. We addressed this instead by applying baseline subtraction using data obtained from a protein-only sample.

3) The magnetization transfer, and hence the sensitivity of the experiment, will be higher when the excitation frequency of the saturation pulse is close to a maximum in the protein NMR spectrum. If any ligand resonances are outside of the bandwidth of the pulse, and if a "ligand-only" subtraction is applied, the magnetization transfer can be maximized. With this condition, the response for a given protein-ligand system in fact becomes invariant to the excitation frequency used (figs. S9 and S10).

4) In complex molecules, such as sialosides, NMR spectra are crowded and overlapped. To reliably obtain magnetization transfer measurements at all points in the ligand, we developed a peak-picking algorithm based on earlier work ([31]) that can automate the process, returning a list of peak locations and a simulated NMR experiment that can be directly compared to the data. The locations of the peaks are in excellent agreement with the locations for multiplets determined using standard multidimensional approaches used for resonance assignment (see Fig. 1, B and D, Fig. 2D, and Fig. 3A for examples; see also overlaps in all subsequent uSTA analyses and table S7).

5) The uSTA software allowed the combination of intensity from scalar coupled multiplets following a user-input assignment, to provide "per resonance" measures of saturation transfer. These are provided as-is, and also as $\langle 1/r^6 \rangle$ interpolated "binding maps" that represent the interaction on nearby heteroatoms, thereby allowing ready visual inspection of the binding pose of a molecule.

Testing of uSTA in model systems

The uSTA method (Fig. 1, A to C) was tested first in an archetypal, yet challenging, ligand-to-protein interaction (Fig. 1, D to G). Implementation in an automated manner through software governing the uSTA workflow reduced artifacts arising from subjective, manual analyses (fig. S6). The binding of L-tryptophan (Trp) to bovine serum albumin (BSA, Fig. 1D) is a long-standing benchmark ([25]) because of the supposed role of hydrophobicity in the plasticity of this interaction as well as a lack of corresponding fully determined, unambiguous three-dimensional (3D; e.g., crystal) structures. This is also a simpler amino acid–protein interaction system (less-modified protein, small ligand) that classical NMR/STD methods are perceived ([24], [32]) to have already delineated well.

As for a standard STD experiment, 1D ^{1}H-NMR spectra were determined for both ligand and protein. In addition, mixed spectra containing both protein and excess ligand ($P + L$) were determined with and without excitation irradiation at frequencies corresponding to prominent resonance within the protein but far from any ligand (pulse "on") or where the center of the pulse was moved to avoid ligand and protein (pulse "off," labeled "1D" in the figures). Deconvolved spectra for ligand determined in the presence of protein were matched with high accuracy by uSTA (Fig. 1E). Moreover, uSTA generated highly consistent binding "heatmaps" comprising atom-specific magnetization transfer efficiencies (proton data mapped onto heteroatoms by taking a local $1/r^6$ average to enable visual comparison) that described the pose of ligand bound to protein (Fig. 1F; see also figs. S8 and S11). These were determined over a range of ligand concentrations even as low as 40 μM [Fig. 1, E(iii) and F(iii)] where the ability of uSTA to extract accurate signal proved unprecedented and critical to quantitation of binding (see below). Binding maps were strikingly consistent across concentrations, indicating a single, consistent pose driven by the strongest interaction of protein with the heteroaromatic indole side chain of Trp. This not only proved consistent with x-ray crystal structures of BSA with other hydrophobic ligands, ([33]) it also revealed quantitative subtleties of this interaction at high precision: Protein "grip" is felt more at the distal edge of the indole moiety.

Next, with indications of expanded capability of uSTA in a benchmark system, we moved to first analyses of sugar-protein interactions. Sugar ligand trehalose (Tre) binds only weakly to trehalose repressor protein TreR and so proves challenging in ligand-to-protein interaction analysis ([34]). Nonetheless, the uSTA workflow again successfully and rapidly determined atom-specific transfer efficiencies with high precision and resolution (fig. S3F). Atom-precise subtleties were revealed in this case as well: Hotspots of binding occur around OH-3/OH-4 and graduate to reduced binding around both sugar rings, with only minimal binding of the primary OH-6 hydroxyl (fig. S3F). Once more, this uSTA-mapped $P + L$ interaction proved consistent with prior x-ray crystal structures ([35]).

Direct determination of ligand-protein K_D using uSTA

The precision of signal determination in uSTA critically allowed variation of ligand/protein concentrations even down to low levels (see above), enabling direct determination of binding constants in a manner not possible by classical methods. Following measurement of magnetization transfer between ligand and protein, variation with concentration (Fig. 1E) was quantitatively analyzed using modified Bloch-McConnell equations ([36]) (see Methods). These accounted for intrinsic relaxation, cross-relaxation, and protein-ligand binding (Fig. 1A) to directly provide measurements of equilibrium binding K_D and associated kinetics (k_{ex}). In the Trp/BSA system, this readily revealed $K_D = 38 \pm 15$ μM, $k_{on} = 1.6$ (± 0.6) $\times 10^5$ M^{-1} s^{-1}, and $k_{off} = 6.0 \pm 2.0$ s^{-1} (Fig. 1G), consistent with prior determinations of K_D by other solution-phase methods [$K_D = 30 \pm 9$ μM by isothermal calorimetry ([32])]. Note that this direct method proved to be possible only because of the ability of the uSTA method to deconvolute a true signal with sufficient precision, even at the lower concentrations used and consequently lower signal (Fig. 1E). Thus, uSTA enabled atom-mapping and quantitation for ligand binding that were improved over previous methods. Critically, these values were fully consistent with all observed NMR data and independently obtained measures of K_D.

uSTA allows interrogation of designed crypticity in influenza HA virus attachment

Having validated the uSTA methodology, we next used it to interrogate sugar binding by viral attachment protein systems that have proved typically intractable to classical methods. The hemagglutinin (HA) trimer of influenza
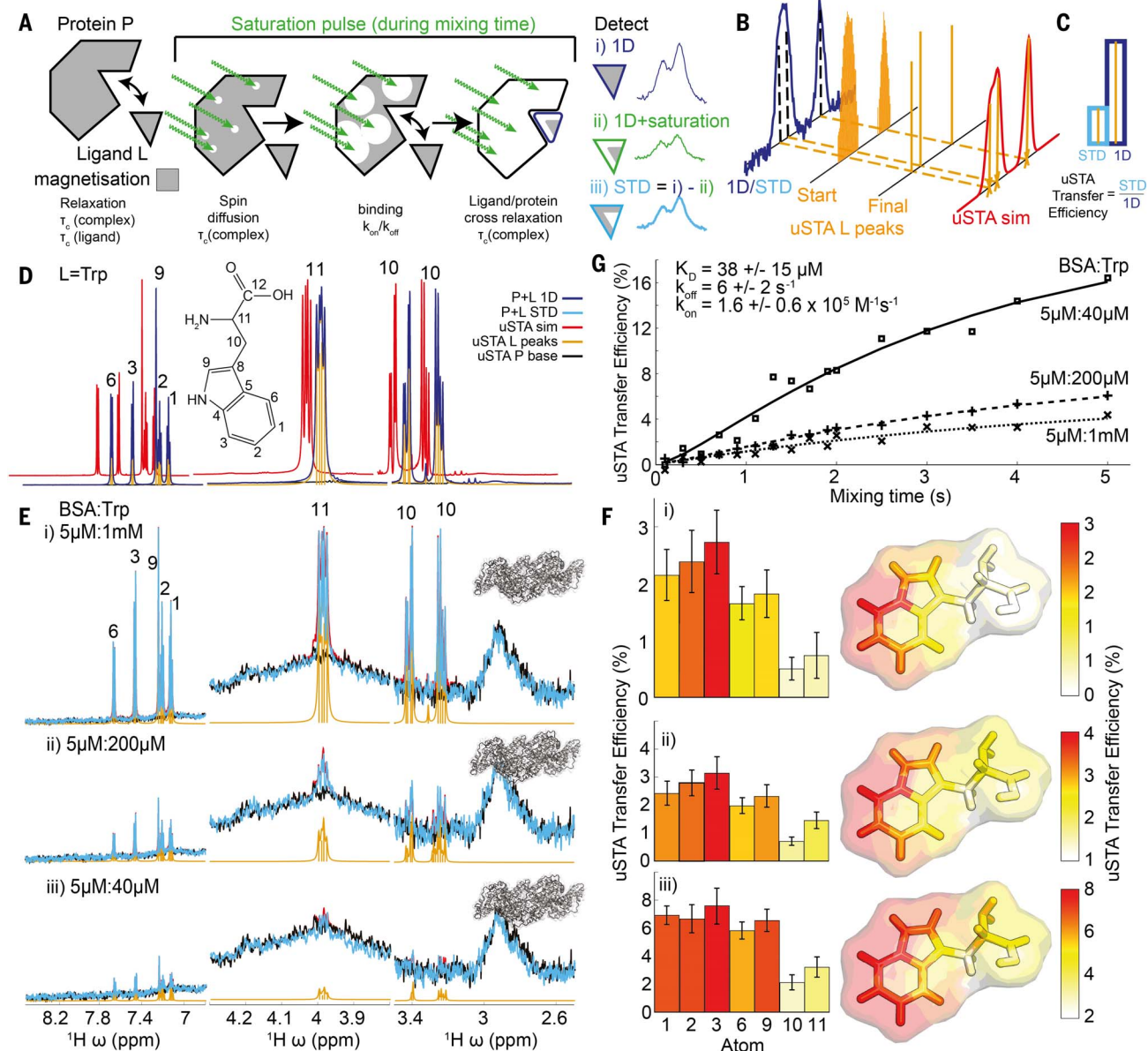
Fig. 1. Development of the uSTA method. (A to C) Schematic of the process for uSTA that exploits comprehensive numerical analysis of relaxation and ligand binding kinetics (A) using full and automatically quantified signal intensities in NMR spectra (B) and calculates per-resonance transfer efficiencies (C). In (B), signal analysis determined the number of peaks that can give rise to the signal, and returned simulated spectrum by convolving these with peak shape function. The precise peak positions returned are in excellent agreement with the known positions of resonances identified using conventional means. Magnetization transfer NMR experiments compare two 1D NMR spectra, where the second involves a specific saturation pulse that aims to "hit" the protein but "miss" the ligand in its excitation. This is accomplished by acquiring the 1D spectrum with the saturation pulse held off resonance at −35 ppm such that it will not excite protons in either ligand or protein [labeled "1D" in (B) and (C)]. The uSTA method requires these two spectra to be analyzed as described in (B), 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. (**D** to **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, yet the ligand (L) and protein (P) can still be deconvolved by 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 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. Although signal-to-noise in the STD increases considerably with increasing ligand concentration, the measured atom-specific transfer efficiencies as determined by uSTA are consistent [(F), left, bar charts; right, transfer efficiency binding "maps"], showing that the primary contact between protein and ligand occurs on the distal edge (C-1, 2, 3, 4; N-7 and C-9 using the numbering shown) of the indole aromatic ring. Application of the same uSTA workflow also allowed precise determination of even weakly binding sugar ligand trehalose (Glc-α1, 1α-Glc) to *E. coli* trehalose repressor TreR. Again, uSTA allows determination of transfer efficiencies with atom-specific precision (see fig. S3F). (**G**) Quantitative analysis of the STD build-up curves using a modified set of Bloch-McConnell equations that account for binding and cross-relaxation allows us to determine thermodynamic and kinetic parameters that describe the BSA-Trp interaction, K_D, k_{on}, and k_{off}. The values obtained are indicated and are in excellent accord with those obtained by other methods (*25, 32*). Errors come from a bootstrapping procedure (see Methods).

