ALLOSTERIC HOTSPOTS IN THE MAIN PROTEASE OF SARS-COV-2

A PREPRINT

Léonie Strömich
Department of Chemistry
Imperial College London

Nan Wu
Department of Chemistry
Imperial College London

Mauricio Barahona
Department of Mathematics
Imperial College London

Sophia N. Yaliraki*
Department of Chemistry
Imperial College London
s.yaliraki@imperial.ac.uk

November 5, 2020

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 analysis, which has been previously successful in identifying allosteric sites without *a priori* knowledge in benchmark data sets, and, Markov transient analysis, which has previously aided in finding novel drug targets in catalytic protein families. We further score the highest ranking sites against random sites in similar distances through statistical bootstrapping and identify four statistically significant putative allosteric sites as good candidates for alternative drug targeting.

1 Introduction

The global pandemic of COVID-19 (coronavirus disease 2019) is caused by the newly identified virus SARS-CoV-2 [1, 2, 3, 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] 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 [8, 9, 10]. 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 [11] (see Fig. S1B) which is responsible for processing the polyproteins translated from the viral RNA [12].

The M^{pro} of the new SARS-CoV-2 shares 96% sequence similarity with that of SARS-CoV, which also extends to a high structural similarity (r.m.s deviation of 0.53 Å between Cα positions) [11]. Moreover, many of the residues which are important for catalytic activity, substrate binding and dimerisation are conserved between these species [13]. Nevertheless, focusing on the mutations from SARS-CoV to SARS-CoV-2, several are located at the dimer interface (for a full list see Table S1) and it has also been suggested that the mutations Thr285Ala and Ile286Leu (see Fig. S1) are responsible for a closer dimer packing [11]. Previous mutational studies on these positions in SARS-CoV M^{pro} have revealed an impact on catalytic activity [14].

Currently, the development of SARS-CoV-2 M^{pro} inhibitors [11, 15, 16, 17], similarly to designing other coronavirus M^{pro} inhibitors [18, 19, 20], focuses on blocking the orthosteric sites to disrupt viral replication (reviewed in Ullrich & Nitsche [21]). Targeting the active site enables high affinity of the drug molecules but could result in off-target-based toxicity when binding to proteins with similar active sites. Drug resistance is another major concern, especially when the active site may potentially alter owing to mutations. Targeting an allosteric site which is distal from the main binding site provides an alternative attractive solution by increasing both the range and selectivity of drugs to fine-tune protein activity without the aforementioned disadvantages. (For reviews and recent successes see Wenthur *et al.* and Cimermancic *et al.* [22, 23]). To the best of our knowledge, there is to date no indication of such putative allosteric sites of the coronavirus M^{pro}s in the literature other than a recent implication of potential allosteric regulation of SARS-CoV-2 M^{pro} [24] and simulated binding events to distant areas of the protein [25]. Encouragingly, however, there have been indications of allosteric processes mediated by the extra domain in the SARS-CoV M^{pro} [26, 27, 28, 14].

Here, we focus on investigating the allosteric properties of the protease and in particular whether there are indeed any strongly connected allosteric sites to the active site that may offer alternative ways to inhibit the virus reproduction. Despite being an attractive drug alternative approach, the identification of allosteric sites remains challenging and is still often done serendipitously. Computational prediction and description of allosteric sites has become an active field of research for allosteric drug design (for reviews see [29, 30]) as it does not require the laborious and time-consuming compound screening process. For example, molecular dynamics (MD) simulations model proteins at the atomic level and the communication pathways detected can be exploited for allosteric residue and site identification [31, 32]. To alleviate the substantial computational resources required by MD simulations and the inability to explore all the required scales involved, variations of normal mode analysis (NMA) of elastic network models (ENM) are widely applied and have achieved moderate accuracy in allosteric site detection when tested on known allosteric proteins [33, 34, 35, 36]. The field of methods for allosteric pathway or site prediction is continuously growing, with new methods ranging from statistical mechanical models [37, 38] to methods based on graph theory [39]. However, even if they overcome the computational resource requirement of atomistic MD, they do so at the cost of resolution by looking at coarse-grained structure representations.

To overcome these limitations, we recently introduced a range of methods based on high resolution atomistic graph analysis which are computationally efficient while at the same time provide insights into the global effects on a protein structure without *a priori* guidance. These computational frameworks retain key physico-chemical details through the derivation of an energy-weighted atomistic protein graph from structural information which incorporates both covalent and weak interactions which are known to be important in allosteric signalling (hydrogen bonds, electrostatics and hydrophobics) through interatomic potentials [40, 41, 42]. Based on this atomistic graph, Bond-to-bond propensity analysis quantitatively shows how an energy fluctuation in a given set of bonds significantly affects any other bond in the graph and provides a measure for instantaneous connectivity. Unlike most graph or network approaches, it is formulated on the bonds or edges of the graph and thus makes a direct link between energy and flow through bonds of the system [43]. It has been shown that Bond-to-bond propensities are capable of successfully predicting allosteric sites in a wide range of proteins without any *a priori* knowledge other than the active site [43]. Of particular relevance to the homodimeric protease studied here, it has been subsequently used to show how allostery and cooperativity are intertwined in multimeric enzymes such as the well studied aspartate carbamoyltransferase (ATCase) [44]. A complementary methodology, Markov transient analysis, further sheds light on the catalytic aspects of allostery and obtains the pathways implicated in allosteric regulation through the transients of the propagation of a random walker on the node space of the atomistic graph [41]. Crucially, while most methods obtain the shortest or optimal path, the method takes into account *all* possible pathways, as allosteric communication is known to involve multiple paths [45]. In doing so, Markov transient analysis has been successful in identifying allosteric paths in caspase-1 [41] as well as previously unknown allosteric inhibitor binding sites in p90 ribosomal s6 kinase 4 (RSK4) which complemented drug repurposing in lung cancer [46]. These two methods are complementary in their application as they have been shown to provide different insights based on the underlying allosteric mechanisms: Bond-to-bond propensity analysis gives insights into the structural connectivity while Markov transient analysis is better suited for the catalytic and time scale dependent aspects of a protein.

