Functional binding dynamics relevant to the evolution of zoonotic spillovers in endemic and emergent *Betacoronavirus* strains

Patrick Rynkiewicz, Miranda L. Lynch, Feng Cui, André O. Hudson & Gregory A. Babbitt

View supplementary material

Published online: 21 Jul 2021.

Submit your article to this journal

View related articles

View Crossmark data

Taylor & Francis
Taylor & Francis Group

Check for updates

Functional binding dynamics relevant to the evolution of zoonotic spillovers in endemic and emergent *Betacoronavirus* strains

Patrick Rynkiewicz[a], Miranda L. Lynch[b], Feng Cui[a], André O. Hudson[a] and Gregory A. Babbitt[a] [iD]

[a]Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester, NY, USA; [b]Hauptman-Woodward Medical Research Institute, Buffalo, NY, USA

Communicated by Ramaswamy H. Sarma

ABSTRACT

Comparative functional analysis of the dynamic interactions between various *Betacoronavirus* mutant strains and broadly utilized target proteins such as ACE2 and CD26, is crucial for a more complete understanding of zoonotic spillovers of viruses that cause diseases such as COVID-19. Here, we employ machine learning to replicated sets of nanosecond scale GPU accelerated molecular dynamics simulations to statistically compare and classify atom motions of these target proteins in both the presence and absence of different endemic and emergent strains of the viral receptor binding domain (RBD) of the S spike glycoprotein. A multi-agent classifier successfully identified functional binding dynamics that are evolutionarily conserved from bat CoV-HKU4 to human endemic/emergent strains. Conserved dynamics regions of ACE2 involve both the N-terminal helices, as well as a region of more transient dynamics encompassing residues K353, Q325 and a novel motif AAQPFLL 386-92 that appears to coordinate their dynamic interactions with the viral RBD at N501. We also demonstrate that the functional evolution of *Betacoronavirus* zoonotic spillovers involving ACE2 interaction dynamics are likely pre-adapted from two precise and stable binding sites involving the viral bat progenitor strain's interaction with CD26 at SAMLI 291-5 and SS 333-334. Our analyses further indicate that the human endemic strains hCoV-HKU1 and hCoV-OC43 have evolved more stable N-terminal helix interactions through enhancement of an interfacing loop region on the viral RBD, whereas the highly transmissible SARS-CoV-2 variants (B.1.1.7, B.1.351 and P.1) have evolved more stable viral binding via more focused interactions between the viral N501 and ACE2 K353 alone.

Introduction

In the worldwide COVID-19 pandemic there has been an unprecedented wealth of high-throughput sequencing surveillance of the SARS-CoV-2 virus and near real-time sequence-based analysis of the evolution of local viral strains. Sequence-based analysis can be highly insightful for detailing the phylogenetic history of viral strains. However, without expensive and time-consuming functional binding assays it is often difficult to predict the rarer but significant mutations with functional public health consequences from among this rapid and largely neutral viral evolution. These assays can be further supplemented by sequence-based phylogenetic tests of selection as well as experimental and computational structural-functional analyses. However, even when employed together, experiments supplemented with analyses of static data (i.e. protein sequences and structures) are still often incapable of fully deciphering the precise molecular functions that underlie important changes in viral transmission. Ultimately, viral transmission is the consequence of soft-matter dynamics responding to weak chemical bonding interactions at many potential sites on both the viral and putative target proteins. The unprecedented pandemic of SARS-CoV-2 brings with it many questions regarding the functional evolution of both emergent and endemic *Betacoronavirus* strains in humans from present and past zoonotic spillover events. It has been demonstrated *in vitro* that ACE2 is the target of SARS-CoV-2 across a broad variety of mammalian taxa (Zhao et al., 2020). With the current surveillance of new variants like the highly transmissible and more lethal strains of SARS-CoV-2 (B.1.1.7, B.1.351, and P.1), the comparative molecular dynamics analysis of functionally conserved binding mechanisms of viral proteins *in silico* can offer new perspectives on the functional consequences of viral evolution during a pandemic (Faria et al., 2021; Rambaut et al., 2020; Tegally et al., 2021) especially when conclusions are derived via modern machine learning methods (Liang et al., 2021; Randhawa et al., 2020).

The SARS-CoV-2 virus belongs to the class of positive-strand RNA viruses (genus *Betacoronavirus*) and is structurally defined by a nucleocapsid interior surrounded by the spike, membrane, and envelope proteins (Fung & Liu, 2019; Haan et al., 2000). Structural membrane and envelope proteins are known to play a key role in virion assembly and budding from infected cells, making them potential targets for

antiviral treatments (Haan et al., 2000). One prominent protein candidate for comparative studies of the propensity for viral transmission is the spike or S glycoprotein, which is responsible for initial viral contact with the target host receptor and subsequent entry into host cells. The spike protein consists of two homotrimeric functional subunits: the S1 subunit consisting of the viral receptor binding domain (RBD) interacting directly with the host, and the S2 subunit which directly fuses the virion to the host cellular membrane (Hoffmann et al., 2018; Shang et al., 2020). The S1 and S2 subunits of coronaviruses are cleaved and subsequently activated by type-II transmembrane serine proteases and the subtilisin-like serine protease furin, the latter of which evidently primes SARS-CoV-2 for cellular entry but not SARS-CoV-1 (Follis et al., 2006; Shang et al., 2020). The boundary between the S1 and S2 subunits forms an exposed loop in human coronaviruses SARS-CoV-2, hCoV-OC43, hCoV-HKU1, and MERS-CoV that is not found in SARS-CoV-1 (Hoffmann et al., 2020) and suggests a complex functional evolutionary landscape for emergent SARS-type coronaviruses.

The spike protein of SARS-CoV-2 has been demonstrated to directly and specifically target the human angiotensin converting enzyme 2 (ACE2) receptor, which has a role in the renin–angiotensin system (RAS) critical for regulating cardiovascular and renal functions (Ou et al., 2017; 2020; Oudit et al., 2003). ACE2 is a glycoprotein consisting of 805 amino acid residues that is the primary target for docking and viral entry in many coronavirus infections, including those belonging to SARS-CoV and human alpha-coronavirus NL63 (Hofmann et al., 2005). ACE2 is classified as a zinc metallopeptidase (carboxypeptidase) involved in cleaving a single amino acid residue from angiotensin I (AngI) to generate Ang1-9 (Kuba et al., 2005). In addition, ACE2 also cleaves a single amino acid residue from angiotensin II (Ang II) to generate Ang1-7 (Kuba et al., 2005). ACE2 receptors reside on the surfaces of multiple epithelial cells, with tissue distribution predominantly in the lung and gut linings, but are also found in heart, kidney, liver and blood vessel tissues. Given the dominant role ACE2 plays in SARS-CoV-2 viral entry, understanding how sequence and structural variations affect the molecular dynamics of ACE2/RBD interactions can provide information to guide the rationale design and development of therapeutics targeting this interaction. But it is important to note that SARS-CoV-2 also targets other human protein receptors including several neurolipins (Cantuti-Castelvetri et al., 2020; Daly et al. 2020) and perhaps other known targets of other viral relatives such as the MERS-CoV target protein CD26. Therefore, it is important to demonstrate at a molecular level, how such promiscuity in binding interaction is achieved by the viral RBD and how it may have contributed to zoonotic spillovers from other species to humans. The comparative dynamic modeling of present, past, and potential progenitor viral strain interactions to ACE2 and other known targets of past outbreaks like protein CD26, can also potentially illuminate the dynamics of molecular interactions that may facilitate zoonotic transmission to humans as well as the molecular evolution toward emergent and endemic human viral strains.

Here, we present a systematic approach to the comparative statistical analysis and machine learning classification of the dynamic motions of proteins generated from all atom accelerated molecular dynamics simulations. This method is subsequently leveraged to characterize the functionally conserved binding signatures of viral strains on specific target proteins at single amino acid site resolution. We employ DROIDS 4.0/maxDemon 3.0, a multi-agent shallow learning assisted software tool designed by our lab to statistically analyze/visualize comparative protein dynamics and then subsequently identify (i.e. classify) protein regions with functionally conserved dynamics that can be discriminated by the machine learning from more random motions caused by thermal noise (Babbitt et al., 2018, 2020). The software also now uses information theoretics to quantify the relative impact of variants on regions of conserved dynamics and to identify amino acid site coordination (i.e. allostery) related to functionally conserved dynamic states. We subsequently analyze molecular dynamics in models of different *Betacoronavirus* strain S spike glycoprotein interactions with their primary human target protein ACE2 as well as a likely progenitor target in bats, CD26. These strains include the endemic human strains of *Betacoronavirus* (hCoV-OC43 and hCoV-HKU1), emergent strains responsible for recent SARS-like outbreaks (i.e. SARS-CoV-1, SARS-CoV-2 and MERS-CoV), as well as models of hypothetical interactions with the bat progenitor strain bat CoV-HKU4. We confirm two regions of binding interaction on ACE2 that appear pre-adapted from two precise and stable binding interactions of the bat progenitor strain with CD26 (Wang et al., 2014) and are functionally conserved across varying evolutionary distances of the human evolving strains. We confirm important key sites of binding interaction at the ACE2 N-terminal helices, K353 and Q325 identified previously by Yan et al. (Yan et al., 2020) and Li et al. (F. Li et al., 2005; W. Li et al., 2005). In addition, we identify a potentially new ACE2 interaction site with the viral N501 that exhibits coordinated dynamic interplay along with K353 and Q325. This indicates that the original outbreak strain of SARS-CoV-2 has quite a significant degree of binding promiscuity in its interactions with ACE2 after its zoonotic spillover into the human population. Importantly, we demonstrate that the new variants (B.1.1.7, B.1.351 and P.1) have lost this more promiscuous binding in favor of a much more static and stable interaction of the N501Y mutation with the ACE2 K353 alone. Lastly, we demonstrate that while all *Betacoronavirus* spillovers both past and present seem to rely upon conserved ACE2 binding in the same regions, the functional evolution of the molecular binding mechanisms in the current emergent SARS-CoV-2 outbreak is progressing quite differently within these conserved sites when compared to the evolution of past human endemic strains (Lubin et al., 2020).

Materials and methods

Molecular dynamic simulation—some background

A powerful tool for *in silico* analysis of protein-protein interactions is all atom accelerated molecular dynamics (aMD) simulation, which attempts to accurately sample the physical trajectories of molecular structures over time via traditional

Newtonian mechanics played out over a modified potential energy landscape that encourages conformational transition states (Pierce et al., 2012). Compared more computationally expensive traditional molecular dynamics (MD) simulations, aMD is increasingly broadening the scope of inquiry by allowing more comprehensive analysis of the rapid time scale dynamics of protein-protein interactions. Protein structures determined via x-ray crystallography have previously been used to provide insights into viral evolution by comparing angstrom scale distances and conformational similarities. MD simulations have already been employed to map the hydrogen bonding topology of the SARS-CoV-2 RBD/ACE2 interaction. The goal here was to explore the merits of drug repurposing, to characterize mutations, and to determine the stability of binding interactions between the SARS-CoV-2 RBD and several contemporary antiviral drugs (Ahamad et al., 2020; Ali & Vijayan, 2020; Al-Khafaji et al., 2020; Mittal et al., 2020; Muralidharan et al., 2020; Wang et al., 2020; Yu et al., 2020). While conventional MD and aMD simulations can provide unprecedented resolution of protein motion over time, interpretations based upon comparisons of single MD trajectories are statistically problematic and difficult to compare. This makes the application of MD to comparative questions regarding functional binding or impacts of mutation quite challenging. Given recent increases in GPU processor power, the statistical comparison of large replicates of a MD simulations is now an available solution to this problem. Because MD trajectories of complex structures like proteins are inherently complex and chaotic, even if they are adequately resampled, they do not often fit traditional model assumptions of many parametric statistical tests. Therefore, reliable comparisons of molecular dynamics need to further be based upon statistical methods that can capture the multimodal distribution space of protein motion and assign statistical significance for divergence in dynamics without over-reliance on the assumptions of normality in the distribution of molecular dynamics over time.

To this end, our lab has developed DROIDS 4.0, a method and software tool designed to address these challenges in the statistical analysis and visualization of comparative protein dynamics (Babbitt et al., 2018, 2020). DROIDS software includes maxDemon 3.0, a machine learning algorithm/pipeline for detecting functionally conserved dynamics in new/novel MD simulations after being trained upon the functional dynamic states defined in the comparison (i.e. bound vs. unbound, hot vs cold, before vs. after mutation etc.). DROIDS 4.0 now also includes information theoretics designed to compare the relative impacts of genetic variation upon functionally conserved dynamics, as well as identify coordination between sites that are involving the functionally conserved or functionally important dynamics. This method has been successfully applied to the evolution of activation loop dynamics relevant to cancer drug hyperactivation in the B-raf kinase protein (Babbitt et al., 2020). Here, we have subsequently applied these methods to analyze hybrid models of different strains of *Betacoronavirus* S spike glycoprotein interactions with their primary human target protein ACE2 as well as a likely bat progenitor strain target CD26.

PDB structure and hybrid model preparation

Structures of the two main human SARS variants of *Betacoronavirus* spike glycoprotein receptor binding domain (RBD) bound to human ACE2 receptor protein were obtained from the Protein Data Bank (PDB). These were SARS-CoV-1 (PDB: 6acg) (Song et al., 2018) and SARS-CoV-2 (PDB: 6m17) (Yan et al., 2020). Three additional hybrid models of viral RBD interaction with ACE2 consisting of human *Betacoronavirus* variants MERS-CoV (PDB: 5x5c, and PDB 4kr0) (Yuan et al., 2017), hCoV-OC43 (PDB: 6ohw) (Tortorici et al., 2019), and hCoV-HKU1 (PDB: 5gnb) (Ou et al., 2017) bound to ACE2 were generated by creating an initial structural alignment to the SARS-CoV-2 RBD/ACE2 model (PDB: 6m17) using UCSF Chimera's MatchMaker superposition tool (Pettersen et al., 2004) followed by deletion of the SARS-CoV-2 RBD in the model and any structures outside of the molecular dynamics modeling space defined in Figure 1(A), leaving only the viral RBD and ACE2 receptor domain. The most important features of the viral RBD-ACE2 interface are shown in Figure 1(B). These hybrid models of ACE2 interaction included viral receptor binding domains from MERS-CoV (PDB:5x5c and 4kr0) (Yuan et al., 2017), HCoV-OC43 (PDB: 6ohw) (Tortorici et al., 2019), and HCoV-HKU1 (PBD: 5gnb) (Ou et al., 2017) bound to the ACE2 protein from the SARS-CoV-1 model. Unbound forms of the ACE2 structure were obtained by deleting the viral structure in PDB: 6m17 and performing energy minimization for 2000 steps and 10 ns of equilibration of molecular dynamics simulation in Amber18 (Case et al., 2005; Götz et al., 2012; Pierce et al., 2012; Salomon-Ferrer et al., 2013) prior to setting up production runs for the sampling regime (described below). Loop modeling and refinement were conducted where needed using Modeller in UCSF Chimera (Fiser et al., 2000; Sali & Blundell, 1993). A hybrid model of the bat HKU4 betacoronavirus interaction with ACE2 was similarly constructed using PDB: 4qzv (Wang et al., 2014), batCoV-HKU4 bound to the MERS target human CD26, as a reference for structural alignment to PDB: 6m17. Our hybrid models of the viral RBD consisted of the largely un-glycosylated region (represented in PDB: 6m17 from site 335-518). Only one glycan was removed using swapaa in UCSF Chimera at ASN 343 on the viral RBD located on the opposite side of the RBD-ACE2 interface (PDB: 6m17). Five glycans on the ACE2 receptor domain were also removed, none occurring near the binding interface (ASN 53, ASN 90, ASN 103, ASN 322, ASN 546). The removal of these post-translational modifications to the protein was justified in our evolutionary comparisons due to their heterogeneity across infected populations, across species boundaries, and across evolutionary timescales. Single mutation models were also similarly constructed to examine the effects of single mutations that potentially have large effect. These models are summarized in Table 1 and are available in a separate supplemental download file titled PDBmodels.zip.