Fig. 2. uSTA allows mapping of a designed cryptic sugar-binding site in H1N1 influenza hemagglutinin (HA). (A) HA presents on the surface of the viral membrane and has been shown to bind with sialic acid surface glycans to mediate host cell entry. A designed (*38*) ΔRBS variant is generated by the creation of an N-linked glycosylation site via the creation of needed sequon NQT from wild-type NQR by the R205T point mutation in HA adjacent to the sialic acid–binding site. In this way, disruption of sialic acid binding through designed "blocking" in HA-ΔRBS is intended to ablate the binding of HA wild-type to sialosides in this synthetic variant. **(B and C)** Notably, wild-type and ΔRBS variants of H1N1 HA in fact show a remarkably similar overall binding pattern (B) for the 2,3-sialo-trisaccharide **2** (focused through engagement with the sialoside) but a significant intensity moderation (C) for the ΔRBS variant, indicative of a partial (but not complete) loss of binding consistent with design (*38*). **(D)** Raw spectral data demonstrate that atom-specific differences in intensity in the 1D versus the difference spectra can be discerned using uSTA. Note that atom numbering shown and used here is generated automatically by uSTA.

A virus is known to be essential for its exploitation of sialoside binding (*37*); H1N1 has emerged as one of the most threatening variants in recent years. We took the H1 HA in both native form and a modified form, containing a non-natural sequon specific for N-glycosylation that was previously designed (*38*) to block intermolecular (in trans) sugar binding. This designed blocking in a so-called HA-ΔRBS variant (*38*) also notably creates an additional glycan beyond the existing, potentially confounding, glycosylation background. It therefore provided another test of uSTA's ability to delineate relevant sialoside ligand interactions in another important pathogen protein (Fig. 2A). Despite this intended blocking, the precision of uSTA was such that residual in trans binding of sialotrisaccharide **2** could still be detected in HA-ΔRBS, albeit at a lower, modulated level [as expected by design (*38*)]. Although H1 HA is known to bind both sialo-trisaccharides **2** and **3**, **2** is the less preferred (2,6- over 2,3-linked) ligand (*39*), and yet its binding could still be mapped (here, to a mode mediated primarily by the sialoside moiety). In this way, the sensitivity of uSTA to detect even lower sialoside binding to relevant proteins was confirmed.

uSTA reveals natural, cryptic sialoside binding by SARS-CoV-2 spike

We next probed putative, naturally cryptic sialoside binding sites in SARS-CoV-2. Our analysis of the 1D protein ¹H-NMR spectrum of the purified prefusion-stabilized ectodomain construct (*10*) of intact trimeric SARS-CoV-2 spike attachment protein (fig. S7) revealed extensive protein glycosylation with sufficient mobility to generate a strong ¹H-NMR resonance in the region 3.4 to 4.0 ppm (Fig. 3A). Although 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 expression in human cells (*40*). As such, these mobile glycans on SARS-CoV-2 spike contain sialoside glycan residues that not only confound analyses by classical NMR methods but are also potential competing, "internal" (in cis) ligands for any putative attachment (in trans) interactions, as well as possible direct ligands for in trans interactions in their own right (*41*). Therefore, their presence in the protein NMR analysis presented clear confounding issues for typical classical STD analyses. As such, SARS-CoV-2 spike represented a stringent and important test of the uSTA method.

We used uSTA to evaluate a representative panel of both natural and site-specifically modified unnatural sialosides as possible ligands of spike (Fig. 3 and figs. S3 and S12). Use of classical methods provided an ambiguous assessment (fig. S8), but use of uSTA immediately revealed binding and nonbinding sugar ligands (Fig. 3, figs. S8, S10, and S11, and table S7). Initially, the simplest sialoside, *N*-acetyl-neuraminic acid (**1**), was tested as a mixture of its mutarotating anomers (**1α ⇌ 1β**). When analyzed using uSTA, these revealed [Fig. 3, E and F(i, ii), table S7, and fig. S11] clear "end-on" interactions by **1α** as a ligand [Fig. 3F(i)], mediated primarily by the acetamide NHAc-5, but no reliably measurable interactions by **1β** [Fig. 3F(ii)]. This detection of selective α-anomer interaction, despite the much greater dominance of the β-anomer in solution, provided yet another demonstration of the power of the uSTA method, here operating in the background of dominant alternative sugar (Fig. 3E and fig. S11). The α selectivity correlates with the near-exclusive occurrence of sialosides on host cell surfaces as their α- but not β-linked conjugates (see also below).

Having confirmed simple, selective monosaccharide α-sialoside binding, we explored extended α-sialoside oligosaccharide ligands (compounds **2** and **3**; Fig. 3, C and D) that would give further insight into the binding of natural endogenous human cell-surface sugars as well as unnatural variants (compounds **4** to **6**; Fig. 3, F and G) that could potentially interrupt such binding. Sialosides are often found appended to galactosyl (Gal/GalNAc) residues in either α2,3-linked (**2**) or α2,6-linked form (**3**). Both were tested (Fig. 3, C and D) and exhibited "end-on" binding consistent with that seen for *N*-acetyl-neuraminic acid (**1**) alone, but with more extended binding surfaces (Fig. 3D), qualitatively suggesting a stronger binding affinity (see below for quantitative analysis). Common features of all sialoside binding modes were observed: The NHAc-5

Fig. 3. uSTA reveals interaction of sialosides with SARS-CoV-2 spike protein. A panel of natural, unnatural, and hybrid variant sialoside sugars **1–6** (see fig. S12) was used to probe interaction between sialosides and spike. (**A**) The 1D ¹H-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 (fig. S8), the ligand cannot be effectively studied. (**B**) Application of the uSTA workflow (fig. S6) to SARS-CoV-2 spike protein (shown in detail for **2**). The uSTA process of ligand peak assignment and deconvolution → P + L peak assignment and deconvolution → application to P + L STD yields precise atom-specific transfer efficiencies (fig. 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). 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 pinpoints ligand peaks and signal intensities. Spectral atom numbering shown and used here is generated automatically by uSTA; all other numbering in sugars follows carbohydrate nomenclature convention. (**C** and **D**) Application of the uSTA workflow (fig. S6) reveals atom-specific binding modes to spike protein for both natural (e.g., sialoside) trisaccharides Siaα2,3Galβ1,4Glc (**2**) and Siaα2,6Galβ1,4Glc (**3**). Comparison of the uSTA method focused on the

NHAc methyl resonance shows excellent agreement (C). The uSTA method allowed determination of binding surfaces for both trisaccharides **2** [D(i)] and **3** [D(ii)]. (**E** and **F**) STD spectrum (E) and mapped atom-specific transfer efficiencies (F) for sialic acid (**1**) and 9-N₃ azido variant **4**. Both interconverting α and β anomeric forms could be readily identified. Despite the dominance of the β form [94 %, E(i)], application of the uSTA method following assignment of resonances from the two forms allowed determination of binding surfaces simultaneously [E(ii), F(i, ii)]. Spike shows strong binding preference for the α anomers [F(i, iii) versus F(ii, iv)] despite this strong population difference. Binding surfaces were also highly similar to those of extended trisaccharides **2** and **3**. (**G**) Using these intensities, atom-specific transfer efficiencies can be determined with high precision, shown here for hybrid sialoside **5**. The details of both the unnatural BPC moiety and the natural sialic acid moiety can be mapped; although the unnatural aromatic BPC dominates interaction, uSTA nonetheless delineates the subtleties of the associated contributions from the natural sugar moiety in this ligand (see also figs. S5 and S6).
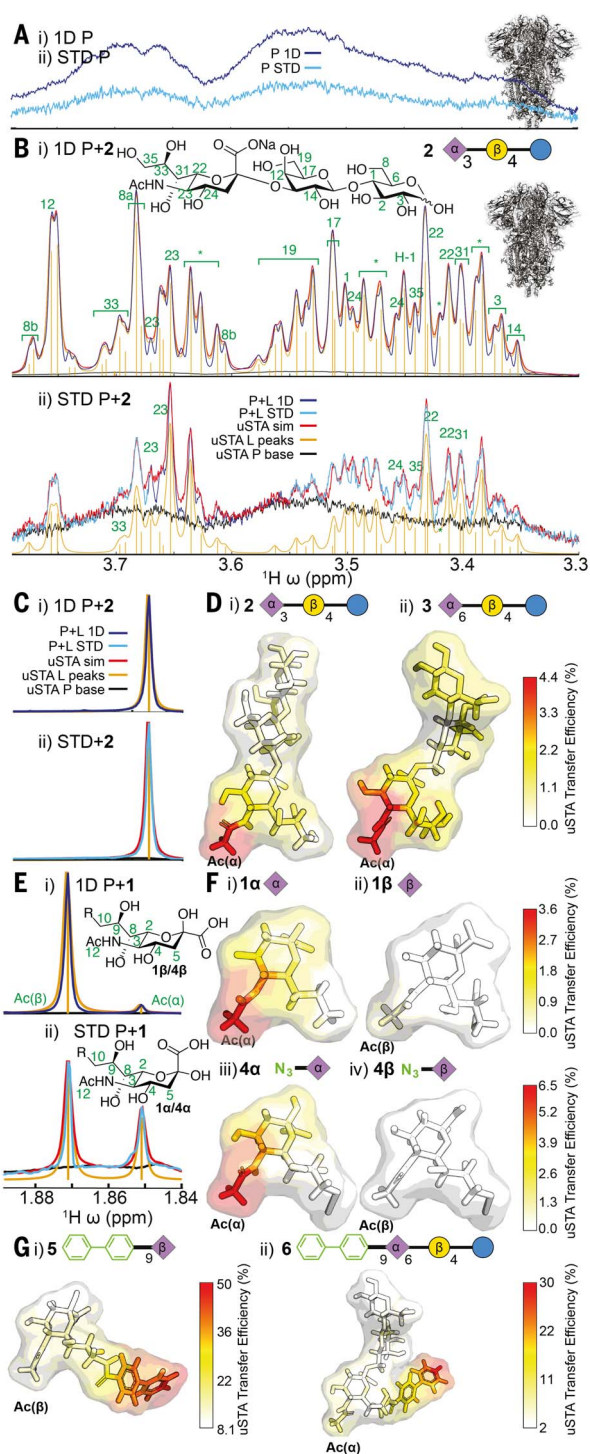
acetamide of the terminal sialic acid (Sia) is a binding hotspot in **1**, **2**, and **3** that drives the "end-on" binding. Differences were also observed: The α2,6-trisaccharide (**3**) displayed a more extended binding face yet with less intense binding hotspots (Fig. 3D) engaging additionally the side-chain glycerol moiety (C7–C9) of the terminal Sia acid as well as the OH-4 C4 hydroxyl of the Gal residue. The interaction with α2,3-trisaccharide (**2**) was tighter and more specific to NHAc-5 of the Sia.

These interactions of the glycerol C7–C9 side chain detected by uSTA were probed further through construction (fig. S12) of unnatural modified variants (**4** to **6**; Fig. 3, F and G). Replacement of the OH-9 hydroxyl group of sialic acid with azide N₃-9 in **4** [Fig. 3F(iii) and table S7] was well tolerated, but larger changes (replacement with aromatic group biphenylcarboxamide BPC-9) in **5** and **6** led to an apparently abrupt shift in binding mode that was instead dominated by the unnatural hydrophobic aromatic modification (Fig. 3G and Fig. 4C). As for native sugar **1**, azide-modified sugar **4** also interacted with spike in a stereochemically specific manner with only the α-anomer displaying interaction [Fig. 3F(iii)], despite dominance of the β-anomer in solution [Fig. 3F(iv)]. uSTA allowed precise dissection of interaction contributions in these unusual hybrid (natural-unnatural) sugar ligands that could not have been determined using classical methods (see text S5).

Using variable concentrations of the most potent natural ligand α2,3-trisaccharide **2** [6 μM spike, **2** at 60 μM, 200 μM, 1 mM, and 2 mM excitation at 5.3 ppm] and variable concentrations of spike protein, we used the uSTA method to directly determine solution-phase affinities (Fig. 4A): $K_D = 32 \pm 12$ μM, $k_{on} = 6300 \pm 2300$ M⁻¹ s⁻¹, and $k_{off} = 0.20 \pm 0.08$ s⁻¹. We also probed binding in a different mode by measuring the affinity of spike to **2** when displayed on a modified surface (fig. S13) using surface plasmon resonance (SPR) analysis. The latter generated a corresponding $K_D = 23.7 \pm 3.6$ μM ($k_{on} = 1004 \pm 290$ M⁻¹ s⁻¹). Such similar values for sialoside ligand in solution (by uSTA) or when displayed at a solid-solution interface (by SPR) suggested no substantial avidity gain from display of multiple sugars on a surface.