We here showcase the application of these 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. Our results shed light on the allosteric communication patterns in the M^{pro} dimer. They further highlight the role of the interface and capture how the subtle structural changes between SARS-CoV and SARS-CoV-2 affect their dimerisation properties. By applying a rigorous scoring procedure to our results, we identify four statistically significant hotspots on the protein which are strongly connected to the active site and propose that they hold potential for allosteric regulation of the main protease. By providing guidance for allosteric drug design we hope to open a new chapter for drug targeting efforts to combat COVID-19.

2 Results

The first step in our graph analysis approach is the construction of an atomistic graph from a protein data bank (PDB) [47] structure. This process takes into account strong and weak interactions like hydrogen bonds, electrostatic and hydrophobic interactions (see Methods and Fig. 4). Additionally, we can incorporate water molecules, which in the case of the M^{pro} are catalytically important and known to expand the catalytic dyad to a triad [11] (see Fig. S1B). In this analysis, we use the structures of the apo form of the SARS-CoV-2 and SARS-CoV main proteases which are deposited with PDB identifier 6Y2E [11] and 2DUC [48], respectively. Once the atomistic graph is constructed, we use Bond-to-bond propensities and Markov transients to complementary explore the connectivity within the proteins when sourced from relevant residues. To achieve this, Bond-to-bond propensity explores the instantaneous strength of communication of a perturbation to every bond in the protein which allows to identify allosteric sites [43] and investigate concepts like cooperativity in multimeric proteins [44]. Markov transients exploit the time evolution of a diffusion process on the atomistic graph to identify groups of atoms which are reached the fastest (i.e. allosteric sites) or form a communication pathway [41]. By applying quantile regression we are able to quantitatively rank all bonds, atoms and subsequently residues. This allows to score the hotspots we identified and statistically prove their significance.

2.1 Bond-to-bond propensities validate molecular mechanism of M^{pro}.

Figure 1 provides detailed insights into the Bond-to-bond propensity analysis of the SARS-CoV-2 M^{pro} when sourced from the active site residues histidine 41 and cysteine 145 in both monomers. The top scoring residues (see Table S2) reveal two main areas of interest in the M^{pro}. The hotspot on the back of the monomer opposite to the active site (Fig. 1A) is described in more detail in the paragraphs below. Hotspot two is located in the dimer interface and contains four residues which 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 one, respectively. Interestingly, these bonds have been found to be essential for dimer formation which in turn is required for M^{pro} activity [49, 27].

2.2 Protease dimerisation is under influence of mutated residues.

To further clarify the interactions between the dimer halves (Fig. S1A) and how the dimer connectivity changed for the new SARS-CoV-2 protease, we ran Bond-to-bond propensity analysis sourced from two mutated residues. Alanine 285 and leucine 286 are involved in the dimer interface and have been shown to lead to a closer dimer packing when mutated from threonine 285 and isoleucine 286 in SARS-CoV [14, 11].

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

Hence, we chose these residues as source when looking into protease dimer connectivity in comparison between SARS-CoV-2 and SARS-CoV. Table 1 shows the top 20 residues in both structures when sorted by quantile score. We can report a strong connectivity towards dimer interface residues which is more apparent in the SARS-CoV-2 protease than in the SARS-CoV one. This can be attributed to a closer dimer packing due to the two smaller side chains of 285/286 in the new protease [11]. In a mutational study in SARS-CoV, this closer dimer packing led to an increased activity [14], however this could not be confirmed in the SARS-CoV-2 protease [11]. This was further validated when we calculated the average residue quantile score of the active site in these runs. For the active site in SARS-CoV-2 M^{pro} the score 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 in this analysis. In SARS-CoV M^{pro} we detect a higher connectivity with a score of 0.50 for the active site which is nevertheless slightly above a random site score of 0.48 (95% CI: 0.47- 0.48). Although we could not identify the direct link between the extra domain and the active site on an

