Genome-wide Bioinformatics Analysis of Human Protease Specificity Identified Potential Cathepsin L Cleavage Site at K790 Position of the SARS-CoV-2 Spike Glycoprotein

Eugenii V. Matveev [1, 2, 3], Gennady V. Ponomarev [1, 2], Marat D. Kazanov [1, 2, 3, 4, *]

[1] Skolkovo Institute of Science and Technology, Moscow 121205, Russia.
[2] A.A.Kharkevich Institute for Information Transmission Problems, Moscow 127051, Russia.
[3] Dmitry Rogachev National Medical Research Center of Pediatric Hematology, Oncology and Immunology, Moscow 117998, Russia.
[4] Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul 34956, Turkiye
[*] Correspondence: mkazanov@gmail.com

Abstract: The severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) primarily enters the cell by binding the virus's spike (S) glycoprotein to the angiotensin-converting enzyme 2 (ACE2) receptor on the cell surface, followed by proteolytic cleavage by host proteases. Studies have identified furin and TMPRSS2 proteases in priming and triggering cleavages of the S glycoprotein, converting it into a fusion-competent form and initiating membrane fusion, respectively. Alternatively, SARS-CoV-2 can enter the cell through the endocytic pathway, where activation is triggered by lysosomal cathepsin L. However, other proteases are also suspected to be involved in both entry routes. In this study, we conducted a genome-wide bioinformatics analysis to explore the capacity of human proteases in hydrolyzing peptide bonds of the S glycoprotein. Predictive models of sequence specificity for 169 human proteases were constructed and applied to the S glycoprotein together with the method for predicting structural susceptibility to proteolysis of protein regions. After validating our approach on extensively studied S2' and S1/S2 cleavage sites, we applied our method to each peptide bond of the S glycoprotein across all 169 proteases. Our results indicate that various members of the PCSK, TTSP, and kallikrein families, as well as specific coagulation factors, are capable of cleaving S2' or S1/S2 sites. We have also identified a potential cleavage site of cathepsin L at the K790 position within the S2' loop. Structural analysis suggests that cleavage of this site induces conformational changes similar to the cleavage at the R815 (S2') position, leading to the exposure of the fusion peptide and subsequent fusion with the membrane. Other potential cleavage sites and the influence of mutations in common SARS-CoV-2 variants on proteolytic efficiency are discussed.

Introduction

Viruses typically hijack host cell proteins to perform essential tasks necessary for their lifecycle, such as cell entry, replication, and assembly [1]. Therefore, viruses generally lack genes in their genome that encode functions exploited from the host cellular machinery. Since specific functions are not encoded in viral genomes, it is challenging to identify which host factors are required for the lifecycle of virus. Cataloging host proteins involved in a viral life cycle is crucial for the development of antiviral drugs and therapeutic strategies [2] The experiments aimed at identifying these host factors are labor-intensive, making bioinformatics predictions beneficial for guiding experimental studies.

The COVID-19 pandemic underscored the importance of virology research and highlighted the need for a comprehensive understanding of the viral life cycle, including the identification of cellular host factors for SARS-CoV-2 infection [3]. SARS-CoV-2 is a positive-stranded RNA virus from the Coronaviridae family [4]. SARS-CoV-2 is thought to involve several host factors [5], among which proteases play an important role [6]. They play a key role in activating the fusion of the SARS-CoV-2 particle with the cell membrane and altering the conformation of the viral receptor, referred to as the spike (S) glycoprotein, which consists of the receptor-binding domain S1 and the fusion domain S2 [7–9].

Previous studies have identified two crucial cleavage sites on the SARS-CoV-2 S glycoprotein, processed by host proteases [10]. The first cleavage site, known as S2', is located in the S2 domain and becomes accessible after the S glycoprotein binds to the angiotensin-converting enzyme 2 (ACE2) cell receptor [11,12]. According to current understanding, this site undergoes cleavage by the membrane-anchored transmembrane protease serine 2 (TMPRSS2), initiating the process of coronavirus particle fusion with the cell membrane [13]. The second one, designated as S1/S2, is located at the linking of the S1 and S2 domains and is thought to be cleaved by furin in Golgi organelles before virus packaging [14]. This cleavage induces conformational changes in the S glycoprotein, transitioning it from a so-called 'close' to 'open' conformation that facilitates binding to the ACE2 receptor [15]. This site is absent in closely related coronaviruses, and the sequence insertion at S1/S2 in SARS-CoV-2 has sparked discussions regarding its artificial or natural origin [16–19].

It has been demonstrated that, even in the absence of TMPRSS2, SARS-CoV-2 can still enter the cell through an alternative endocytic pathway called 'the late pathway' [20,21]. Studies suggest that cathepsin L activates fusion of the coronavirus particle with the endolysosomal membrane via cleavage around the S1/S2 site [22], while other studies propose cleavage near the S2′ site [23]. This entrance pathway is still relatively less studied compared to the ACE2-mediated 'the early' pathway. Currently, it is the subject of ongoing research [24], particularly in light of recent reports indicating that the SARS-CoV-2 Omicron BA.1 variant strain shows increased dependence on the endocytic cell entry route [25].

