Allosteric Hotspots in the Main Protease of SARS-CoV-2

Léonie Strömich [1], **Nan Wu** [1], **Mauricio Barahona** [2] and **Sophia N. Yaliraki** [1]*

1 - *Department of Chemistry Imperial College London,* United Kingdom
2 - *Department of Mathematics Imperial College London,* United Kingdom

Correspondence to Sophia N. Yaliraki: s.yaliraki@imperial.ac.uk (S.N. Yaliraki) @CMPHImperial 🐦 (S.N. Yaliraki)
https://doi.org/10.1016/j.jmb.2022.167748
Edited by Igor Berezovsky

Abstract

Inhibiting the main protease of SARS-CoV-2 is of great interest in tackling the COVID-19 pandemic caused by the virus. Most efforts have been centred on inhibiting the binding site of the enzyme. However, considering allosteric sites, distant from the active or orthosteric site, broadens the search space for drug candidates and confers the advantages of allosteric drug targeting. Here, we report the allosteric communication pathways in the main protease dimer by using two novel fully atomistic graph-theoretical methods: Bond-to-bond propensity, which has been previously successful in identifying allosteric sites in extensive benchmark data sets without *a priori* knowledge, and Markov transient analysis, which has previously aided in finding novel drug targets in catalytic protein families. Using statistical bootstrapping, we score the highest ranking sites against random sites at similar distances, and we identify four statistically significant putative allosteric sites as good candidates for alternative drug targeting.

Introduction

The global pandemic of COVID-19 (coronavirus disease 2019) is caused by SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2),[1–4] a member of the coronavirus family of enveloped, single-stranded ribonucleic acid (RNA) viruses that also includes the virus responsible for the severe acute respiratory syndrome (SARS) epidemic of 2003.[5] Since coronaviruses have been known to infect various animal species and share phylogenetic similarity to pathogenic human coronaviruses, the potential of health emergency events had already been noted.[6] However, their high mutation rate, similarly to other RNA viruses[7] and particularly of the SARS-CoV-2 Spike protein,[8] made the development of long lasting drugs challenging. Developing therapeutics against coronaviruses is of renewed interest due to the ongoing global health emergency.

One of the main approaches for targeting coronaviruses is to inhibit the enzymatic activity of their replication machinery. The main protease (M^{pro}), also known as 3C-like protease (3CL^{pro}), is the best characterised drug target owing to its crucial role in viral replication.[9–11] The M^{pro} is only functional as a homodimer and the central part of the active (or orthosteric) site is composed of a cysteine-histidine catalytic dyad[12] (see Figure 1 (B)) which is responsible for processing the polyproteins translated from the viral RNA.[13]

The M^{pro} of SARS-CoV-2 is very similar to that of SARS-CoV: they share 96 % sequence similarity, and exhibit high structural similarity (the root mean square deviation between Cα positions is only 0.53 Å).[12] Indeed, many of the residues that are important for catalytic activity, substrate binding and dimerisation are conserved between both proteases.[14] However, mutations in SARS-CoV-2 M^{pro} (for a full list see Table S1) would indeed
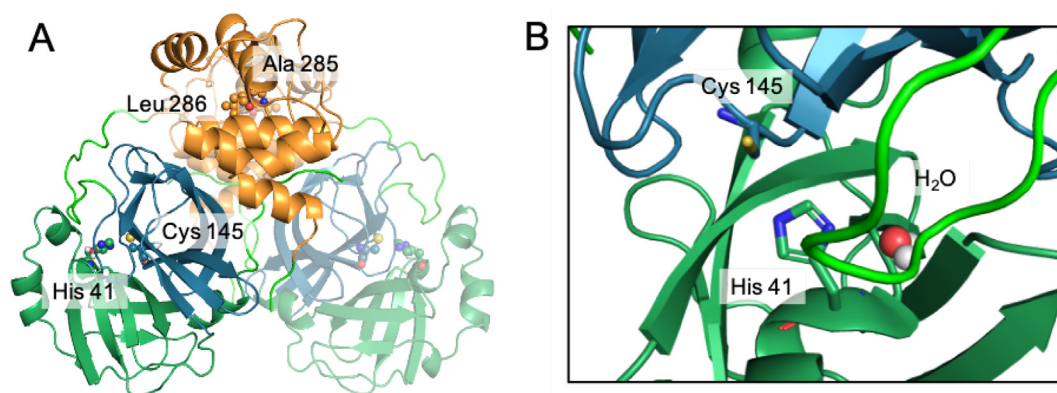
Figure 1. Overview of the SARS-CoV-2 main protease dimer. Atomic coordinates are obtained from the PDB file (PDB ID: 6Y2E). (A) shows the full dimer with the active site residues and residues 285/286 on both monomers shown as spheres. The second monomer is shown with increased transparency to visualise where the monomers interact. Colours are according to domain: Domain I residues 10 to 99 - dark green, domain II residues 100 to 182 - dark blue, domain III residues 198 to 303 - orange, loops in light green. (B) Zoom-in of the active site with histidine 41 and cysteine 145 forming a catalytic dyad which is extended to a triad by a water molecule in close proximity.

modulate distant regions via allosteric signalling. Notably, two of those mutations at the dimer interface (Thr285Ala and Ile286Leu, see Figure 1) have been the object of particular interest: on one hand, it has been suggested that they could be responsible for closer dimer packing in SARS-CoV-2;[12] on the other hand, previous mutational studies on those positions have revealed an impact on catalytic activity in SARS-CoV M[pro].[15] Currently, the development of inhibitors for the M[pro] of SARS-CoV-2[12,16–18] focuses on blocking the active sites to disrupt viral replication,[19] similarly to the strategy followed for the design of other inhibitors for coronavirus proteases.[20–22]

Targeting the active site enables high affinity of the drug molecules, but can also result in off-target toxicity through binding to proteins with similar active sites.[23,24] Drug resistance is another major concern, especially when the active site may potentially change owing to mutations. Targeting an allosteric site distal from the main binding site provides an alternative strategy by increasing the range and selectivity of drugs that can fine-tune protein activity, yet circumventing some of the aforementioned disadvantages. For reviews and recent successes of allosteric drug design, see Wenthur *et al*. and Cimermancic *et al*.[25,26] This approach also holds additional potential for drug repurposing targeting binding at allosteric sites to reduce time and costs for bringing new drugs to the market,[27] an important consideration in the time-sensitive setting of COVID-19.[28]

Encouragingly, following the very recent surge of research around the M[pro] of SARS-CoV-2, some experimental studies have found drugs and small fragments that bind to sites other than the substrate binding site on this protein, and that might have implications for allosteric regulation.[29,30] Preliminary studies have also simu-

lated binding events to distant areas of the protein by using docking and molecular dynamics (MD),[31–33] MD[34–38] or elastic network models (ENMs).[39,40] Moreover, there have already been indications of allosteric processes mediated by the extra domain in the protease of the old SARS-CoV.[41–43,15]

We focus here on the allostericity of the SARS-CoV-2 main protease, and specifically whether there are potential allosteric sites strongly connected to the active site that may offer alternative ways to inhibit virus reproduction. The identification of allosteric sites in enzymes remains challenging and is still often done serendipitously. Computational prediction of allosteric sites has become an active field of research for drug design (for reviews see[44,45]) as it promises to help reduce the laborious and time-consuming process of compound screening. Thermodynamic models, notably the classic Monod-Wyman-Changeux[46] and Koshland-Némethy-Filmer[47] models, offer insights into the conformational changes of protein structure upon ligand binding. However, they do not explain the underpinning molecular signal transmission within the protein, which has been argued as a key structural component of allostery,[48] nor the idea of allosteric signals being propagated over bond paths within the structure.[49] Other studies have used MD simulations, which model the dynamics of proteins at the atomic level, to detect communication pathways in the protein structure that can be exploited for allosteric residue and site identification.[50,51] To alleviate the substantial computational resources required by MD simulations, as well as their inability to explore all the required time and length scales, variations of normal mode analysis (NMA) or ENM are widely employed and have achieved moderate accuracy in allosteric site detection when tested on known allosteric proteins.[52–55]

The toolbox for allosteric site prediction is continuously growing, and new methods range from statistical mechanical models[56,57] to methods based on graph theory.[58] However, most of them overcome the computational requirements of MD at the cost of resolution by looking at coarse-grained representations of protein structures.[59]

To overcome some of these limitations, we have recently introduced a suite of methods for the analysis of high-resolution atomistic protein graphs derived from structural data. The methods are computationally efficient and can span across scales in an unsupervised manner. The graphs have atoms as nodes and retain key physico-chemical detail through energy-weighted edges obtained from structural information and interatomic potentials of covalent and weak interactions (hydrogen bonds, electrostatics and hydrophobics) which are known to be important in allosteric signalling.[60–62] Here we apply two techniques that take full advantage of this detailed atomistic graph: bond-to-bond propensity (B2B-prop)[63] and Markov Transients (MT).[61] Both methods share a common foundation, namely, a Markov process sourced at atoms or bonds of interest diffusing on the atomistic protein graph is used to explore the structure and reveal important protein regions regarding signal propagation. Yet, each method evaluates different and complementary properties.