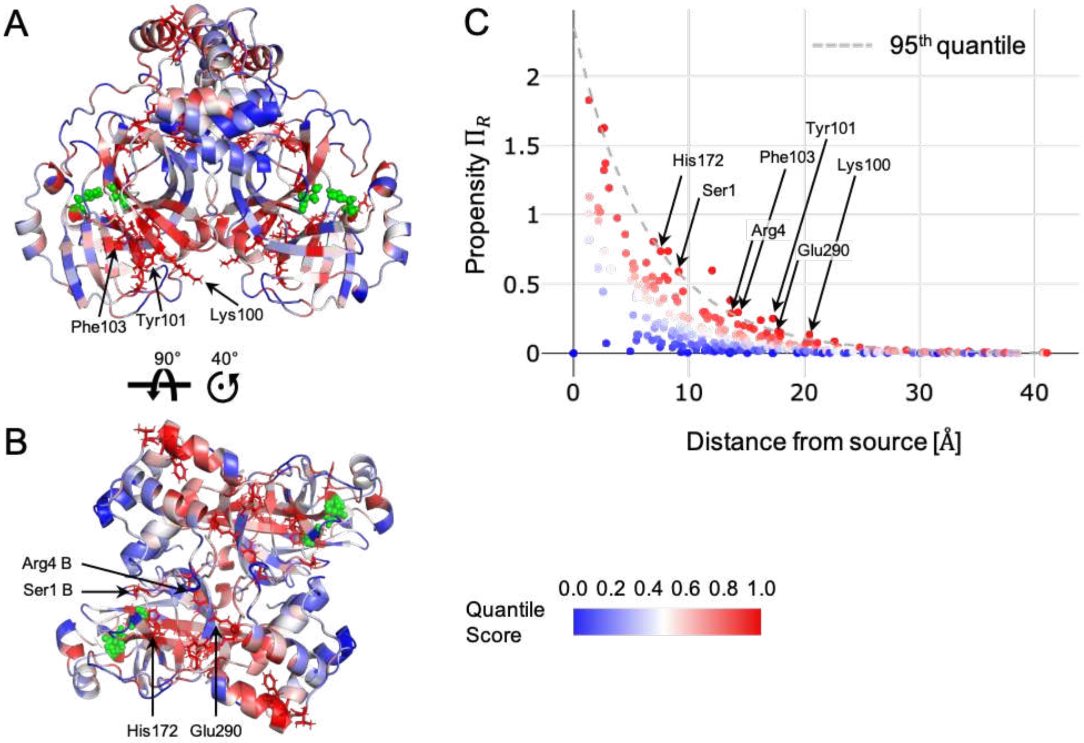

Figure 1: Bond-to-bond propensities of M^{pro} sourced from the orthosteric sites. The source sites have been chosen as the catalytically active residues His41 and Cys145 in both chains of the homodimer and are shown in green (front A) and top B) view). All other residues are coloured by quantile score as shown in the legend and reveal two main areas of interest with important residues labelled. C) The propensity of each residue, Π_R, is plotted against the residue distance from the orthosteric site. The dashed line indicates the quantile regression estimate of the 0.95 quantile cutoff used for identifying relevant residues.

atomistic level here, we assume that studying the dimer interface residues in a systematic manner would help elucidate the link between domain III and the catalytic activity of the M^{pro}.

2.3 Identification and scoring of putative allosteric sites.

Bond-to-bond propensities have been shown to successfully detect allosteric sites on proteins [43] and we here present the results in the SARS-CoV-2 M^{pro} to that effect. By choosing the active site residues histidine 41 and cysteine 145 as source, we can detect areas of strong connectivity towards the active centre which allows us to reveal putative allosteric sites. We could detect two hotspots on the protease which might be targetable for allosteric regulation of the protease (Fig. 2). Most of the residues present in the two putative sites are amongst the highest scoring residues which are listed in Table S2. Site 1 (Fig. 2A shown in yellow) which is located on the back of the monomer in respect to the active site and is formed by nine residues from domain I and II (full list in Table S4). The second hotspot identified with Bond-to-bond propensities is located in the dimer interface and contains 6 residues (Tab. S5) which are located on both monomers (Fig. 2B shown in pink). Two of these residues, Glu290 and Arg4 of the respective second monomer, are forming a salt bridge which is essential for dimerisation [27]. Quantile regression allows us to rank all residues in the protein and thus we can score both sites with an average residue quantile score as listed in Table 2. Site 1 and 2 have a high score of 0.97 and 0.96, respectively and score much higher than a randomly sampled site would score with 0.53 (95% CI: 0.53-0.54) for a a site of the size of site 1 or 0.52 (95% CI: 0.51-0.53) for a site of the size of site 2.

Our methodologies further allow to investigate the reverse analysis to assess the connectivity of the predicted allosteric sites. For this purpose, we defined the source as all residues within the respective identified sites (Tables S4 and S5). After a full Bond-to-bond propensity analysis and quantile regression to rank all residues, we are able to score the active site to obtain a measure for the connectivity towards the catalytic center (Tab. S8). For site 1 the active site score is 0.64 which is above a randomly sampled site score of 0.47 (95% CI:0.47-0.48). However, for site 2 the active site score is 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 in the dimer interface, this is in line with the above described suggestion that the allosteric effect is not directly conferred from the dimer interface towards the catalytic centre. Nonetheless, this site might provide scope for inhibiting the M^{pro} by disrupting the dimer formation at these sites.