Structural insights from uSTA delineate binding to SARS-CoV-2 spike

uSTA analyses consistently identified binding hotspots in sugars **1** to **4** providing the highest transfer efficiencies in an atom-specific manner, particular the "end" NHAc-5 acetamide methyl group of the tip sialic acid residue in all. A combination of uSTA with so-called high ambiguity–driven docking (HADDOCK) methods (*42*, *43*) was then used to probe likely regions in SARS-CoV-2 spike for this "end-on" binding mode via uSTA data-driven atomistic
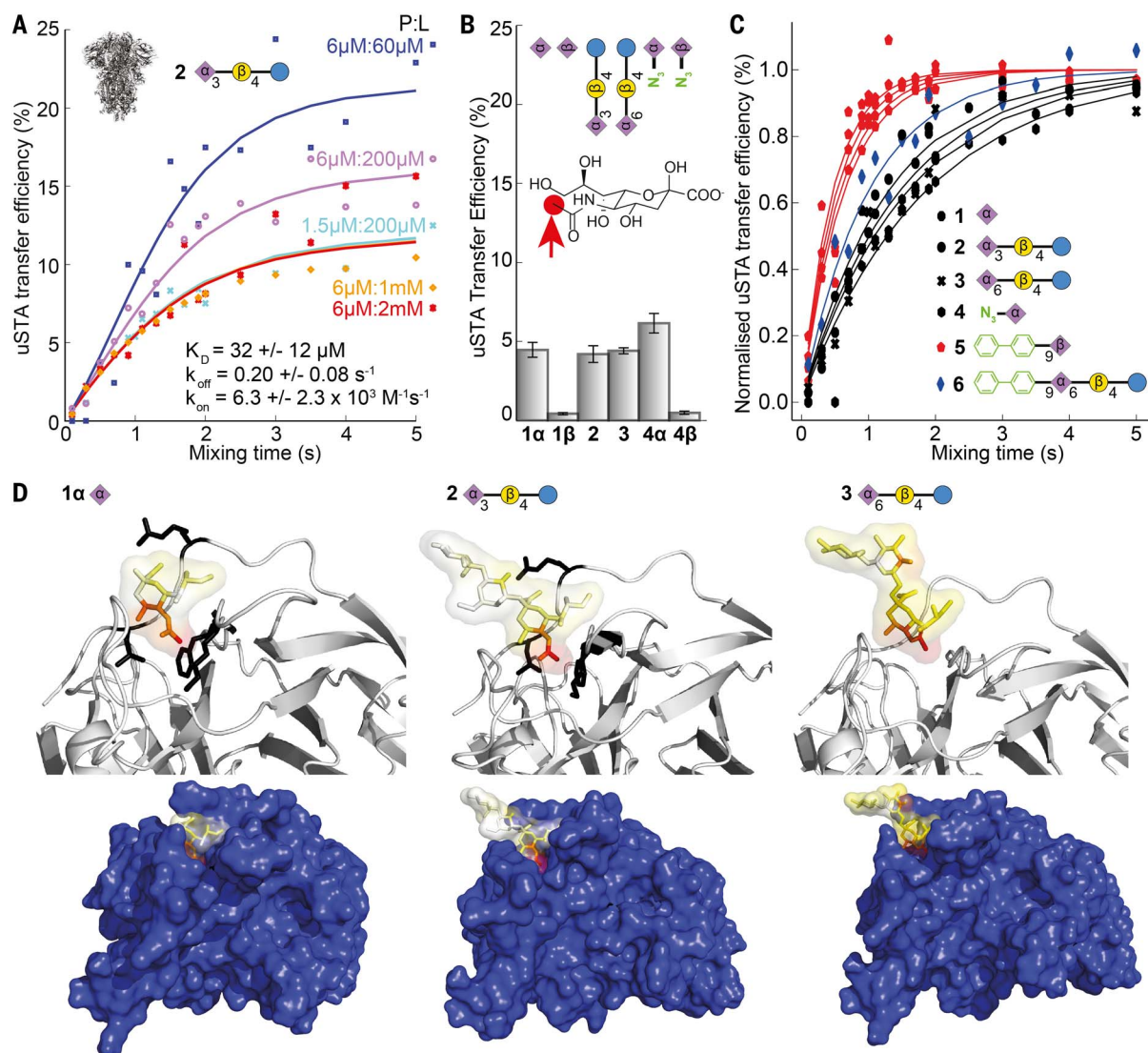
Fig. 4. Quantitative uSTA analyses allow comparison and predictive restraints for protein ligand-binding prediction. (**A**) Quantitative analysis of the STD build-up curves using a modified set of Bloch-McConnell equations that account for binding and cross-relaxation allow us to determine thermodynamic and kinetic parameters that describe the SARS-CoV-2•**2** interaction, K_D, k_{on}, and k_{off}. The values obtained are indicated. Errors come from a bootstrapping procedure (see Methods). The lower ligand concentrations yield data that are of lower sensitivity than higher concentrations. The transfer efficiencies will be higher in this case, as more molecules are effectively involved in the binding. Thus, data at lower concentrations will in general have more scatter and higher transfer efficiencies. These data points are desirable for the analysis, as it is here where we expected the greatest variation of transfer efficiency with ligand concentration. The analysis is applied globally and so the uncertainties in the final fitted parameters from the bootstrapping analysis (see Methods) provide a direct and confident measure of the goodness of fit. (**B**) Normalized uSTA transfer efficiencies of the NAc-5 methyl protons can be determined for each ligand studied here. This allowed relative contributions to "end on" binding to be assessed via uSTA in a "mode-specific" manner. This confirmed strong α- over β-sialoside selectivity. Errors were determined through a bootstrapping procedure where mixing times were sampled with replacement, allowing for the construction of histograms of values in the various parameters that robustly reflect their fitting errors. (**C**) Normalized build-up curves for the most intense resonances allowed two distinct modes of binding

to be identified in natural (**1–4**) and hybrid (**5**, **6**) sugars. Data are shown at constant protein and ligand concentrations. With the BPC moiety present, the build-up of magnetization occurs significantly faster than when not; various such hybrid ligands give highly similar curves. By contrast, natural ligands have a much slower build-up of magnetization. This, together with the absolute transfer efficiencies being very different, and the overall pattern on the interaction map combine to reveal that the ligands are most likely binding via two different modes and possibly locations on the protein. (**D**) Coupling uSTA with an integrative modeling approach such as HADDOCK (*42*, *43*) allowed 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 binding "map" onto modeled poses. uSTA mapping the interaction between SARS-CoV-2 spike [based on RCSB 7c2l (*19*)] 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 describe the interaction between ligand and protein (fig. 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 a pocket in the spike NTD. Although sequence and structural homology are low (fig. S1), MERS spike protein possesses a corresponding NHAc-binding pocket characterized by an aromatic (Phe[39])–hydrogen-bonding (Asp[36])–hydrophobic (Ile[132]) triad (*5*).

models. In each case, a cluster of likely poses emerged (Fig. 4D) for **1α**, **2**, and **3** (see fig. S14 for details) consistent with "end-on" binding where the acetamide NHAc-5 methyl group of the sialic acid moiety was held by the unusual β sheet–rich region of the NTD of SARS-CoV-2 spike. Under the restraints of uSTA and homology, a glycan-binding pocket was delineated by a triad of residues (Phe[79], Thr[259], and Leu[18]) mediating aromatic, carbonyl hydrogen–bonding, and hydrophobic interactions, respectively. However, the sequence and structural homology to prior (i.e., MERS) coronavirus spike proteins in this predicted region was low; the MERS spike protein uses a corresponding NHAc-binding pocket characterized by an aromatic (Phe[39]), hydrogen-bonding (Asp[36]), and hydrophobic (Ile[132]) triad to bind the modified sugar 9-O-acetyl-sialic acid (5).

SARS-CoV-2 glycan attachment mechanisms have to date only identified a role for spike RBD in binding rather than NTD (15, 17). We used uSTA to compare the relative potency of the sialoside binding identified here to previously identified (17) heparin binding motifs (Fig. 5, A and B). Heparin sugars **7** and **8** of similar size to natural sialosides **3** and **4** were selected so as to allow a near ligand-for-ligand comparison based on similar potential binding surface areas. **7** and **8** also differed from each other only at a single glycan residue (residue 2) site to allow possible dissection of subtle contributions to binding. Unlike the "end-on" binding seen for sialosides **3** and **4**, uSTA revealed an extended, nonlocalized binding interface for **7** and **8** consistent instead with "side-on" binding [Fig. 5, A(ii) and B(ii)].

Next, we examined the possible evolution of sialoside binding over lineages of SARS-CoV-2 (44). Four notable variants of concern—alpha/B1.1.7, beta/B1.351, delta/B.1.617.2, and omicron/B.1.1.529—emerged in later phases of the pandemic. When these corresponding spike protein variants were probed by uSTA, all displayed

ablated binding toward sialoside **2** as compared to first-phase B-origin-lineage spike (Fig. 5D and fig. S15).

Finally, to explore the possible role of sialoside binding in relation to ACE2 binding, we also used uSTA to probe the effects upon binding of the addition of a known, potent neutralizing antibody of ACE2-spike binding, C5 (Fig. 5, D and E, and fig. S16) (45, 46). 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 and near-identical binding maps (Fig. 5, E and F) were consistent with a maintained sialoside-binding pocket with undisrupted topology and mode of binding.

Together, these findings allow us to conclude that the sialoside binding observed with uSTA involves a previously unidentified "end-on" mechanism/mode that operates in addition to and potentially cooperatively with
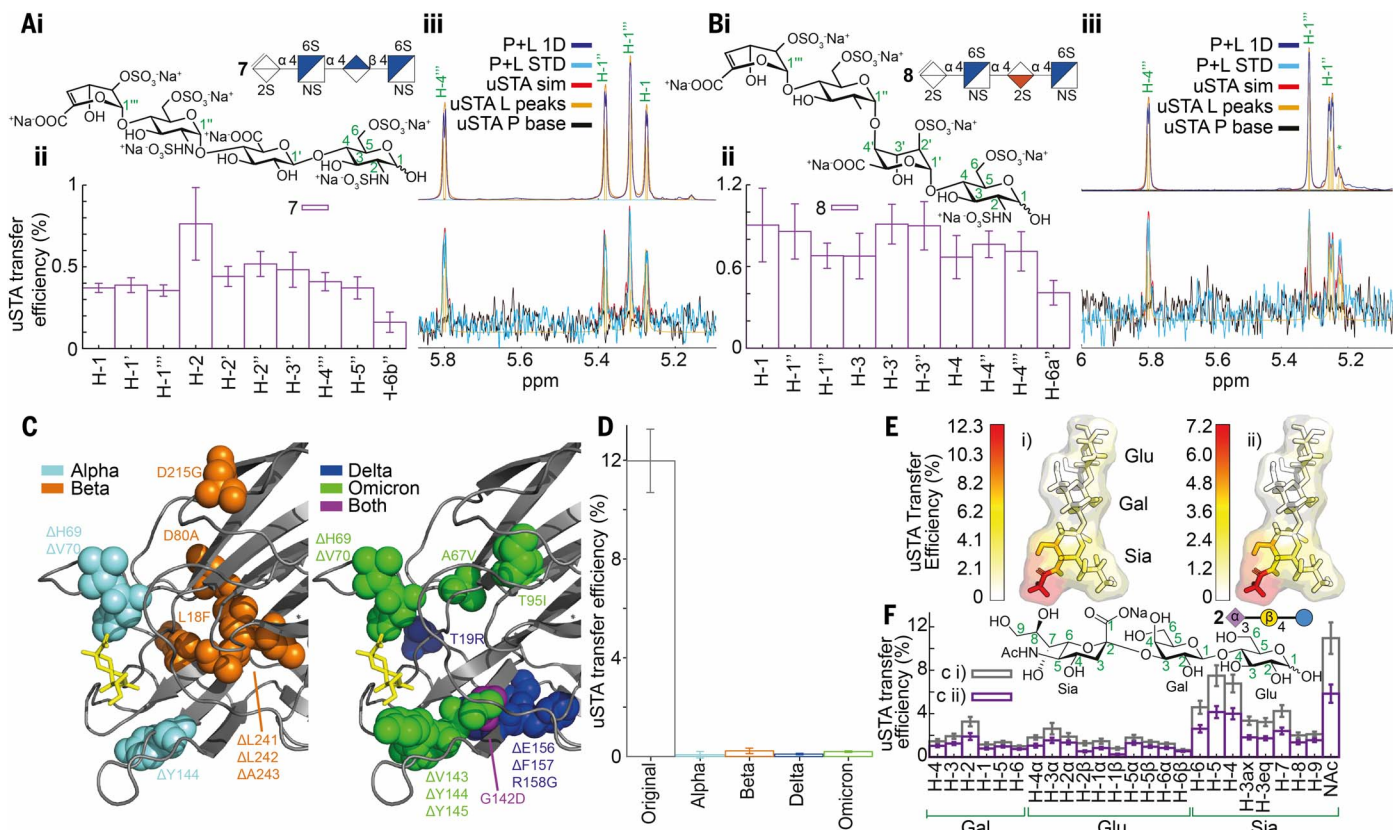
Fig. 5. Comparison of SARS-CoV-2 glycan attachment mechanisms and variant evolution via uSTA suggests binding away from the RBD that is lost. (**A** and **B**) Two heparin tetrasaccharides (**7**, **8**) are shown by uSTA [A(iii), B(iii)] to bind B-origin-lineage SARS-CoV-2 spike protein ("original" spike) in a "side-on" mode [A(ii), B(ii)]. Atom specific binding is shown in A(ii) and B(ii). Assignments shown in green use conventional glycan numbering. (**C**) Substantial numbers of mutations arise in the NTD region identified by uSTA in the B.1.1.7/alpha (cyan), B.1.351/beta (orange), B.1.617.2/delta (dark blue) and B.1.1.529/omicron (green) lineage variants of SARS-CoV-2 spike. (**D**) Ablated binding of the α-2,3-sialo-trisaccharide **2**, as measured by the transfer efficiency of the NHAc protons, is identified by uSTA in the lineage variants of SARS-CoV-2 spike [see (C) for colors]. This proves consistent with mutations that appear in the sialoside-binding site of NTD identified in this study [see (C) and Fig. 6]. (**E** and **F**) uSTA of sialoside **2** with B-origin-lineage SARS-CoV-2 spike in the presence of the potent RBD-neutralizing nanobody C5 [spike E(i), nanobody-plus-spike E(ii)] shows essentially similar binding patterns with uniformly modulated atomic transfer efficiencies (F).