Mutant model construction and rationale

To generate *in silico* site-directed mutagenesis computations, we created mutant models using the swapaa function in UCSF Chimera 1.13 by first swapping amino acids using
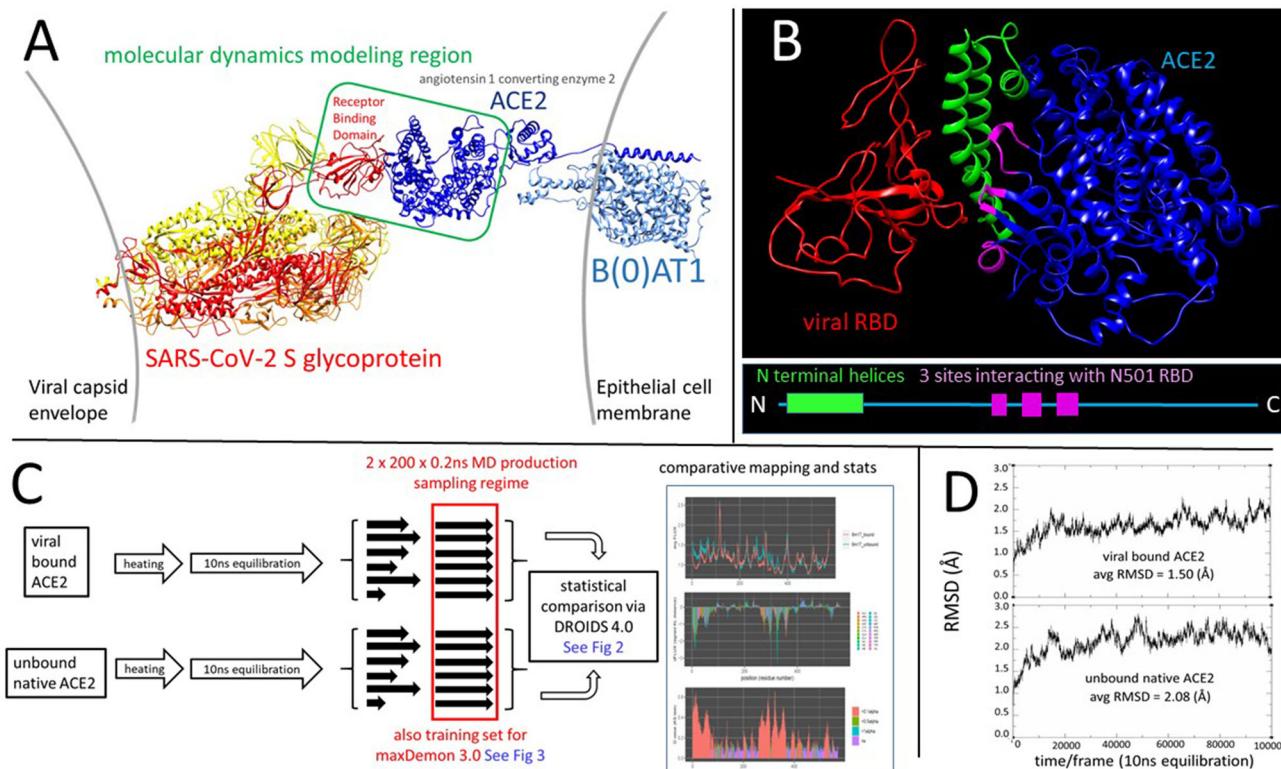
Figure 1. An overview of the molecular system and workflow to generate the comparative protein dynamics data sets. (A) The open conformation SARS-CoV-2 spike glycoprotein (PDB: 6vyb) interaction with its human target cell receptor, angiotensin 1 converting enzyme 2 (ACE2 PDB: 6m17). The green box in the center shows our modeling region that bounds the subsequent molecular dynamics (MD) simulations in the study. (B) The modeling region of the viral receptor binding domain (RBD) in red and ACE2 target protein in blue is shown along with the most relevant features of the binding interface; the N-terminal helices of ACE2 shown in green and the three main sites of the RBD N501 site interactions shown in magenta. (C) An overview of the workflow for the MD simulation pipeline for binding signature analysis used throughout this study. The statistical comparison of dynamics is defined by the difference in atom fluctuation or 'dFLUX' upon ACE2 binding by the virus, measured as the signed/symmetric KL divergence in the distribution of viral bound vs. unbound ACE2 atom fluctuations (Figure 2) collected over 200 production runs and averaged to each amino acid site on the protein backbone (i.e. N, C, Cα, and O atoms). (D) RMSD plots and average RMSD (angstrom) over time are shown for the 10 ns equilibration runs of the ACE2 model in both its SARS-CoV-2 bound and unbound states.

optimal configurations in the Dunbrack rotamer library, then using 2000 steepest gradient descent steps of energy minimization to relax the regions around the amino acid replacement. Mutant models were chosen to test the predicted sites functional role mitigating local binding between the viral RBD and ACE2 by comparing amino acid replacements with either very similar or very different sidechain properties. This is summarized in Table 2. The SARS-CoV-2 N501Y mutant model was constructed using PDB:6m17. This model of the viral RBD only extended from site 335-518 and therefore only included this single mutation (N501Y) from among the seven other amino acid replacements and two deletions found to occur on the B.1.1.7 lineage (The total set of mutations defining this variant are del 69-70, del 144, N501Y, A570D, D614G, P681H, T716I and S982A). Because of its proximity to the binding interface, the N501Y mutation is believed to be one of the most functional. In the B.1.351 and P.1 lineages, we were able to model three mutations in the viral RBD. These were N501Y, E484K and B417T/N.

Molecular dynamics simulation protocols

To facilitate the proper statistical confidence of the differences in rapid vibrational states in our model comparisons, many individual graphic processing unit (GPU) aMD simulations were produced to build large statistical replicate sets of

MD trajectories. aMD simulations were prepared with explicit solvation and conducted using the particle mesh Ewald method employed by pmemd.cuda in Amber18 (Case et al., 2005; Darden et al., 1993; Ewald, 1921; Pierce et al., 2012; Salomon-Ferrer et al., 2013) via the DROIDS 4.0 interface (Detecting Relative Outlier Impacts in Dynamic Simulation) (Babbitt et al., 2018, 2020). Simulations were run on a Linux Mint 19 operating system mounting two Nvidia Titan Xp graphics processors. Explicitly solvated protein systems were prepared using tLeap (Ambertools18) using the ff14SB protein force field (Maier et al., 2015) and protonated at all available sites using pdb4amber in Ambertools18. Solvation was generated using the Tip3P water model (Mao & Zhang, 2012) in a 12 nm octahedral water box and subsequent charge neutralization with $Na+$ and $Cl-$ ions. All comparative molecular dynamics binding analyses utilized the following protocol. After energy minimization, heating to 300 K, and 10 ns equilibration, an ensemble of 200 MD production runs each lasting 0.2 ns was created for both viral bound and unbound ACE2 receptor (Figure 1C). RMSD plots for the 10 ns equilibration runs in our primary models are also shown (Figure 1D). Each MD production run was preceded by a single random length short spacing run selected from a range of 0 to 0.1 ns to mitigate the effect of chaos (i.e. sensitivity to initial conditions) in the driving ensemble differences in the MD production runs. All MD was conducted using an

Table 1. Hybrid models comparatively analyzed in this study.

Effect of … on MD	Bound structure	Unbound structure	PDB IDs utilized to create model	Figure panel(s)
Viral binding	SARS-CoV-2 RBD (wildtype and N501Y) and ACE2	ACE2	6m17	Figure 5(A)–(C)
Viral binding	SARS-CoV-1 RBD and ACE2	ACE2	6acg	Figure 5(D)–(E)
Viral binding	hCoV-HKU1 RBD and ACE2	ACE2	5gnb and 6m17	Figure 7(D) and 4(E)
Viral binding	hCoV-OC43 RBD and ACE2	ACE2	6ohw and 6m17	Figure 7(A)–(C)
Viral binding	bat CoV-HKU4 and CD26	CD26	4qzv	Figure 8(A)–(C)
Viral binding	bat CoV-HKU4 and ACE2	ACE2	4qzv and 6m17	Figure 8(D) and 3(E)
Viral binding	MERS and ACE2/CD26	CD26	5x5c and 4qzv	Figure S4

Structural details, PDB IDs and relevant figures in the manuscript are also given.

Table 2. *In silico* targeted mutagenesis studies of the SARS-CoV-2 receptor-spike interaction with various amino acid mutations.

Protein	*In silico* mutation	N-terminal helix	Q325	K353	386-AAQPFLL-392
	Wild type	D	NC	D	D
ACE2	K353A	NC	NC	D	D
ACE2	386-AA/EE-387	D	D	D	D
ACE2	390-FLL/EEE-392	D	NC	D	D
ACE2	V739E	NC	NC	D	NC
ACE2	V739L	D	D	D	NC
ACE2	K408A	D	NC	D	D
RBD	V185E	NC	NC	D	NC
RBD	V185L	D	D	D	NC

Mutations were conducted with the *swapaa* command in UCSF Chimera, and bound structures were minimized with 2000 steps of steepest descent. D (blue) signifies dampened molecular motion, NC (yellow) indicates no significant change in molecular motion as a result of binding, and A (orange) indicates an amplification in molecular motion after RBD binding. Significance criteria were determined as having $<=-1$ KL divergence or less (dampened motion during binding), or $>=1$ KL divergence or more (amplified motion during binding). N-terminal helix and AAQPFLL motif categorization was determined as dampened or amplified when most of their constituent amino acids had a KL divergence of $<=-1$ or $>=1$, respectively.

Andersen thermostat under a constant pressure of 1 atmosphere (Andersen, 1980). Root mean square atom fluctuations (rmsf) and atom correlations were calculated using the atomicfluct and atomiccorr functions in CPPTRAJ (Roe & Cheatham, 2013).

Comparative dynamics of bound/unbound and mutant/wild type functional states

Computational modeling of bound and unbound protein structure dynamics can identify both proximal and distal differences in molecular motion resulting from the binding interaction. In the interest of studying functional dynamics, we focus on regions with highly dampened molecular motion at the interface between the viral RBDs and ACE2. Given the chaotic nature of individual molecular dynamics simulations, ensembles on the order of hundreds of samples are required for comparing functional states and deriving statistically significant differences between them. Furthermore, binding interactions can potentially be disrupted or prevented entirely by dynamic thermal noise that exists inherently in the system. DROIDS software employs a relative entropy (or Kullback-Leibler or KL divergence) calculation on atom fluctuation distributions derived from large numbers of replicates of MD trajectories to quantify the local differences in both the magnitude and shape of the distribution of atom fluctuations in two different functional states of

a protein (NOTE: this refers to Shannon entropy in the context of information theory, not entropy in the context of thermodynamics). In this application, we compare viral bound versus unbound dynamics of target proteins to ascertain a computational signature of the local degree of dampened target protein motions during viral binding (Figure 2). DROIDS software also employs a non-parametric two sample Kolmogorov-Smirnov statistical test of significance, generating a p-value for each amino acid site on ACE2 that indicates whether differences in atom fluctuation are significant when comparing the viral bound and unbound states. Because this test is conducted over multiple sites it must be subsequently corrected for false discovery rate via the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995). This procedure identifies where these dampened dynamics (i.e. negative signed/symmetric KL divergence) upon viral binding are truly and locally significant. In our study, the dampened atom fluctuations of target proteins ACE2 and CD26 upon binding are indicative of molecular recognition, in the sense that weak bonding interactions overcome baseline thermal motion and lead to a persistent functional state.