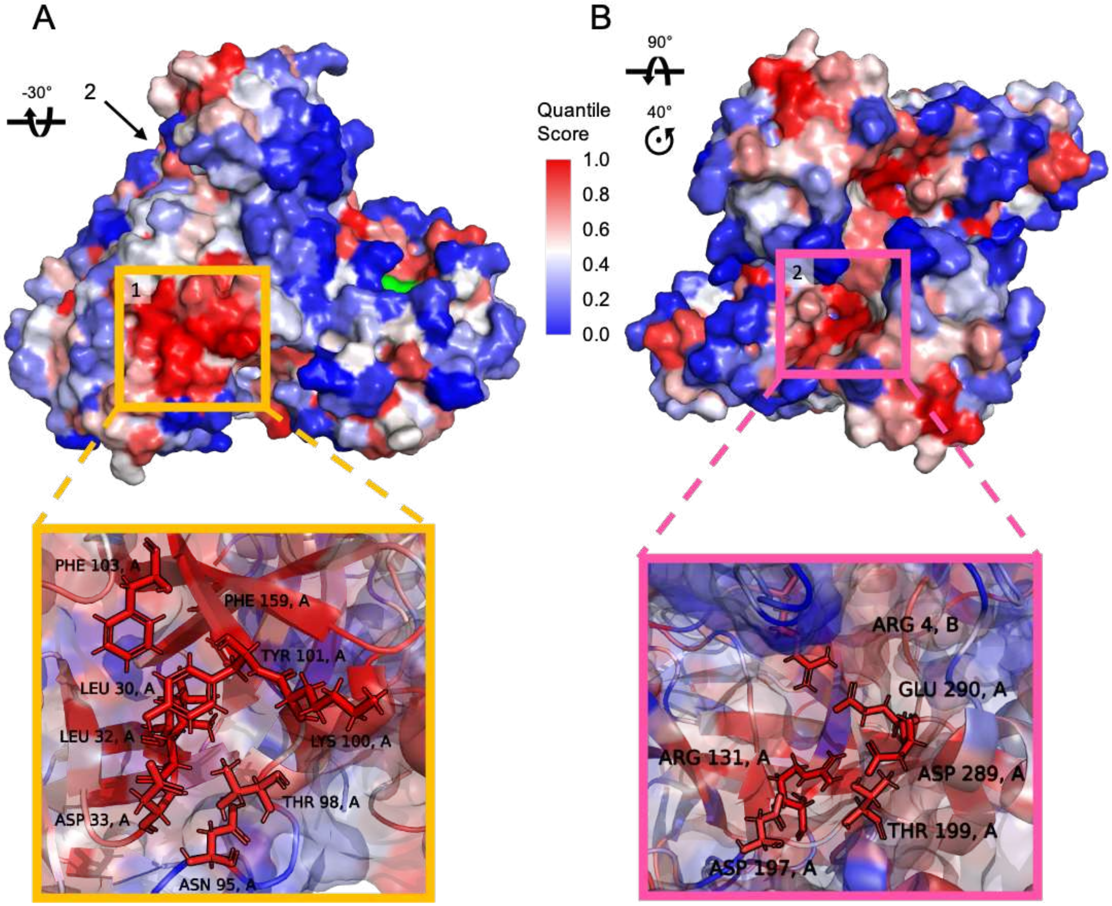

Figure 2: Putative allosteric sites identified by Bond-to-bond propensities. Surface representation of the M^{pro} dimer coloured by quantile score (as shown in the legend). A) Rotated front view with site 1 (yellow) which is located on the opposite of the orthosteric site (coloured in green). B) Top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

Overall, this missing bi directional connectivity hints to a more complex communication pattern in the protein and gave us reason to utilize another tool which has been shown to be effective in catalytic frameworks [41] like the protease. Markov transients reveal fast signal propagation which happens often along allosteric communication pathways within the protein structure. The top scoring residues with a QS > 0.95 in a Markov transient analysis sourced from the active site residues are shown in Figure 3A and a full list can be found in Table S3. In the SARS-CoV-2 M^{pro}, this analysis subsequently led to the discovery of two more putative sites as shown in Figure 3C. Both hotspots are located on the back of the monomer in relation to the active site. Site 3 (shown in turquoise in Figure 3C) is located solely in domain II and consists of ten residues as listed in Table S6. One of which is a cysteine at position 156 which might provide a suitable anchor point for covalent drug design. Site 4 (orange in Figure 3C) is located further down the protein in domain I with 11 residues as listed in Table S7. Both sites were scored as described above and in the Methods section. Both sites have high average residue quantile scores of 0.87 (Tab. 2) which are significantly higher than the 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 described for site 1 and 2, we can investigate the protein connectivity from the opposite site by sourcing our runs from the residues in site 3 and 4. We then score the active site to measure the impact of the putative sites on the catalytic centre (Tab. S8). For site 3, the active site has an average residue quantile score of 0.66 in comparison to a random site score of 0.53 (95% CI: 0.52-0.53) which indicates a significant catalytic link between site 3 and the active site. For site 4 (as for site 2) the scores are similar to a randomly sampled score, which means that we do not detect a significant connectivity from this site to the active site. Judging from previous experience