ACE2 binding in SARS-CoV-2. The primary sialoside glycan-binding site SARS-CoV-2 spike is distinct from that of heparin ("end-on" versus "side-on"), not in the RBD (not neutralized by RBD-binding antibody), and found instead in an unusual NTD region that has become altered in emergent variants (loss of binding in alpha and beta variants of concern).

Cryo-EM pinpoints the sialoside-binding site in B-origin-lineage spike

Structural analysis of the possible binding of sialosides has been hampered to date by the moderate resolution, typically less than 3 Å, of most SARS-CoV-2 spike structures. Currently deposited cryo-EM–derived coulombic maps show that the NTD of spike is often the least well-resolved region. In our initial attempts with native protein, large stretches of amino acids within the NTD were not experimentally located ([45]); the most disordered regions occur in the NTD regions that contribute to the surface of the spike. A stabilized closed mutant form ([47]) of the spike was examined and gave improvements, but the coulombic map was still too weak and noisy to permit tracing of many of the loops in the NTD. However, with a reported fatty acid–bound form of the spike ([48]), which has shown prior improved definition of the NTD, we were able to collect a 2.3 Å dataset in the presence of the α2,3-sialo-trisaccharide 9. The map was clear for almost the entire structure including the previously identified linoleic acid (fig. S17D); only 13 N-terminal residues and two loops (residues 618 to 632 and 676 to 689) were not located. Although the density is weaker at the outer surface of the NTD than at the core of the structure (fig. S17B), the map was of sufficient quality to model N-glycosylation at site Asn[149], which is in a flexible region, and the fucosylation state of N-linked glycans at Asn[165] (Fig. 6A).

We observed density in a pocket at the surface of the NTD lined by residues His[69], Tyr[145], Trp[152], Gln[183], Leu[249], and Thr[259] (Fig. 2B and fig. S17). This density is absent in other spike structures of higher than 2.7 Å resolution (PDB IDs 7jji, 7a4n, 7dwy, 6x29, 6zge, 6xlu, 7n8h, 6zb5, and 7lxy), even those (such as PDB IDs 7jji, 6zb5, and 6zge) that have a well-ordered NTD. The density when contoured at 2.6σ is fitted by an α-sialoside consistent with the terminal residue of 9, with the distinctive glycerol and N-acetyl groups clear. To further strengthen our confidence in the identification of the sialic acid, we determined a native (unsoaked) structure to 2.4 Å using the same batch of protein (fig. S17D). This structure showed no density in the sialic acid binding site supporting our assignment. The sialoside was therefore included in the refinement and the thermal factors (108 Å²) were comparable to those for the adjacent protein residues (95 to 108 Å²). In this position, the glycerol moiety

makes a hydrogen bond with the side chain of Tyr[145], the N-acetyl group with Ser[247], and the carboxylate with Gln[183], and there are hydrophobic interactions with Trp[152]. Lowering the map threshold to 1.6σ would be consistent with a second pyranoside (e.g., galactoside) residue (fig. S17C). At this contour level, the map clearly covers the axially configured carboxylate of the sialic acid (fig. S17). The middle galactoside residue of 9 positioned in this density would make contacts with Arg[248] and Leu[249].

Structural superposition of the NTD with that of the MERS spike (RCSB 6NZK) shows that although the sialic acid–binding pockets of both are on the outer surface, these pockets are 12 Å apart (as judged by the C2 atom of respective sialic acids; Fig. 6D). In MERS spike, sialic acid is bound at the edge of the central β sheet, whereas in SARS-CoV-2 the sugar is bound at the center of the sheet; thus, the pockets use different elements of secondary structure. Because of distinct changes in the structure of the loops connecting the strands, the sialic acid pocket from one protein is not present in the other protein. Several regions of additional density were not fitted by the model (fig. S18).

Disclosure of sialoside trisaccharide as a ligand for B-origin-lineage SARS-CoV-2 correlates with clinical genetic variation in early-phase pandemic

A distinctive mode of sialoside binding by spike confirms a potential attachment point for SARS-CoV-2 found commonly on cell surfaces (sialosides are attached both as glycolipid and glycoprotein glycoconjugates), thus raising the question of whether glycosylation function in humans affects infection by SARS-CoV-2 and hence the presentation and pathology of COVID-19 disease. Analysis of whole-exome sequencing data of an early 2020 cohort of 533 COVID-19–positive patients (see table S1) identified two glycan-associated genes within the top five that were most influential upon disease severity. Specifically, recursive feature elimination applied to a LASSO (least absolute shrinkage and selection operator)–based ([49]) logistic regression model identified *LGALS3BP* (fourth of >18,000 analyzed genes) and *B3GNT8* (fifth of >18,000) (Fig. 7A and fig. S19). Variants in these two genes were beneficially associated with less severe disease outcome (Fig. 7, B and C; see also tables S2 to S6 for specific *B3GNT8* and *LGALS3BP* genetic variants, *B3GNT8* χ^2 five categories, *B3GNT8* χ^2 2×2, *LGALS3BP* χ^2 five categories, *LGALS3BP* χ^2 2×2, respectively).

LGALS3BP encodes for a secreted protein, galectin-3–binding protein (Gal-3-BP, also known as Mac-2-BP), that is a partner and blocker of a specific member (Gal-3) of the galectin class of carbohydrate-binding proteins ([50]). Galectins are soluble and are typically secreted and implicated in a wide range of cellular functions ([51]). Notably, Gal-3 binds the

so-called poly-*N*-acetyl-lactosamine [polyLacNAc or (Gal-GlcNAc)$_n$] chain-extension variants found in tetraantennary N-linked glycoproteins (Fig. 7D), including those displaying sialyl-Gal-GlcNAc sialoside motifs ([52], [53]). Variants in *LGALS3BP* were present in 9 of 114 a/pauci-symptomatic subjects or mildly affected patients (~8%) compared to 8 of the remaining 419 patients who required more intensive care: oxygen support, CPAP/BiPAP, or intubation (<2%); none of the 69 most seriously affected patients (intubated) carried variants of *LGALS3BP* (Fig. 7B). Intact *LGALS3BP* gene product Gal-3-BP therefore appears correlated with more severe COVID-19 outcome. The other implicated gene, *B3GNT8*, encodes a protein glycosyltransferase, β-1,3-*N*-acetyl-glucosaminyltransferase-8 (GlcNAcT8 or β3GnT8), that is responsible for the creation of the anchor point of poly-*N*-acetyl-lactosamine (polyLacNAc) in such tetraantennary N-linked glycoproteins (Fig. 7D) ([54]). Again, rare variants in *B3GNT8* were present in 11 of 114 a/pauci-symptomatic subjects or mildly affected patients (~10%) compared to 10 of the remaining 419 patients who required more intensive care (~2%) (Fig. 7C).

Discussion

Experimentally, there are still few, if any, orthogonal approaches to the useful surface display methods (e.g., "glycoarrays") currently used for readily surveying ligands that might be exploited by pathogens. Following validation in model ligand-protein systems, uSTA provided a ready method for identifying sugar ligands bound by pathogens, as well as their binding parameters and poses, even in posttranslationally modified (e.g., glycosylated) protein systems.

In an influenza virus HA protein variant designed to abolish binding through competition by an added glycan site on HA ([55]), uSTA was nonetheless able to unambiguously reveal and "map" residual sialoside binding despite the presence of an added protein-linked glycan as "internal blocker." This is a protein type that has been well-studied in array formats ([39]); we showed here that, even with a "blocked" HA (in a glycosylated state) and a non-preferred 2,3-sialoside ligand, binding could still be mapped by uSTA.

In the spike trimer of the B-origin lineage of SARS-CoV-2, despite the presence of mobile, protein-linked glycans, uSTA clearly revealed sialoside binding and, through mapping, revealed that this binding is more potent when the sialosyl moiety terminates galactosyl oligosaccharides. This pose is in agreement with our cryo-EM structures, which show that the NHAc-5 N-acetyl group at the sugar's tip is buried, a mode of binding we refer to as "end-on." Prior modeling was partly misled by use of lower-resolution structures of spike, because the NTD is highly disordered in these

initial structures. uSTA and cryo-EM also identified a second, differing mode of binding in SARS-CoV-2 spike only by hybrid, aromatic sugars (e.g., **5**, **6**, **9**) driven by aromatic engagement, but the physiological relevance of this binding pocket is currently unclear. We cannot exclude the presence of another sialic acid–binding site (*15*), but there is no structural support for it.

The spike sialoside-binding site in the NTD is also coincident or near to numerous mutational and deletion hotspots in, for example, alpha and omicron (His[69], Val[70], Tyr[145]) and beta (β strand Leu[241]-Leu[242]-Ala[243]-Leu[244]) variants of concern (Fig. 5C). Changes remove either key interacting residues (alpha/omicron variant: residues 69, 70, 145) or perturb structurally important residues (beta variant: β strand) that form the pocket. The "end-on" binding we observed is quite different to the "side-on" binding observed for heparin (*17*). Heparins are often bound in a non–sequence-specific, charge-mediated manner, consistent with such a "side-on" mode. The location of three binding sites in the trimer, essentially at the extreme edges of the spike (Fig. 6C), also imposes substantial geometric constraints for avidity enhancement through multivalency (*56*, *57*).

We also find a clear link between our data and genetic analyses of patients that correlate with the severity of their disease. This association suggests potential roles in infection and disease progression for cell-surface glycans and the two glycan-associated genes that we have identified. Despite their independent identification here, both gene products interact around a common glycan motif: the polyLacNAc chain-extension variants found in tetraantennary N-linked glycoproteins. Consistent with the sialoside ligands found here, these glycoproteins contain Sia-Gal-GlcNAc motifs within N-linked polyLacNAc chains. These motifs have recently been identified in the deeper human lung (*58*).