The molecular dynamics of the viral bound and unbound models of ACE2 were generated and statistically compared using the DROIDS 4.0 comparative protein dynamics software interface for Amber18 (Babbitt et al., 2018, 2020). The signed/symmetric Kullback-Leibler (KL) divergence or relative entropy (Kullback & Leibler, 1951) between the distributions of atom fluctuation (i.e. root mean square fluctuation or rmsf taken from 0.01 ns time slices of total MD simulation time) on viral bound and unbound ACE2 were computed using DROIDS and color mapped to the protein backbone with individual amino acid resolution to the bound structures using a temperature scale (i.e. where red is hot or amplified fluctuation and blue is cold or dampened fluctuation). The rmsf value is thus

$$rmsf = \frac{1}{4}\sum_{i=N,C,C_\alpha,O}^{4}\sqrt{\frac{1}{n}*\sum_{j=1}^{n}(v_{jx}-w_x)^2+(v_{jy}-w_y)^2(v_{jz}-w_z)^2} \quad (1)$$

where v represents the set of XYZ atom coordinates for i backbone atoms (C, N, O, or Cα) for a given amino acid residue over j time points and w represents the average coordinate structure for each MD production run in a given ensemble. The Kullback Leibler divergence (i.e. relative entropy) or similarity between the root mean square fluctuation (rmsf) of two homologous atoms moving in two molecular dynamics simulations representing a functional binding interaction (i.e. where 0 = unbound state and 1 = viral bound state) can be described by
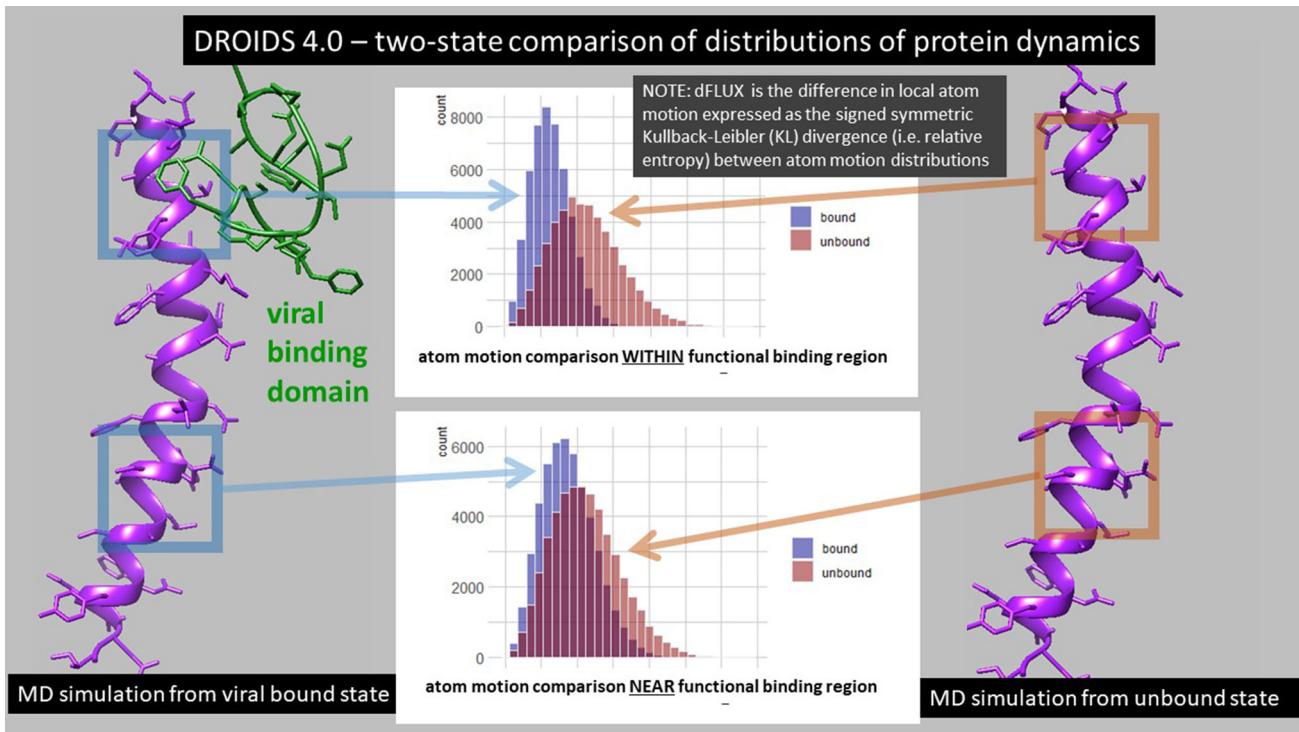
Figure 2. An overview of the comparative protein dynamics analysis with DROIDS 4.0. Two comparative functional protein states are defined and subjected to repeated sampling of molecular dynamics (MD) simulations (Figure 1C). The differences in local atom motions (dFLUX) are defined by Kullback-Leibler (KL) divergences in protein backbone specific distributions in root mean square atom fluctuations (rmsf) collected over the MD sampling of each functional state. In this study, we compare the target protein human ACE2 in both its viral bound (blue-left) and its native unbound states (red-right). The reduced energy of motion at or near functional binding sites of the various viral strains will negatively shift the velocity of atoms creating a corresponding negative KL divergence in statistical distribution of rmsf (i.e. negative dFLUX).

$$KL_{divergence} = \sum_{t=10ps}^{T} \left[\left(rmsf_0 * \log \frac{rmsf_0}{rmsf_1} \right) + \left(rmsf_1 * \log \frac{rmsf_1}{rmsf_0} \right) \right] / T$$

(2)

where rmsf represents the average root mean square deviation of a given atom over time. More specifically, the rmsf is a directionless root mean square fluctuation sampled over an ensemble of MD runs with similar time slice intervals. Because mutational events at the protein level are typically amino acid replacements, this calculation is more useful if applied to resolution of single amino acids rather than single atoms. Because only the 4 protein backbone atoms (N, Cα, C and O) are homologous between residues, the R groups or side chains are ignored in the calculation and equation 2 can be applied. Since the sidechain atoms always attach to this backbone, rmsf still indirectly samples the dynamic effect of amino acid sidechain replacement as they are still present in the simulation. The reference state of the protein is unbound while the query state is viral bound. Therefore, this pairwise comparison represents the functional impact of viral binding on the ACE2 protein's normal unbound motion, where it is expected that viral contact would typically dampen the fluctuation of atoms at the sites of binding to some degree. This calculation in equation 2 is used to derive all the color maps and KL divergence (i.e. dFLUX) values presented in the Figures 5–8. In these maps, a negative dFLUX, indicated on the PDB surface map in blue, represents dampened atom fluctuation occurring due to protein-protein binding interaction. Multiple test-corrected two sample Kolmogorov-

Smirnov tests are used to determine the statistical significance of local site-wise differences in the rmsf distributions in viral bound and unbound ACE2 models. The test was applied independently to each amino acid site. The Benjamini-Hochberg method was used to all the adjust p-values for the false discovery rate generated from the multiple sites of a given protein structure (Benjamini & Hochberg, 1995).

Machine learning to detect functionally conserved binding dynamics

We employed maxDemon 3.0 (Figure 3) for the machine learning detection of the functionally conserved dynamics of the binding interaction between SAR-CoV-2 and ACE2, the viral bound and unbound ACE2 structures modeled from PDB ID: 6m17 were each subjected to MD sampling consisting of 100 production runs each for 1.0 ns time, after 1 ns heating and 10 ns equilibration. Each production run was subdivided into short 50 frame time slices, from which subsequent atom fluctuation and atom correlations were computed using the atomicfluct and atomiccorr functions in cpptraj (Roe & Cheatham, 2013) with a single residue mask applied only to the 4 atom types on protein backbone (N. Cα, C, O). The atom correlations for a given amino acid site were parsed only to include correlations to downstream sites at $i+1$, $i+3$, $i+5$, and $i+9$ positions distant. The pre-classifications of each time slice were simply determined from whether they originated from the viral bound or unbound
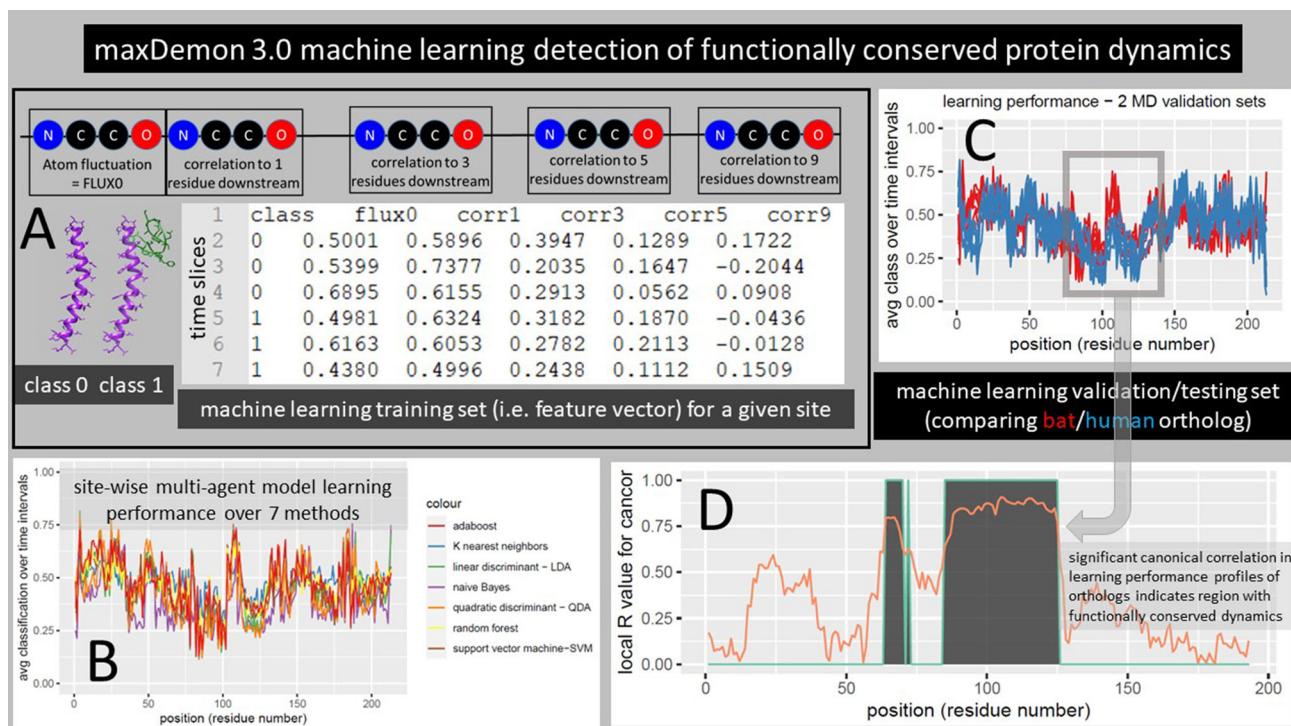
Figure 3. Detecting functionally conserved dynamics with the maxDemon 3.0 machine learning algorithm. (A) The MD simulation sets generated above are now the source of a pre-classified machine learning training set where class 0 and class 1 represent the binary functional states of the protein interaction (unbound vs. bound; Figure 1C). A feature vector is constructed comprising the atom fluctuations of any given amino acid (AA) site and its correlations to motions of downstream sites located 1, 3, 5, and 9 AA sites distant taken at every 50 frame intervals or 'time slices' in the MD simulations. This rich feature vector thus captures both the local energy on the protein backbone, as well as potential sequence-based information that coordinates specific motions with nearby AA sites. (B) A multi-agent classifier comprised of up to seven learning methods is individually deployed on each site and a learning performance profile across all sites on the protein is generated as the time average classification of the site-specific dynamics defined by the feature vector. Thus, successful learning of functional dynamic features will deviate from an average classification of 0.5 on the profile. (C) This learning profile is then modeled for two evolutionary homologs (e.g. human ACE2 bound to the recent emergent strain SARS-CoV-2 and bound to bat CoV-HKU4, a potential zoonotic precursor). (D) A local canonical correlation of the learning performance profiles from both homologs is taken over 20 adjacent AA sites from a sliding window moved across the protein. Significant correlation of the homologous learning profiles identifies regions (shown in gray) where the functional binding dynamics defined by the pre-classification is functionally conserved from the time of divergence of the viral homologs.

simulation of ACE2. Thus, a feature vector for training the machine learning model on a given site (Figure 3A) was comprised of the pre-classifications (0 = unbound, 1 = viral bound), the atom fluctuations at the given site (flux0) and the downstream correlations (corr1, corr3, corr5 and corr9). Two new single MD simulations were conducted (1 ns heating, 10 ns equilibration and 3 ns production) on the zoonotic progenitor model bat CoV-HKU4 bound to human ACE2, and new runs of SARS-CoV-2 bound to ACE2. These MD runs were subjected to time slicing and feature pre-classification as in the training set and then were used as validation runs for the machine learning. Because they represent the dynamic states of evolutionary orthologs, any local similarity of learning performance shared between them will represent functionally important binding dynamics, defined in the training set, that has persisted and been conserved over the time of evolutionary divergence of the virus. A multi-agent or 'stacked' model of shallow learners was individually applied to each amino acid site in the ortholog validation runs and learning performance profiles for each learning method were calculated as the average classification state over all the time slices (NOTE: a learning performance = 0.5 indicates no functional dynamics is detected at that position). An example of the learning performance profiles for all seven learners is shown for one of the validation runs in Figure 3(B). The classification methods included K-nearest

neighbors, naïve Bayes, linear and quadratic discriminant analysis, support vector machine with radial basis function kernel, random forest with 500 decision trees, and adaptive gradient boosting. All machine learning classifications are conducted in R using the class, MASS, kernlab, randomForest, and ada packages (Breiman, 2001; Culp et al., 2016; Karatzoglou et al., 2004; Liaw & Wiener, 2002; Venables & Ripley, 2010). Functionally conserved viral binding dynamics between the orthologs was captured using a canonical correlation analysis (Härdle & Simar, 2007) applied to the 14 learning profiles (i.e. 7 for each ortholog in red and blue; Figure 3C) using a sliding window of 30 sites and an alpha level of significance of 0.01. An example of the local R-value and regions of significance (dark grey) are shown in Figure 3(D). For more details, see our software publications (Babbitt et al., 2018, 2020).

Information theoretics to detect variant impact and amino acid site coordination in conserved dynamics

After the functionally or evolutionary conserved dynamic regions are determined, the DROIDS/maxDemon software offers two new information theoretical calculations that can be applied to subsequent MD simulations conducted upon genetic or drug class/ligand binding variants of interest. The first metric is an information gain or relative entropy of correlation
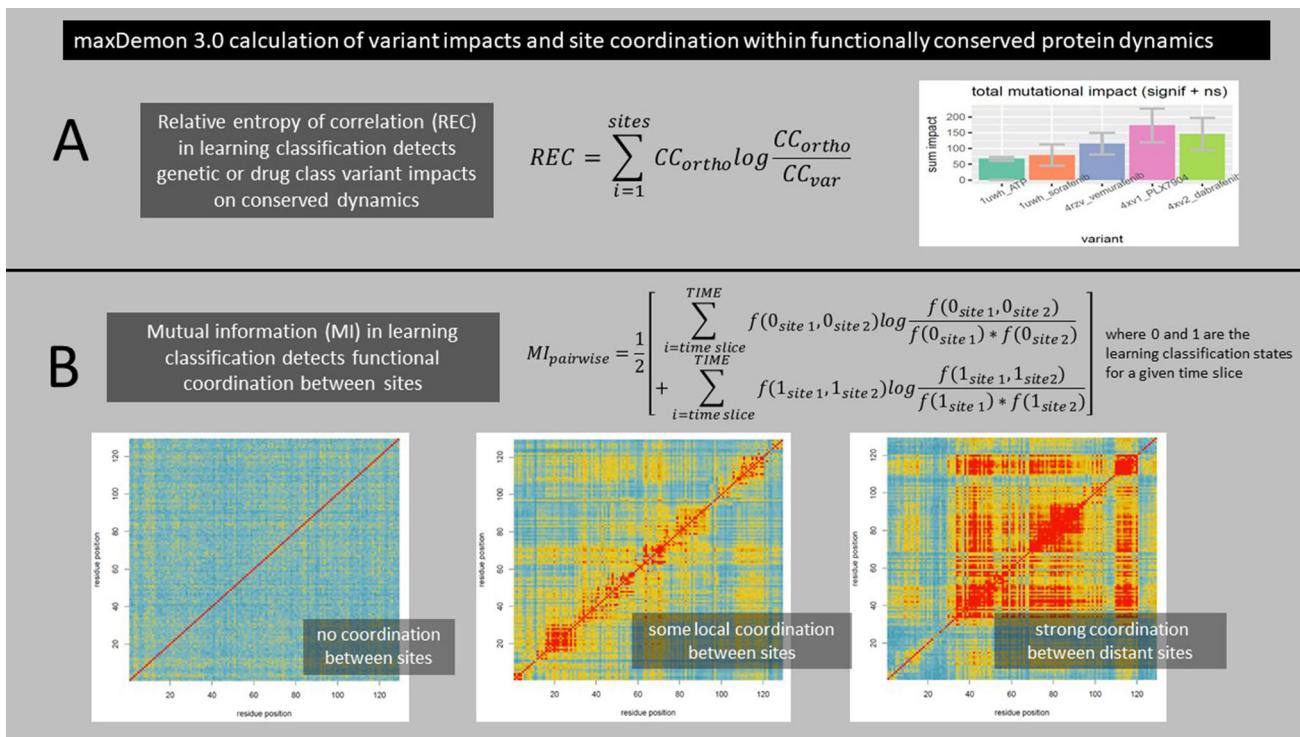
Figure 4. Information theoretics for measuring variant impacts and functional coordination between sites. Further information theoretics applied to the predicted binary classifications (0 or 1) of new MD runs can be used to compare relative impacts of different variants upon the signal of the conserved dynamics and to detect specific sites of functional coordination of conserved dynamics on the protein. (A) A relative entropy of the correlation (REC) of learning performance profiles can be used to measure both the local and total impact of genetic variation in the viral receptor binding domain on the functionally conserved binding dynamics of the bat/human orthologs. (B) Mutual information between the predicted binary states of learning classification (0 or 1) between any two amino acid sites, collected over the time of the simulations, can be used to detect functionally relevant coordination of the binding dynamics between the two sites. The mutual information is reported for all pairwise comparisons on the ACE2 protein as a heatmap with a scale blue/yellow/red indicating increasing coordination of functional binding dynamics.