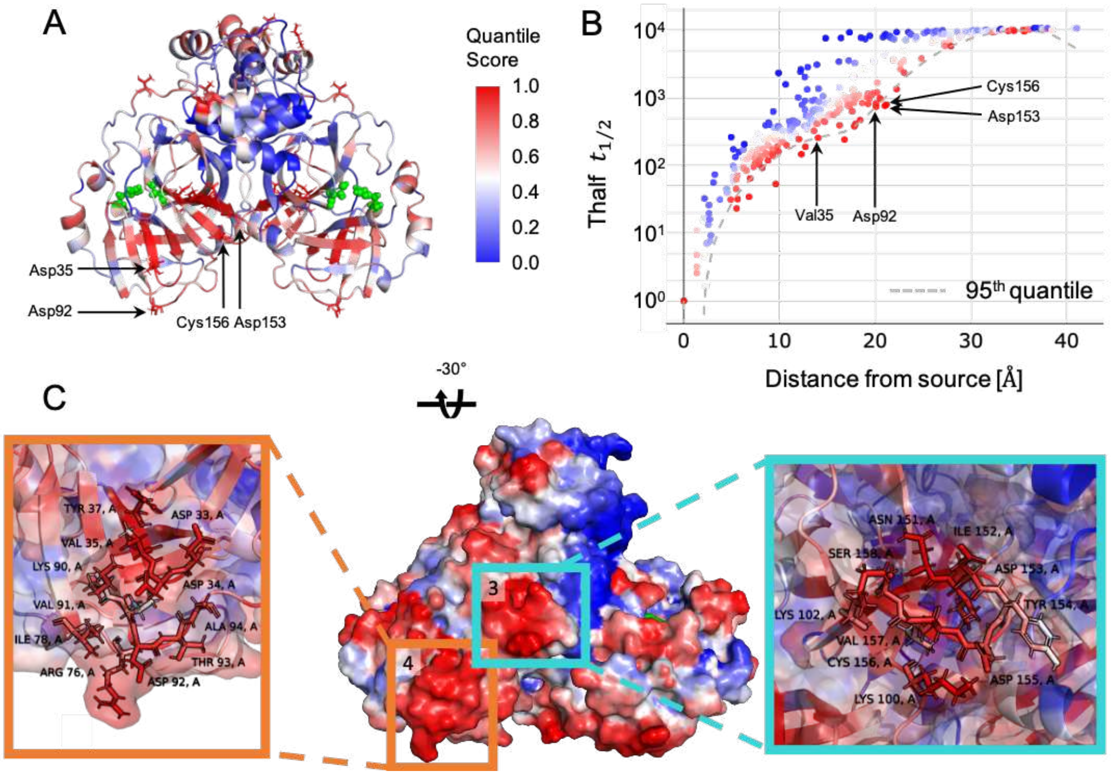

Figure 3: Markov transient analysis of M^{pro} sourced from the orthosteric sites. The orthosteric sites are shown in green and include His41 and Cys145 in both chains of the homodimer (front A) view). B) The $t_{1/2}$ values of each residue are plotted against their distance from the orthosteric site. 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 surface of the M^{pro} dimer (front A) view), coloured as shown in the legend. C) Surface representation of a rotated front view the M^{pro} dimer coloured by quantile score. Site 3 (turquoise) and 4 (orange) are located on the opposite site of the active site (coloured in green). A detailed view of both sites is provided with important residues labelled.

Table 2: Scoring of the 4 identified putative allosteric sites. Included is a structural bootstrap score of 1,000 randomly sampled sites with 95% confidence interval (CI).

Site	Average Residue Quantile Score	Random Site Score [95% CI]
Site 1	0.97	0.53 [0.53, 0.54]
Site 2	0.96	0.52 [0.51, 0.53]
Site 3	0.87	0.50 [0.49, 0.51]
Site 4	0.87	0.49 [0.49, 0.51]

in multimeric proteins this might be due to another structural or dynamic factor which we did not yet uncover between site 4 and the active site.

Overall we see a similar pattern of hot and cold spots in the SARS-CoV M^{pro} (results not shown). We find a high overlap for the identified four sites which gives us confidence, that a potential drug effort would find applications in COVID-19 as well as SARS. To provide a first indication of the druggability of the identified sites, we chose to align the fragments identified in the Diamond Light Source XChem fragment screen [50] with our sites. The screen identified 25 fragments which bind outside of the active site and 15 of these bind within 4 Å of any of the four putative allosteric sites. Due to the computational efficiency of our methodologies we were able to conduct a full analysis of all 15 structures and ran our methods from the fragments as source sites. We subsequently scored the active sites in each run (full data in Table S9) and found that the fragment deposited with the PDB identifier 5RE8 might be of particular interest as it has the highest connectivity to the active site. Moreover, one of the fragments within 4 Å of site 1 with the PDB identifier 5RGJ, has been shown to inhibit the proteolytic activity of the M^{pro} [24] and possesses a relatively high connectivity to the active site.

3 Discussion

During the global pandemic of COVID-19 that has started in January 2020, we have seen an increase of research activities 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 [11, 16], high-throughput docking approaches to the active site [15] or re-purposing approved drugs [51] and protease inhibitors [52] 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 provide a full computational analysis of the protease structure which gives insights into allosteric signalling and identifies potential putative sites. Our methodologies are based on concepts from graph theory and the propagation of perturbations and fluctuations on a protein graph. We have previously demonstrated the applications of Bond-to-bond propensities and Markov transients in identifying allosteric sites and communication pathways in a range of biological settings [41, 43, 44, 46]. Applying Bond-to-bond propensities on the SARS-CoV-2 M^{pro} gave us important insights into connectivity of the protein and highlighted residues at the dimer interface. We further explored the interface residues in comparison with the SARS-CoV protease as dimerisation is known to be essential for the proteolytic activity [14] and might provide scope for inhibitor development [53]. Important for the dimer packing and mutated in SARS-CoV-2 are residues 285 and 286 [11]. When sourced from these residues, we find a higher proportion of dimer interface residues within the top 20 scoring residues for SARS-CoV-2 which confirms a stronger dimer connectivity as described in literature [11]. Although we could not identify the direct link between the mutated residues and the active site on an atomistic level here, we assume that further systematic studies of the residues at the dimer interface would provide clarity.

This gave us confidence to further explore the SARS-CoV-2 protease with our methodologies. Using the above described approaches we have identified four allosteric binding sites on the protease. We describe the location of the sites and possible implications for the proteolytic activity of the protein. Site 1 and 2 have been identified using Bond-to-bond propensities and hence have a strong instantaneous connectivity to the active site. Sourced from both sites, we noticed that site 1 is directly connected to the active site, which is detected with a score above a randomly sampled site score ($0.64 > 0.47$) while site 2 is indirectly connected to the active site with a active site score only slightly above that of a random site ($0.49 > 0.48$). 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. Indeed, a fragment near site 1 has been shown to exhibit some inhibitory effect on the M^{pro} in a recent study [24]. Notably, site 2, although not directly coupled to the active site as a functional site, is located in the dimer interface (Fig. 2B) and provides a deep pocket for targeting the protease and maybe disrupting dimer formation. Targeting site 2 could result in a conformation change of the protease and inhibition of dimerisation.

The sites identified with Markov transients are reached the fastest by a signal sourced from the active site and are both located at the back of each monomer in relation to the active site. Site 3 is assumed to be directly coupled to the active site as seen from the score of the active site ($0.66 > 0.53$) and perturbation at site 3 would thus affect the catalytic activity of M^{pro}. Besides, Site 3 (Fig. 3C) contains a cysteine residue (Cys156) which provides an anchor point for covalently binding inhibitors [54]. Similar to site 2, site 4 is not directly connected to the active site. Effects exerted at site 4 could affect other parts of the protein which in turn lead to an altered activity of M^{pro}.