These data lead us to suggest that B-origin-lineage SARS-CoV-2 virus may have exploited glycan-mediated attachment to host cells (Fig. 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 (*59*) or glycolipid ligands (*15*). This analysis of the influence of genetic variation upon susceptibility to virus was confined to "first wave," early-pandemic patients infected with B-origin-lineage SARS-CoV-2. Our discovery here also that in B-lineage virus such
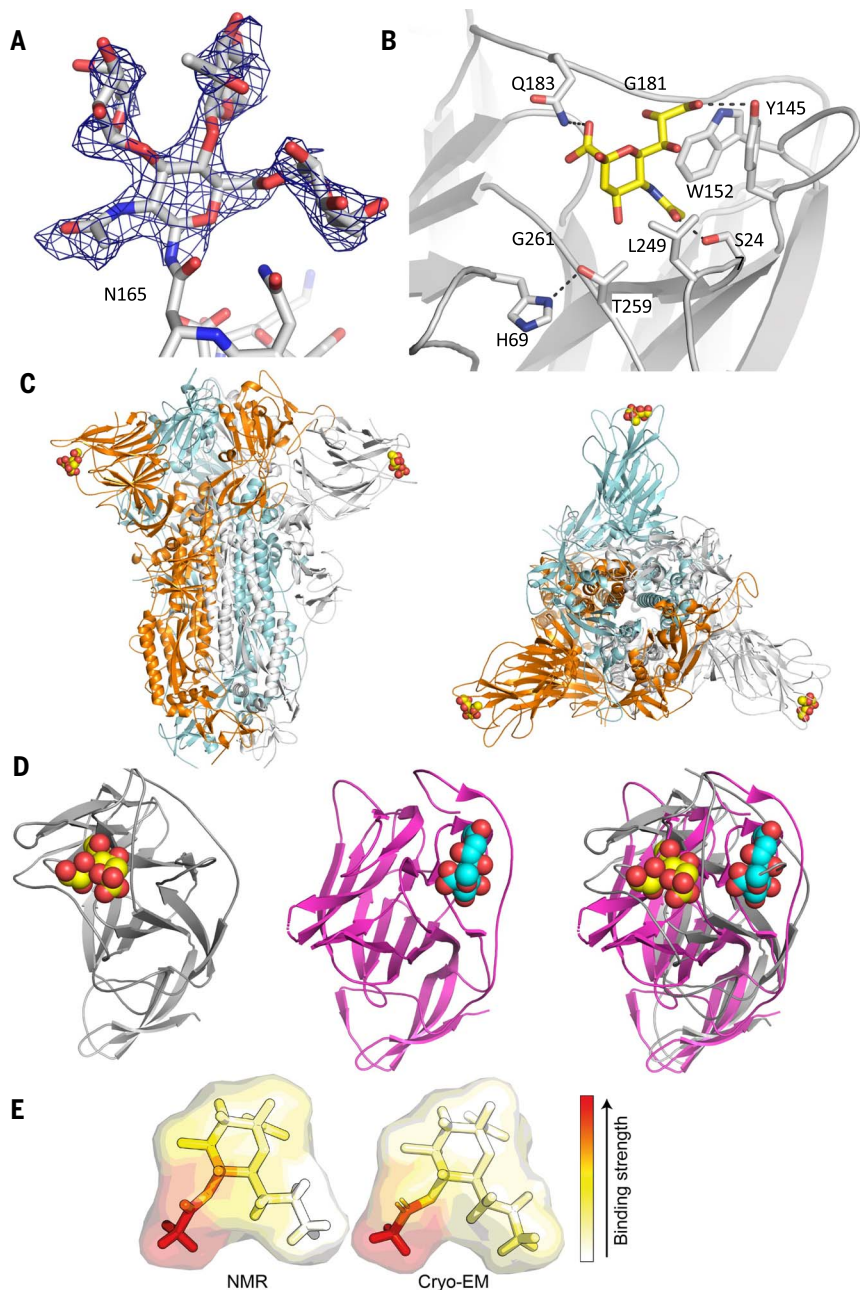
Fig. 6. Cryo-EM analysis of sialoside binding in B-origin-lineage SARS-CoV-2 spike. (**A**) In the presence of sialoside **9** (fig. S18), a well-ordered structure with resolution (2.3 Å) sufficient to identify even glycosylation states of N-linked glycans within spike was obtained. Here they reveal a paucimannosidic base bis-fucosylated chitobiose core structure GlcNAcβ1,4(Fucα1,3-)(Fucα1,6-)GlcNAc–Asn. (**B**) The sialoside-binding site in the NTD is bounded by His[69], Tyr[145], Trp[152], Ser[247], and Gln[183]. Gln[183] controls stereochemical recognition of α-sialosides by engaging COOH-1. Tyr[145] engages the C7-C9 glycerol side chain of sialoside. Ser[247] engages the NHAc-5. Notably, Tyr[145] and His[69] are deleted in the alpha variant of SARS-CoV-2 that loses its ability to bind sialoside (see Fig. 5). See also fig. S17 for coulombic maps. (**C**) Spike's sialoside binding site is found in a distinct region of the NTD (left, from side, right from above) that is at the "edge" of the spike (see text). (**D**) Superposition (right) of SARS-CoV-2-NTD (left, gray) with MERS-NTD (middle, magenta) shows distinct sites 12 Å apart. (**E**) A comparison of the ligand-binding mode measured by uSTA NMR and a map calculated from the cryo-EM structure. For each hydrogen environment in the resolved sialic acid, an array of distances to all protein hydrogens, *r*, was calculated. The interaction of each ligand environment with the spike protein was defined as the summation of all $1/r^6$ environment-protein distances. Values were interpolated in a $1/r^6$ manner onto heteroatoms following the same procedure according to the described NMR methods, and the color bar was scaled from zero to the maximum value.
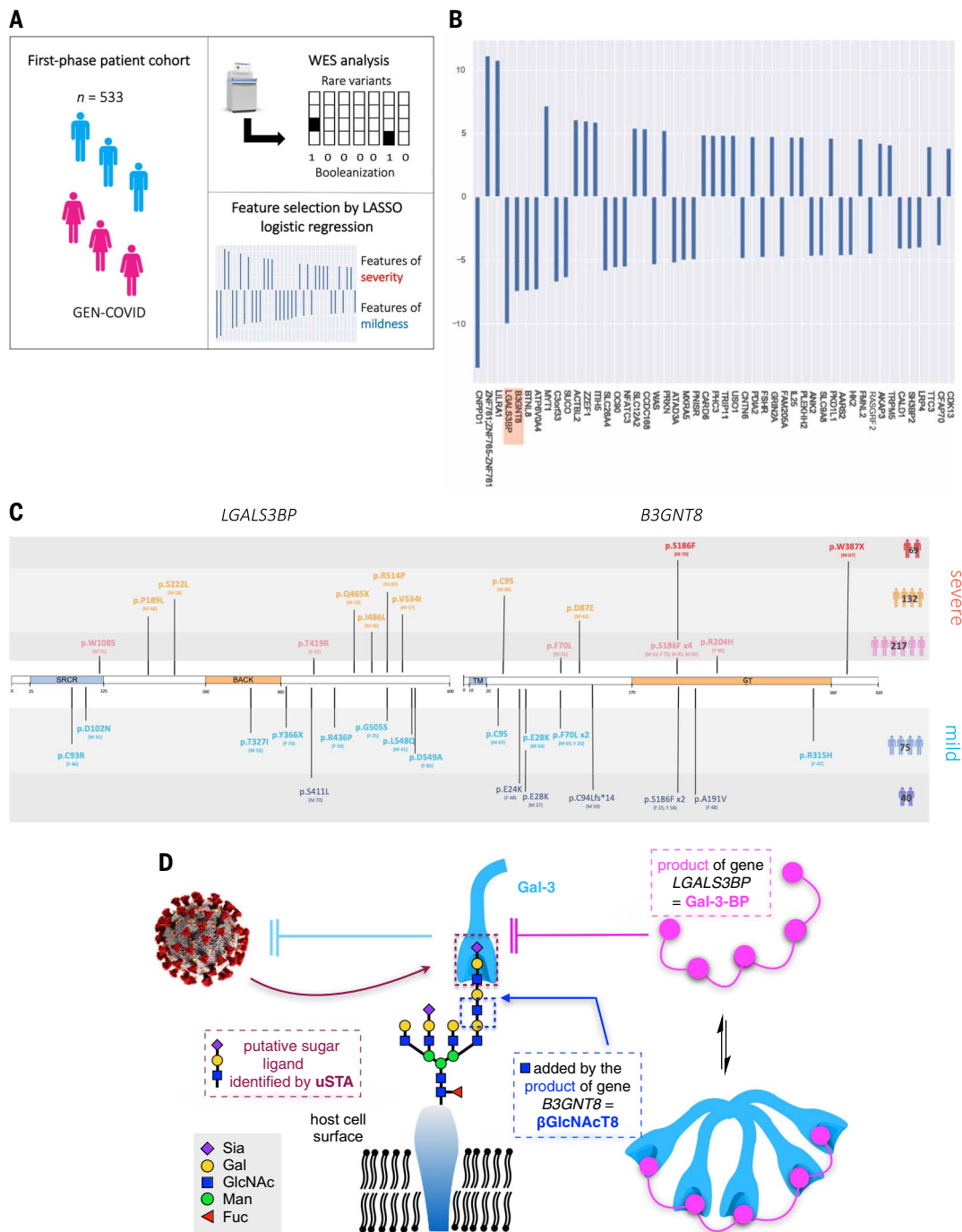
Fig. 7. Analyses of early 2020 first-phase SARS-CoV-2 PCR-positive patients reveals glycan-associated genes suggesting a model of glycan interaction consistent with uSTA observations of sialoside binding in B-origin-lineage SARS-CoV-2. (**A**) GEN-COVID workflow. Left: The GEN-COVID Multicenter Study cohort, of 533 SARS-CoV-2 PCR-positive subjects of different severity from phase one of the pandemic, was used for rare variant identification. Upper right: Whole-exome sequencing (WES) data were analyzed and binarized into 0 or 1 depending on the presence (1) or the absence (0) of variants in each gene. Lower right: LASSO logistic regression feature selection using a Boolean representation of WES data leads to the identification of final sets of features divided according to severity or mildness of disease, contributing to COVID-19 variability. See also (79, 80) for further details of background methodology. (**B**) 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 COVID-19 disease, whereas negative weights confer protective action through variance. Variation in glycan-associated genes B3GNT8 and LGALS3BP score second and third out of all (>18,000) genes as the most protective, respectively (highlighted red). (**C**) Distribution of rare variants in B3GNT8 and LGALS3BP. Left: Rare beneficial mutations distributed along the Gal-3-BP protein product of LGALS3BP, divided into the SRCR (scavenger receptor cysteine-rich) domain (light blue) and the BACK domain (light orange). Right: Rare beneficial mutations distributed 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 colors of the mutation bands (top to bottom) refer to the severity grading of the PCR-positive 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/pauci-symptomatic). (**D**) A proposed coherent model consistent with observation of implicated *B3GNT8* and *LGALS3BP* genes and the identification of sialosides as ligands for spike by uSTA and cryo-EM. 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.

binding to certain sialosides was ablated in later phases of the pandemic (after September 2020) in variants of concern further highlights the dynamic role that sugar binding may play in virus evolution and may be linked, as has been previously suggested for H5N1 influenza A virus, to the "switching" of sugar-binding preferences by pathogens during or after zoonotic transitions (*1*). The focused "end-on" binding of the N-acetyl group in the N-acetyl-sialosides, which are found as the biosynthetically exclusive form of sialosides in humans (*60*), might have been a contributing factor in driving zoonosis.