B2B-prop quantifies how fluctuations at a given set of bonds (the 'source') get redistributed to any other bond in the protein graph and provides a measure (with statistical significance) of instantaneous connectivity as mediated by the graph structure at stationarity. Unlike most network approaches, B2B-prop is formulated on the edges of the graph and thus makes a direct link between energy and flow through bonds, i.e., physico-chemical interactions.[63] B2B-prop is capable of successfully predicting allosteric sites in a wide range of proteins without any *a priori* knowledge, other than the active site.[63] Of particular relevance to the obligate homodimeric protease studied here, B2B-prop has been subsequently used to show how allostery and cooperativity are intertwined in multimeric enzymes such as the well-studied aspartate carbamoyltransferase (ATCase).[64] B2B-prop has been benchmarked against two extensive allosteric protein databases,[65,66] and shown to outperform methods that use simple distance cutoff for interactions or coarse-grained descriptions.[52,54,55]

MT provides additional information by shedding light on the catalytic aspects of allostery. MT extract pathways implicated in allosteric regulation by analysing the dynamical transients of propagation from the active site as the diffusion progresses on the atomistic graph.[61] Specifically, MT compute a statistical measure that highlight atoms and residues that are reached significantly fast by fluctuations propagating from the source. Crucially, MT analysis takes into account *all* possible pathways, not just the shortest or optimal paths — an important feature since allosteric communication is known to involve multiple paths across the protein.[67] MT analysis has been successful in identifying allosteric paths in caspase-1,[61] as well as previously unknown allosteric inhibitor binding sites in p90 ribosomal s6 kinase 4 (RSK4) which complemented and helped guide experimental and clinical studies in drug repurposing for lung cancer.[68]

Both B2B-prop and MT share a common foundation (namely, the use of diffusive processes on the atomistic protein structure), yet they evaluate different and complementary properties: B2B-prop finds bonds and residues that accumulate a disproportionate amount of fluctuations injected at the 'source' at stationarity, whereas MT highlights atoms and residues that are reached particularly fast by the propagation of fluctuations from the source. Mathematically, bond-to-bond propensity reveals properties of the stationary distribution of fluctuations, whereas Markov Transients reflects properties of pathways for signal propagation by concentrating on the transient approach to stationarity. Therefore, each method reveals different aspects of the underlying allosteric mechanisms: B2B-prop analysis gives insights into the effects of structural connectivity, whereas MT analysis is better suited to capture the time scales and catalytic effects of the enzyme. As M^{pro} is a catalytic protein, we expected both methods to prove valuable in revealing regions of the protein (hotspots) that can affect different aspects of allosteric regulation (i.e., sites and pathways). In both approaches we employ an energy-weighted atomistic protein graph and their low computational demands make them suitable to be applied to large proteins and complexes while retaining physico-chemical and atomistic detail. Both methods have been recently built into a web server[65] that creates atomistic graphs from PDB structures and analyses them using B2B-prop and MT, thus facilitating their application to user-provided biomolecular structures.

In this paper, we apply these two methodologies in the setting of COVID-19. We analysed the SARS-CoV-2 main protease and obtained bond-to-bond propensities for all bonds as well as Markov transient half-times $t_{1/2}$ for all atoms in the protein. Our results shed light on the allosteric communication patterns in the M^{pro} dimer, highlighting the role of the dimer interface. We use our methods to show how the subtle structural changes between SARS-CoV and SARS-CoV-2 affect the dimer properties.

By applying a rigorous scoring procedure, we identify four statistically significant hotspots on the protein that are strongly connected to the active site and propose that they hold potential for allosteric regulation of the main protease. Aligning

our results to hits from the Diamond Light Source XChem fragment screen,[69] we find molecules that could be a first starting point for allosteric drug design. The inhibitory effect of some of these molecules has been proven by mass spectrometry based assays.[29] By providing guidance for allosteric drug design we hope to help drug targeting efforts to combat COVID-19.

Results

Exploiting bond-to-bond propensity to provide insights into the M^pro dimer at atomistic resolution

We analysed the resolved apo structure of the main protease M^pro of SARS-CoV-2 (PDB ID: 6Y2E).[12] Although we concentrate our exposition on 6Y2E, we have also repeated our analysis for another structure (PDB ID: 7JP1) with the same results, see SI Table S7, Figure S1 and S2.

In its active form, the protease forms a homodimer, and each monomer has three domains (Figure 1(A)). The active site in each monomer forms a catalytic dyad, which is expanded to a triad by the presence of a water molecule[12] (Figure 1(B)).

Our analysis starts with the PDB file, from which we construct an atomistic graph that includes both strong (covalent) bonds and weak interactions (hydrogen bonds, electrostatic and hydrophobic interactions) as well as structural water molecules, which are known to be catalytically important (see Methods and Figure 5).

We then employ B2B-prop and MT to characterise the propagation of perturbations emanating from the source residues across the atomistic protein graph. To quantify these effects, we use quantile regression to score all bonds and atoms, and consequently all residues. This allows us to identify statistically significant 'hotspots', i.e., regions of the protein that are affected more strongly (using B2B-prop) or reached more quickly (using MT) by perturbations emanating from the source (see Methods).

We use these techniques in two ways: firstly, in a forward step, we source perturbations at the two active sites of the dimer and identify hotspots in the rest of the protein, which we mark as putative allosteric sites; secondly, in a reverse step, we source perturbations at the obtained hotspots and analyse the pattern of propagation back to the active sites and to other regions of interest in the dimer (e.g., the dimer interface).

We start with an exploration of the structure of the M^pro of SARS-CoV-2 using bond-to-bond propensity analysis. Figure 2 shows the forward step of B2B-prop using the active sites in the homodimer (specifically, the catalytically active residues histidine 41 and cysteine 145 in both monomers) as sources. The propensity of every residue in the protein is regressed against their distance to the active site using quantile regression (Figure 2(C)), and the resulting quantile score (QS) of every residue is shown on the protein structure using a colour map (Figure 2(A-B)). Quantile regression allows us to rank all residues in the protein. The list of residues with QSs above 95% (a total of 40 residues) is given in Table S2.

This initial B2B-prop analysis reveals two main areas of interest: a hot region at the back of the monomer opposite to the active site (see Figure 2 (A)), which is a main focus of our study in the sections below; and a hot region overlapping with part of the dimer interface (see Figure 2(B)).

Protease dimerisation is under influence of mutated residues. Given that M^pro is an obligate dimer, the detection of a hot region at the dimer interface points at the importance of cooperative effects[64] and we first explore further some of its features.

The hot region at the dimer interface contains four residues that form salt bridges between the two monomers (serine 1 and arginine 4 from one monomer connect to histidine 172 and glutamine 290 from the other monomer). These bonds have been found experimentally to be essential for dimer formation[70,42] in the original SARS-CoV, an obligate homodimer.

Furthermore, a comparison of the M^pro of SARS-CoV-2 and SARS-CoV (see Table S1) shows that there are two mutated residues at the dimer interface: threonine 285 and isoleucine 286 (SARS-CoV protease) mutated to alanine 285 and leucine 286 (SARS-CoV-2 protease). These two residues have been shown to lead to closer dimer packing in the M^pro of both SARS-CoV and SARS-CoV-2.[15,12]

To further clarify the interactions between the dimer halves (Figure 1(A)), and how the dimer connectivity is changed in the SARS-CoV-2 protease, we carried out a comparative analysis of the apo structures of the M^pro of SARS-CoV-2 (PDB ID: 6Y2E[12]) and SARS-CoV (PDB ID: 2DUC[71]). Specifically, we ran bond-to-bond propensity analysis sourced from the two mutated residues (285 and 286) in both structures. Table 1 shows the top 20 residues by B2B-prop QS (sourced from 285/286) for both structures. We find that 16 out of the top 20 residues are at the dimer interface for 6Y2E compared to 8/20 for 2DUC. Hence, although we find strong connectivity towards dimer interface residues in both structures, there is an increased connectivity in the SARS-CoV-2 protease, including to important residues such as serine 1 and arginine 4. This can be attributed to a closer dimer packing due to the two smaller side chains of 285/286 in the new protease.[12]