Multiple studies showed that other proteases beyond furin, TMPRSS2, and cathepsin L may be involved in proteolytic processing of the S glycoprotein [26]. Thus, it has been demonstrated that coagulation factors, such as thrombin and factor Xa, can directly cleave the S glycoprotein [27]. Another study reported the potential involvement of metalloproteinases in this proteolytic event [28]. It was established earlier for other coronaviruses that members of type II transmembrane family serine proteases (TTSPs), including HAT and matriptase, can cleave the spike glycoprotein [10]. The human genome contains more than 600 proteases, classified into five catalytic groups - aspartic, metallo, cysteine, serine, and threonine peptidases [29,30], so this wide repertoire of proteases boosts the chance that some of them can be involved in activating S glycoprotein. Despite these individual cases, to our knowledge, there is no systematic studies of the role of other human proteases in SARS-CoV-2 infection. To fill this gap in the field, we conducted a genome-wide bioinformatics analysis to explore the potential involvement of human proteases in the cleavage of the S glycoprotein. This analysis has generated high-confidence predictions that can guide future experimental studies.

Methods

Sequence and structural data, data on protease localization and expression

The amino acid sequence of the S glycoprotein from the Wuhan strain was obtained from UniProt (ID: P0DTC2) [29]. The three-dimensional structures of the S glycoprotein in both the 'open' (PDB ID: 6VYB) and 'closed' (PDB ID: 6VXX) conformations, as well as its complex with the ACE2 receptor (PDB ID: 7DF4) were downloaded from the Protein Data Bank (PDB) [30]. Loops were modeled using SWISS-MODEL with the default parameters [31]. Data on SARS-CoV-2 mutations were obtained from the CoVariants database [32]. Data on protease localization was taken from the COMPARTMENTS database [33]. Protease expression in human tissues was obtained from the TISSUES database [34].

Models of protease sequence specificity

Data on proteolytic events was retrieved from the MEROPS database [35]. Models of protease sequence specificity were constructed in the form of positional-weighted matrices (PWM) as described in [36]. Briefly, for each human protease from the MEROPS database with at least eight proteolytic events, the frequencies of amino acid occurrences at P3-P3' positions f_{ij} were calculated (Schechter–Berger notation [37]), where i represents the amino

acid and j represents the position. Background frequencies of amino acids b_j were calculated for the whole human proteome. Each cell of the PWM, with columns representing P3-P3' positions rows representing 20 amino acids, was assigned a value r_{ij} calculated as the logarithm of the ratio of the amino acid frequency at cleavage sites to the amino acid frequency at the background $r_{ij} = log f_{ij}/b_j$. By considering the PWM positions as independent, the cleavage score (CS) for particular protein peptide bond was calculated as the sum of the relevant values from each P3-P3' column, corresponding to the amino acids observed at those positions. The P1 position is used as a reference point for specifying the positions of the cleavage sites throughout the paper.

Model of the structural susceptibility to proteolysis

The method for estimation of the structural susceptibility to proteolysis was previously described in [38]. Briefly, proteolytic events extracted from CutDB [39] were mapped to the corresponding three-dimensional structures of protease substrates. A set of predictors was composed from conventional structural features, such as solvent accessibility, temperature factor, secondary structure, as well as additional features like loop length and terminal regions. A training set was curated to exclude data from experiments with potentially denatured substrates, which were identified based on the presence of high numbers of cleavages in the hydrophobic protein core. The final model was trained using linear discriminant analysis, which demonstrated better performance compared to other conventional machine learning methods from the scikit-learn library [40].

Results

Models of protease sequence specificity

To elucidate a capacity of human proteases in proteolytic processing of the SARS-CoV-2 spike glycoprotein, we initially constructed predictive models of sequence specificity for 169 human proteases (Supplemental File 1). We defined this subset of human proteases based on the availability of information on identified protease substrates taken from the MEROPS database [35]. To construct predictive models of protease sequence specificity in the form of positional-weight matrices (PWMs) [36], we filtered the MEROPS data to identify human proteases with a sufficient number of substrates. We applied the obtained models to the amino acid sequence of the SARS-CoV-2 S glycoprotein from the original Wuhan strain to estimate the probability of cleavage for each protein peptide bond by individual proteases. To facilitate the interpretation of the obtained cleavage scores (CSs), we first compared CS distributions for known cleaved and uncleaved positions using MEROPS data on known protease substrates (Fig.1a). The results showed that proteolytic sites have Q1-Q3 CS values in the interval [1.64, 5.02], with mean and median values of 3.36 and 3.16, respectively, while uncleaved positions have Q1-Q3 CS values in the interval [-5.89, -1.03], with mean and median values of -3.57 and -3.42, respectively.