We also include the analysis of 15 structures containing small fragments from a recent Diamond Light Source XChem fragment screen [50] which bind in proximity to the putative sites. We scored the active site (His41 and Cys145) using these fragments as the source. The active site score is analysed rigorously with a structural bootstrap to compare the effect of each fragment on the protease. Some fragments have a direct link to the active site and have been recently investigated in experimental studies [24] and might provide a first starting point for rational drug design.

Together our methods provide in depth insights into the global connectivity of the main protease. By taking our results into consideration we hope to broaden the horizon for targeting the main protease of SARS-CoV-2. This will aid in the development of effective medications for COVID-19.

4 Methods

Protein Structures. We analysed the X-ray crystal structures of the apo conformations of the SARS-CoV-2 (PDB ID: 6Y2E [11]) and the SARS-CoV (PDB ID: 2DUC [48]) main proteases (M^{pro}). 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 [55] (for a full list of the resulting dimer interface residues see https://doi.org/10.6084/m9.figshare.12815903).

Atomistic Graph Construction. In contrast to most network methods for protein analysis, we derive atomistic protein graphs obtained from the three-dimensional protein structure and parameterise with physico-chemical energies, where the nodes of the graph are the atoms and the weighted edges represent interactions, both covalent bonds and weak interactions, including hydrophobic, hydrogen bonds and salt bridges (See Fig. 4). Details of this approach can be found in Refs [40, 41, 43]. We summarise the main features below and note three additional improvements, namely, in the stand-alone detection of edges without need of third-party software, the many-body detection of hydrophobic edges across scales, and, the computational efficiency of the code. For further details for the atomistic graph construction used in this work see [56, 42].

Figure 4 gives an overview of the workflow where we start from atomistic cartesian coordinates from PDB files. Since X-ray structures do not include hydrogen atoms and NMR structures may not report all of them , we use *Reduce* [57] to add any missing hydrogens. Hydrophobic interactions and hydrogen bonds are identified with a cutoff of 8 Å and 0.01 kcal/mol respectively. The edges are weighted by their energies: covalent bond energies from their bond-dissociation energies [58], hydrogen bonds and salt bridges by the modified Mayo potential [59, 60] and hydrophobic interactions are calculated using a hydrophobic potential of mean force [61].

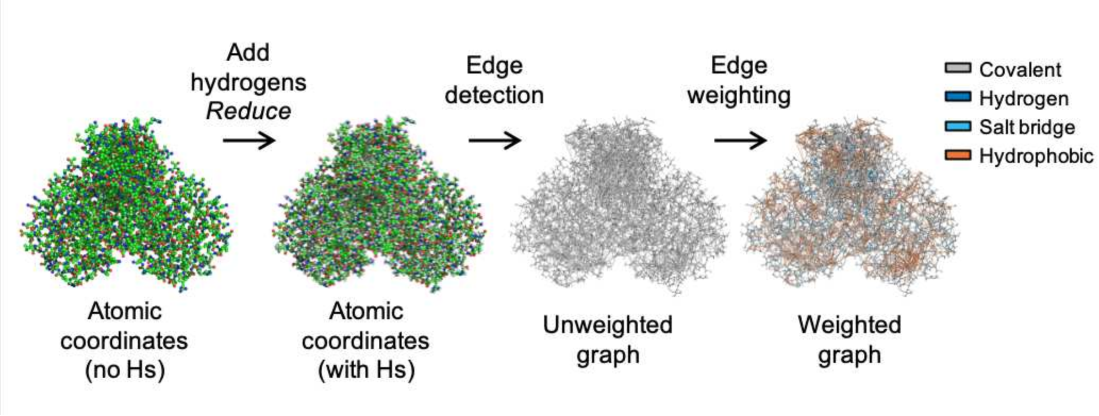

Figure 4: Atomistic Graph Construction. We showcase the general procedure here on the main protease of SARS-Cov-2: Atomic coordinates are obtained from the PDB (ID: 6Y2E [11] and hydrogens are added by Reduce [57]. Edges are identified and the weights are assigned, as described in the methods section, by taking into account covalent bonds as well as weak interactions: hydrogen bonds, electrostatic interactions and the hydrophobic effect which are coloured as indicated.

Bond-to-bond Propensities. Bond-to-bond propensity analysis was first introduced in Ref. [43] and further discussed in Ref. [44], hence we only briefly summarise it here. This edge-space measure examines and exhibits the instantaneous communication of a perturbation at a source towards every bond in the protein. The edge-to-edge transfer matrix M was introduced to study non-local edge-coupling and flow redistribution in graphs [62] and an alternative interpretation of M as a Green 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

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

where B is the $n \times m$ incidence matrix for the atomistic protein graph with n nodes and m edges; W = diag(w_{ij}) is an $m \times m$ diagonal matrix which possesses all the edge interaction energies with w_{ij} as the weight of the edge connecting nodes i and j, i.e. the bond energy between the atoms; and $L^{\dagger}$ is the pseudo-inverse of the weighted graph Laplacian matrix L [63] and defines the diffusion dynamics on the energy-weighted graph [64].