A mutational study showed experimentally that closer dimer packing led to increased activity in SARS-CoV,[15] yet this increase was not confirmed
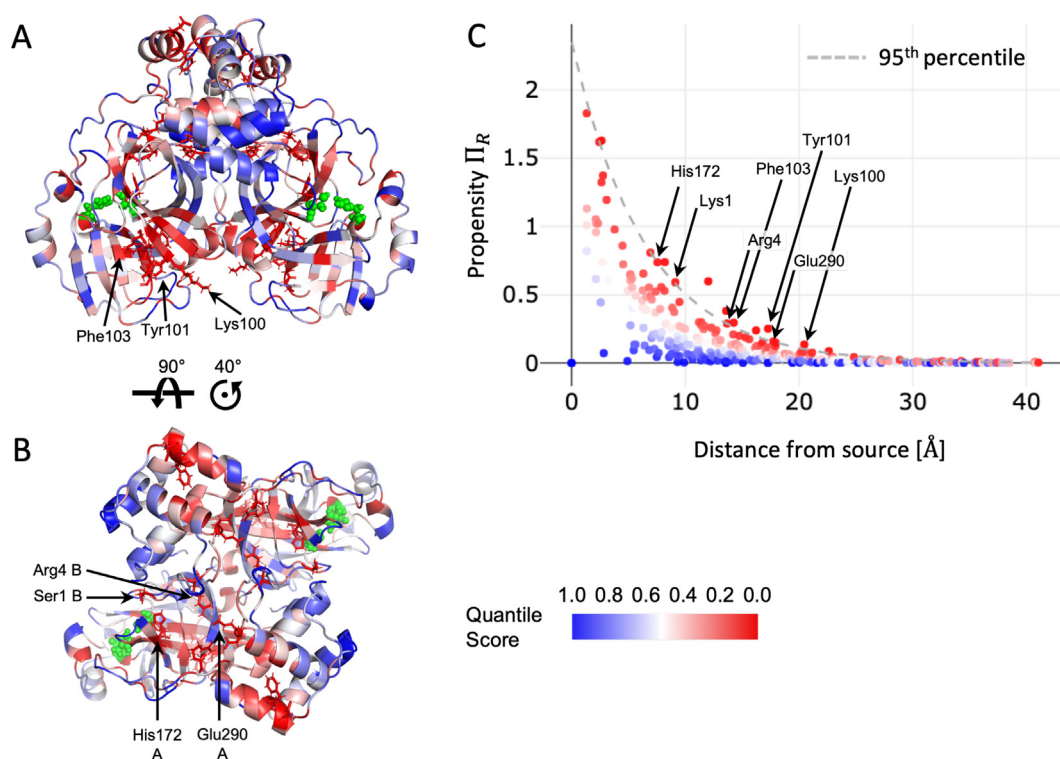
Figure 2. B2B-prop analysis of the SARS-CoV-2 M^{pro} sourced from the orthosteric sites. The residues of the protease (PDB ID: 6Y2E[12]) viewed from the front (A) and top (B) are coloured according to their propensity value. The source sites (shown in green) are the catalytically active residues His41 and Cys145 in both chains of the homodimer. All other residues are coloured by their QS as per the colourbar. There are two main areas of interest with high propensity (Hotspot 1 indicated in A; Hotspot 2 indicated in (B) with important residues labelled. (C) The propensity of each residue, Π_R, is plotted against the distance of the residue from the orthosteric site. The dashed line indicates the quantile regression estimate of the 0.95 quantile cutoff used to identify the significant residues in Table S2.

experimentally in the SARS-CoV-2 protease.[12] Our computations reveal this effect: the average quantile score of the active site of SARS-CoV-2 M^{pro} for B2B-prop sourced at 285/286 is 0.26, which is below a randomly sampled site score of 0.48 (95% CI: 0.47–0.49) and makes the active site a coldspot. On the other hand, the average quantile score of the active site of SARS-CoV M^{pro} for B2B-prop sourced at 285/286 is 0.50, slightly above a random site score of 0.48 (95% CI: 0.47–0.48).

Identification and scoring of putative allosteric sites

Predicted sites using bond-to-bond propensity. Based on our B2B-prop analysis we detected two main hot regions on the protease, each of which contains a 'hotspot' or site with contiguous high-scoring residues (see Table S2) which could be targetable for allosteric regulation of the protease (Figure 3):

- Site 1 (Figure 3(A) framed in yellow) is located at the back of the monomer with respect to the active site and is formed by 9 residues from domain I and II (full list in Table S3).

- Site 2 is located at the dimer interface and contains 6 residues (listed in Table S3) which are located on both monomers (Figure 3(B) framed in pink). Two of these residues, glutamine 290 and arginine 4 of the respective second monomer form a salt bridge which is essential for dimerisation.[42]

Sites 1 and 2 have a high average residue quantile score of 0.97 and 0.96, respectively, which is much higher than random as quantified by a statistical comparison to 1000 random sites with 10,000 bootstrap resamples (see Methods) which gives a QS of 0.53 (95% CI: 0.53–0.54) for a random site of the size of Site 1, and a QS of 0.52 (95% CI: 0.51–0.53) for a site of the size of site 2.

Next, we investigated the connectivity of the putative allosteric sites using B2B-prop in its reverse step, i.e., we carry out a full B2B-prop analysis using as source all the residues within each of the identified sites and rank all residues using quantile regression. The average residue quantile score of the active site for the reverse B2B-prop sourced at site 1 is 0.64, which is above a randomly sampled site score of 0.47 (95% CI:0.47–0.48) obtained by our statistical bootstrap.
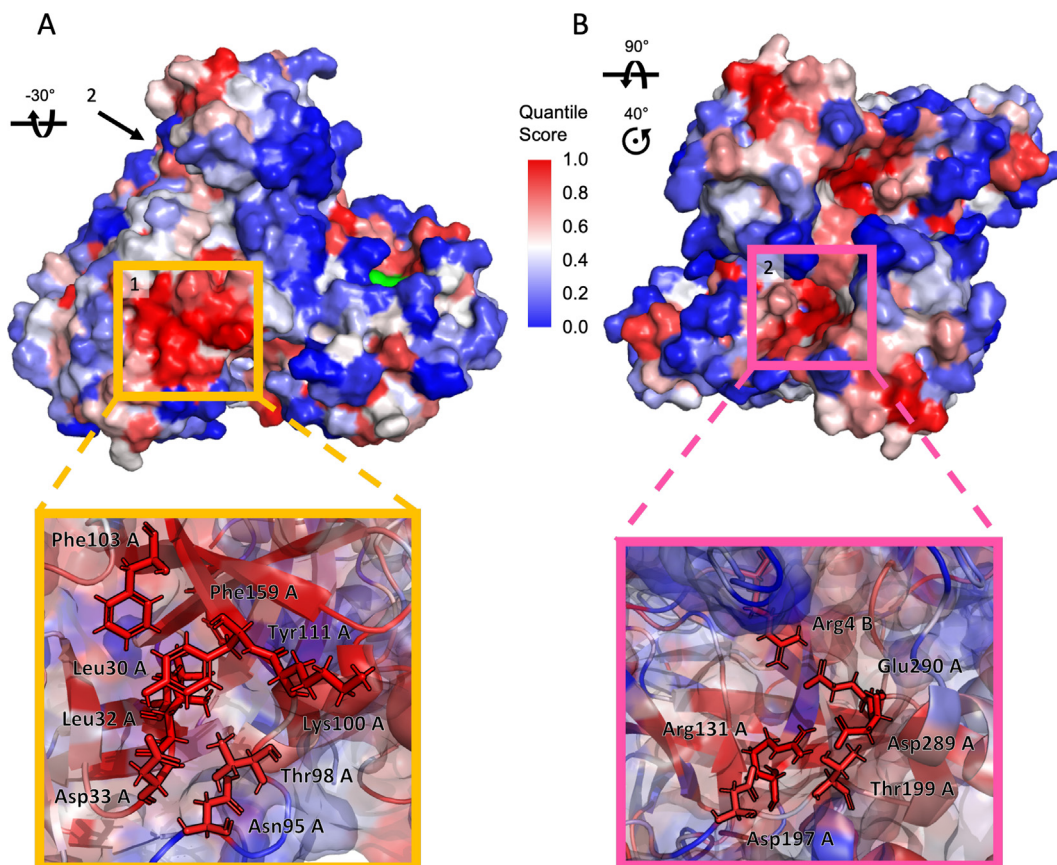
Table 1 **Comparison of Top 20 residues between Covid-19 and SARS main protease.** Highlighted in blue are residues which are in the dimer interface.

SARS-CoV-2	SARS-CoV
SER1 A	ARG40 A
ARG4 A	SER123 A
ARG40 A	GLU166 A
PRO122 A	ASP187 A
GLN306 A	PHE305 A
SER1 B	ARG40 B
ARG4 B	ASN95 B
ARG40 B	PRO122 B
PRO122 B	ARG131 B
GLN306 B	ASP187 B
PHE3 A	ILE281 B
SER10 A	TYR54 A
GLU14 A	ILE281 A
ASN95 A	SER1 B
GLU166 A	PHE3 B
PHE305 A	ARG4 B
SER10 B	SER10 B
ASN95 B	ASP56 B
GLU166 B	ARG60 B
PHE305 B	TRP207 B

Therefore there is a significant bi-directional coupling between the active site and site 1.

However, the same score for site 2 is only 0.49, which is only marginally above a randomly sampled site score of 0.48 (95% CI:0.47–0.48). As site 2 is located at the dimer interface, this finding is in line with the above described suggestion that the allosteric effect is not conferred by direct signalling from the dimer interface towards the catalytic centre but rather indirectly through strengthening of cooperative effects. Nonetheless, site 2 might provide scope for inhibiting the M^{pro} by disrupting dimer formation, and could help elucidate the link between domain III and the catalytic activity of the M^{pro}.

Predicted sites using Markov Transients. Overall, the observed asymmetry, from and to the active site, in the B2B-prop connectivity hints to complex communication patterns in this catalytic protein. This motivated our use of MT to reveal fast signal propagation which happens often along allosteric communication pathways. MT have

Figure 3. Putative allosteric sites identified by bond-to-bond propensities. Surface representation of the M^{pro} dimer coloured by quantile score (as shown in the legend). (A) Rotated front view with site 1 (yellow) which is located on the opposite of the orthosteric site (coloured in green). (B) Top view with site 2 (pink) located in the dimer interface. A detailed view of both sites is provided with important residues labelled.

been effective in predicting relevant sites in catalytic enzymes such as caspase-1[61] and RSK4.[68]

We performed a full MT study in the forward step (i.e., source at the active site) including quantile regression of Markov half-times against distance to find residues that are reached significantly faster than expected by perturbations emanating from the active site (see Methods and Figure 4). The QSs of all residues are shown in Figure 4(A) (as a colourmap), and the top-scoring residues (30 in total with QS > 0.95) are listed in Table S2.