(REC) which can be used to compare the relative impacts of different variants on the functionally conserved dynamics defined in the section above (Figure 4A). This is a relative entropy calculation comparing the R values of the canonical correlations of each variant to that of the homologous functional validation described in the section above. REC calculations are returned locally in a sliding window as a line plot for each variants, and it is calculated globally over the whole protein backbone for each variant and returned as a bootstrapped bar plot with a one-way ANOVA test comparing all variants.

$$REC = \sum_{i=1}^{sites} CC_{ortho} \log \frac{CC_{ortho}}{CC_{var}} \quad (3)$$

A second information theoretic used is a mutual information (MI) for all pairwise site comparisons calculated using the learning classifications of all the time slices in a subsequently deployed MD simulation (Figure 4B). This calculation captures the frequency of the co-occurrence of learning classifications at any two given sites compared to the independent frequencies of the learning classifications at the two sites.

$$MI_{pairwise} = \frac{1}{2} \left[\begin{array}{c} \sum_{i=timeslice}^{TIME} f(0_{site1}, 0_{site2}) \log \frac{f(0_{site1}, 0_{site2})}{f(0_{site1}) * f(0_{site2})} \\ + \sum_{i=timeslice}^{TIME} f(1_{site1}, 1_{site2}) \log \frac{f(1_{site1}, 1_{site2})}{f(1_{site1}) * f(1_{site2})} \end{array} \right] \quad (4)$$

All site-wise comparisons are presented in a heatmap with blue/yellow/red scale that indicates the level of

co-occurrence of functional dynamic states. This provides a useful way of quantifying the amount of site coordination in functional dynamics which can be caused by allostery and/or cooperative binding interactions.

Results

Characterization of the SARS-CoV-2 to ACE2 binding signatures in the original and N501Y mutant strain

The modeled region of binding interaction consisted of the extracellular domain region of human ACE2, and the viral RBD specified in Figure 1(A). The main key sites of the ACE2 interface with the viral RBD are shown in Figure 1(B), with the N-terminal helices labeled in green and the three main sites that can potentially interact with the viral RBD N501 site are shown in magenta. The DROIDS pipeline (Babbitt et al., 2018, 2020) used to characterize the changes in ACE2 molecular dynamics due to viral binding is summarized Figure 1(C). More details about this can be found in the methods section here as well as in our two software notes cited above. All the binding models developed and analyzed in this study are listed in Table 1. Comparisons of the molecular dynamics of SARS-CoV-2 RBD bound ACE2 to unbound ACE2 after long range equilibration revealed extensive dampening of atom fluctuations (shown color mapped in blue) at the ACE2 N-terminal helices as well as three downstream sites on ACE2 focused at Q325, K353, and a
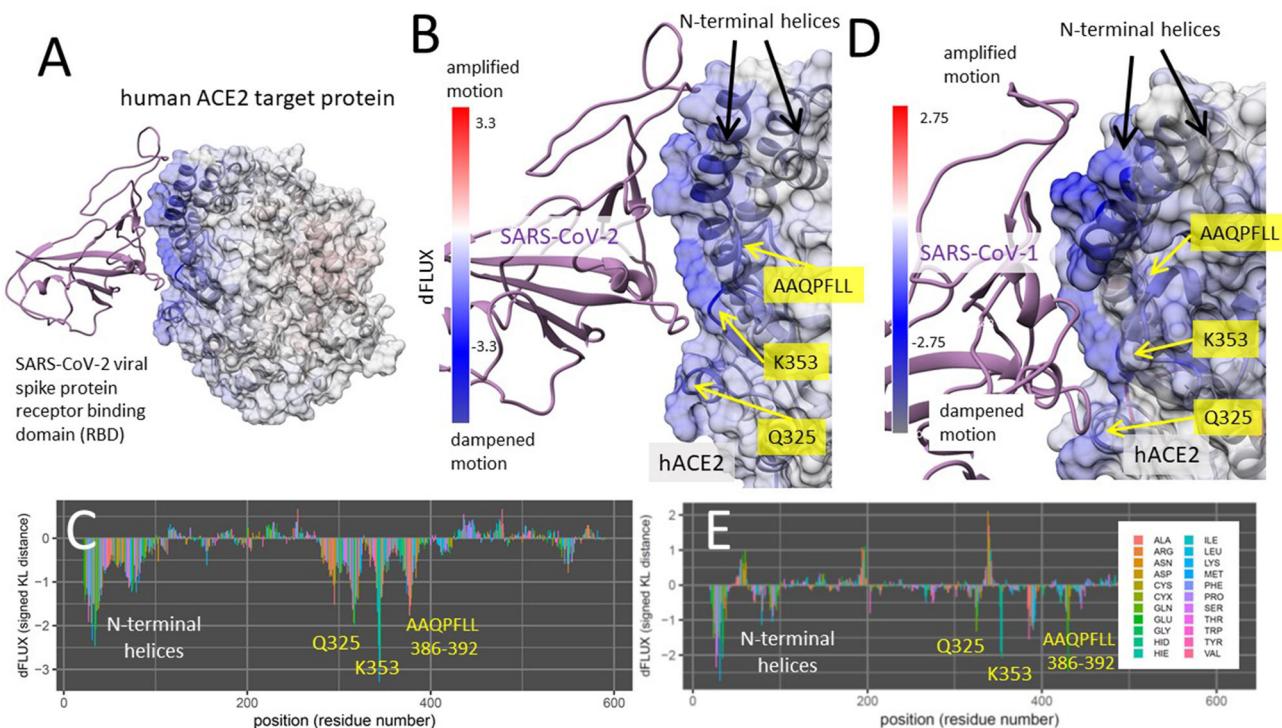
Figure 5. DROIDS binding signature of dampened atom fluctuations in human ACE2 receptor protein upon interaction with both SARS-CoV-2 spike glycoprotein (modeled from PDB: 6m17) and the upon interaction with the past human outbreak strain SARS-CoV-1 spike glycoprotein (modeled from PDB: 6acg). Shift in atom fluctuations were calculated as the signed symmetric Kullback-Leibler (KL) divergence (i.e. relative entropy or dFLUX) of distributions of the root mean square time deviations (rmsf) for residue averaged protein backbone atoms (i.e. N, Cα, C and O) for ACE2 in its spike bound (PDB: 6m17/6acg) versus unbound dynamic state (PDB: 6m17/6acg without the viral receptor binding domain or RBD). Here we show color mapping (A, B, D) and sequence positional plotting (C, E) of dampening of atom motion on viral RBD-protein target interface in blue for (A–C) the original strain of SARS-CoV-2 and (D, E) its predecessor, SARS-CoV-1. The sequence profile of the KL divergence between SARS-CoV-2 viral bound and unbound ACE2 produces strong negative peaks indicating key residue binding interactions with the N-terminal helices (white) and the N501 interactions at Q325, K353 and the AAQPLL 386-392 motif (yellow). The interactions at these sites are more moderated in the SARS-CoV-1 interaction with the ACE2 target protein.

novel motif AAQPFLL 386-92 (Figure 5A–C). These color mapped structures show the local site-wise mathematical divergence of protein backbone atom fluctuation (i.e. root mean square fluctuation averaged over N, C, Cα, and O) when comparing viral bound ACE2 dynamics to unbound ACE2 dynamics. They are presented for both the whole structural model (Figure 5A) and the viral binding interface with ACE2 (Figure 5B) along with sequence-based positional plotting of the signed symmetric Kullback-Leibler or KL divergence (Figure 5C). The site-wise positional profiles of average atom fluctuation in both the viral bound and unbound states (Figure S1), as well as the multiple test-corrected statistical significance of these KL divergences determined by a two-sample Kolmogorov-Smirnov or KS test is also given (Figure S2). The p-value given at each site is adjusted for the probability of false discovery that is governed by the length of the polypeptide.

Binding interaction models of past human outbreak strain SARS-CoV-1 in complex with ACE2 (Figure 5D and 5E) indicates that the binding signature of SARS-CoV-1 and SARS-CoV-2 is nearly identical with both having stable interaction with N-terminal helices of ACE2 and promiscuous interactions with K353 and the same novel sites Q325 and AAQPFLL 386-92. However, the SARS-CoV-1 interactions at these sites appear more moderate than in SARS-CoV-2. Again, the site-wise atom fluctuation profiles and KS tests of

significance of the differences in these fluctuations are also given (Figures S1 and S2).

Characterization of the SARS-CoV-2 to ACE2 binding signatures in the N501Y variant strains B.1.1.7 and B.1.351

An identical characterization of the ACE2 binding signature created by the reportedly more highly transmissible SARS-CoV-2 B.1.1.7 (UK) variant indicated a similar yet stronger binding profile at the N-terminal helices, with more attraction to the C helix (Figure 6A–C). The binding to the ACE2 K353 by the viral RBD Y501 was also far more stable while the former promiscuous interactions with Q325 and AAQPFLL 386-92 were markedly reduced (Figure 6B and 6C, with point of the viral RBD mutation N501Y marked by the green star in Figure 3B). In the B.1.351 (South African) variant, we also find a more stable binding pattern overall that involves both the N-terminal helices and K353 site in ACE2 (Figure 6D and 6E). In this variant, we observed a retention of promiscuous binding of Y501 with the K353, Q325 and the AAQPFLL motif as well (Figure 6E). The three mutations prevalent in the RBD of the B.1.351 and P.1 variant are shown with a green star in Figure 6(D). These are N501Y, E484K and K417N/T. Both variants display an extraction of the K353 side chain away from the body of ACE2 and
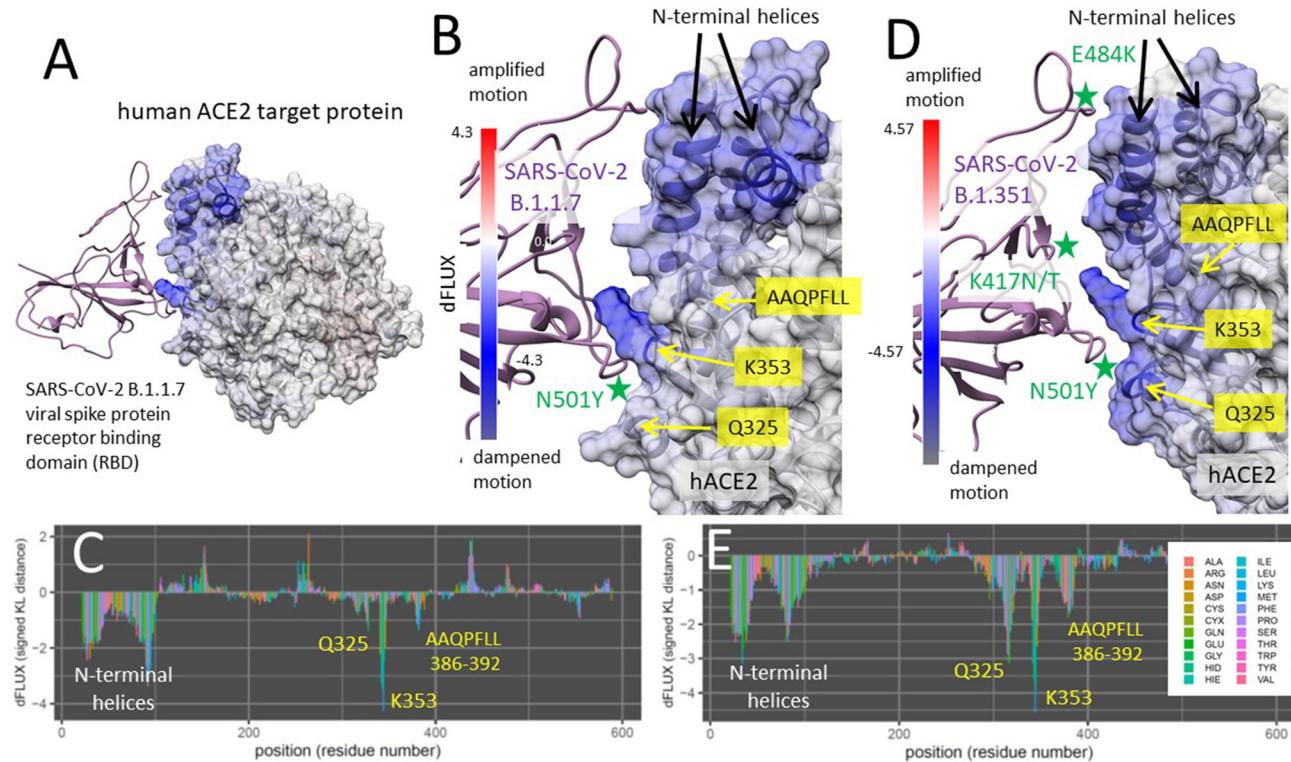
Figure 6. DROIDS binding signature of dampened atom fluctuations in human ACE2 receptor protein upon interaction with both the highly transmissible B.1.1.7 (UK) variant and the B.1.351 (South African and Brazilian) variant of SARS-CoV-2 spike glycoprotein (modeled from PDB: 6m17). Shift in atom fluctuations were calculated as the signed symmetric Kullback-Leibler (KL) divergence (i.e. relative entropy or dFLUX) of distributions of the root mean square time deviations (rmsf) for residue averaged protein backbone atoms (i.e. N, Cα, C and O) for ACE2 in its spike bound (PDB: 6m17) versus unbound dynamic state (PDB: 6m17 without the viral receptor binding domain or RBD). Here we show color mapping (A, B, D) and sequence positional plotting (C, E) of dampening of atom motion on viral RBD-protein target interface in blue for (A–C) the B.1.1.7 variant and (D, E) the B.1.351 variant of SARS-CoV-2. The sequence profile of the KL divergence between viral bound and unbound ACE2 produces negative peaks indicating key residue binding interactions with the N-terminal helices (white) and the N501 interactions at Q325, K353 and the AAQPLL 386-392 motif (yellow). The N501Y, E484K and K417N/T mutations are marked by the green stars (2B and 2D).

towards the viral RBD Y501 (Figure 6B and 6D) that is generally maintained throughout the MD simulations.