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:

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

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

$$\prod_{R} = \sum_{b \in R} \prod_{b} \tag{3}$$

Markov Transient Analysis (MTA). A complementary, node-based method, Markov Transient analysis (MTA) identifies areas of the protein that are significantly connected to a site of interest, the source, such as the active site, and obtains the signal propagation that connects the two sites at the atomistic level. The method has been introduced and discussed in detail in Ref. [41] and has successfully identified allosteric hotspots and pathways without any *a priori* knowledge [41, 46]. 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 $t_{1/2}$,

$$t_{1/2}^{(i)} = \arg\min_{t} \left[p_t^{(i)} \geq \frac{\pi^{(i)}}{2} \right] \tag{4}$$

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 the stationary distribution 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 (QR). 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) [65], a robust statistical measure widely used in different areas [66]. In contrast to standard least squares regressions, QR provides models for conditional quantile functions. 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 while in the case of transient times which do not follow a particular parametric dependence on distance, we use cubic splines to retain flexibility. From the estimated quantile regression functions, we can then compute the quantile score for each atom or bond. To obtain residue quantile scores, we use the minimum distance between each atom of a residue and those of the source. Further details of this approach for Bond-to-bond propensities can be found in Ref. [43] and for Markov Transient Analysis in Ref. [67].

Site scoring with structural bootstrap sampling. To allow an assessment of 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 quantile score of the site of interest is calculated. After sampling 1000 random sites on the protein, the average residue quantile scores are calculated. By performing a bootstrap with 10,000 resamples with replacement on the random sites average residue quantile scores, we are able to provide a 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. For scoring the active site as a measure of the connectivity towards the main binding site, we use all non-covalent hits bound in the active site from the XChem fragment screen against the SARS-CoV-2 M$^{\mathrm{pro}}$ [50] . The 22 found structures were further investigated using PyMol v.2.3 [68] for residues which have atoms within 4Å 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 and constitute 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$^{\mathrm{pro}}$ [50], 25 hits were found at regions other than the active site. The 15 fragments which contain atoms that are within 4Å 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.

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

Visualisation and Solvent Accessible Surface Area. We use PyMol (v.2.3) [68] for structure visualisation and presentation of Markov transient and Bond-to-bond propensity results directly on the structure. The tool was also used to calculate the residue solvent accessible surface area (SASA) reported here, 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.

Acknowledgements

We acknowledge helpful discussions with Florian Song, Francesca Vianello, Ching Ching Lam and Jerzy Pilipczuk. This work was funded by a Wellcome Trust studentship to L.S. [grant number 215360/Z/19/Z]. M.B. and S.N.Y. acknowledge funding from the EPSRC award EP/N014529/1 supporting the EPSRC Centre for Mathematics of Precision Healthcare.

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.

Competing interests

The authors declare no competing interests.

Materials & Correspondence

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

References

[1] Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020). URL https://doi.org/10.1038/s41586-020-2012-7.

[2] Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020). URL https://doi.org/10.1038/s41586-020-2008-3.

[3] Zhu, N. *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine* **382**, 727–733 (2020). URL https://doi.org/10.1056/NEJMoa2001017.

[4] Gorbalenya, A. E. *et al.* The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology* **5**, 536–544 (2020). URL https://doi.org/10.1038/s41564-020-0695-z.

[5] Peiris, J. S. M., Guan, Y. & Yuen, K. Y. The severe acute respiratory syndrome. *Nature Medicine* **10**, S88–S97 (2004). URL https://doi.org/10.1038/nm1143.

[6] Graham, R. L., Donaldson, E. F. & Baric, R. S. A decade after SARS: strategies for controlling emerging coronaviruses. *Nature Reviews Microbiology* **11**, 836–848 (2013). URL https://doi.org/10.1038/nrmicro3143.

[7] Steinhauer, D. A. & Holland, J. J. Direct method for quantitation of extreme polymerase error frequencies at selected single base sites in viral RNA. *Journal of Virology* **57**, 219–228 (1986). URL https://doi.org/10.1128/JVI.57.1.219-228.1986.

[8] Anand, K. *et al.* Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain. *The EMBO journal* **21**, 3213–3224 (2002). URL https://doi.org/10.1093/emboj/cdf327.

[9] Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R. & Hilgenfeld, R. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science* **300**, 1763–1767 (2003). URL https://doi.org/10.1126/science.1085658.

[10] Yang, H. *et al.* The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 13190–13195 (2003). URL https://doi.org/10.1073/pnas.1835675100.

[11] Zhang, L. *et al.* Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science* **368**, 409–412 (2020). URL https://doi.org/10.1126/science.abb3405.

[12] Hilgenfeld, R. From SARS to MERS: crystallographic studies on coronaviral proteases enable antiviral drug design. *FEBS Journal* **281**, 4085–4096 (2014). URL http://doi.org/10.1111/febs.12936.

[13] Chen, Y. W., Yiu, C. P. B. & Wong, K. Y. 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** (2020). URL https://doi.org/10.12688/f1000research.22457.2.

[14] Lim, L., Shi, J., Mu, Y. & Song, J. 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** (2014).

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

[16] Jin, Z. *et al.* Structural basis for the inhibition of SARS-CoV-2 main protease by antineoplastic drug carmofur. *Nature Structural & Molecular Biology* **27**, 529–532 (2020). URL https://doi.org/10.1038/s41594-020-0440-6.

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

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

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

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

[21] Ullrich, S. & Nitsche, C. The SARS-CoV-2 main protease as drug target. *Bioorganic and Medicinal Chemistry Letters* **30**, 127377 (2020). URL https://doi.org/10.1016/j.bmcl.2020.127377.

[22] Wenthur, C. J., Gentry, P. R., Mathews, T. P. & Lindsley, C. W. Drugs for Allosteric Sites on Receptors. *Annual Review of Pharmacology and Toxicology* **54**, 165–184 (2014). URL https://doi.org/10.1146/annurev-pharmtox-010611-134525.