This MT analysis led us to the detection of two more putative sites in the SARS-CoV-2 M^{pro}, which are located at the back of the monomer relative to the active site, as shown in Figure 4(C):

- Site 3 (Figure 4(C), framed in turquoise) is located solely in domain II and consists of 10 residues, as listed in Table S4. One of the residues is a cysteine at position 156 which might provide a suitable anchor point for covalent drug design.

- Site 4 (Figure 4(C), framed in orange) has 11 residues (list in Table S4) and and is located further down the protein in domain I.

Both sites have high average residue MT QSs of 0.87, significantly higher than the bootstrapped random site scores of 0.50 (95% CI: 0.49–0.50) and 0.49 (95% CI: 0.49–0.50), respectively.

Following the same thought process as for sites 1 and 2, we investigated the reverse step by sourcing our computations from the residues in each of sites 3 and 4 and scoring the active site to measure the impact of the putative sites on the catalytic centre.

With source at site 3, the active site has an average residue MT quantile score of 0.66, well above a random site score of 0.53 (95% CI: 0.52–0.53) thus indicating a significant reciprocal link between site 3 and the active site. For site 4 (as for site 2), on the other hand, the average residue MT quantile score of the active site is 0.52, similar to a randomly sampled site score of 0.50 (95% CI: 0.50–0.51), hence we do not detect a significant
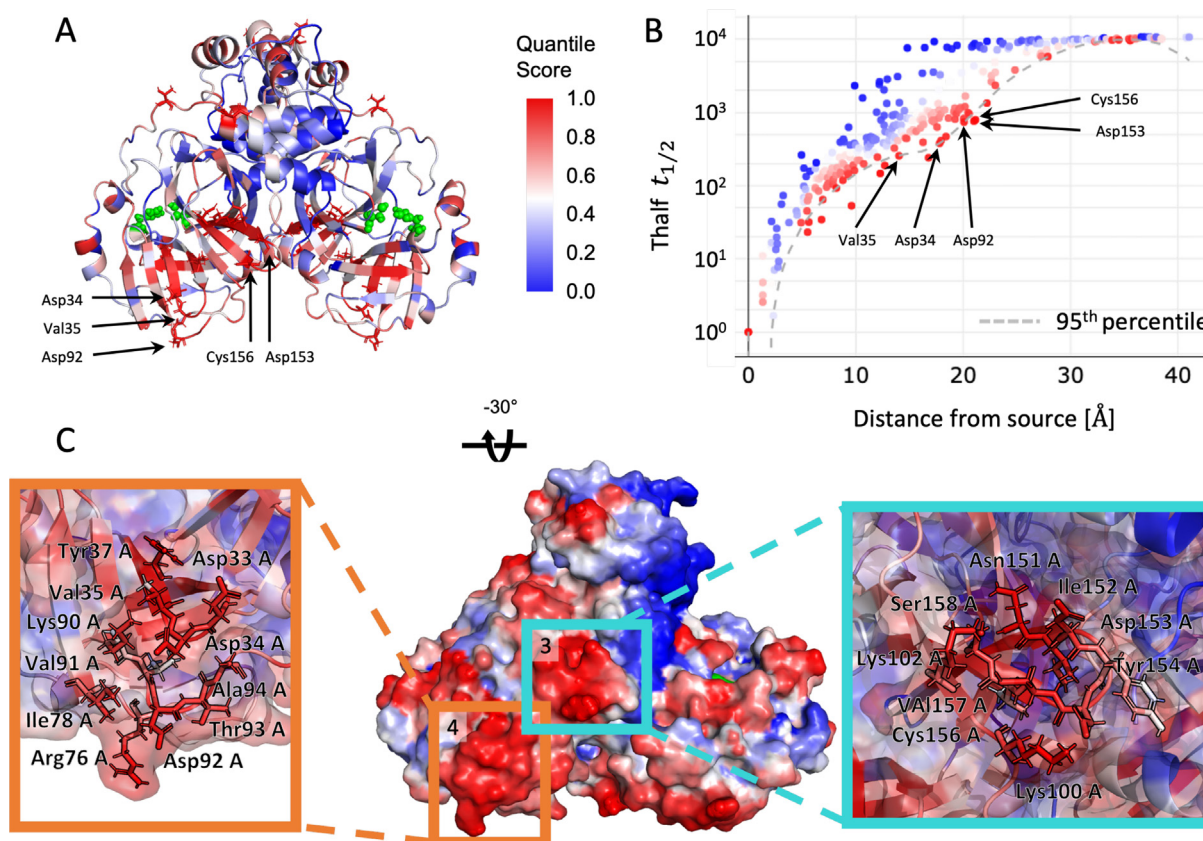

Figure 4. MT analysis of M^{pro} sourced from the orthosteric sites. (A) The orthosteric sites are shown in green and include His41 and Cys145 in both chains of the homodimer (front view). Residues with QS > 0.95 are shown as sticks. (B) The $t_{1/2}$ values of each residue are plotted against their distance from the orthosteric site. The dashed line indicates the quantile regression estimate of the 0.95 quantile cutoff used for identifying significant residues. The quantile scores of all residues are mapped onto the structure of the M^{pro} dimer (front A) view), coloured as shown in the legend. (C) Surface representation of a rotated front view of the M^{pro} dimer coloured by QS. Site 3 (turquoise) and 4 (orange) are located on the opposite side of the active site (coloured in green). A detailed view of both sites is provided with important residues labelled.

connectivity from this site back to the active site. Judging from previous experience in multimeric proteins[64] this might be due to another structural or dynamic factor not yet uncovered between site 4 and the active site.

Summary of predicted sites and evaluation of small fragment binding. We summarise the information of the predicted sites in several tables. The list of residues is presented in Tables 1, which also include the solvent accessible surface area (SASA) as a coarse indicator of targetability. Table 2 presents the average residue quantile scores for the four sites obtained with the forward step of either B2B-prop or MT compared against the statistical bootstrap. The average residue quantile scores of the active site for the reverse step (i.e., source at the identified sites) is presented in Table S5 with the corresponding statistical bootstrap values.

To provide a first indication of the druggability of the identified sites, we used the experimental results from the Diamond Light Source XChem fragment screen.[69] This screen identified 25 small fragments that bind outside of the active site of and 15 of these bind within 4 Å of at least one of the four putative allosteric sites predicted here.

Due to the computational efficiency of our methodologies, we were able to conduct a full analysis of all 15 structures using both our methods. Specifically, we used B2B-prop and MT analyses using the small fragments as sources and scored the connectivity to the active site. The results are presented in Table S6. We found that several fragments have high connectivity to the active site according to one or both of our methods. The fragment deposited with PDB dentifier5RE8 might be of particular interest as it has the highest connectivity to the active site taking both methods together. Moreover, one fragment (PDB ID: 5RFA), which is located at the dimer interface, has been found experimentally to act as a destabiliser of dimerisation and an inhibitor of M^{pro}.[29] Fragment 5RFA is only 5.8 Å away from site 2 and overlaps spatially with another fragment (PDB ID: 5RGQ) that is less than 4 Å away from site 2.

Taken together, this indicates that disrupting dimerisation provides scope for inhibition of the M^{pro}. Furthermore, in the same experimental study,[29] one of the fragments (PDB ID: 5RGJ) binds within 4 Å of site 1 and has been shown to inhibit the proteolytic activity of the M^{pro}. Fragment 5RGJ has a relatively high connectivity to the active site (Table S6).

Since the posting of our original preprint, potential allosteric pockets have been published by other research groups. Several have identified binding pockets at the dimer interface, which partially overlap or are in close proximity to our site 2.[31,32,36,38] This is in line with our conclusion above that disrupting dimerisation could inhibit M^{pro}. Shi et al.[42,43] found regulation of SARS-CoV M^{pro} by its extra domain, and cryptic sites involving the extra domain have been proposed for SARS-CoV-2 M^{pro}.[31,34–36,39] In this regard, B2B-prop highlights five residues, Asp229, Phe230, Tyr239, Lys269 and Leu272 (all with QS > 0.95), which are within the extra domain and align with pockets proposed by others. On the other hand, our sites 1, 3 and 4 all lie within domain I, and only Komatsu et al.[34] have suggested a putative site in domain I, although on the opposite side of the sites 1,3,4 found here. Therefore, our finding of sites 1, 3 and 4 suggests that more attention would be required to examine the potential of domain I to modulate M^{pro} activity.

Discussion

During the global pandemic of COVID-19 that started in January 2020, we have seen an increase of research to develop new drugs against the disease-causing virus SARS-CoV-2. A wide range of approaches from chemistry, structural biology and computational modelling have been used to identify potential protease inhibitors. However, most of these initiatives focus on investigating the active site as a drug target,[12,17] high-throughput docking approaches to the active site,[16] or re-purposing approved drugs[72] and protease inhibitors[73] which bind at the active site.