Characterization of binding signatures in human endemic strains and past emergent outbreaks

Binding interaction models of human endemic strains hCoV-OC43 and hCoV-HKU1 in complex with ACE2 were also analyzed identically to previous models. The interaction with the most pathologically benign and longest co-evolved strain hCoV-OC43 (Figure 7A–C) shows highly enhanced dampened motion due to binding only at the N-terminal helices of ACE2, while the interaction with hCoV-HKU1 demonstrates this as well, but also adds moderate promiscuous interactions of N501 with K353 and Q325 (Figure 7D and 7E). The enhanced binding at the N-terminal helix is clearly associated with the evolution of extended loop structure of the viral RBD at this region of the interface (Figure 7A). The site-wise atom fluctuation profiles and KS tests of significance of the differences in these fluctuations are also given (Figures S1C, S1D, S2C, and S2D). Binding interaction models of past human outbreak strain MERS-CoV in complex with both CD26 and ACE2 (Figure S4) indicate a two-touch interaction with CD26 sites SAMLI 291-5 and SS 333-4 (Figure S4B and S4C). The MERS-CoV/ACE2 interaction indicates a strong

interaction with the N-terminal helices and only a weak interaction with K353 (Figure S4D and S4E). In the MERS-CoV MD simulations, we observe a similar maintenance of beta sheet over the SAMLI 291-5 region in CD26, however this is partially lost over the N-terminal helices when the MERS-CoV is presented to ACE2.

We also examined the dampening of atom fluctuations in the viral receptor binding domains (RBD's) of SARS-CoV-2 and the two human endemic strains (Figure S5). We find that in general, all of the viral RBD's show stronger negative dFLUX than their ACE2 targets, indicative of their thermodynamic favoring of the target-bound state as one would expect in a viral system. Interestingly though, while we find that the SARS-CoV-2 RBD shows strong universal dampening across the whole RBD (Figure S5A), the human strains demonstrate the strongest change in dynamics at the enhanced loop regions of the RBD (Figure S5B and S5C), perhaps suggestive of the evolution of a disordered structure that becomes ordered only as it binds its target. Such a mechanism could obviously help the virus evade immune recognition, and we note that this has evolved in the same location in the endemic strains as the site of the E484K mutation that is potentially linked to the SARS-CoV-2 variants that appear to have partially escaped the Pfizer Comirnaty vaccine (Planas et al., 2021). Also of interest, is a common small
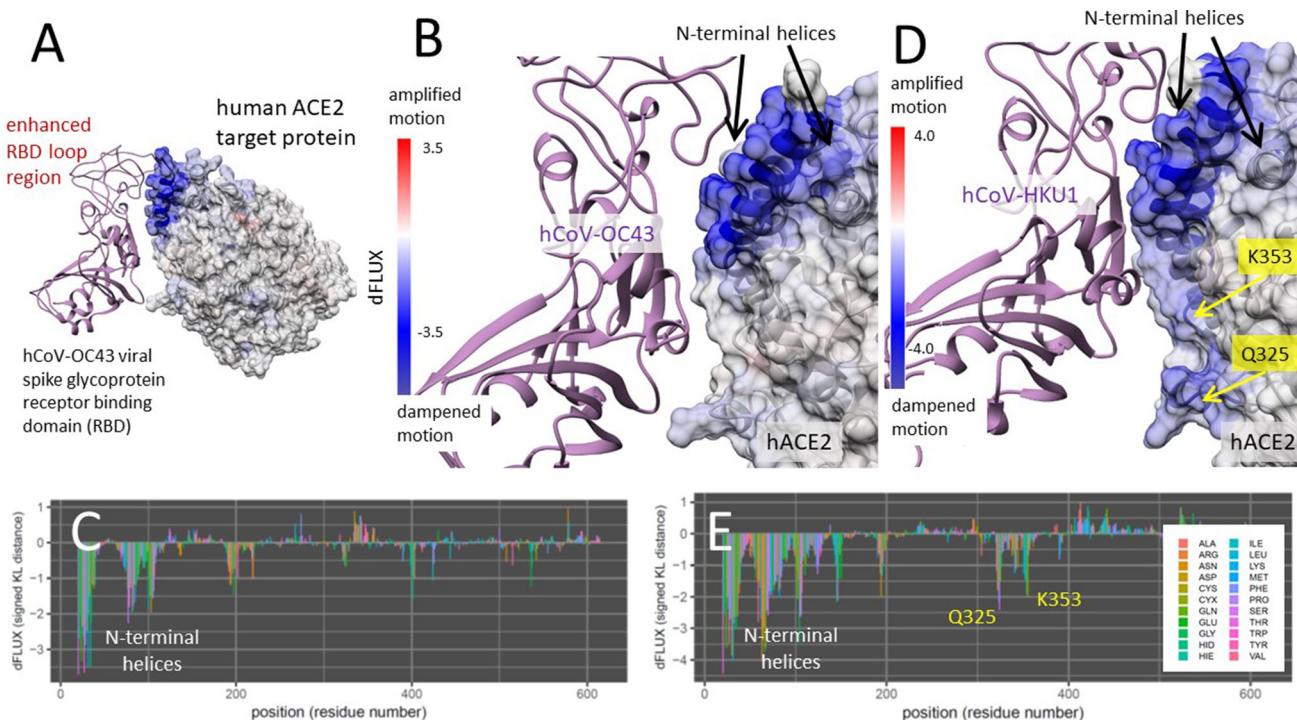
Figure 7. DROIDS binding signature of dampened atom fluctuations in human ACE2 receptor proteins upon interaction with two human endemic strains hCoV-OC43 and hCoV-HKU1 glycoprotein (modeled from PDB: 6ohw, 5gnb and 6m17). Here we show color mapping (A, B, D) and sequence positional plotting (C, E) of dampening of atom motion on the viral RBD-protein target interface in blue for (A–C) the targeting of ACE2 by the most benign strain hCoV-OC43 and (D, E) its less benign counterpart hCoV-HKU1. The sequence profile of the KL divergence between viral bound and unbound target proteins produces strong negative peak indicating key residue binding interactions (C, E) with the N-terminal helices on ACE2 and (E) moderate interactions with K353 and Q325 observed only in hCoV-HKU1.
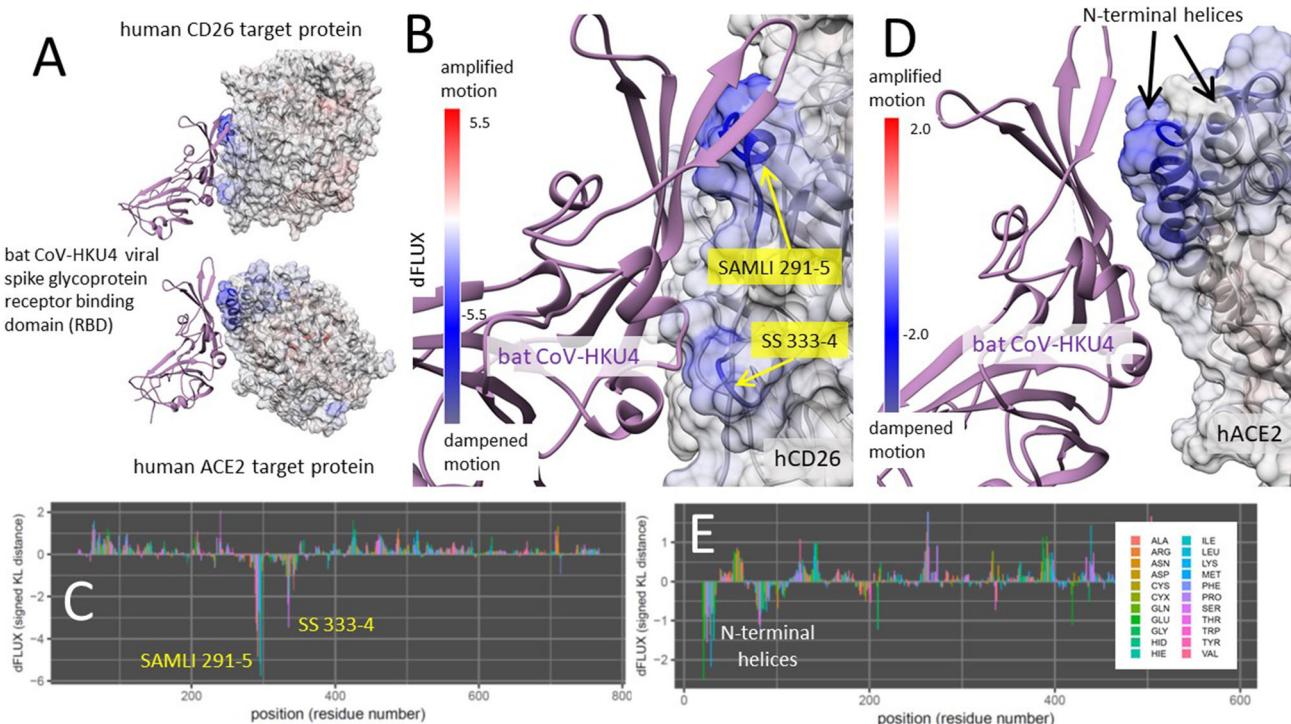

Figure 8. DROIDS binding signature of dampened atom fluctuations in human CD26 and ACE2 receptor proteins upon interaction with the bat progenitor strain batCoV-HKU4 glycoprotein (modeled from PDB: 4qzv and 6m17). Here we show color mapping (A, B, D) and sequence positional plotting (C, E) of dampening of atom motion on the viral RBD-protein target interface in blue for (A–C) its known targeting of CD26 and (A, D, E) its hypothetical targeting of ACE2. The sequence profile of the KL divergence between viral bound and unbound target proteins produces two strong negative peaks indicating key residue binding interactions with the (C) the SAMLI and double serine motifs (shown in yellow) on CD26 and (E) the N-terminal helices on ACE2.

region of amplified atom fluctuation that occurs upon binding in all three RBD's (and labelled with orange question mark in Figure S5A).

Characterization of binding signatures of the bat progenitor strain HKU4 with its primary target CD26 and potential zoonotic spillover target ACE2

Binding interaction models of *Tylonycteris* bat progenitor strain batCoV-HKU4 RBD in complex with CD26 (its normal target) and human ACE2 were also analyzed in the same manner as SARS-CoV-2 (Figure 8). The bat coronavirus HKU4 demonstrated strongly and significantly dampened molecular motion in two precise regions of CD26, the SAMLI 291-295 motif, and the double serine motif at SS-333-4 (Figure 8A–C). However, upon interacting with ACE2, the bat coronavirus RBD only dampened the atom fluctuation at a broader region of the ACE2 N-terminal helices (Figure 8D and 8E) utilizing the same region that interacts with the CD26 SAMLI motif. Thus, the interaction with the ACE2 N-terminal helix appears pre-adapted by the evolution of the interaction with SAMLI on CD26. The site-wise atom fluctuation profiles and KS tests of significance of the differences in these fluctuations are also given (Figure S3). In both comparative MD simulations, we observe that the bat CoV-HKU4 viral RBD maintains more secondary structure in the form of anti-parallel beta sheet over the both the SAMLI 291-5 stie in CD26 and the N-terminal helices in ACE2.

Summary of the comparative protein dynamic surveys

In comparisons of molecular dynamics simulations between four strains of human-pathogenic coronaviruses, SARS-CoV-2, SARS-CoV-1, hCoV-HKU1, and hCoV-OC43, all four RBDs were associated with statistically significant dampening of molecular motion in the N-terminal helices of human ACE2, indicated by blue color mapping corresponding to a negative Kullback-Leibler (KL) divergence on the ACE2 PDB structure mapped at single amino acid resolution (Figures 2–6). With the exceptions of bat CoV-HKU4 and hCoV-OC43, all RBDs were also associated with dampened molecular motion of ACE2 at residue K353. The absence of a secondary interaction between hCoV-OC43 and ACE2 is further supported by multiple sequence alignment of the RBD loop region proximal to the K353 site, where hCoV-OC43 is the only surveyed strain with a polar uncharged threonine residue in place of a small aliphatic residue (Figure S6). A significant dampening effect is also present in two novel sites near to K353, the Q325 and 386-AAQPFLL-392 motif, when interacting with SARS-CoV-1 and SARS-CoV-2 (Figure 5), indicating a transiently dynamic or promiscuous interaction of the same part of the viral RBD with these three key ACE2 binding sites. In the recent evolution of the far more transmissible variants (Figure 6), the stabilization of RBD binding interaction with the N-terminal helices and sites Q325, K353, and 386-AAQPFLL-392 in ACE2 while favoring a more stable K353 interaction highlights an important molecular explanation that linking population transmission rate with the evolution of increased binding specificity in these new variants.

In silico mutagenesis study of hACE2 and SARS-CoV-2 bound dynamics

To confirm the relative importance of the three promiscuous sites in the binding interaction of coronavirus RBDs to human ACE2, a total of eight *in silico* targeted mutagenesis studies were performed, on ACE2 and the viral RBD alone (Table 2). K353A, a mutation previously identified to abolish SARS-CoV-1 RBD interaction with ACE2, neutralized dampening of the N-terminal helices during binding with the SARS-CoV-2 RBD (Figure S7) supporting the observation of decreased infectivity in the mutant (W. Li et al., 2005). Within the newly identified 386-AAQPFLL-392 motif, a mutation of the nonpolar N-terminal double alanine to the charged polar glutamic acid resulted in an additional strong dampening of molecular motion in the ancestral Q325 site, with a KL divergence of less than -3 between the unbound and bound states. Mutating the C-terminal FLL residues to EEE in the 386-AAQPFLL-392 motif resulted in no overall change to the binding characteristics between the SARS-CoV-2 RBD and ACE2. Upon mutating V739 in ACE2 to hydrophilic glutamate, the N-terminal helix region and 386-AAQPFLL-392 motif showed no significant change in molecular motion between bound and unbound states (Figure S8). This same pattern of binding characteristics could be seen in the mutation of V185 to glutamate in the loop region of the SARS-CoV-2 RBD, proximal to ACE2 during binding (Figure S9A). Mutating V185 to leucine in the RBD resulted in the same binding characteristics, resulting in a pattern of dampening in molecular motion that was the same as seen in SARS-CoV-1 (Figure S9B). All surveyed mutations maintained significant dampening in K353 of ACE2 during binding. The mutation of V739 in ACE2 to leucine resulted in dampening of Q325 in the interaction, as is seen in SARS-CoV-1, and a lack of dampening in the AAQPFLL motif as seen in the SARS-CoV-2 wild type. However, mutating V739 in ACE2 to glutamate resulted in pronounced dampening only at the N-terminal helices and K353. Upon mutating polar K408 in the SARS-CoV-2 RBD to nonpolar alanine, the Q325 interaction disappeared and only interactions with N-terminal helices, K353, and 382-AAQPFLL-392 remained (Figure S10).