Finally, our data also raise the question of why binding might be ablated in later variants of SARS-CoV-2. Again by comparison with influenza, which uses neuraminidases for the purpose of "release" when budding from a host cell (*61*), we speculate that in the absence of its own encoded neuraminidase, SARS-CoV-2 must walk a tight balance between the ability to bind human host glycans (potentially useful in a zoonotic leap) and cell-to-cell transmission (where release could become rate-limiting). One answer to this problem would be to ablate N-glycan binding via the sialoside motif subsequent to a successful zoonotic leap. This solution also has the advantage of removing a potential site for antibody neutralization for an interaction that might prove pivotal or critical in the context of zoonosis as a potentially global driver of virus fitness. Our combined data and models may therefore support decades-old hypotheses (*20*) proposing the benefit of cryptic sugar binding by pathogens that may be "switched on and off" to drive fitness in a different manner (e.g., in virulence or zoonosis) as needed.

Methods

Protein expression and purification: SARS-CoV-2 spike

The templates for wild type, alpha, and beta spike were kindly provided by P. Supasa and G. Screaton (University of Oxford). The gene encoding amino acids 1 to 1208 of the wild type, alpha, and beta 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 trimerization domain] was cloned into the pOPINTTGneo-BAP vector using the forward primer (5′-GTCCAAG-TTTATACTGAATTCCTCAAGCAGGCCACCAT-

GTTCGTGTTCCTGGTGCTG-3′) and the reverse primer (5′-GTCATTCAGCAAGCTTAAAAAGG-TAGAAAGTAATAC-3′), resulting in an aviTag/Bap sequence plus 6His in the 3′ terminus of the construct. The template for B-lineage-origin (wild type) spike is previously described (*62*). The templates for the alpha and beta spike are in the supplementary materials.

For B-lineage-origin, alpha, and beta spike, 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 1.5×10^6 to 2×10^6 cells per ml. Following transfection, cells were cultured at 37°C and 5% CO_2 at 120 rpm for 17 hours. Enhancers (6 mM valproic acid, 6.5 mM sodium propionate, 50 mM glucose, all from Sigma) were then added and protein was expressed at 30°C for 5 days before purification.

For delta and omicron spike, cDNA was synthesized (IDT) as gBlock, flanked by KpnI and XhoI restriction sites based on the HexaPro spike sequence (*63*). HexaPro delta spike was made based on the B-lineage-origin HexaPro spike with these additional mutations: T19R, G142D, E156G, del157/158, L452R, T478K, D614G, P681R, D950N. HexaPro Omicron BA.1 spike is made based on the original Wuhan HexaPro spike with these additional mutations: A67V, HV69-70 deletion, T95I, G142D, VYY143-145 deletion, N211 deletion, L212I, ins214EPE, G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y, N679K, P681H, N764K, D796Y, F817P, N856K, A892P, A899P, A942P, Q954H, N969K, L981F, K986P, V987P. Mutations for delta and omicron spike were guided by the following databases: https://covariants.org/variants/21K. Omicron and https://viralzone.expasy.org/9556. The cDNA fragment was digested, cleaned up, and ligated to the pαH vector backbone using T4 Ligase (the same backbone as the original HexaPro Spike (B-lineage-origin) Addgene plasmid # 154754). Ligated plasmids were transformed in NEB DH5-alpha cells and plated on agar with ampicillin. Colonies were picked, cultured, and purified using the Qiagen HiSpeed MaxiPrep Kit and sent for Sanger sequencing to confirm the identity. Expi293F

cells were transfected using ExpiFectamine293 transfection reagent (ThermoFisher) according to the manufacturer's instructions. The cells were cultured at 37°C, 8% CO_2, 125 rpm (25 mm throw) for 5 to 6 days before purification.

For the purification of wild type, alpha, beta, delta, and omicron spike, the medium in which the spike protein was secreted was supplemented with 1× PBS buffer at pH 7.4 (1:1 v/v) and 5 mM $NiSO_4$. The pH was adjusted with NaOH to pH 7.4 and filtered using a 0.8-μm 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 5-ml His trap FF GE Healthcare column in PBS, 40 mM imidazole, pH 7.4, and eluted in PBS, 300 mM imidazole, pH 7.4. The protein was then injected onto either a Superdex 200 16/600 or 10/300 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 (prewashed multiple times with deuterated PBS) to a concentration of roughly 1 mg/ml.

Protein expression and purification: Influenza HA

Freestyle 293-F cells were cultured in Freestyle expression media (Life Technologies) (37°C, 8% CO_2, 115 rpm orbital shaking). Cells were transfected at a density of 10^9 cells/liter with pre-incubated expression vector (300 μg/liter) and polyethyleneimine (PEI) MAX (Polysciences) (900 μg/liter). Expression vectors encoded terminally His-tagged wild-type influenza A virus (IAV) NC99 (H1N1) HA or a ΔRBS mutant previously described (*55*). After 5 days, supernatant was harvested and protein was purified via immobilized metal chromatography.

Protein expression and purification: C5 anti-spike nanobody

C5-Nanobody was purified as described (*46*). Purified C5 nanobody was then dialyzed into deuterated PBS buffer using 500 μl of Slide-A-Lyzer cassette (3.5 kDa cutoff).

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, which further is in accord with values obtained from independent repeated measurements.

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). 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. See supplementary materials for syntheses of key compounds.

Protein NMR experiments

All NMR experiments in table S8 were conducted at 15°C on a Bruker AVANCE NEO 600 MHz spectrometer with CPRHe-QR-1H/19F/13C/15N-5mm-Z helium-cooled cryoprobe. 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 was adjusted. Typical acquisition parameters were sweep width of 9615.39 Hz, 16 scans per transient (NS), with four dummy scans, 32,768 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 samples were acquired similarly with 16,384 scans per transient with a total acquisition time of 12.5 hours.

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 50-ms 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 200 Hz, which was calculated internally through scaling the power of the high-power 90° pulse.

The total relaxation delay was set to 5 s, during which the saturation pulse was applied. The data were acquired in an interleaved fashion, with each individual excitation frequency being repeated eight times (L4) until the total desired number of scans was achieved. Again, the spectrum was centered on the water peak, and the receiver gain was optimized. After recording of the free induction decay (FID), and prior to the recycle delay, a pair of water-selective pulses wee applied to destroy any unwanted magnetization. For all gradients (excitation sculpting and spoil), the duration was 3 ms using the smooth-square shape SMSQ10.100.

In a typical experiment, two excitation frequencies were required, one exciting protein, and one exciting far from the protein (+20,000 Hz, +33 ppm 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 0.1 s, 0.3 s, 0.5 s, 0.7 s, 0.9 s, 1.1 s, 1.3 s, 1.5 s, 1.7 s, 1.9 s, 2.0 s, 2.5 s, 3.0 s, 3.5 s, 4.0 s, and 5.0 s.

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

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

For STD 10- to 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 5-mm TCI CryoProbe, running TopSpin 3.2.6, recorded in table S9, and a 950-MHz spectrometer with Bruker Avance III HD console and 5-mm TCI CryoProbe, running TopSpin 3.6.1, recorded in table S11. The 950-MHz spectrometer 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-MHz spectrometer, typical acquisition parameters were sweep width of 9615.39 Hz with typically 128 scans per transient (NS = 16 * L4 = 8), 32,768 complex points in the direct dimension and two dummy scans, executed prior to data acquisition.

On the 950-MHz spectrometer, typical acquisition parameters were sweep width of 15,243.90 Hz with typically 128 scans per transient (NS = 16 * L4 = 8), 32,768 complex points in the direct dimension and 2 dummy scans, executed prior to data acquisition.

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 (fig. S6).

To analyze 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 1D "pulse off" spectra to identify peak positions and intensities. Having identified possible peak positions, the algorithm then analyzes 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 (fig. S8).

The ligand-only spectra were analyzed similarly and 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 analyzed, which returned results very similar to the ligand-only case. Contributions from the protein, although small, were typically evident in the spectra, justifying the explicit inclusion of the protein-only baseline during the analysis. When analyzing the mixture, we included the protein-only background as a peak shape whose contribution to the spectrum could 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 analyze 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 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, signals 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 figs. S21 and S22).

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 (figs. S9 and S10). Depending on the precise relationship among 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 (figs. S9 and 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 these data (figs. S6 and S8). The invariance of relative transfer efficiency with excitation frequency suggests that the internal evolution of magnetization within the protein during saturation (likely on the micro-/millisecond time scales) is much faster than the effective cross-relaxation rate between protein and ligand (occurring on the seconds time scale).

Having identified the relevant resonances of interest and performed both a protein and residual ligand subtraction, we reanalyzed the spectra 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; 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 direct measurements of the heteroatom.

In practice, raw unformatted FIDs are submitted to the uSTA pipeline, and the various steps described 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 reconvoluted spectrum and the raw data, and input manual atomic assignments for each observed multiplet. The uSTA 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 (Fig. 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 set up 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:

$$\rho(t) = \rho(0)\exp(-Rt)$$

The challenge comes from the number of spins that must 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 (*64*) that takes a static structure of a protein/ligand complex and estimates STD transfers. The CORCEMA calculations performed to arrive at cross-relaxation rates assume the complex is rigid, which is often 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 fit 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 that the system returns to thermal equilibrium at long times. As such, our evolution matrix **R** will be a square matrix with 13×13 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 the free ligand chemical shift is set to a value that matches 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 dipole-coupled spins separated by a distance R with similar Larmor frequency. In addition, cross-relaxation between ligand and protein is allowed only when the two are bound. The relaxation rates are characterized by an effective distance and an effective correlation time,

$$R_1 = \frac{1}{4}K[J(0) + 3J(\omega) + 6J(2\omega)]$$

$$R_2 = \frac{1}{4}K\left[\frac{5}{2}J(0) + \frac{9}{2}J(\omega) + 3J(2\omega)\right]$$

$$\sigma = \frac{1}{4}K[6J(2\omega) - J(0)]$$

which are each parameterized in terms of an interaction constant (depending on effective distance)

$$K = \left(\frac{\mu_0 \hbar \gamma_H^2}{4\pi r^3}\right)^2$$

and a spectral density function (depending on effective distance)

$$J(\omega) = \frac{2}{5}\frac{\tau}{1 + \omega^2\tau^2}$$

The longitudinal and transverse relaxation rates R_1 and R_2 describe auto-relaxation of diagonal z elements and xy elements, respectively. The cross-relaxation rates σ describe cross-relaxation and couple 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_1M_0 or σM_0 linking the identify element and the z matrix elements. Overall, the relaxation part of the model is parameterized 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

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

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

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

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

The density matrix is initialized 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 eight 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, and care is needed using optimization to avoid local minima. By obtaining data at various protein and ligand concentrations, however, it is possible to break this degeneracy and obtain well-described values as in the text.

In practical terms, the initial rate of the build-up curve is predominantly affected by the cross-relaxation rate and the off rate, and the final height of the build-up 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 parameterized by a set of values that characterize the intrinsic and cross relaxation. From τ_G and r_{IS}(ligand) we estimate R_1 and R_2 of the ligand; from τ_E and r_{IS}(protein) we obtain R_1 and R_2 of the protein; and from τ_E and r_{IS}(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 eight parameters that specify our model. The distances should be considered "effective" values that parameterize the relaxation rates, although in principle it should be possible to obtain physical insights from their interpretation. The concentration-independent relaxation rates can be separated from the exchange rates by comparing the curves as a function of ligand and protein concentration. By treating the system as comprising 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 seconds, which is sufficient for near-saturating spin diffusion within a protein. Because of the complexity of the model, optimization of the parameters via a gradient descent method can get stuck in local minima. In practical applications, it is advisable to start the optimization over a range of initial conditions, particularly in the rates, to ensure that the lowest possible χ^2 is achieved.

Thermostability assays

Thermal stability assays were performed using a NanoTemper Prometheus NT.48 (Membrane Protein Laboratory, Diamond Light Source). To 11 μl of 2 μM spike (deuterated PBS), 2 μl 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 immobilization of SiaLac onto the sensor chip, a flow rate of 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) reagent **10** was injected over 10 min and a final 2-min wash step was performed (see fig. S13 and supplementary materials for further details).

For analysis of spike binding, a flow rate of 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 analyze the SPR data, we assume an equilibrium of the form $PL \leftrightarrows P + L$ characterized by a dissociation constant

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

To follow the kinetics of binding and dissociation, we assume that 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 that the free protein concentration is constant, which leads to the following:

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

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 immobilized onto a chip surface (*65*). Nonetheless, 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.