To increase the targetable space of the SARS-CoV-2 main protease and allow a broader approach to inhibitor discovery, we have provided a full computational analysis of the protease structure which gives insights into allosteric signalling and identifies potential putative target sites. Our methodologies are based on concepts from graph theory and the propagation of perturbations on a protein graph. We have previously demonstrated the applications of B2B-prop and MT analyses for the identification of allosteric sites and communication pathways in a range of biological settings[61,63,64,68] and have been benchmarked on extensive databases.[65,66] Applying B2B-prop to the SARS-CoV-2 M^{pro} revealed a hot region of connectivity at the dimer interface. Since dimerisation is known to be essential for the

Table 2 Scoring of the 4 identified putative allosteric sites (source at the active site). Included is a statistical score computed from 1,000 randomly sampled sites (with 10,000 bootstrap resamples) to obtain the 95 % confidence interval (CI).

	Quantile Score from active site	Quantile Score of random site [95 % CI]
Site 1 🟡	0.97 (*B2B-prop*)	0.53 [0.53, 0.54]
Site 2 🟣	0.96 (*B2B-prop*)	0.52 [0.51, 0.53]
Site 3 🔵	0.87 (*MT*)	0.50 [0.49, 0.51]
Site 4 🟠	0.87 (*MT*)	0.49 [0.49, 0.51]

proteolytic activity of the SARS-CoV M[pro],[15] we carried out a comparative analysis with the SARS-CoV-2 protease. Important for dimer packing and mutated in SARS-CoV-2 are residues 285 and 286.[12] When sourced from these residues, we find a higher proportion of dimer interface residues within the top 20 scoring residues for SARS-CoV-2, confirming a stronger dimer connectivity as described in literature.[12]

Therefore targeting sites at the dimer interface might provide scope for inhibitor development.[74] It is also worth remarking that, beyond the study of the dimer interface, we see a similar overall pattern of hot and cold regions in the SARS-CoV M[pro]. In particular, we find high overlap for the four identified sites (Figure S3) which gives us confidence that a potential drug effort on allosteric sites would find applications both in COVID-19 as well as SARS.

Using our approaches we identified four allosteric binding sites on the protease: Sites 1 and 2 were identified using B2B-prop (forward step) and hence have a strong connectivity to the active site at stationarity. Using our reverse step from both sites, we found that site 1 displays reciprocal connectivity to the active site, whereas site 2 (which sits at the dimer interface) is indirectly connected to the active site.

This suggests that site 1 might be a functional site and any perturbation at site 1 would induce a structural change of the protease thereby impacting the active site directly. Our methods measure connectivity through perturbations. Yet proteins can be both inhibited or activated over allosteric mechanisms, and we can not tell from our results alone whether a perturbation would lead to an up- or down-regulation of activity. However, a fragment near site 1 has been shown experimentally to exhibit some inhibitory effect on the M[pro] by El-baba *et al.*[29]

Notably, site 2, although not directly coupled to the active site as a functional site, is located at the dimer interface (Figure 3(B)) and provides a deep pocket for targeting and potentially disrupting dimer formation. Targeting site 2 could thus result in a conformational change of the protease and inhibition of dimerisation.[29]

We identified two further sites using MT. These sites are reached the fastest by a signal sourced from the active site and are both located at the back of each monomer relative to the active site. Using the reverse step, Site 3 is found to display reciprocally fast propagation back to the active site; hence perturbations at site 3 would thus potentially affect the catalytic activity of M[pro]. Site 3 (Figure 4(C)) contains a cysteine residue (Cys156) which provides an anchor point for covalently binding inhibitors.[75] Similar to site 2, site 4 does not display a reciprocally fast connection to the active site; hence actions exerted at site 4 could affect other parts of the protein which in turn could lead to an altered activity of M[pro].

We also include the analysis of structures containing small fragments bound to the structure of SARS-CoV-2 M[pro] from the Diamond Light Source XChem experimental fragment screen.[69] We analysed 15 PDB structures with fragments that bind in proximity (less than 4 Å to the putative sites, and we scored the active site using the fragments as the source. We find that several fragments display strong connectivity to the the active site, as measured by B2B-prop and MT, and some have been investigated in experimental studies[29] where their binding leads to a decrease in protease activity. Moreover, the X-ray screening study by Günther *et al.*[30] identified an allosteric site which overlaps with our allosteric hotspot 3 around residues 151 to 153. Taken together, these results show the promise of our approach, which might provide a starting point for rational drug design.

Our methods provide in depth insights into the global connectivity and have been used to propose putative allosteric sites on the main protease which could be combined with drug repurposing for approved and investigational drugs to target potential allosteric sites.[27,68] Although B2B-prop and MT are based on static structures, they can be readily applied to ensembles of structures obtained from solution nuclear magnetic resonance (NMR) or MD.[63] We hope our methods can help broaden the space of druggable targets in proteins, and aid in the development of effective medications for COVID-19 that can interfere with the main protease of SARS-CoV-2. Additionally, the high mutation rate of SARS-CoV-2 adds a further motivation to study the effect exerted by residue mutations on allostery. The computational efficiency of our methods allows further in-depth exploration of mutational analyses,[61,76,77] a direction that we leave for further research.

Methods

Protein Structures. We analysed the X-ray crystal structures of the apo conformations of the SARS-CoV-2 (PDB ID: 6Y2E[12]) and the SARS-CoV (PDB ID: 2DUC[71]) main proteases (M[pro]). The dimeric structure of the SARS-CoV-2 with PDB ID 6Y2E was constructed using the MakeMultimer.py webserver based on the BIOMT records contained in PDB files. All residues of the M[pro] proteins that are mutated between the two viruses are listed in Table S1. Both structures contained a water molecule in proximity to the catalytic dyad formed by histidine 41 and cysteine 145. These water molecules were kept while all other solvent molecules were removed. Atom and residue, secondary structural names and numberings are in accordance with the original PDB files. The dimer interface was investigated using the online tool PDBePISA[78] (for a full list of the resulting dimer interface residues see https://doi.org/10.6084/m9.figshare.12815903). We have checked for the robustness

of our results by analysing a second published structure of the SARS-Cov-2 M^{pro} with PDB ID 7JP1. The results of B2B-prop and MT analyses for both wild type structures are highly coincident as shown in Figure S1. No major differences in the spatial patterns of hot and cold regions were observed, especially in relation to the identified hot-spots as seen in Figure S2 and Table S7.

Atomistic Graph Construction. Instead of the coarse-grained descriptions typical of most network methods for protein analysis, we use protein data bank (PDB)[79] structure files to derive fully atomistic protein graphs from the three-dimensional protein structures. In our graph, the nodes are atoms and the weighted edges represent interactions, both covalent bonds and weak interactions, including hydrophobic, hydrogen bonds and salt bridges (See Figure 5). The edges are weighted with physico-chemical energies obtained from well studied potentials, as discussed below. Details of earlier versions of this approach can be found in Refs.[60,61,63] We summarise briefly the main features below and we note three further improvements in the current version: (i) the stand-alone detection of edges without need of third-party software; (ii) the many-body detection of hydrophobic edges across scales; and (iii) the improved computational efficiency of the code. For further details of the updated atomistic graph construction used in this work see.[65,62]

Figure 5 gives an overview of the workflow. We start from atomistic cartesian coordinates of a PDB file. Since X-ray structures do not include hydrogen atoms and NMR structures may not report all of them, we use the software *Reduce*[80] to add any missing hydrogen atoms. Hydrophobic interactions and hydrogen bonds are identified with a cutoff of 9 Å and 0.01 kcal/mol respectively. In addition, hydrogen bonds are also identified based on the angles related to the hybridisation of the donor - acceptor atoms. The edges are weighted by their energies: covalent bond energies from their

bond-dissociation energies;[81] hydrogen bonds and salt bridges by the modified Mayo potential;[82,83] hydrophobic interactions by using a hydrophobic potential of mean force.[84]

Bond-to-bond propensity. Bond-to-bond propensity (B2B-prop) analysis was first introduced in Ref.[63] and further discussed in Ref.[64], hence we only briefly summarise it here. This edge-space measure evaluates the redistribution of perturbations introduced at the source towards every bond in the protein graph at stationarity. The edge-to-edge transfer matrix *M* was introduced to study non-local edge-coupling and flow redistribution in graphs[85] and an alternative interpretation of *M* as a Green's function is employed to analyse the atomistic protein graph. The element M_{ij} describes the effect that a perturbation at edge *i* has on edge *j*. *M* is given by

$$M = \frac{1}{2} W B^T L^\dagger B. \tag{1}$$

Here *B* is the $n \times m$ incidence matrix for the atomistic protein graph with *n* nodes and *m* edges; $W = \text{diag}(w_{ij})$ is an $m \times m$ diagonal matrix where w_{ij} is the weight of the edge connecting nodes *i* and *j*, i.e. the bond energy between those atoms; and $L^\dagger$ is the pseudo-inverse of the weighted graph Laplacian matrix L[86] and defines the diffusion dynamics on the energy-weighted graph.[87]

To evaluate the effect of perturbations from a group of bonds b' (i.e., the source), on bond *b* of other parts of the protein, we define the bond propensity as:

$$\Pi_b = \sum_{b' \in \text{source}} |M_{bb'}| \tag{2}$$

and then calculate the residue propensity of a residue *R*:

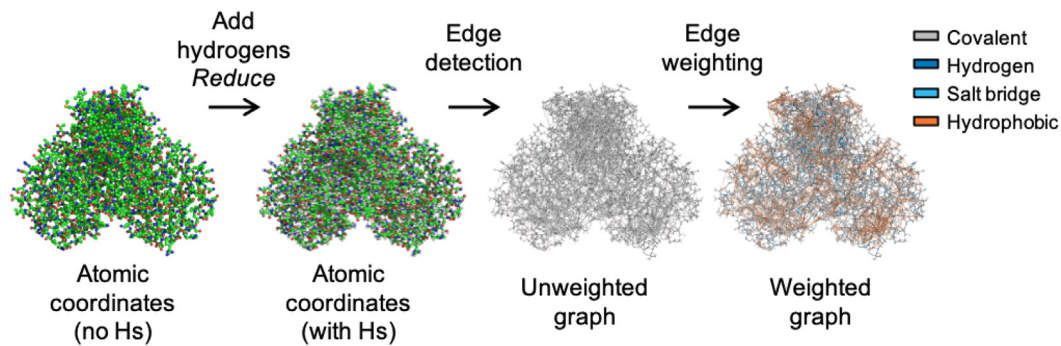$$\Pi_R = \sum_{b \in R} \Pi_b. \tag{3}$$

Figure 5. Atomistic Graph Construction. We showcase the general procedure here on the main protease of SARS-Cov-2: Atomic coordinates are obtained from the PDB (ID: 6Y2E[12]) and hydrogens are added by Reduce.[80] Edges are identified and the weights are assigned, as described in the methods section, by taking into account covalent bonds as well as weak interactions: hydrogen bonds, electrostatic interactions and the hydrophobic effect which are colour.ed as indicated.