Machine learning-based detection of functionally conserved ACE2 binding dynamics at sites shared between human/bat Betacoronavirus orthologs

The maxDemon algorithm detected significant regional canonical correlations in the machine learning classification performance profiles of bat/human and SARS/SARS variant orthologs, indicating regions of functionally conserved dynamics that was recognizable from thermal noise and shared between the molecular dynamics modeling of even distant orthologs (Figure 9A–C). The most distant evolutionary comparison of bat CoV-HKU4 to SARS-CoV-2 identified functionally conserved dynamics at 4 highly localized sites including the region surrounding K353 (Figure 9A). The comparison of the bat CoV-HKU4 to human hCoV-HKU1 identified functionally conserved dynamics at 2 of the 3 N-terminal helices, as well as two of the same highly localized regions
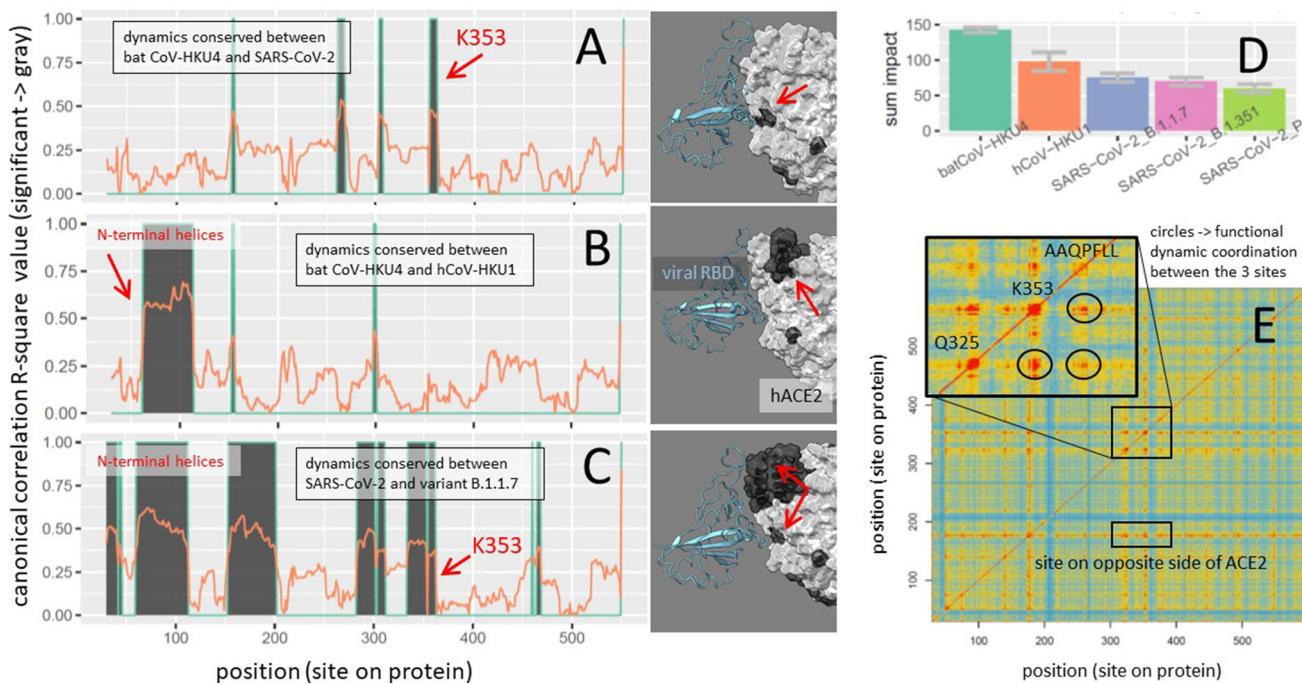
Figure 9. Functionally conserved and coordinated binding dynamics. (A–C) Maps of evolutionary conserved functional binding dynamics (shown in dark gray) of viral RBD on both ACE2 sequence and structure. Three different evolutionary distances were considered including the most distant comparison between SARS-CoV-2 and (A) the hypothetical bat CoV-HKU4 progenitor, (B) the human endemic strain hCoV-HKU1, and (C) the UK variant, SARS-CoV-2 B.1.1.7. (D) These evolutionary distances are also represented as the total sum of the mutational impacts on conserved dynamics. (E) The coordination of the functionally conserved binding dynamics of ACE2 upon interactions with SARS-CoV-2 receptor binding domain is shown as a matrix with red/blue indicating high/low pairwise coordination (i.e. mutual information in learning classifications) between sites. The inset highlights the high degree of coordination occurring between K353, Q325 and the AAQPFLL motif in their dynamic interactions with the viral N501 site.

observed in the comparison of the bat progenitor to SARS-CoV-2 (Figure 9B). However, the dynamics of the K353 site was not conserved in the comparison with the human endemic strain (i.e. the K353 dynamics observed in the emergent strains was not identifiable by machine learning in the endemic strain). In the closest evolutionary comparison of the dynamics of SARS-CoV-2 and its B.1.1.7 variant, we observed larger regions of functionally conserved dynamics that includes all the regions observed in the more distant comparisons (Figure 9C). These regions included all the N-terminal helices, as well as the regions connecting the key sites at K353 and Q325. The protein dynamics from position 150–200 and 280 to 308 were also significantly conserved. The consistency of the detection algorithm across these different simulations, as well as the decrease in functionally conserved dynamics with increased evolutionary distance, very reminiscent of DNA sequence-based methods, was highly encouraging despite the 'black box' nature of machine learning approaches to complex identification problems such as those posed when trying to compare molecular dynamic states of protein function.

Information theoretics for comparing variants and detecting functional coordination in ACE2 sites of interest

The total impact (in terms of relative entropy of correlation) of the genetic differences among the *Betacoronovirus* strains upon the functionally conserved dynamics detected by the machine learning model followed a quite predictable pattern,

with more evolutionary distant comparisons exhibiting more total impact on dynamics (Figure 9D). The mutual information in learning classifications between sites also captured the predicted coordination in functional binding that we expected to see in the dynamics at K353, Q325 and 382-AAQPFLL-392 (Figure 9E). Not only were the functional dynamics coordinated across amino acid sites near each of these binding events (red regions on the diagonal of the heatmap) but they were coordinated across these three sites as well (at circled intersections off the diagonal). A two-dimensional representation of the machine learning pre-classifications and discriminant learning space for the training data at each of the three key sites and a site on the N-terminal helices is also shown (Figure 10). Here, we can observe that the K353 site clearly has the highest signal to noise ratio regarding the functional binding of SARS-CoV-2 RBD and human ACE2 (i.e. observable in the contours and stronger point separation of the time-sliced dynamics of the viral bound and unbound states).

A model of the functional evolution of ACE2 binding interactions during zoonotic spillover

Based on our overall results comparing regions of functionally conserved dampened molecular motion in ACE2 and CD26 target proteins across zoonotic, emergent, and endemic *Betacoronavirus* strains, we propose a model of protein interaction for coronavirus RBD evolution and human specificity highlighting the role of the two distinct touch points between the viral RBD and the human target proteins
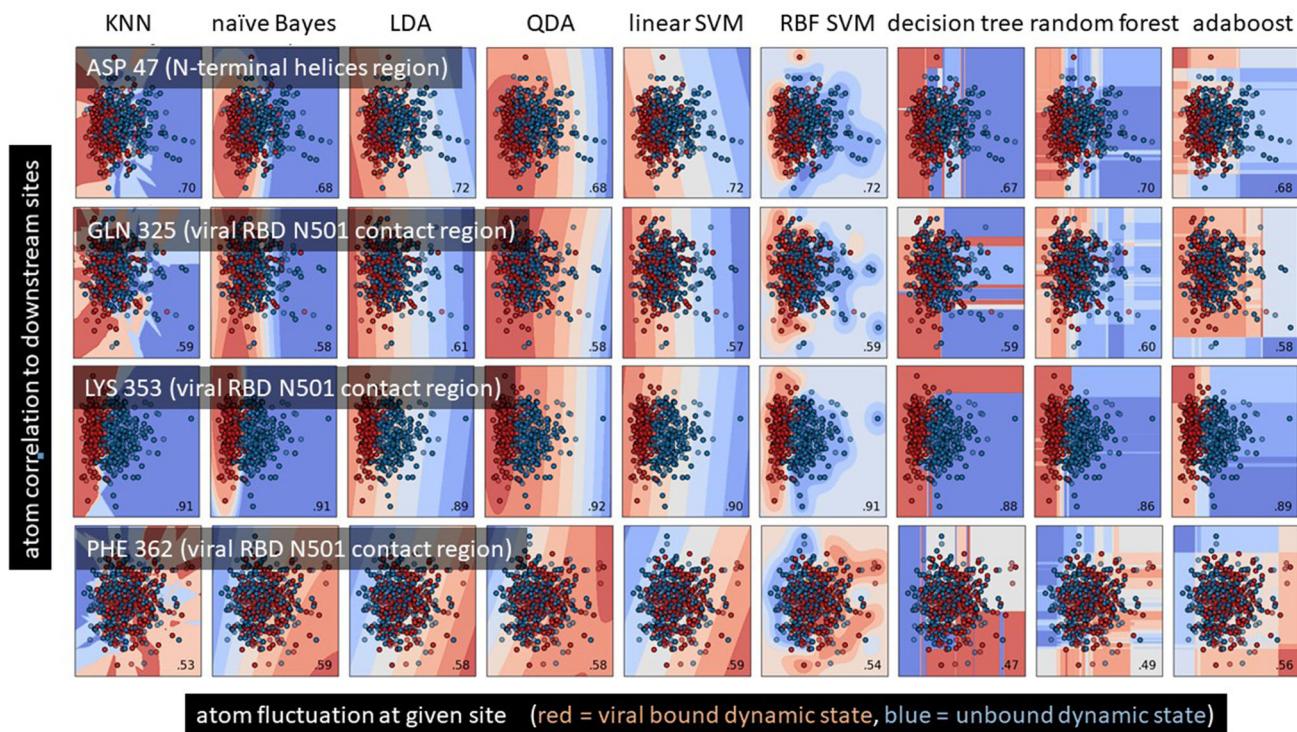
Figure 10. Feature vector and learning classification spaces for key binding sites on ACE2. The machine learning training data and the optimized learning classification spaces is shown at four key residue sites on the binding interface of SARS-CoV-2 and ACE2. The X axis is the atom fluctuation at the given site and the Y axis is the correlations of the fluctuations with nearby downstream sites. The data points represent individual 50 frame time slices in the simulations with dark red indicating those time slices pre-classified to the viral bound state of ACE2 and dark blue points indicating those time slices pre-classified to the unbound state of ACE2. The colors of the background indicate the probability state of the machine learning classification model after training on the data points. The average correct classification (i.e. machine learning performance) for the specified amin acid is shown in the lower right corner of each plot.

(Figure 11A). The black arrowhead in the model signifies a relatively conserved region of sustained interaction between the coronavirus RBDs and the N-terminal helix region of ACE2. The red arrowhead indicates more promiscuous interactions with Q325 and K353, present individually across SARS-CoV-2 as well as SARS-CoV-1, MERS, and HKU1, as well as the newly identified 386-AAQPFLL-392 motif prominent mainly in ACE2 interactions with the wild-type SARS-type spike protein RBD. This two-touch model highlights a common interaction point of the viral RBD with the N-terminal helices of ACE2 that is conserved in all known strains of potentially human-interacting *Betacoronavirus*, demonstrated by the universal dampening of ACE2 molecular motions upon viral binding. This interaction corresponds structurally with the viral bat progenitor strain RBD interaction with the SAMLI 291-5 site in CD26. The second and more complex dynamic touch point in the model involves the contact of the viral RBD N501 with the previously described contact sites K353 and Q325, as well as the nearby novel site of interaction at the AAQPFLL motif. This second interaction corresponds with a strong and precise binding interaction of the viral bat progenitor strain RBD with SS 333-4 in CD26. This promiscuous interaction site seems considerably apparent in the two pathogenic SARS-type strains, suggesting that the binding interactions during zoonotic spillover may follow a common functional evolutionary path in SARS-type emergent outbreaks (Figure 11B), perhaps utilizing a preadapted

interactions with the SAMLI and SS motifs in CD26 in bat reservoir populations to establish loose and promiscuous binding with the ACE2 N-terminal helices and the loop regions near the beta sheet turn at K353. In SARS-CoV-2, this has most recently evolved to higher specificity for ACE2 particularly through the strengthening of the ACE2 K353 and the N-terminal interactions. This appears quite different from the evolutionary path of the binding interactions in human endemic strains like hCoV-OC43 (Figure 11B), where an insertion event appears to have enhanced the RBD loop structures near the N-terminal helices of ACE2 and subsequently strengthen binding interactions in this region.