To validate our approach, we analyzed the calculated CSs for S2' (position 815) and S1/S2 (position 685) cleavage sites across a set of 169 considered human proteases (Fig.1b,1c). For the S2' position, the highest scores were observed for TMPRSS2, which is consistent with the current understanding of the activation of the virus-cell membrane fusion process. Likewise, at the S1/S2 cleavage position, the highest CS were obtained for furin, which aligns with the current knowledge of SARS-CoV-2 biology [41].

S2' cleavage site

Besides TMPRSS2, other members of the type II transmembrane family serine proteases (TTSPs) [42] – hepsin, DESC1, matriptase and matriptase-2 – showed sufficiently high CS values at the S2' position (Fig. 1b, Fig. S1). Additionally, high scores for the S2' site were obtained for furin and other members of the proprotein convertases subtilisin/kexin type (PCSK) family, including PCSK1, PCSK2, PCSK4, PCSK5, PCSK6, and PCSK7. The third protease family whose members showed sufficiently high CS values are the kallikreins. Thus, CS values were elevated for several members of this family – KLKB1, KLK1, KLK2, KLK4, KLK6, KLK14, and KLK15. A little lower but still significant CSs were observed for several coagulation factors.

To investigate the potential interactions between the mentioned proteases and the S glycoprotein during the coronavirus's life cycle inside a host cell, we analyzed data on protease localization (Fig. 1b, Fig. S1). According to the data on the localization of proteins in human cells (see Methods), the discussed above members of the TTSP family are localized in the plasma membrane, similar to the TMPRSS2 protease. Some members of the PCSK family, including furin, PCSK1, PCSK5, PCSK6, and PCSK7, are found in the Golgi apparatus, and all of them, excluding PCSK7 and including PCSK2, are localized in the extracellular space. The localization of the mentioned members of the kallikreins family is primarily in the extracellular space, while coagulation factors were found in the plasma