Modeling of the N-terminal domain of SARS-CoV-2 with glycans

We modeled the structure of the NTD on Protein Data Bank (PDB) entry 7c2l (*19*) because it provided much better coverage of the area of interest when compared to the majority of the templates available at PDB as of 15 July 2020. The models were created with Modeller (*66*),

using the "automodel" protocol without refining the "loop." We generated 10 models and ranked them by their DOPE score (*67*), selecting the top five 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 (*42*, *43*). The binding site on NTD was defined by comparison with PDB entry 6q06 (*5*), a complex of MERS-CoV spike protein and 2,3-sialyl-*N*-acetyl-lactosamine. The binding site could not be directly 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. Because only one restraint was used, we disabled the random removal of restraints. Following our small-molecule docking recommended settings (*68*), 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 300 K, respectively ("tadinit2_t" and "tadinit3_t" to 500 and 300, respectively). Clustering was performed based on "RMSD" with a distance cutoff of 2 Å, and the scoring function was modified to

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

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

Cryo-EM analysis

SARS-CoV-2 spike protein, generated and purified as described (*48*), in 1.1 mg/ml was incubated with 10 mM ethyl(triiodobenzamide) sialyllactoside overnight at 4°C. A 3.5-μl sample was applied to glow-discharged Quantifoil gold R1.2/1.3 300-mesh grids and blotted for ~3 s at 100% humidity and 6°C before vitrification in liquid ethane using Vitrobot (FEI). Two datasets were collected on Titan Krios equipped with a K2 direct electron detector at the cryo-EM facility (OPIC) in the Division of Structural Biology, University of Oxford. Both datasets were collected by SerialEM at a magnification of 165,000× with a physical pixel size of 0.82 Å per pixel. Defocus range was −0.8 μm to −2.4 μm. Total doses for the

two datasets were 60 e/Å² (5492 movies) and 61 e/Å² (8284 movies), respectively.

Motion correction was performed by MotionCor2 (*69*). The motion-corrected micrographs were imported into cryoSPARC (*70*) and contrast transfer function values were estimated using Gctf (*71*) in cryoSPARC. Templates were produced by 2D classification from 5492 micrographs with particles auto-picked by Laplacian-of-Gaussian (LoG)–based algorithm in RELION 3.0 (*72*, *73*). Particles were picked from all micrographs using Template picker in cryoSPARC. Multiple rounds of 2D classification were carried out and the selected 2D classes (372,157 particles) were subjected to 3D classification (Heterogeneous Refinement in cryoSPARC) using six classes. One class was predominant after 3D classification. Nonuniform refinement (*74*) was performed for this class (312,018 particles) with *C*1 and *C*3 symmetry, respectively, yielding a 2.27 Å map for *C*3 symmetry and a 2.44 Å map for *C*1 symmetry. See also fig. S23 and table S12 for cryo-EM data collection, refinement, and validation statistics.

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 sample where the gene is not mutated, and a score of 1 was assigned when at least one variant is present on the gene.

The cohort was distributed as follows. Ethnicity: 504 white, 4 Black, 5 Asian, 16 Hispanic ethnicity, 4 patients for which this information was not available. Sex: 317 male, 216 female. Age: minimum age 19 years, maximum age 99 years, mean age 62.5 years.

Gene 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 logistic

regression model. Specifically, for a set of n samples $\{x_i, y_i\}$ (i = 1, …, n), each of which consists of p input features $x_{i,k} \in \chi_i$, k = 1, …, 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 $\chi = \chi_1 \times \chi_2 \ldots \times \chi_p$ was denoted "input space," whereas the "hypothesis space" was the space of all the possible functions $f: \chi \to Y$ mapping the inputs to the output. Given that the number of features (p) is substantially higher than the number of samples (n), LASSO regularization (*49*) 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:

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

where $\mathbf{x}$ 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

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

where $\hat{y} = \Pr(Y = 1|X = \mathbf{x})$ 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 importance of the subset of most relevant genes for the task. Because a feature-ranking criterion can become suboptimal when the subset of removed features is large (*75*), 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 hyperparameter 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 misclassification of minority class with a multiplicative factor inversely proportional to the class frequencies. For the RFE, the number of included features at each step of the algorithm as well as the final number of features was fixed at 100. All data preprocessing and the RFE procedure were coded in Python; the LR model was used, as included, in the scikit-learn module with the liblinear coordinate descent optimization algorithm.

REFERENCES AND NOTES

1. S. Yamada et al., Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type receptors. Nature 444, 378–382 (2006). doi: 10.1038/nature05264; pmid: 17108965

2. W. Li et al., Identification of sialic acid-binding function for the Middle East respiratory syndrome coronavirus spike glycoprotein. Proc. Natl. Acad. Sci. U.S.A. 114, E8508–E8517 (2017). doi: 10.1073/pnas.1712592114; pmid: 28923942

3. C. Schwegmann-Wessels, G. Herrler, Sialic acids as receptor determinants for coronaviruses. Glycoconj. J. 23, 51–58 (2006). doi: 10.1007/s10719-006-5437-9; pmid: 16575522

4. R. J. G. Hulswit 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. Proc. Natl. Acad. Sci. U.S.A. 116, 2681–2690 (2019). doi: 10.1073/pnas.1809667116; pmid: 30679277

5. Y.-J. Park et al., Structures of MERS-CoV spike glycoprotein in complex with sialoside attachment receptors. Nat. Struct. Mol. Biol. 26, 1151–1157 (2019). doi: 10.1038/s41594-019-0334-7; pmid: 31792450

6. E. Qing, M. Hantak, S. Perlman, T. Gallagher, Distinct Roles for Sialoside and Protein Receptors in Coronavirus Infection. mBio 11, e02764-19 (2020). doi: 10.1128/mBio.02764-19; pmid: 32047128

7. W. Li et al., Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. Nature 426, 450–454 (2003). doi: 10.1038/nature02145; pmid: 14647384

8. K. Kuba et al., A crucial role of angiotensin converting enzyme 2 (ACE2) in SARS coronavirus-induced lung injury. Nat. Med. 11, 875–879 (2005). doi: 10.1038/nm1267; pmid: 16007097

9. F. Li, W. Li, M. Farzan, S. C. Harrison, Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. Science 309, 1864–1868 (2005). doi: 10.1126/science.1116480; pmid: 16166518

10. D. Wrapp et al., Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 367, 1260–1263 (2020). doi: 10.1126/science.abb2507; pmid: 32075877

11. R. Yan et al., Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science 367, 1444–1448 (2020). doi: 10.1126/science.abb2762; pmid: 32132184

12. D. Morniroli, M. L. Gianni, A. Consales, C. Pietrasanta, F. Mosca, Human Sialome and Coronavirus Disease-2019 (COVID-19) Pandemic: An Understated Correlation? Front. Immunol. 11, 1480 (2020). doi: 10.3389/fimmu.2020.01480; pmid: 32655580

13. A. N. Baker et al., The SARS-COV-2 Spike Protein Binds Sialic Acids and Enables Rapid Detection in a Lateral Flow Point of Care Diagnostic Device. ACS Cent. Sci. 6, 2046–2052 (2020). doi: 10.1021/acscentsci.0c00855; pmid: 33269329

14. W. Hao et al., Binding of the SARS-CoV-2 spike protein to glycans. Sci. Bull. 66, 1205–1214 (2021). doi: 10.1016/j.scib.2021.01.010; pmid: 33495714

15. L. Nguyen et al., Sialic acid-containing glycolipids mediate binding and viral entry of SARS-CoV-2. Nat. Chem. Biol. 18, 81–90 (2022). doi: 10.1038/s41589-021-00924-1

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

17. T. M. Clausen et al., SARS-CoV-2 Infection Depends on Cellular Heparan Sulfate and ACE2. Cell 183, 1043–1057.e15 (2020). doi: 10.1016/j.cell.2020.09.033; pmid: 32970989

18. N. Behloul, S. Baha, R. Shi, J. Meng, Role of the GTNGTKR motif in the N-terminal receptor-binding domain of the SARS-CoV-2 spike protein. Virus Res. 286, 198058 (2020). doi: 10.1016/j.virusres.2020.198058; pmid: 32531235

19. X. Chi et al., A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. Science 369, 650–655 (2020). doi: 10.1126/science.abc6952; pmid: 32571838

20. M. G. Rossmann, The canyon hypothesis. J. Biol. Chem. 264, 14587–14590 (1989). doi: 10.1016/S0021-9258(18)63732-9; pmid: 2670920

21. Y. Watanabe et al., Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. Nat. Commun. 11, 2688 (2020). doi: 10.1038/s41467-020-16567-0; pmid: 32461612

22. M. Mayer, B. Meyer, Characterization of Ligand Binding by Saturation Transfer Difference NMR Spectroscopy. Angew. Chem. Int. Ed. 38, 1784–1788 (1999). doi: 10.1002/(SICI)1521-3773(19990614)38:12<1784::AID-ANIE1784>3.0.CO;2-Q; pmid: 29711196

23. J. L. Wagstaff, S. L. Taylor, M. J. Howard, Recent developments and applications of saturation transfer difference nuclear magnetic resonance (STD NMR) spectroscopy. Mol. Biosyst. 9, 571–577 (2013). doi: 10.1039/C2MB25395J; pmid: 23232937

24. J. Angulo, P. M. Enríquez-Navas, P. M. Nieto, Ligand-receptor binding affinities from saturation transfer difference (STD) NMR spectroscopy: The binding isotherm of STD initial growth rates. Chem. Eur. J. 16, 7803–7812 (2010). doi: 10.1002/chem.200903528; pmid: 20496354

25. R. H. McMenamy, J. L. Oncley, The specific binding of L-tryptophan to serum albumin. J. Biol. Chem. 233, 1436–1447 (1958). doi: 10.1016/S0021-9258(18)49353-2; pmid: 13610854

26. P. Vallurupalli, G. Bouvignies, L. E. Kay, Studying "invisible" excited protein states in slow exchange with a major state conformation. J. Am. Chem. Soc. 134, 8148–8161 (2012). doi: 10.1021/ja3001419; pmid: 22554188

27. T. Xie, T. Saleh, P. Rossi, C. G. Kalodimos, Conformational states dynamically populated by a kinase determine its function. Science 370, eabc2754 (2020). doi: 10.1126/science.abc2754; pmid: 33004676

28. N. L. Fawzi, J. Ying, D. A. Torchia, G. M. Clore, Probing exchange kinetics and atomic resolution dynamics in high-molecular-weight complexes using dark-state exchange saturation NMR spectroscopy. Nat. Protoc. 7, 1523–1533 (2012). doi: 10.1038/nprot.2012.077; pmid: 22814391

29. G. Bouvignies et al., Solution structure of a minor and transiently formed state of a T4 lysozyme mutant. Nature 477, 111–114 (2011). doi: 10.1126/science.abc2754; pmid: 33004676

30. M. Mayer, B. Meyer, Group epitope mapping by saturation transfer difference NMR to identify segments of a ligand in direct contact with a protein receptor. J. Am. Chem. Soc. 123, 6108–6117 (2001). doi: 10.1021/ja0100120; pmid: 11414845

31. M. T. Marty et al., Bayesian deconvolution of mass and ion mobility spectra: From binary interactions to polydisperse ensembles. Anal. Chem. 87, 4370–4376 (2015). doi: 10.1021/acs.analchem.5b00140; pmid: 25799115

32. L. Fielding, S. Rutherford, D. Fletcher, Determination of protein-ligand binding affinity by NMR: Observations from serum albumin model systems. Magn. Reson. Chem. 43, 463–470 (2005). doi: 10.1002/mrc.1574; pmid: 15816062

33. A. Bujacz, K. Zielinski, B. Sekula, Structural studies of bovine, equine, and leporine serum albumin complexes with naproxen. Proteins 82, 2199–2208 (2014). doi: 10.1002/prot.24583; pmid: 24753230

34. I. Pérez-Victoria et al., Saturation transfer difference NMR reveals functionally essential kinetic differences for a sugar-binding repressor protein. Chem. Commun. 2009, 5862–5864 (2009). doi: 10.1039/b913489a; pmid: 19787122