Markov Transients. A complementary, node-based method, Markov Transients (MT) identifies areas of the protein that are significantly connected to the source, usually a site of interest such as the active site, by computing the speed of signal propagation at the atomistic level. The method was introduced and discussed in detail in Ref.[61] and has successfully identified allosteric hotspots and pathways without *a priori* knowledge.[61,68] Importantly, it captures *all* paths that connect the two sites. The contribution of each atom in the communication pathway between the active site and all other sites in a protein or protein complex is measured by the characteristic transient time $t_{1/2}$ (or t_{half}):

$$t_{1/2}^{(i)} = \arg\min_t \left[p_t^{(i)} \geqslant \frac{\pi^{(i)}}{2} \right], \tag{4}$$

where $t_{1/2}^{(i)}$ is the number of time steps in which the probability of a random walker to be at node i reaches half its stationary value. This provides a measure of the speed by which perturbations originating from the active site diffuse into the rest of the protein by a random walk on the above described atomistic protein graph. To obtain the transient time $t_{1/2}$ for each residue, we take the average $t_{1/2}$ over all atoms of the respective residue.

Quantile Regression. To determine the significant bonds with high bond-to-bond propensity and atoms with fast transient times $t_{1/2}$ at the same geometric distance from the source, we use conditional quantile regression (QR),[88] a robust statistical measure widely used in different areas of science.[89] In contrast to standard least squares regressions, QR provides models for conditional quantile functions which are obtained by solving an optimisation problem (a linear program). This is significant here because it allows us to identify not the "average" atom or bond but those that are outliers from all those found at the same distance from the active site and because we are looking at the tails of highly non-normal distributions.

As the distribution of propensities over distance follows an exponential decay, we use a linear function of the logarithm of propensities when performing QR for B2B-prop, whereas in the case of the $t_{1/2}$ of MT, which do not follow a particular parametric dependence on distance, we use cubic splines to retain flexibility. From the estimated quantile regression functions, we then compute the quantile score (QS) for each atom or bond. To obtain residue QSs, we use the minimum distance between each atom of a residue and those of the source. Further details of this approach for B2B-prop analysis can be found in Ref.[63] and for MT analysis in Ref.[76].

Site scoring with structural bootstrap sampling. To assess the statistical significance of a site of interest, we score the site against 1000 randomly sampled sites of the same size. For this purpose, the average residue QS of the site of interest is calculated. After sampling 1000 random sites on the protein, the average residue QSs of these sites are calculated. By performing a bootstrap with 10,000 resamples with replacement on the random sites average residue QSs, we are able to provide a 95% confidence interval to assess the statistical significance of the site of interest score in relation to the random site score.

Residues used when scoring the active site (reverse step). To identify the residues that are used as the active site to be scored in the reverse step of both B2B-prop and MT analyses, we proceed as follows. First, we found all structures with non-covalent hits bound in the active site from the XChem fragment screen against the SARS-CoV-2 M^pro.[69] There were 22 such structures. These structures were further investigated using PyMOL (v.2.3)[90] to identify residues that have atoms within 4 Å of any of the bound fragments. These residues are Thr25, Thr26, His41, Cys44, Thr45, Ser46, Met49, Tyr54, Phe140, Leu141, Asn142, Ser144, Cys145, Met162, His163, His164, Met165, Glu166, Leu167, Pro168, Asp187, Arg188, Gln189, Thr190. This list constitutes the active site as a site of interest in all scoring calculations.

XChem fragment screen hits selection. From the above mentioned XChem fragment screen against the SARS-CoV-2 M^pro,[69] 25 hits were found at regions other than the active site. The 15 fragments which contain atoms that are within 4 Å from any of the putative allosteric site residues we obtained were selected as candidates for further investigation as shown in Table 3.

For each of these fragment-bound structures, we performed bond-to-bond propensity and Markov transient analyses to evaluate the connectivity to the active site. The active site was scored as described above.

Visualisation and Solvent Accessible Surface Area. We use PyMOL (v.2.3)[90] for structure visualisation, and presentation of Markov Transients and bond-to-bond propensity scores on the structure. PyMOL was also used to calculate the residue solvent accessible surface area (SASA) with a rolling probe radius of 1.4 and a sampling density of 2.

Data Availability

All data presented in this study are available at figshare with DOI: 10.6084/m9.figshare.12815903.

Table 3 XChem fragments in 4 Å proximity to the identified allosteric sites.

Site	Fragment PDB ID
Site 1	5RGJ, 5RE8, 5RF4, 5RF9, 5RFD, 5RED, 5REI, 5RF5, 5RGR
Site 2	5RF0, 5RGQ
Site 3	5RF9
Site 4	5RGG, 5RE5, 5RE7, 5RFC, 5RE8, 5RF4, 5RFD

Author Contributions

L.S., N.W., M.B and S.N.Y. conceived the study. L.S and N.W. performed the computations, L.S. created the figures and all authors analysed the data and wrote the manuscript.

CRediT authorship contribution statement

Léonie Strömich: Conceptualisation, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – Original draft preparation, Writing – Reviewing and Editing, Visualisation. **Nan Wu:** Software, Formal analysis, Investigation, Data curation, Writing – Original draft preparation, Writing – Reviewing and Editing, Visualisation. **Mauricio Barahona:** Methodology, Writing – Reviewing and Editing. **Sophia N. Yaliraki:** Conceptualisation, Methodology, Writing – Reviewing and Editing, Supervision.

Materials & Correspondence

All requests for data and code shall be directed to s.yaliraki@imperial.ac.uk.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jmb.2022.167748.

References

1. Zhou, P. et al, (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273. https://doi.org/10.1038/s41586-020-2012-7.
2. Wu, F. et al, (2020). A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269. https://doi.org/10.1038/s41586-020-2008-3.
3. Zhu, N. et al, (2020). A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733. https://doi.org/10.1056/NEJMoa2001017.
4. Gorbalenya, A.E. et al, (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiol.* **5**, 536–544. https://doi.org/10.1038/s41564-020-0695-z.
5. Peiris, J.S.M., Guan, Y., Yuen, K.Y., (2004). The severe acute respiratory syndrome. *Nat. Med.* **10**, S88–S97. https://doi.org/10.1038/nm1143.
6. Graham, R.L., Donaldson, E.F., Baric, R.S., (2013). A decade after SARS: strategies for controlling emerging coronaviruses. *Nat. Rev. Microbiol.* **11**, 836–848. https://doi.org/10.1038/nrmicro3143.
7. Steinhauer, D.A., Holland, J.J., (1986). Direct method for quantitation of extreme polymerase error frequencies at selected single base sites in viral RNA. *J. Virol.* **57**, 219–228. https://doi.org/10.1128/JVI.57.1.219-228.1986.
8. Tan, Z.W. et al, (2022). Allosteric perspective on the mutability and druggability of the SARS-CoV-2 Spike protein. *Structure* **30**, 590–607.e4. https://doi.org/10.1016/j.str.2021.12.011.
9. Anand, K. et al, (2002). Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain. *EMBO J.* **21**, 3213–3224. https://doi.org/10.1093/emboj/cdf327.
10. Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J.R., Hilgenfeld, R., (2003). Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science* **300**, 1763–1767. https://doi.org/10.1126/science.1085658.
11. Yang, H. et al, (2003). The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proc. Nat. Acad. Sci. USA* **100**, 13190–13195. https://doi.org/10.1073/pnas.1835675100.
12. Zhang, L. et al, (2020). Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science* **368**, 409–412. https://doi.org/10.1126/science.abb3405.
13. Hilgenfeld, R., (2014). From SARS to MERS: crystallographic studies on coronaviral proteases enable antiviral drug design. *FEBS J.* **281**, 4085–4096. https://doi.org/10.1111/febs.12936.
14. Chen, Y.W., Yiu, C.P.B., Wong, K.Y., (2020). Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CLpro)

structure: Virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. *F1000Research* **9** https://doi.org/10.12688/f1000research.22457.2.