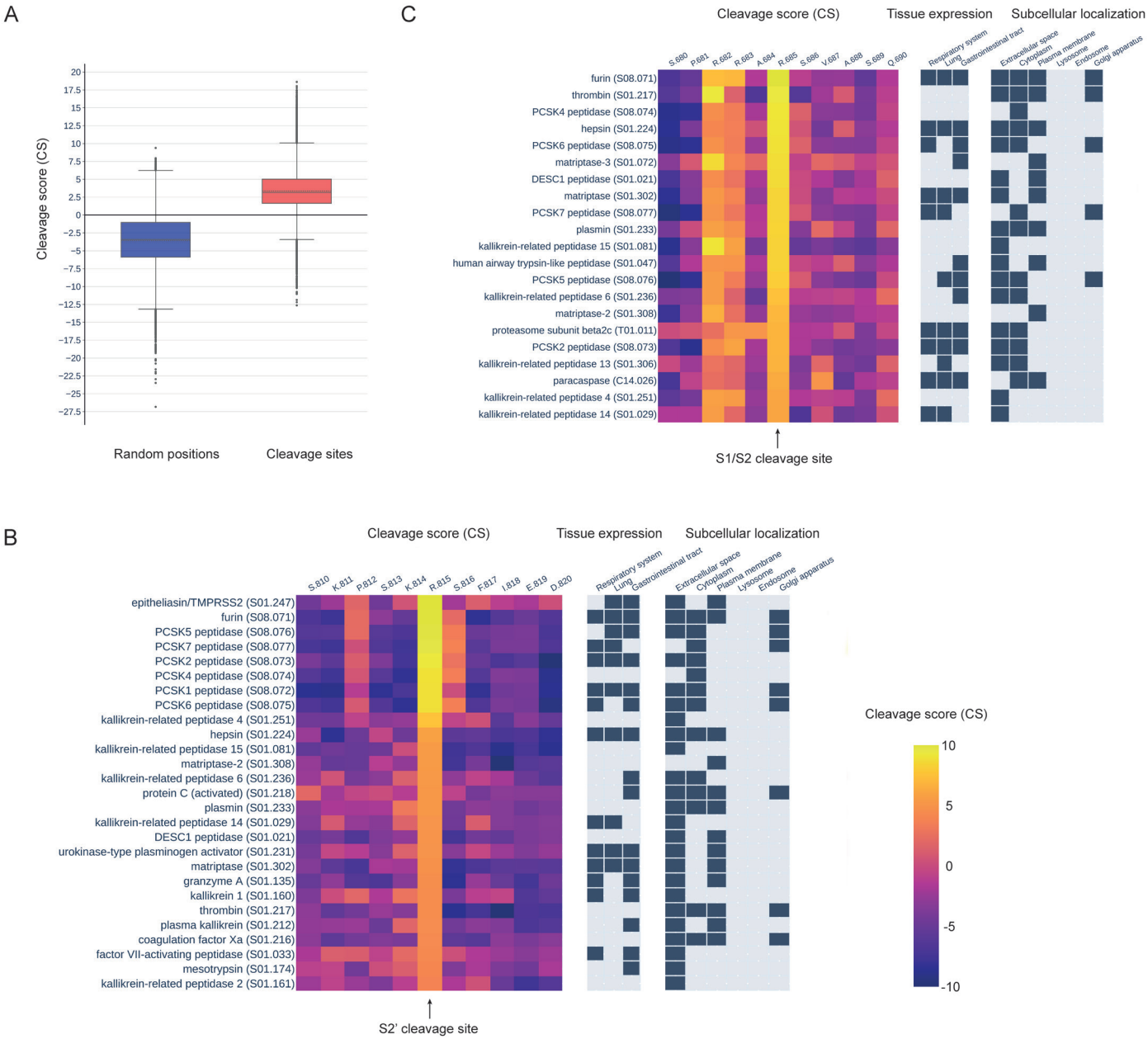

Figure 1. (**A**) *A comparison of cleavage scores generated by protease sequence specificity models for known cleavage sites and random protein positions.* (**B**) *Cleavage scores, protease cellular localization, and protease tissue expression for the top 20 proteases with the highest cleavage scores at the S2' (R815) position.* (**C**) *Cleavage scores, protease cellular localization, and protease tissue expression for the top 20 proteases with the highest cleavage scores at the S1/S2 (R685) position.*

membrane, extracellular space and cytoplasm. Thus, many members of these protease families theoretically have the potential to proteolytically activate the process of SARS-CoV-2 fusion with the membrane of an infected cell.

We have also analyzed data on protease expression in human tissues (Fig. 1b, Fig. S1). Available data points to the expression of three out of the five mentioned members of the TTSP family - TMPRSS2, hepsin, and matriptase - in the respiratory system, lungs, and gastrointestinal tract, i.e. three tissue types associated with SARS-CoV-2 spreading in the human body. Most of the aforementioned members of the PCSK family, including furin, PCSK1, PCSK2, PCSK5, PCSK6, and PCSK7, also have evidence of expression in at least one of these tissue types. Among specified kallikreins only KLK14 has data indicating expression in the respiratory system and lungs, while KLKB1 and KLK6 have data showing expression in the gastrointestinal tract. Urokinase is the only coagulation factor with expression data present in all three tissue types.