[23] Cimermancic, P. *et al.* CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *Journal of Molecular Biology* **428**, 709–719 (2016). URL http://doi.org/10.1016/j.jmb.2016.01.029.

[24] El-baba, T. J. *et al.* Allosteric inhibition of the SARS-CoV-2 main protease - insights from mass spectrometry-based assays. *Angewandte Chemie International Edition* (2020). URL https://doi.org/10.1002/anie.202010316.

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

[26] Shi, J., Wei, Z. & Song, J. 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. *The Journal of biological chemistry* **279**, 24765–24773 (2004). URL https://doi.org/10.1074/jbc.M311744200.

[27] Shi, J. & Song, J. The catalysis of the SARS 3C-like protease is under extensive regulation by its extra domain. *FEBS Journal* **273**, 1035–1045 (2006). URL https://doi.org/10.1111/j.1742-4658.2006.05130.x.

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

[29] Greener, J. G. & Sternberg, M. J. Structure-based prediction of protein allostery. *Current Opinion in Structural Biology* **50**, 1–8 (2018). URL https://doi.org/10.1016/j.sbi.2017.10.002.

[30] Lu, S., He, X., Ni, D. & Zhang, J. Allosteric Modulator Discovery: From Serendipity to Structure-Based Design. *Journal of Medicinal Chemistry* **62**, acs.jmedchem.8b01749 (2019). URL http://doi.org/10.1021/acs.jmedchem.8b01749.

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

[32] Penkler, D., Sensoy, , Atilgan, C. & Tastan Bishop, Perturbation-Response Scanning Reveals Key Residues for Allosteric Control in Hsp70. *Journal of Chemical Information and Modeling* **57**, 1359–1374 (2017). URL https://doi.org/10.1021/acs.jcim.6b00775.

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

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

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

[36] Song, K. *et al.* Improved Method for the Identification and Validation of Allosteric Sites. *Journal of Chemical Information and Modeling* **57**, 2358–2363 (2017). URL https://doi.org/10.1021/acs.jcim.7b00014.

[37] Guarnera, E. & Berezovsky, I. N. Structure-Based Statistical Mechanical Model Accounts for the Causality and Energetics of Allosteric Communication. *PLoS computational biology* **12**, e1004678–e1004678 (2016). URL https://doi.org/10.1371/journal.pcbi.1004678.

[38] Tee, W.-V., Guarnera, E. & Berezovsky, I. N. Reversing allosteric communication: From detecting allosteric sites to inducing and tuning targeted allosteric response. *PLOS Computational Biology* **14**, e1006228 (2018). URL https://doi.org/10.1371/journal.pcbi.1006228.

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

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

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

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

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

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

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

[46] Chrysostomou, S. *et al.* Abstract 1775: Targeting RSK4 prevents both chemoresistance and metastasis in lung cancer. *Cancer Research* **79**, 1775 (2019). URL https://doi.org/10.1158/1538-7445.AM2019-1775.

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

[48] Muramatsu, T. *et al.* SARS-CoV 3CL protease cleaves its C-terminal autoprocessing site by novel subsite cooperativity. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 12997–13002 (2016). URL https://doi.org/10.1073/pnas.1601327113.

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

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

[51] Mahanta, S. *et al.* Potential anti-viral activity of approved repurposed drug against main protease of SARS-CoV-2: an in silico based approach. *Journal of Biomolecular Structure and Dynamics* (2020). URL https://doi.org/10.1080/07391102.2020.1768902.

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

[53] Goyal, B. & Goyal, D. Targeting the Dimerization of the Main Protease of Coronaviruses: A Potential Broad-Spectrum Therapeutic Strategy. *ACS Combinatorial Science* **22**, 297–305 (2020). URL https://doi.org/10.1021/acscombsci.0c00058.

[54] Hallenbeck, K., Turner, D., Renslo, A. & Arkin, M. Targeting Non-Catalytic Cysteine Residues Through Structure-Guided Drug Discovery. *Current Topics in Medicinal Chemistry* **17**, 4–15 (2017). URL https://doi.org/10.2174/1568026616666160719163839.

[55] Krissinel, E. & Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *Journal of Molecular Biology* **372**, 774–797 (2007). URL https://doi.org/10.1016/j.jmb.2007.05.022.

[56] Mersmann, S. *et al.* ProteinLens: a web-based application for the analysis of allosteric signalling on atomistic graphs of biomolecules (2020). URL https://doi.org/10.6084/m9.figshare.12369125.v1.

[57] Word, J., Lovell, S. C., Richardson, J. S. & Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of Molecular Biology* **285**, 1735–1747 (1999). URL https://doi.org/10.1006/jmbi.1998.2401.

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

[59] Mayo, S. L., Olafson, B. D. & Goddard, W. A. DREIDING: A generic force field for molecular simulations. *Journal of Physical Chemistry* **94**, 8897–8909 (1990). URL https://doi.org/10.1021/j100389a010.

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

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

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

[63] Biggs, N. *Algebraic graph theory*, vol. 67 (Cambridge university press, 1993).

[64] Lambiotte, R., Delvenne, J. & Barahona, M. Random Walks, Markov Processes and the Multiscale Modular Organization of Complex Networks. *IEEE Transactions on Network Science and Engineering* **1**, 76–90 (2014). URL https://doi.org/10.1109/TNSE.2015.2391998.

[65] Koenker, R. & Hallock, K. F. Quantile Regression. *Journal of Economic Perspectives* **15**, 143–156 (2001). URL https://doi.org/10.1257/jep.15.4.143.

[66] Koenker, R. quantreg: Quantile Regression. R package version 5.52 (2019). URL https://cran.r-project.org/package=quantreg.

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

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