15. Lim, L., Shi, J., Mu, Y., Song, J., (2014). Dynamically-driven enhancement of the catalytic machinery of the SARS 3C-like protease by the S284–T285-I286/A mutations on the extra domain. *PLoS ONE* **9**

16. Ton, A.-T., Gentile, F., Hsing, M., Ban, F., Cherkasov, A., (2020). Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds. *Mol. Informat.* **39**, 2000028. https://doi.org/10.1002/minf.202000028.

17. Jin, Z. et al, (2020). Structural basis for the inhibition of SARS-CoV-2 main protease by antineoplastic drug carmofur. *Nature Struct. Mol. Biol.* **27**, 529–532. https://doi.org/10.1038/s41594-020-0440-6.

18. Jin, Z. et al, (2020). Structure of M(pro) from SARS-CoV-2 and discovery of its inhibitors. *Nature* **582**, 289–293. https://doi.org/10.1038/s41586-020-2223-y.

19. Ullrich, S., Nitsche, C., (2020). The SARS-CoV-2 main protease as drug target. *Bioorg. Med. Chem. Lett.* **30**, 127377. https://doi.org/10.1016/j.bmcl.2020.127377.

20. Yang, H. et al, (2005). Design of Wide-Spectrum Inhibitors Targeting Coronavirus Main Proteases. *PLoS Biol.* **3**, e324. https://doi.org/10.1371/journal.pbio.0030324.

21. Pillaiyar, T., Manickam, M., Namasivayam, V., Hayashi, Y., Jung, S.-H., (2016). An Overview of Severe Acute Respiratory Syndrome-Coronavirus (SARS-CoV) 3CL Protease Inhibitors: Peptidomimetics and Small Molecule Chemotherapy. *J. Med. Chem.* **59**, 6595–6628. https://doi.org/10.1021/acs.jmedchem.5b01461.

22. Dyall, J. et al, (2017). Middle East Respiratory Syndrome and Severe Acute Respiratory Syndrome: Current Therapeutic Options and Potential Targets for Novel Therapies. *Drugs* **77**, 1935–1966. https://doi.org/10.1007/s40265-017-0830-1.

23. Rudmann, D.G., (2012). On-target and Off-target-based Toxicologic Effects. *Toxicol. Pathol.* **41**, 310–314. https://doi.org/10.1177/0192623312464311.

24. Guengerich, F.P., (2011). Mechanisms of drug toxicity and relevance to pharmaceutical development. *Drug Metabol. Pharmacokinet.* **26**, 3–14. https://doi.org/10.2133/dmpk.dmpk-10-rv-062.

25. Wenthur, C.J., Gentry, P.R., Mathews, T.P., Lindsley, C. W., (2014). Drugs for Allosteric Sites on Receptors. *Annu. Rev. Pharmacol. Toxicol.* **54**, 165–184. https://doi.org/10.1146/annurev-pharmtox-010611-134525.

26. Cimermancic, P. et al, (2016). CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *J. Mol. Biol.* **428**, 709–719. https://doi.org/10.1016/j.jmb.2016.01.029.

27. Drug repurposing: progress, challenges and recommendations (2019, journal = Nature Reviews Drug Discovery, author = Pushpakom, Sudeep and Iorio, Francesco and Eyers, Patrick A and Escott, K Jane and Hopper, Shirley and Wells, Andrew and Doig, Andrew and Guilliams, Tim and Latimer, Joanna and McNamee, Christine and Norris, Alan and Sanseau, Philippe and Cavalla, David and Pirmohamed, Munir, number = 1, pages = 41–58, volume = 18, url = https://doi.org/10.1038/nrd.2018.168, issn = 1474-1784).

28. Sultana, J. et al. (2020). Challenges for Drug Repurposing in the COVID-19 Pandemic Era. https://www.frontiersin.org/article/10.3389/fphar.2020.588654.

29. El-baba, T.J. et al, (2020). Allosteric inhibition of the SARS-CoV-2 main protease - insights from mass spectrometry-based assays. *Angew. Chem. Int. Ed.*. https://doi.org/10.1002/anie.202010316.

30. Günther, S. et al, (2021). X-ray screening identifies active site and allosteric inhibitors of SARS-CoV-2 main protease. *Science* **372**, 642–646. https://doi.org/10.1126/science.abf7945.

31. Bhat, Z.A., Chitara, D., Iqbal, J., Sanjeev, B.S., Madhumalar, A., (2021). Targeting allosteric pockets of SARS-CoV-2 main protease Mpro. *J. Biomol. Struct. Dyn.*. https://doi.org/10.1080/07391102.2021.1891141.

32. Novak, J. et al, (2021). Proposition of a new allosteric binding site for potential SARS-CoV-2 3CL protease inhibitors by utilizing molecular dynamics simulations and ensemble docking. *J. Biomol. Struct. Dyn.*, 1–14. https://doi.org/10.1080/07391102.2021.1927845.

33. Amamuddy, O.S., Boateng, R.A., Barozi, V., Nyamai, D. W., Bishop, Ö.T., (2021). Novel dynamic residue network analysis approaches to study allosteric modulation: SARS-CoV-2 Mpro and its evolutionary mutations as a case study. *Comput. Struct. Biotechnol. J.* **19**, 6431–6455. https://doi.org/10.1016/j.csbj.2021.11.016.

34. Komatsu, T.S. et al, (2020). Drug Binding Dynamics of the Dimeric SARS-CoV-2 Main Protease, determined by Molecular Dynamics Simulation. *Sci. Reports* **10**, 16986. https://doi.org/10.1038/s41598-020-74099-5.

35. Carli, M., Sormani, G., Rodriguez, A., Laio, A., (2021). Candidate Binding Sites for Allosteric Inhibition of the SARS-CoV - 2 Main Protease from the Analysis of Large-Scale Molecular Dynamics Simulations. *J. Phys. Chem. Lett.* **12**, 65–72. https://doi.org/10.1021/acs.jpclett.0c03182.

36. Sztain, T., Amaro, R., McCammon, J.A., (2021). Elucidation of Cryptic and Allosteric Pockets within the SARS-CoV-2 Main Protease. *J. Chem. Inf. Model.*.

37. Verma, S., Pandey, A.K., (2021). Factual insights of the allosteric inhibition mechanism of SARS-CoV-2 main protease by quercetin: an in silico analysis. *Biotech* **11** https://doi.org/10.1007/s13205-020-02630-6.

38. Amamuddy, O.S., Boateng, G.M., Verkhivker, Bishop, Ö. T., (2020). Impact of Early Pandemic Stage Mutations on Molecular Dynamics of SARS-CoV-2 Mpro. *J. Chem. Informat. Model.* **60**, 5080–5102. https://doi.org/10.1021/acs.jcim.0c00634.

39. Dubanevics, I., McLeish, T.C., (2021). Computational analysis of dynamic allostery and control in the SARS-CoV-2 main protease. *J. Roy. Soc. Interface* **18** https://doi.org/10.1098/rsif.2020.0591rsif20200591.

40. DasGupta, D., Chan, W.K.B., Carlson, H.A., (2022). Computational Identification of Possible Allosteric Sites and Modulators of the SARS-CoV-2 Main Protease. *J. Chem. Inf. Model.* **62**, 618–626. https://doi.org/10.1021/acs.jcim.1c01223.

41. Shi, J., Wei, Z., Song, J., (2004). Dissection study on the severe acute respiratory syndrome 3C-like protease reveals the critical role of the extra domain in dimerization of the enzyme: defining the extra domain as a new target for design of highly specific protease inhibitors. *J. Biol. Chem.* **279**, 24765–24773. https://doi.org/10.1074/jbc.M311744200.

42. Shi, J., Song, J., (2006). The catalysis of the SARS 3C-like protease is under extensive regulation by its extra domain.

FEBS J. **273**, 1035–1045. https://doi.org/10.1111/j.1742-4658.2006.05130.x.

43. Shi, J. et al, (2011). Dynamically-Driven Inactivation of the Catalytic Machinery of the SARS 3C-Like Protease by the N214A Mutation on the Extra Domain. *PLOS Comput. Biol.* **7**, e1001084. https://doi.org/10.1371/journal.pcbi.1001084.

44. Greener, J.G., Sternberg, M.J., (2018). Structure-based prediction of protein allostery. *Curr. Opin. Struct. Biol.* **50**, 1–8. https://doi.org/10.1016/j.sbi.2017.10.002.

45. Lu, S., He, X., Ni, D., Zhang, J., (2019). Allosteric Modulator Discovery: From Serendipity to Structure-Based Design. *J. Med. Chem.* **62** https://doi.org/10.1021/acs.jmedchem.8b01749. acs.jmedchem.8b01749.

46. Monod, J., Changeux, J.-P., Jacob, F., (1963). Allosteric proteins and cellular control systems. *J. Mol. Biol.* **6**, 306–329. URL https://www.sciencedirect.com/science/article/pii/S0022283663800911.

47. Koshland, D.E., Némethy, G., Filmer, D., (1966). Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits*. *Biochemistry* **5**, 365–385. URL http://pubs.acs.org/doi/abs/10.1021/bi00865a047.

48. Tsai, C.J., Nussinov, R., (2014). A Unified View of How Allostery Works. *PLoS Comput. Biol.* **10** https://doi.org/10.1371/journal.pcbi.1003394.

49. Ribeiro, A.A., Ortiz, V., (2016). A Chemical Perspective on Allostery. *Chem. Rev.* **116**, 6488–6502. https://doi.org/10.1021/acs.chemrev.5b00543.