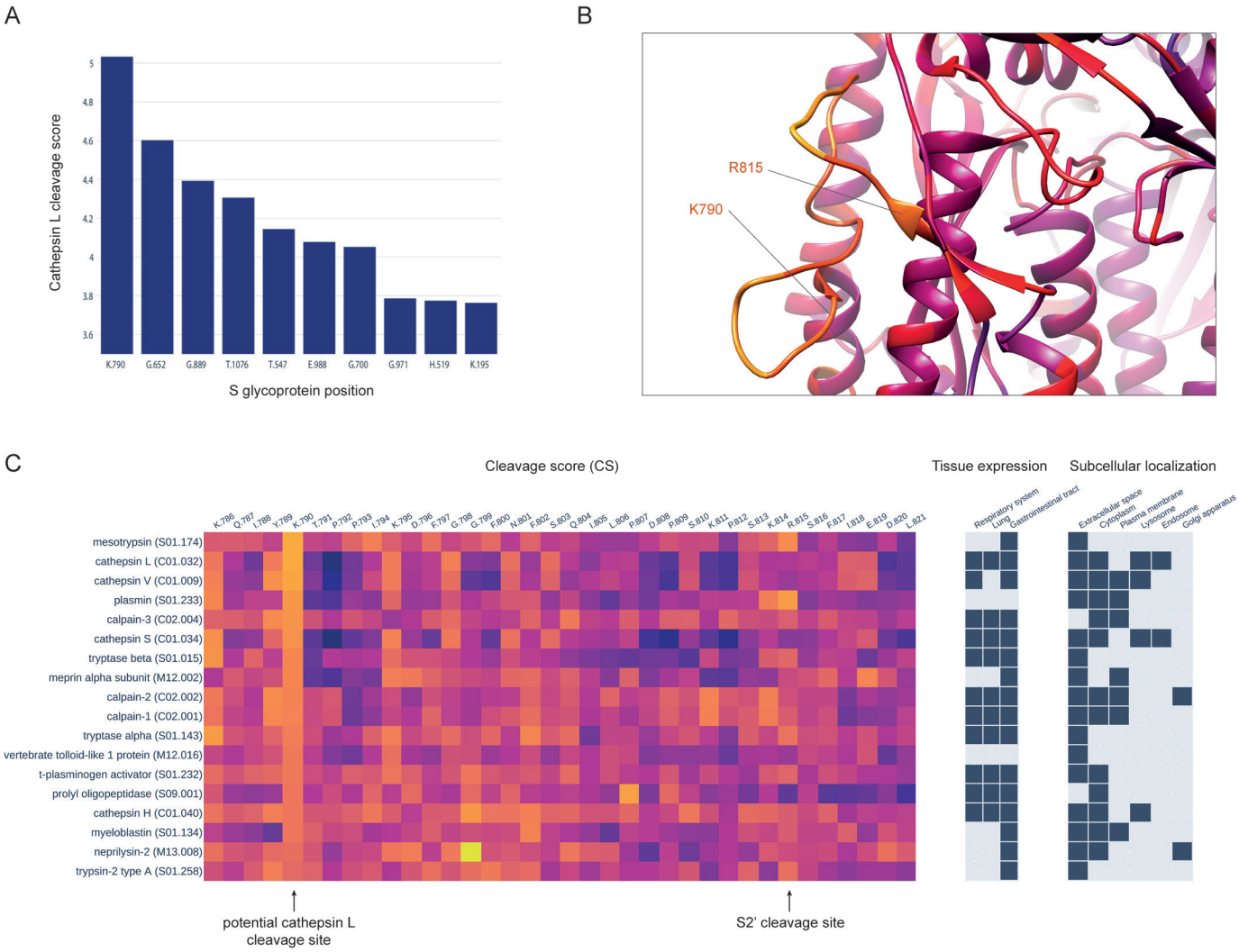

Figure 2. (**A**) *Top ten positions of the S glycoprotein with the highest cleavages scores, generated by the sequence specificity model of cathepsin L.* (**B**) *The location of the presumable cathepsin L cleavage site K790 on the 3D structure of the SARS-CoV-2 spike glycoprotein. The K790 cleavage site is located within the S2' loop, in close proximity to the known TMPRSS2 cleavage site R815. It is known that cleavage at the S2' site leads to the exposure of the fusion peptide and initiates the fusion process between the coronavirus particle and the cell membrane.* (**C**) *Cleavage scores, protease cellular localization, and protease tissue expression for the top 20 proteases with the highest cleavage scores at the K790 position.*

S1/S2 cleavage site

We conducted the same analysis for the S1/S2 cleavage site as we did for the S2' site (Fig. 1c, Fig. S2). Alongside furin, which obtained the highest CS value at the S1/S2 position, we observed high cleavage scores for other members of PCSK family, namely PCSK2, PCSK4, PCSK5, PCSK6, PCSK7. Additionally, we found sufficiently high CS values for coagulation factors such as thrombin and plasmin, as well as for hepsin, HAT, DESC1, and other members of the TTSP family. Slightly lower scores were observed for kallikreins. The known localization of HAT, a member of the