35. U. Hars, R. Horlacher, W. Boos, W. Welte, K. Diederichs, 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 Sci. 7, 2511–2521 (1998). doi: 10.1002/pro.5560071204; pmid: 9865945

36. H. M. McConnell, Reaction Rates by Nuclear Magnetic Resonance. J. Chem. Phys. 28, 430–431 (1958). doi: 10.1063/1.1744152

37. J. E. Stencel-Baerenwald, K. Reiss, D. M. Reiter, T. Stehle, T. S. Dermody, The sweet spot: Defining virus-sialic acid interactions. Nat. Rev. Microbiol. 12, 739–749 (2014). doi: 10.1038/nrmicro3346; pmid: 25263223

38. D. Lingwood et al., Structural and genetic basis for development of broadly neutralizing influenza antibodies. Nature 489, 566–570 (2012). doi: 10.1038/nature11371; pmid: 22932267

39. H.-Y. Liao et al., Differential receptor binding affinities of influenza hemagglutinins on glycan arrays. J. Am. Chem. Soc. 132, 14849–14856 (2010). doi: 10.1021/ja104657b; pmid: 20882975

40. Y. Watanabe, J. D. Allen, D. Wrapp, J. S. McLellan, M. Crispin, Site-specific glycan analysis of the SARS-CoV-2 spike. Science 369, 330–333 (2020). doi: 10.1126/science.abb9983; pmid: 32366695

41. M. P. Lenza et al., Structural Characterization of N-Linked Glycans in the Receptor Binding Domain of the SARS-CoV-2 Spike Protein and their Interactions with Human Lectins. Angew. Chem. Int. Ed. 59, 23763–23771 (2020). doi: 10.1002/anie.202011015; pmid: 32915505

42. C. Dominguez, R. Boelens, A. M. J. J. Bonvin, HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. J. Am. Chem. Soc. 125, 1731–1737 (2003). doi: 10.1021/ja026939x; pmid: 12580598

43. G. C. P. van Zundert et al., The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. J. Mol. Biol. 428, 720–725 (2016). doi: 10.1016/j.jmb.2015.09.014; pmid: 26410586

44. A. Rambaut et al., A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. 5, 1403–1407 (2020). doi: 10.1038/s41564-020-0770-5; pmid: 32669681

45. J. Huo et al., Neutralizing nanobodies bind SARS-CoV-2 spike RBD and block interaction with ACE2. Nat. Struct. Mol. Biol. 27, 846–854 (2020). doi: 10.1038/s41594-020-0469-6; pmid: 32661423

46. J. Huo et al., A potent SARS-CoV-2 neutralising nanobody shows therapeutic efficacy in the Syrian golden hamster model of COVID-19. Nat. Commun. 12, 5469 (2021). doi: 10.1038/s41467-021-25480-z; pmid: 34552091

47. M. McCallum, A. C. Walls, J. E. Bowen, D. Corti, D. Veesler, Structure-guided covalent stabilization of coronavirus spike glycoprotein trimers in the closed conformation. Nat. Struct. Mol. Biol. 27, 942–949 (2020). doi: 10.1038/s41594-020-0483-8; pmid: 32753755

48. C. Toelzer et al., Free fatty acid binding pocket in the locked structure of SARS-CoV-2 spike protein. Science 370, 725–730 (2020). doi: 10.1126/science.abd3255; pmid: 32958580

49. R. Tibshirani, Regression Shrinkage and Selection Via the Lasso. J. R. Stat. Soc. B 58, 267–288 (1996). doi: 10.1111/j.2517-6161.1996.tb02080.x

50. K. Koths, E. Taylor, R. Halenbeck, C. Casipit, A. Wang, 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. J. Biol. Chem. 268, 14245–14249 (1993). doi: 10.1016/S0021-9258(19)85233-X; pmid: 8390986

51. L. Johannes, R. Jacob, H. Leffler, Galectins at a glance. J. Cell Sci. 131, jcs208884 (2018). doi: 10.1242/jcs.208884; pmid: 29717004

52. S. R. Stowell et al., Galectin-1, -2, and -3 exhibit differential recognition of sialylated glycans and blood group antigens. J. Biol. Chem. 283, 10109–10123 (2008). doi: 10.1074/jbc.M709545200; pmid: 18216021

53. N. A. Kamili et al., Key regulators of galectin-glycan interactions. Proteomics 16, 3111–3125 (2016). doi: 10.1002/pmic.201600116; pmid: 27582340

54. A. Togayachi et al., β3GnT2 (B3GNT2), a major polylactosamine synthase: Analysis of B3GNT2-deficient mice. Methods Enzymol. 479, 185–204 (2010). doi: 10.1016/s0076-6879(10)79011-x

55. M. Kanekiyo et al., Self-assembling influenza nanoparticle vaccines elicit broadly neutralizing H1N1 antibodies.

Nature **499**, 102–106 (2013). doi: 10.1038/nature12202; pmid: 23698367

56. M. Mammen, S.-K. Choi, G. M. Whitesides, Polyvalent Interactions in Biological Systems: Implications for Design and Use of Multivalent Ligands and Inhibitors. *Angew. Chem. Int. Ed.* **37**, 2754–2794 (1998). doi: 10.1002/(SICI)1521-3773 (19981102)37:20<2754::AID-ANIE2754>3.0.CO;2-3; pmid: 29711117

57. R. T. Lee, Y. C. Lee, Affinity enhancement by multivalent lectin-carbohydrate interaction. *Glycoconj. J.* **17**, 543–551 (2000). doi: 10.1023/A:1011070425430; pmid: 11421347

58. N. Jia *et al.*, The Human Lung Glycome Reveals Novel Glycan Ligands for Influenza A Virus. *Sci. Rep.* **10**, 5320 (2020). doi: 10.1038/s41598-020-62074-z; pmid: 32210305

59. M. A. Toscano *et al.*, Differential glycosylation of TH1, TH2 and TH-17 effector cells selectively regulates susceptibility to cell death. *Nat. Immunol.* **8**, 825–834 (2007). doi: 10.1038/ni1482; pmid: 17589510

60. P. Tangvoranuntakul *et al.*, Human uptake and incorporation of an immunogenic nonhuman dietary sialic acid. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12045–12050 (2003). doi: 10.1073/pnas.2131556100; pmid: 14523234

61. J. L. McAuley, B. P. Gilbertson, S. Trifkovic, L. E. Brown, J. L. McKimm-Breschkin, Influenza Virus Neuraminidase Structure and Functions. *Front. Microbiol.* **10**, 39 (2019). doi: 10.3389/fmicb.2019.00039; pmid: 30761095

62. J. Huo *et al.*, Neutralization of SARS-CoV-2 by Destruction of the Prefusion Spike. *Cell Host Microbe* **28**, 445–454.e6 (2020). doi: 10.1016/j.chom.2020.06.010; pmid: 32585135

63. C.-L. Hsieh *et al.*, Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science* **369**, 1501–1505 (2020). doi: 10.1126/science.abd0826; pmid: 32703906

64. V. Jayalakshmi, N. R. Krishna, Complete relaxation and conformational exchange matrix (CORCEMA) analysis of intermolecular saturation transfer effects in reversibly forming ligand-receptor complexes. *J. Magn. Reson.* **155**, 106–118 (2002). doi: 10.1006/jmre.2001.2499; pmid: 11945039

65. K. Nakamura *et al.*, Immobilized glycosylated Fmoc-amino acid for SPR: Comparative studies of lectin-binding to linear or biantennary diLacNAc structures. *Carbohydr. Res.* **382**, 77–85 (2013). doi: 10.1016/j.carres.2013.10.003; pmid: 24211369

66. A. Sali, Comparative protein modeling by satisfaction of spatial restraints. *Mol. Med. Today* **1**, 270–277 (1995). doi: 10.1016/S1357-4310(95)91170-7; pmid: 9415161

67. M. Y. Shen, A. Sali, Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507–2524 (2006). doi: 10.1110/ps.062416606; pmid: 17075131

68. P. I. Koukos, L. C. Xue, A. M. J. J. Bonvin, Protein-ligand pose and affinity prediction: Lessons from D3R Grand Challenge 3. *J. Comput. Aided Mol. Des.* **33**, 83–91 (2019). doi: 10.1007/s10822-018-0148-4; pmid: 30128928

69. S. Q. Zheng *et al.*, MotionCor2: Anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017). doi: 10.1038/nmeth.4193; pmid: 28250466

70. A. Punjani, J. L. Rubinstein, D. J. Fleet, M. A. Brubaker, cryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017). doi: 10.1038/nmeth.4169; pmid: 28165473

71. K. Zhang, Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016). doi: 10.1016/j.jsb.2015.11.003; pmid: 26592709

72. S. H. W. Scheres, RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012). doi: 10.1016/j.jsb.2012.09.006; pmid: 23000701

73. J. Zivanov *et al.*, New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7**, e42166 (2018). doi: 10.7554/eLife.42166; pmid: 30412051

74. A. Punjani, H. Zhang, D. J. Fleet, Non-uniform refinement: Adaptive regularization improves single-particle cryo-EM reconstruction. *Nat. Methods* **17**, 1214–1221 (2020). doi: 10.1038/s41592-020-00990-8; pmid: 33257830

75. I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.* **46**, 389–422 (2002). doi: 10.1023/A:1012487302797

76. C. Buchanan *et al.*, Pathogen-sugar interactions revealed by universal saturation transfer analysis. Zenodo, doi: 10.5281/zenodo.6299883 (2022).doi: 10.5281/zenodo.6299883

77. P. I. Koukos, A. M. J. J. Bonvin, Modelling of the binding of various sialic acid-containing oligosaccharides to the NTD domain of the spike protein of SARS-CoV2. Zenodo, doi: 10.5281/zenodo.4271288 (2022).

78. C. J. Buchanan, A. J. Baldwin, Universal Saturation Transfer Analysis software. Zenodo, doi: 10.5281/zenodo.6303964 (2022).

79. N. Picchiotti *et al.*, Post-Mendelian genetic model in COVID-19. *Cardiol. Cardiovasc. Med.* **5**, 673–694 (2021). doi: 10.26502/fccm.92920232

80. C. Fallerini *et al.*, Common, low-frequency, rare, and ultra-rare coding variants contribute to COVID-19 severity. *Hum. Genet.* **141**, 147–173 (2022). doi: 10.1007/s00439-021-02397-7; pmid: 34889978

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIALS

Science

Pathogen-sugar interactions revealed by universal saturation transfer analysis

Charles J. BuchananBen GauntPeter J. HarrisonYun YangJiwei LiuAziz KhanAndrew M. GiltrapAudrey Le BasPhilip N. WardKapil GuptaMaud DumouxTiong Kit TanLisa SchimaskiSergio DagaNicola PicchiottiMargherita BaldassarriElisa BenettiChiara FalleriniFrancesca FavaAnnarita GilibertiPanagiotis I. KoukosMatthew J. DavyAbirami LakshminarayananXiaochao XueGeorgios PapadakisLachlan P. DeimelVirgínia Casablancas-AntràsTimothy D. W. ClaridgeAlexandre M. J. J. BonvinQuentin J. SattentauSimone FuriniMarco GoriJiandong HuoRaymond J. OwensChristiane SchaffitzelImre BergerAlessandra RenieriJames H. NaismithAndrew J. BaldwinBenjamin G. Davis

Glycans in the spotlight

Viral infection of a cell requires a complex series of molecular recognition events, often mediated by glycoproteins and cell-surface glycans. Buchanan *et al.* developed a nuclear magnetic resonance analysis method to better study such interactions and applied it to influenza and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) proteins binding sialoside glycans. For SARS-CoV-2 in particular, they found evidence for a sialoside-binding site in the N-terminal domain of the original B-origin lineage spike protein that was lost in subsequent variants. These results were corroborated by cryo–electron microscopy structures of the glycan-bound spike protein and genetic variation analysis from patients early in the pandemic, which uncovered host factors involved in glycosylation that potentially contributed to variation in disease severity. —MAF

View the article online
https://www.science.org/doi/10.1126/science.abm3125
Permissions
https://www.science.org/help/reprints-and-permissions

Use of this article is subject to the Terms of service