Discussion

Here we conducted machine-learning assisted statistical comparisons of hundreds of replicate sets of nanosecond scale GPU accelerated molecular dynamics simulations on viral bound and unbound target proteins to survey the evolutionarily conserved functional similarity, as well as the comparative functional differences between various emergent, endemic, and potentially zoonotic *Betacoronavirus* strains in humans. We present evidence that strongly suggests a common route of functional molecular evolution occurring at the binding interface of the viral receptor binding domain and their primary protein targets ACE2 and CD26. We propose a two-touch contact model of viral evolution that may be
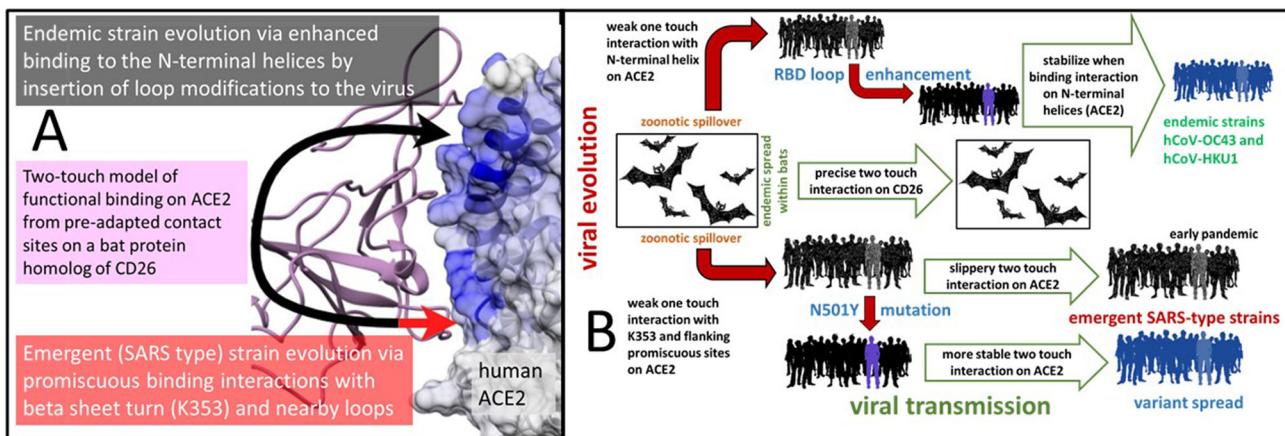
Figure 11. A functional evolutionary model of SARS-type and endemic viral protein binding interactions during zoonotic spillover. (A) The two-touch interaction model consists of a combination of evolutionary conserved ACE2 interactions at the N-terminal helices likely preadapted from its interaction with the SAMLI 291-5 motif on CD26 (shown with black arrow) and a promiscuous, dynamically complex interaction with K353 and nearby sites at Q325 and AAQPFLL 386-392 (shown with red arrow), likely preadapted from its interaction with SS 333-334 on CD26. (B) Given the conserved nature of viral receptor binding across the diversity of strains we surveyed, we hypothesize these regions of the virus are most likely responsible for past and present instances of zoonotic spillover to humans. The more pathogenic SARS-like outbreaks associated with ACE2 seem to develop transient dynamics of N501 that involve promiscuous interactions at three co-located sites mentioned above. This may create a temporary 'slippery' two touch interaction with the ACE2 target that stabilizes over evolutionary time. In the much more transmissible new variant strains (B.1.1.7, B.1.351, and P.1), the promiscuous interactions are mostly lost, favoring a much more stable binding interaction with K353 alone. In both endemic strains, we observe the evolution of enhanced binding to the N-terminal helices of ACE2 through the evolution of enhanced loop structure on the corresponding interface of the viral receptor binding domain.

evolving similarly during SARS-type *Betacoronavirus* zoonotic spillover events from either *Tylonycteris* (bamboo bat) or *Rhinolophus* (horseshoe bat) to human populations. We present computationally derived signatures of receptor binding domain interactions with target human proteins, in the bat progenitor strain batCoV-HKU4 and five human strains, the emergent MERS-CoV, SARS-CoV-1 and SARS-CoV-2 strains and the endemic hCoV-OC43 and hCoV-HKU1 strains. We use a novel machine-learning based approach to detect functionally conserved ACE2 binding dynamics among viral ortholog comparisons. Our studies of the biophysics of viral binding at single amino acid site resolution demonstrate that the coronavirus spike glycoprotein receptor binding domain (RBD) has a strong, static, and well-defined binding interaction occurring in two distinct regions, or 'touch points' of human CD26 and also the CD26 ortholog in bats as well. In its interactions with the human ACE2 target protein, the spike glycoprotein RBD appears pre-adapted through its binding interactions involving two of these sites on CD26 (SAMLI 291-5 and SS 333-4) to allow broad and somewhat promiscuous binding interactions with the N-terminal helical region of ACE2 as well as regions connecting K353, Q325 and 382-AAQPFLL-392.

We find that the functional binding dynamics of these two pre-adapted contact regions on ACE2 have been conserved differently between the bat progenitor and the human endemic and recent SARS-type emergent strains. The human endemic strains hCoV-OC43 and hCoV-HKU1 appear to have evolved towards enhanced stability on ACE2 via strong broad interaction with the ACE2 N-terminal helices that is facilitated by the evolution of enhanced loop structures on the viral RBD that markedly stabilized upon binding the interface with N-terminal helices on ACE2. The slightly more pathogenic human strain hCoV-HKU1 also shows shares this feature with hCoV-OC43, but still also shows

weak interactions with K353. Thus, while the binding of the ACE2 N-terminal helices appears common to all of our *Betacoronavirus* strain dynamics models, this interaction appears especially conserved in the functional evolution of past endemic human strains from the bat progenitor. However, in the more emergent SARS-type strains, we observe evolutionary conserved binding dynamics mainly at and near the K353 site on ACE2, where we also observe a high degree of mutual information (i.e. coordination of functional dynamic states) between this site and Q325 and 382-AAQPFLL-392. This transient viral RBD interaction with multiple sites on ACE2 in emergent SARS-type strains suggests that the interaction with K353 reported by (Yan et al., 2020), is not as stable in the early evolution of SARS-CoV-2 binding as their initial structural analysis might imply. We hypothesize that the binding interactions at these three sites are perhaps more evolvable in SARS-type strains than the interactions at the N-terminal helices. The recent appearance of the N501Y mutation in the highly transmissible variants of SARS-CoV-2 appears to further support the functional evolution of more stable binding at this same location on the ACE2 target protein. When we characterized the effect of the mutation on the binding dynamics in this region, we found that this one mutation has greatly stabilized the Y501 interaction with the ACE2 K353 site, also now reported by (Ali et al., 2021), and has reduced the transient and promiscuous interactions with the other two flanking binding sites we have described here. Thus, the stabilization the dynamics at of this point of contact has enhanced the binding, and by extension perhaps also the transmissibility of the SARS-CoV-2 virus, all the while bringing it much closer to the molecular binding profile of the bat progenitor strain batCoV-HKU4 with its primary target protein, CD26. It has been recently hypothesized and demonstrated that promiscuity in protein-protein interactions is related to a lack of protein stability

(Cohen-Khait et al., 2017) and it would also appear, through our work on this system, that a lack of protein stability may also be related to the molecular facilitation of zoonotic spill-overs of certain viral pathogens as well. We believe the complex dynamic interaction at the second 'slippery or transient touch point' in our two-touch model may represent the molecular manifestation of a lack of binding specificity that may characterize many viral binding interactions during emergent SARS-type outbreaks when the co-evolution between a viral binding domain and a potential host target receptor is still historically very recent. Our simulations suggest several alternative paths of viral evolution in emergent versus endemic strains that have both favored a more precise and specific targeting of human ACE2, which may also be associated with enhanced transmissibility as well.

While our approach offers the considerable advantage of combining a comparative statistical method and a physics-based modeling approach towards addressing functional molecular evolution, it is not without some pitfalls. Some potential limitations of MD simulations as a probative method for functional molecular evolution are its many implicit simplifying computational assumptions, its complex and inherently stochastic nature, and high computational expense which can limit its conformational sampling. Specifically, the sampling of even the accelerated MD method employed here can have significant runtime limitations, and even on modern graphics cards our simulations can typically have a cumulative runtime of several weeks to generate the proper statistical replication to compare physical time frames of only several hundreds of nanoseconds. In addition, MD simulations always involve some simplification of the biophysics within the system being studied as it can ignore some aspects of atomic charge regulation, bond motions in the solvent, charge screening during interaction, and other macromolecular crowding effects. Insight into long-term microsecond to millisecond dynamics in large explicit solvent systems are still limited by currently available hardware, even when advanced algorithms for accelerating MD simulations are used. Glycosylation, mannosylation, sialy-lation and other potential post-translational modifications of the viral bound ACE2 model is another aspect of coronavirus spike protein biology that is not fully captured by our MD simulations. While this is mainly due to lack of glycosylation in the functional binding interface of most of our key starting structures, we also acknowledge that there are recent studies demonstrating the functional importance of these modifications and the effect they have upon binding and other interactions between spike glycoprotein and ACE2 (Allen et al., 2021; Sanda et al., 2021; Yang et al., 2020). However, given the potential heterogeneity of post-transla-tional modifications within infected populations and across species boundaries, and the fact that most mammalian protein structures are expressed out of bacterial systems that lack some and/or all the enzymes responsible for these post-translational modifications, we have no grounded way of investigating their evolution in relation to functional binding at this time. Recent studies have demonstrated considerable impermanence of glycosylation in the SARS-CoV-2 spike glycoprotein (Sanda et al., 2021; Zhang et al., 2021). And while glycosylation probably does affect the flexibility of ACE2 (Barros et al., 2021), we would argue that while these post-translational modifications can have immediate impacts on dynamics, their probable lack of genetic heritability justi-fies our decision to omit their potential role in long term viral evolution outside of specific outbreaks. However, their potential role in both short term and long term viral evolu-tion should be investigated in future studies when more reli-able information regarding the heterogeneity of these chemical modifications becomes available.

Despite these limitations, we conclude that our identifica-tion of additional key residues in the binding interaction between the SARS-CoV-2 RBD and human ACE2 receptor, as well as our evolutionary exploration of the two-touch model of RBD evolution provide a conceptual framework for future functional mutagenesis studies of this system. This will be especially important for understanding the functional evolu-tion of transmissibility in new SARS-CoV-2 variants, recently responsible for the large community lockdowns during the COVID-19 pandemic and recently implicated in reduced vac-cine efficacy (Planas et al., 2021) as well as possible reduction in neutralizing antibody efficacy caused by changes in the molecular dynamics of prefusion conformations of the spike glycoprotein (Corbett et al., 2020; Melero et al., 2020). Future surveys of *Betacoronavirus* circulating in past, present, and future human populations as well as molecular and clinical investigations of SARS-CoV-2 infection will likely continue to be further informed by the model interpretation of binding interaction that we present in this study. In future studies of molecular models derived from potential zoonotic corona-virus strains, our functional evolutionary binding study can lend greater interpretability to observations regarding evolu-tionary diversity in coronaviruses infecting reservoir species like bats, birds, and small mammals. Finally, as recent work has called out the importance of specific molecular motions in forming novel therapeutic targets for intervention in emerging zoonotic spillover events (Pierri, 2020), tools for highlighting functionally conserved dynamics of conform-ational alterations of the interaction host and pathogen pro-teins will prove a valuable addition to the arsenal of modeling approaches available for drug development.

Acknowledgements

Disclosure statement

The authors declare no competing interests.

ORCID

Gregory A. Babbitt http://orcid.org/0000-0003-4421-6527

Data availability statement

Structural models, images, movies and data sets for the post-processed molecular dynamics presented in this work are deposited at Zenodo CERN under the DOI 10.5281/zenodo.4702398 https://zenodo.org/record/4702398.

A movie file comparing SARS-CoV-2 RBD interaction with ACE2, in both wildtype and N501Y mutant strains is also available at https://youtu.be/ptn1_BBJi70.

Author contributions

PR, FC and GAB designed and executed the biological study. GAB designed the statistical methods and wrote computer code to implement the methods. PR, GAB, FC, AOH, and MLL interpreted the results and wrote the manuscript.

References

Ahamad, S., Kanipakam, H., & Gupta, D. (2020). Insights into the structural and dynamical changes of spike glycoprotein mutations associated with SARS-CoV-2 host receptor binding. *Journal of Biomolecular Structure and Dynamics*, 1–13.

Ali, A., & Vijayan, R. (2020). Dynamics of the ACE2-SARS-CoV-2/SARS-CoV spike protein interface reveal unique mechanisms. *Scientific Reports*, 10(1), 14214. https://doi.org/10.1038/s41598-020-71188-3

Ali, F., Kasry, A., & Amin, M. (2021). The new SARS-CoV-2 strain shows a stronger binding affinity to ACE2 due to N501Y mutant. *Medicine in Drug Discovery*, 10, 100086. https://doi.org/10.1016/j.medidd.2021.100086

Al-Khafaji, K., Al-Duhaidahawi, D., & Tok, T. T. (2020). Using integrated computational approaches to identify safe and rapid treatment for SARS-CoV-2. *Journal of Biomolecular Structure and Dynamics*, 39(9), 1–9.

Allen, J. D., Watanabe, Y., Chawla, H., Newby, M. L., & Crispin, M. (2021). Subtle influence of ACE2 glycan processing on SARS-CoV-2 recognition. *Journal of Molecular Biology*, 433(4), 166762. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7744274/

Andersen, H. C. (1980). Molecular dynamics simulations at constant pressure and/or temperature. *Journal of Chemical Physics*, 72(4), 2384–2393. https://doi.org/10.1063/1.439486

Babbitt, G. A., Fokoue, E. P., Evans, J. R., Diller, K. I., & Adams, L. E. (2020). DROIDS 3.0-detecting genetic and drug class variant impact on conserved protein binding dynamics. *Biophysical Journal*, 118(3), 541–551. https://doi.org/10.1016/j.bpj.2019.12.008

Babbitt, G. A., Lynch, M. L., McCoy, M., Fokoue, E. P., & Hudson, A. O. (2020). Function and evolution of B-Raf loop dynamics relevant to cancer recurrence under drug inhibition. *Journal of Biomolecular Structure and Dynamics*, 1–16.

Babbitt, G. A., Mortensen, J. S., Coppola, E. E., Adams, L. E., & Liao, J. K. (2018). DROIDS 1.20: A GUI-based pipeline for GPU-accelerated comparative protein dynamics. *Biophysical Journal*, 114(5), 1009–1017. https://doi.org/10.1016/j.bpj.2018.01.020

Barros, E. P., Casalino, L., Gaieb, Z., Dommer, A. C., Wang, Y., Fallon, L., Raguette, L., Belfon, K., Simmerling, C., & Amaro, R. E. (2021). The flexibility of ACE2 in the context of SARS-CoV-2 infection. *Biophysical Journal*, 120(6), 1072–1084. https://doi.org/10.1016/j.bpj.2020.10.036

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cantuti-Castelvetri, L., Ojha, R., Pedro, L. D., Djannatian, M., Franz, J., Kuivanen, S., van der Meer, F., Kallio, K., Kaya, T., Anastasina, M., Smura, T., Levanov, L., Szirovicza, L., Tobi, A., Kallio-Kokko, H., Österlund, P., Joensuu, M., Meunier, F. A., Butcher, S. J., … Simons, M. (2020). Neuropilin-1 facilitates SARS-CoV-2 cell entry and infectivity.