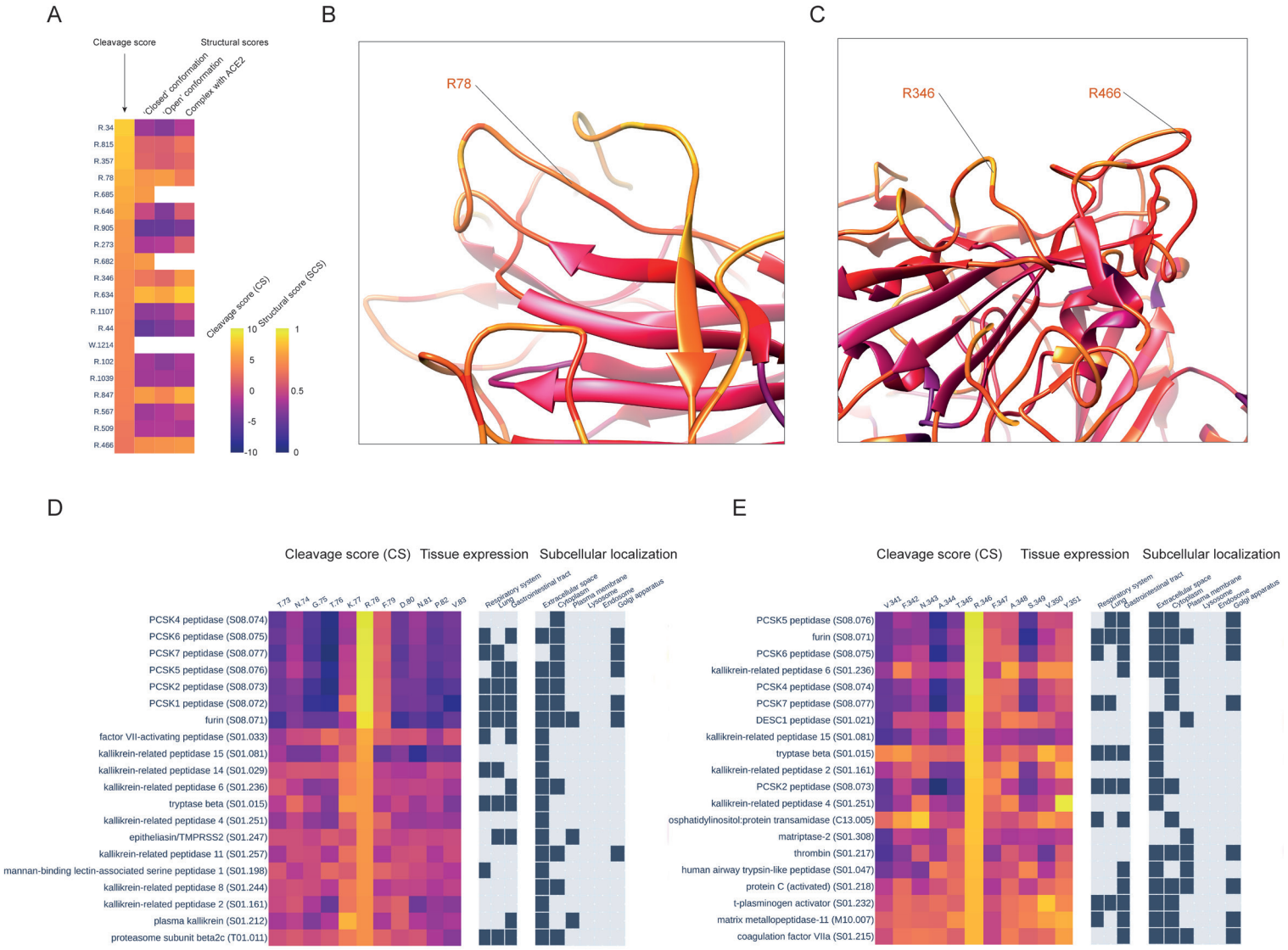

Figure 3. (**A**) *Top 20 positions of S glycoprotein with the highest mean cleavage scores values across PCSK, TTSP, and kallikreins protease families. For each position, the scores of the structural susceptibilities to proteolysis were also calculated using 3D structures of the S glycoprotein in both "closed" and "open" conformations, as well as in complex with the ACE2 receptor.* (**B**) *The location of the presumable cleavage site R78 within the N-terminal domain (NTD) of the SARS-CoV-2 spike glycoprotein.* (**C**) *The location of the presumable cleavage sites R346 and R466 within the receptor-binding domain (RBD) of the SARS-CoV-2 spike glycoprotein.* (**D**) *Cleavage scores, protease cellular localization, and protease tissue expression for the top 20 proteases with the highest cleavage scores at the R78 position.* (**E**) *Cleavage scores, protease cellular localization, and protease tissue expression for the top 20 proteases with the highest cleavage scores at the R346 position.*

TTSP family not described above, is in the extracellular space and plasma membrane. Data on tissue expression shows that the HAT protease is present in the gastrointestinal tract, a tissue also linked to the spread of SARS-CoV-2.

Cathepsin L cleavage site

Although the role of cathepsin L in activating the virus's entrance from the endolysosome to the cytoplasm is established for many coronaviruses [43], including SARS-CoV-2, the specific cathepsin L cleavage site is yet to be identified. There are contradictions in cleavage site positions according to the bioinformatics and experimental studies [44,45]. Our results for cathepsin L indicated that the K790 position has the highest CS value (Fig. 2a).