50. Shukla, D., Meng, Y., Roux, B., Pande, V.S., (2014). Activation pathway of Src kinase reveals intermediate states as targets for drug design. *Nature Commun.* **5**, 3397. https://doi.org/10.1038/ncomms4397.

51. Penkler, D., Sensoy, Ö., Atilgan, C., Tastan Bishop, Ö., (2017). Perturbation-Response Scanning Reveals Key Residues for Allosteric Control in Hsp70. *J. Chem. Inf. Model.* **57**, 1359–1374. https://doi.org/10.1021/acs.jcim.6b00775.

52. Panjkovich, A., Daura, X., (2012). Exploiting protein flexibility to predict the location of allosteric sites. *BMC Bioinformat.* **13**, 273. https://doi.org/10.1186/1471-2105-13-273.

53. Panjkovich, A., Daura, X., (2014). PARS: a web server for the prediction of Protein Allosteric and Regulatory Sites. *Bioinformatics* **30**, 1314–1315. https://doi.org/10.1093/bioinformatics/btu002.

54. Greener, J.G., Sternberg, M.J.E., (2015). AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. *BMC Bioinformatics* **16**, 335. https://doi.org/10.1186/s12859-015-0771-1.

55. Song, K. et al, (2017). Improved Method for the Identification and Validation of Allosteric Sites. *J. Chem. Inf. Model.* **57**, 2358–2363. https://doi.org/10.1021/acs.jcim.7b00014.

56. Guarnera, E., Berezovsky, I.N., (2016). Structure-Based Statistical Mechanical Model Accounts for the Causality and Energetics of Allosteric Communication. *PLoS Comput. Biol.* **12** https://doi.org/10.1371/journal.pcbi.1004678. e1004678–e1004678.

57. Tee, W.-V., Guarnera, E., Berezovsky, I.N., (2018). Reversing allosteric communication: From detecting allosteric sites to inducing and tuning targeted allosteric response. *PLOS Comput. Biol.* **14**, e1006228. https://doi.org/10.1371/journal.pcbi.1006228.

58. Wang, J. et al, (2020). Mapping allosteric communications within individual proteins. *Nature Commun.* **3862** https://doi.org/10.1038/s41467-020-17618-2.

59. Collier, G., Ortiz, V., (2013). Emerging computational approaches for the study of protein allostery. *Arch. Biochem. Biophys.* **538**, 6–15. https://doi.org/10.1016/j.abb.2013.07.025.

60. Delmotte, A., Tate, E.W., Yaliraki, S.N., Barahona, M., (2011). Protein multi-scale organization through graph partitioning and robustness analysis: application to the myosin-myosin light chain interaction. *Phys. Biol.* **8**, 055010. https://doi.org/10.1088/1478-3975/8/5/055010.

61. Amor, B., Yaliraki, S.N., Woscholski, R., Barahona, M., (2014). Uncovering allosteric pathways in caspase-1 using Markov transient analysis and multiscale community detection. *Mol. BioSyst.* **10**, 2247–2258. https://doi.org/10.1039/C4MB00088A.

62. Song, F., Barahona, M. & Yaliraki, S.N. (2020). BagPyPe: A Python package for the construction of atomistic, energy-weighted graphs from biomolecular structures. Manuscript in preparation.

63. Amor, B.R.C., Schaub, M.T., Yaliraki, S.N., Barahona, M., (2016). Prediction of allosteric sites and mediating interactions through bond-to-bond propensities. *Nature Commun.* **7**, 12477. https://doi.org/10.1038/ncomms12477.

64. Hodges, M., Barahona, M., Yaliraki, S.N., (2018). Allostery and cooperativity in multimeric proteins: bond-to-bond propensities in ATCase. *Sci. Reports* **8**, 11079. https://doi.org/10.1038/s41598-018-27992-z.

65. Mersmann, S.F. et al, (2021). ProteinLens: a web-based application for the analysis of allosteric signalling on atomistic graphs of biomolecules. *Nucleic Acids Res.*. https://doi.org/10.1093/nar/gkab350.

66. Wu, N., Strömich, L., Yaliraki, S.N., (2021). Prediction of allosteric sites and signaling: Insights from benchmarking datasets. *Patterns* **3**, 100408.

67. del Sol, A., Tsai, C.-J., Ma, B., Nussinov, R., (2009). The origin of allosteric functional modulation: multiple pre-existing pathways. *Structure* **17**, 1042–1050. https://doi.org/10.1016/j.str.2009.06.008.

68. Chrysostomou, S. et al, (2021). Repurposed floxacins targeting RSK4 prevent chemoresistance and metastasis in lung and bladder cancer. *Sci. Translat. Med.* **13**, eaba4627. URL http://stm.sciencemag.org/content/13/602/eaba4627.abstract.

69. Douangamath, A. et al, (2020). Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. *Nature Commun.* **11**, 5047. https://doi.org/10.1038/s41467-020-18709-w.

70. Chou, C.Y. et al, (2004). Quaternary structure of the severe acute respiratory syndrome (SARS) coronavirus main protease. *Biochemistry* **43**, 14958–14970. https://doi.org/10.1021/bi0490237.

71. Muramatsu, T. et al, (2016). SARS-CoV 3CL protease cleaves its C-terminal autoprocessing site by novel subsite cooperativity. *Proc. Nat. Acad. Sci. USA* **113**, 12997–13002. https://doi.org/10.1073/pnas.1601327113.

72. Mahanta, S. et al, (2020). Potential anti-viral activity of approved repurposed drug against main protease of SARS-CoV-2: an in silico based approach. *J. Biomol. Struct. Dyn.*. https://doi.org/10.1080/07391102.2020.1768902.

73. Eleftheriou, P., Amanatidou, D., Petrou, A., Geronikaki, A., (2020). In Silico Evaluation of the Effectivity of Approved Protease Inhibitors against the Main Protease of the Novel SARS-CoV-2 Virus. *Molecules* **25**, 2529. https://doi.org/10.3390/molecules25112529.

74. Goyal, B., Goyal, D., (2020). Targeting the Dimerization of the Main Protease of Coronaviruses: A Potential Broad-Spectrum Therapeutic Strategy. *ACS Combinat. Sci.* **22**, 297–305. https://doi.org/10.1021/acscombsci.0c00058.

75. Hallenbeck, K., Turner, D., Renslo, A., Arkin, M., (2017). Targeting Non-Catalytic Cysteine Residues Through Structure-Guided Drug Discovery. *Curr. Top. Med. Chem.* **17**, 4–15. https://doi.org/10.2174/1568026616666160719163839.

76. Amor, B.R.C. (2016). Exploring allostery in proteins with graph theory. Ph.D. thesis, Imperial College London. URL https://doi.org/10.25560/58214.

77. Peach, R.L. et al. (2019). Unsupervised Graph-Based Learning Predicts Mutations That Alter Protein Dynamics. bioRxiv. https://www.biorxiv.org/content/early/2019/11/20/847426.

78. Krissinel, E., Henrick, K., (2007). Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **372**, 774–797. https://doi.org/10.1016/j.jmb.2007.05.022.

79. Berman, H.M. et al, (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242. https://doi.org/10.1093/nar/28.1.235.

80. Word, J., Lovell, S.C., Richardson, J.S., Richardson, D.C., (1999). Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **285**, 1735–1747. https://doi.org/10.1006/jmbi.1998.2401.

81. Huheey, J.E., Keiter, E.A., Keiter, R.L., (1993). Inorganic chemistry: principles of structure and reactivity. HarperCollins College Publishers, New York, NY.

82. Mayo, S.L., Olafson, B.D., Goddard, W.A., (1990). DREIDING: A generic force field for molecular simulations. *J. Phys. Chem.* **94**, 8897–8909. https://doi.org/10.1021/j100389a010.

83. Dahiyat, B.I., Gordon, D.B., Mayo, S.L., (1997). Automated design of the surface positions of protein helices. *Protein Sci.* **6**, 1333–1337. https://doi.org/10.1002/pro.5560060622.

84. Lin, M.S., Fawzi, N.L., Head-Gordon, T., (2007). Hydrophobic Potential of Mean Force as a Solvation Function for Protein Structure Prediction. *Structure* **15**, 727–740. https://doi.org/10.1016/j.str.2007.05.004.

85. Schaub, M.T., Lehmann, J., Yaliraki, S.N., Barahona, M., (2014). Structure of complex networks: Quantifying edge-to-edge relations by failure-induced flow redistribution. *Network Sci.* **2**, 66–89. https://doi.org/10.1017/nws.2014.4.

86. Biggs, N., (1993). *Algebraic Graph Theory*, **vol. 67** Cambridge University Press.

87. Lambiotte, R., Delvenne, J., Barahona, M., (2014). Random Walks, Markov Processes and the Multiscale Modular Organization of Complex Networks. *IEEE Trans. Network Sci. Eng.* **1**, 76–90. https://doi.org/10.1109/TNSE.2015.2391998.

88. Koenker, R., Hallock, K.F., (2001). Quantile Regression. *J. Econ. Perspect.* **15**, 143–156. https://doi.org/10.1257/jep.15.4.143.

89. Koenker, R. (2019). Quantreg: Quantile Regression. R package version 5.52. https://cran.r-project.org/package=quantreg.

90. Schrodinger/pymol-open-source. (2020). Open-source foundation of the user-sponsored PyMOL molecular visualization system. https://github.com/schrodinger/pymol-open-source.