Science (New York, N.Y.), 370(6518), 856–860. https://doi.org/10.1126/science.abd2985

Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., & Woods, R. J. (2005). The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16), 1668–1688. https://doi.org/10.1002/jcc.20290

Cohen-Khait, R., Dym, O., Hamer-Rogotner, S., & Schreiber, G. (2017). Promiscuous protein binding as a function of protein stability. *Structure*, 25(12), 1867–1874.e3. https://doi.org/10.1016/j.str.2017.11.002

Corbett, K. S., Edwards, D. K., Leist, S. R., Abiona, O. M., Boyoglu-Barnum, S., Gillespie, R. A., Himansu, S., Schäfer, A., Ziwawo, C. T., DiPiazza, A. T., Dinnon, K. H., Elbashir, S. M., Shaw, C. A., Woods, A., Fritch, E. J., Martinez, D. R., Bock, K. W., Minai, M., Nagata, B. M., … Graham, B. S. (2020). SARS-CoV-2 mRNA vaccine design enabled by prototype pathogen preparedness. *Nature*, 586(7830), 567–571. https://doi.org/10.1038/s41586-020-2622-0

Culp, M., Johnson, K., & Michailidis, G. (2016). ada: An R package for stochastic boosting.

Daly, J. L., Simonetti, B., Klein, K., Chen, K.-E., Williamson, M. K., Antón-Plágaro, C., Shoemark, D. K., Simón-Gracia, L., Bauer, M., Hollandi, R., Greber, U. F., Horvath, P., Sessions, R. B., Helenius, A., Hiscox, J. A., Teesalu, T., Matthews, D. A., Davidson, A. D., Collins, B. M., Cullen, P. J., & Yamauchi, Y. (2020). Neuropilin-1 is a host factor for SARS-CoV-2 infection. *Science (New York, N.Y.)*, 370(6518), 861–865. https://doi.org/10.1126/science.abd3072

Darden, T., York, D., & Pedersen, L. (1993). Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *Journal of Chemical Physics*, 98(12), 10089–10092. https://doi.org/10.1063/1.464397

Ewald, P. P. (1921). Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen Der Physik*, 369(3), 253–287. https://doi.org/10.1002/andp.19213690304

Faria, N. R., Morales Claro, I., Candido, D., Moyses Franco, L. A., Andrade, P. S., Coletti, T. M., Silva, C. A., Sales, F. C., Manuli, E. R., & Aguiar, R. S. (2021). Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: Preliminary findings. Virological [Internet]. https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586

Fiser, A., Do, R. K., & Sali, A. (2000). Modeling of loops in protein structures. *Protein Science*, 9(9), 1753–1773. https://doi.org/10.1110/ps.9.9.1753

Follis, K. E., York, J., & Nunberg, J. H. (2006). Furin cleavage of the SARS coronavirus spike glycoprotein enhances cell-cell fusion but does not affect virion entry. *Virology*, 350(2), 358–369. https://doi.org/10.1016/j.virol.2006.02.003

Fung, T. S., & Liu, D. X. (2019). Human coronavirus: Host-pathogen interaction. *Annual Review of Microbiology*, 73, 529–557. https://doi.org/10.1146/annurev-micro-020518-115759

Götz, A. W., Williamson, M. J., Xu, D., Poole, D., Le Grand, S., & Walker, R. C. (2012). Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized Born. *Journal of Chemical Theory and Computation*, 8(5), 1542–1555. https://doi.org/10.1021/ct200909j

Haan, C. d., Vennema, H., & Rottier, P. J. M. (2000). Assembly of the Coronavirus envelope: Homotypic interactions between the M proteins. *Journal of Virology*, 74(11), 4967–4978. https://doi.org/10.1128/.74.11.4967-4978.2000

Härdle, W., & Simar, L. (Eds.). (2007). Canonical correlation analysis. In *Applied multivariate statistical analysis* (pp. 321–330). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-72244-1_14

Hofmann, H., Pyrc, K., van der Hoek, L., Geier, M., Berkhout, B., & Pöhlmann, S. (2005). Human coronavirus NL63 employs the severe acute respiratory syndrome coronavirus receptor for cellular entry. *Proceedings of the National Academy of Sciences of the United States of America*, 102(22), 7988–7993. https://doi.org/10.1073/pnas.0409465102

Hoffmann, M., Hofmann-Winkler, H., & Pöhlmann, S. (2018). Priming time: How cellular proteases arm coronavirus spike proteins. In E. Böttcher-Friebertshäuser, W. Garten, & H. Klenk (Eds.), *Activation of Viruses by Host Proteases* (pp. 71–98). Springer.

Hoffmann, M., Kleine-Weber, H., & Pöhlmann, S. (2020). A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for

infection of human lung cells. *Molecular Cell*, 78(4), 779–784.e5. https://doi.org/10.1016/j.molcel.2020.04.022

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab - An S4 package for Kernel methods in R. *Journal of Statistical Software*, 11, 1–20.

Kuba, K., Imai, Y., Rao, S., Gao, H., Guo, F., Guan, B., Huan, Y., Yang, P., Zhang, Y., Deng, W., Bao, L., Zhang, B., Liu, G., Wang, Z., Chappell, M., Liu, Y., Zheng, D., Leibbrandt, A., Wada, T., … Penninger, J. M. (2005). A crucial role of angiotensin converting enzyme 2 (ACE2) in SARS cor-onavirus-induced lung injury. *Nature Medicine*, 11(8), 875–879. https://doi.org/10.1038/nm1267

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Li, F., Li, W., Farzan, M., & Harrison, S. C. (2005). Structure of SARS cor-onavirus spike receptor-binding domain complexed with receptor. *Science (New York, N.Y.)*, 309(5742), 1864–1868. https://doi.org/10.1126/science.1116480

Liang, D., Song, M., Niu, Z., Zhang, P., Rafailovich, M., & Deng, Y. (2021). Supervised machine learning approach to molecular dynamics fore-cast of SARS-CoV-2 spike glycoproteins at varying temperatures. MRS Advances [Internet]. https://doi.org/10.1557/s43580-021-00021-4

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 18–22.

Li, W., Zhang, C., Sui, J., Kuhn, J. H., Moore, M. J., Luo, S., Wong, S.-K., Huang, I.-C., Xu, K., Vasilieva, N., Murakami, A., He, Y., Marasco, W. A., Guan, Y., Choe, H., & Farzan, M. (2005). Receptor and viral determi-nants of SARS-coronavirus adaptation to human ACE2. *The EMBO Journal*, 24(8), 1634–1643. https://doi.org/10.1038/sj.emboj.7600640

Lubin, J. H., Zardecki, C., Dolan, E. M., Lu, C., Shen, Z., Dutta, S., Westbrook, J. D., Hudson, B. P., Goodsell, D. S., & Williams, J. K. (2020). Evolution of the SARS-CoV-2 proteome in three dimensions (3D) dur-ing the first six months of the COVID-19 pandemic. *BioRxiv: The Preprint Server for Biology*.

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., & Simmerling, C. (2015). ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11(8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255

Mao, Y., & Zhang, Y. (2012). Thermal conductivity, shear viscosity and specific heat of rigid water models. *Chemical Physics Letters*, 542, 37–41. https://doi.org/10.1016/j.cplett.2012.05.044

Melero, R., Sorzano, C. O. S., Foster, B., Vilas, J.-L., Martínez, M., Marabini, R., Ramírez-Aportela, E., Sanchez-Garcia, R., Herreros, D., & Del Caño, L. (2020). Continuous flexibility analysis of SARS-CoV-2 spike prefusion structures. *IUCrJ*, 7(6), 1059–1069.

Mittal, A., Manjunath, K., Ranjan, R. K., Kaushik, S., Kumar, S., & Verma, V. (2020). COVID-19 pandemic: Insights into structure, function, and hACE2 receptor recognition by SARS-CoV-2. PLoS Pathogens [Internet], 16. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7444525/

Muralidharan, N., Sakthivel, R., Velmurugan, D., & Gromiha, M. M. (2020). Computational studies of drug repurposing and synergism of lopina-vir, oseltamivir and ritonavir binding with SARS-CoV-2 protease against COVID-19. *Journal of Biomolecular Structure and Dynamics*, 39(7), 2573–2678.

Ou, X., Guan, H., Qin, B., Mu, Z., Wojdyla, J. A., Wang, M., Dominguez, S. R., Qian, Z., & Cui, S. (2017). Crystal structure of the receptor bind-ing domain of the spike glycoprotein of human betacoronavirus HKU1. *Nature Communications*, 8, 15216. https://doi.org/10.1038/ncomms15216

Ou, X., Liu, Y., Lei, X., Li, P., Mi, D., Ren, L., Guo, L., Guo, R., Chen, T., Hu, J., Xiang, Z., Mu, Z., Chen, X., Chen, J., Hu, K., Jin, Q., Wang, J., & Qian, Z. (2020). Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nature Communications*, 11(1), 1620. https://doi.org/10.1038/s41467-020-15562-9

Oudit, G. Y., Crackower, M. A., Backx, P. H., & Penninger, J. M. (2003). The role of ACE2 in cardiovascular physiology. *Trends in Cardiovascular Medicine*, 13(3), 93–101. https://doi.org/10.1016/s1050-1738(02)00233-5

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera-a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13), 1605–1612. https://doi.org/10.1002/jcc.20084

Pierce, L. C. T., Salomon-Ferrer, R., de Oliveira, C. A. F., McCammon, J. A., & Walker, R. C. (2012). Routine access to millisecond time scale events with accelerated molecular dynamics. *Journal of Chemical Theory and Computation*, 8(9), 2997–3002. https://doi.org/10.1021/ct300284c

Pierri, C. L. (2020). SARS-CoV-2 spike protein: Flexibility as a new target for fighting infection. *Signal Transduction and Targeted Therapy*, 5, 1–3.

Planas, D., Bruel, T., Grzelak, L., Guivel-Benhassine, F., Staropoli, I., Porrot, F., Planchais, C., Buchrieser, J., Rajah, M. M., & Bishop, E. (2021). Sensitivity of infectious SARS-CoV-2 B.1.1.7 and B.1.351 variants to neutralizing antibodies. *Nature Medicine*, 27, 917–924.

Rambaut, A., Loman, N., Pybus, O., Barclay, W., Barrett, J., Carabelli, A., Connor, T., Peacock, T., Robertson, D. L., & Volz, E. (2020). Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. Virological [Internet]. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563

Randhawa, G. S., Soltysiak, M. P. M., Roz, H. E., Souza, C. P. E., de Hill, K. A., & Kari, L. (2020). Machine learning using intrinsic genomic signa-tures for rapid classification of novel pathogens: COVID-19 case study. *PLOS ONE*, 16(1), e0246465.

Roe, D. R., & Cheatham, T. E. (2013). PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *Journal of Chemical Theory and Computation*, 9(7), 3084–3095. https://doi.org/10.1021/ct400341p

Sali, A., & Blundell, T. L. (1993). Comparative protein modelling by satis-faction of spatial restraints. *Journal of Molecular Biology*, 234(3), 779–815. https://doi.org/10.1006/jmbi.1993.1626

Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S., & Walker, R. C. (2013). Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *Journal of Chemical Theory and Computation*, 9(9), 3878–3888. https://doi.org/10.1021/ct400314y

Sanda, M., Morrison, L., & Goldman, R. (2021). N- and O-glycosylation of the SARS-CoV-2 spike protein. *Analytical Chemistry*, 93(4), 2003–2009. https://doi.org/10.1021/acs.analchem.0c03173

Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A., & Li, F. (2020). Cell entry mechanisms of SARS-CoV-2. *Proceedings of the National Academy of Sciences of the United States of America*, 117(21), 11727–11734. https://doi.org/10.1073/pnas.2003138117

Song, W., Gui, M., Wang, X., & Xiang, Y. (2018). Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. *PLoS Pathogens*, 14(8), e1007236. https://doi.org/10.1371/journal.ppat.1007236

Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E. J., Msomi, N., Mlisana, K., von Gottberg, A., Walaza, S., Allam, M., Ismail, A., Mohale, T., Glass, A. J., Engelbrecht, S., Van Zyl, G., … de Oliveira, T. (2021). Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*, 592(7854), 438–443. https://doi.org/10.1038/s41586-021-03402-9

Tortorici, M. A., Walls, A. C., Lang, Y., Wang, C., Li, Z., Koerhuis, D., Boons, G.-J., Bosch, B.-J., Rey, F. A., de Groot, R. J., & Veesler, D. (2019). Structural basis for human coronavirus attachment to sialic acid receptors. *Nature Structural & Molecular Biology*, 26(6), 481–489. https://doi.org/10.1038/s41594-019-0233-y

Venables, W. N., & Ripley, B. D. (2010). Modern applied statistics with S. Springer Publishing Company, Incorporated.

Wang, Q., Qi, J., Yuan, Y., Xuan, Y., Han, P., Wan, Y., Ji, W., Li, Y., Wu, Y., Wang, J., Iwamoto, A., Woo, P. C. Y., Yuen, K.-Y., Yan, J., Lu, G., & Gao, G. F. (2014). Bat origins of MERS-CoV supported by bat coronavirus HKU4 usage of human receptor CD26. *Cell Host & Microbe*, 16(3), 328–337. https://doi.org/10.1016/j.chom.2014.08.009

Wang, Y., Liu, M., & Gao, J. (2020). Enhanced receptor binding of SARS-CoV-2 through networks of hydrogen-bonding and hydrophobic

interactions. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(25), 13967–13974. https://doi.org/10.1073/pnas.2008209117

Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., & Zhou, Q. (2020). Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science (New York, N.Y.)*, *367*(6485), 1444–1448. https://doi.org/10.1126/science.abb2762

Yang, Q., Hughes, T. A., Kelkar, A., Yu, X., Cheng, K., Park, S., Huang, W.-C., Lovell, J. F., & Neelamegham, S. (2020). Inhibition of SARS-CoV-2 viral entry upon blocking N- and O-glycan elaboration. McConville MJ, Kana BD, McConville MJ, Thaysen-Andersen M, editors. *eLife*, *9*:e61552.

Yu, J., Qiao, S., Guo, R., & Wang, X. (2020). Cryo-EM structures of HKU2 and SADS-CoV spike glycoproteins provide insights into coronavirus evolution. *Nature Communications*, *11*(1), 3070. https://doi.org/10.1038/s41467-020-16876-4

Yuan, Y., Cao, D., Zhang, Y., Ma, J., Qi, J., Wang, Q., Lu, G., Wu, Y., Yan, J., Shi, Y., Zhang, X., & Gao, G. F. (2017). Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. *Nature Communications*, *8*, 15092. https://doi.org/10.1038/ncomms15092

Zhang, Y., Zhao, W., Mao, Y., Chen, Y., Wang, S., Zhong, Y., Su, T., Gong, M., Du, D., Lu, X., Cheng, J., & Yang, H. (2021). Site-specific N-glycosylation characterization of recombinant SARS-CoV-2 spike proteins. *Molecular & Cellular Proteomics*, *20*, 100058. https://doi.org/10.1074/mcp.RA120.002295

Zhao, X., Chen, D., Szabla, R., Zheng, M., Li, G., Du, P., Zheng, S., Li, X., Song, C., & Li, R. (2020). Broad and differential animal angiotensin-converting enzyme 2 receptor usage by SARS-CoV-2. *Journal of Virology*, *94*(18), e00940-20. https://jvi.asm.org/content/94/18/e00940-20