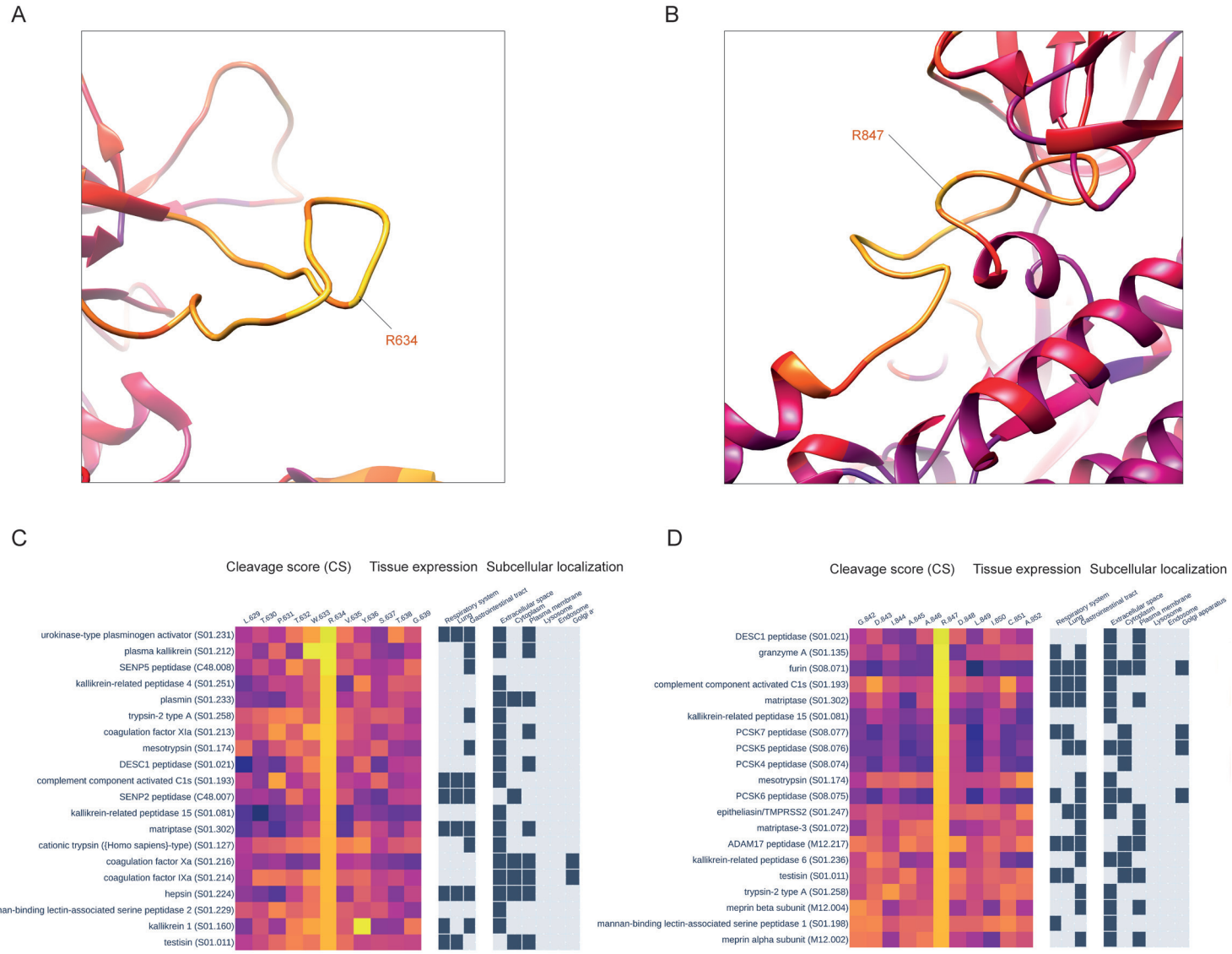

Figure 4. (**A**) *The location of the presumable cleavage site R634 within the so-called 630 loop of the SARS-CoV-2 spike glycoprotein.* (**B**) *The location of the presumable cleavage site R847 on the 3D structure of the SARS-CoV-2 spike glycoprotein.* (**C**) *Cleavage scores, protease cellular localization, and protease tissue expression for the top 20 proteases with the highest cleavage scores at the R634 position.* (**D**) *Cleavage scores, protease cellular localization, and protease tissue expression for the top 20 proteases with the highest cleavage scores at the R847 position.*

This site is located close to the S2', and we suggest that proteolytic cleavage of the K790 could result in similar conformational changes of the S glycoprotein as proteolysis at S2' leading to the exposure of the fusion peptide, activating the coronavirus fusion with the membrane.

To assess the accessibility of K790 to proteolysis from the structural perspective, we applied to the 3D structure of the S glycoprotein a bioinformatics method, previously developed by our group, for predicting structural susceptibility to proteolysis of the protein regions [38]. We mapped the calculated structural cleavage scores (SCSs) onto the 3D structure of the S glycoprotein in complex with the ACE2 receptor (Fig. 2b) and analyzed them specifically for the R815 and K790 sites. Both sites are located within the loop, which is adjacent to the fusion peptide. We observed that the SCSs were sufficiently high and quite comparable at both the K790 and R815 sites, with values of 0.61 and 0.58, respectively. Thus, we speculate that the K790 position could be a cleavage site of cathepsin L, whose proteolytic processing leads to the activation of fusion with the lysosomal membrane during the endocytic route of SARS-CoV-2 entry. We have also analyzed CSs calculated for other proteases at this position (Fig. 2c, Fig. S3). Interestingly, cathepsin L had the second-highest CS value, while the highest CS value was observed for mesotrypsin. The third-highest CS was obtained for another member of the papain family – cathepsin V. However, mesotrypsin is expressed solely in the gastrointestinal tract, whereas cathepsin L's expression data indicates its presence across the respiratory system, lungs, and gastrointestinal tract.

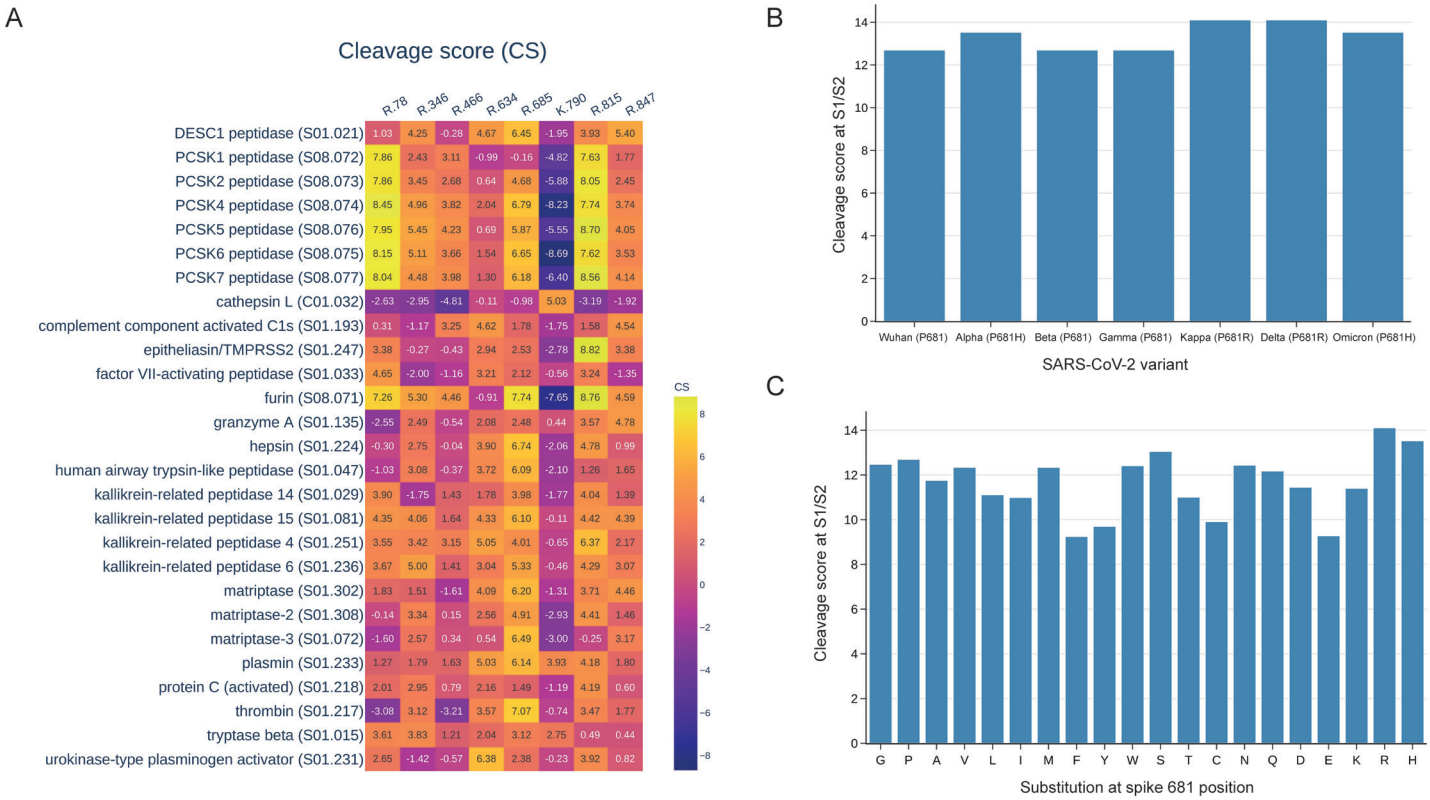

Figure 5. (**A**) *Cleavage scores computed for both known (R815, R685) and potential proteolytic sites using protease specificity models specifically developed for members of the PCSK, TTSP, kallikreins, and coagulation factors protease families.* (**B**) *Cleavage scores for the spike S1/S2 site, calculated using the P5-P5' PWM matrix, among diverse SARS-CoV-2 variants, including those carrying H/R substitutions at the P681 position.* (**C**) *Complete landscape of spike S1/S2 site cleavage scores across substitutions at the 681 position.*

Other potential cleavage sites

To investigate other potential cleavage sites, we further analyzed the entire set of CSs obtained for the S glycoprotein sequence, limiting our analysis to the four mentioned protease families – PCSK, TTSP, kallikreins and coagulation factors. We calculated the mean CS values across these four protease families for each S glycoprotein position and analyzed the top 20 positions with the highest CSs, together with the SCS values (Fig. 3a). Besides the known cleavage sites R685 and R815, only five positions among these 20 showed high values for both sequence and structural scores: R78, R346, R466, R634 and R847. The R78 site is located in N-terminal domain (Fig. 3b), whereas R346 and R466 sites (Fig. 3c) are positioned within the receptor-binding domain (RBD). The R634 site is located within the loop connecting S1 and S2 domains (Fig. 4a), known as the 630 loop [7]. Contrary to R815 site, which is located immediately upstream of the fusion peptide (FP), the R847 site is situated downstream of the FP (Fig. 4b).

We then analyzed cleavage scores at these positions for all 169 proteases. The cleavage scores at the R78 site were most favorable for proteases from the PCSK family, with all of them being approximately equal and significantly higher than scores calculated for other proteases (Fig. 3d, Fig. S4). For the R346 site, the best CSs were obtained for the PCSK family, with PCSK5 peptidase and furin being the first and second top proteases, respectively (Fig. 3e, Fig. S5). Cathepsin C, also known as dipeptidyl peptidase 1, showed highest CS at the R466 site, followed by members of the PCSK family (Fig. S6). Interestingly, for the R634 site, the best four CS values were obtained for proteases from three different classes (Fig. 4c, Fig. S7) – serine (urokinase and plasma kallikrein), cysteine (SENP5), and glutamate (RCE1 peptidase). The highest CS at the R847 site was obtained for a member of TTSP family – DESC peptidase (Fig. 4d, Fig. S8).

To gain insights into the difference between the cleavage scores of presumed cleavage sites and known cleavage sites S2' and S1/S2, we created a heatmap visualizing the cleavage scores of all proteolytic sites mentioned in this study, calculated for proteases from PCSK, TTSP, kallikreins, and coagulation factors families (Fig. 5a). Analysis of the heatmap reveals the distinct preferences of each protease for specific cleavage sites. As anticipated, TMPRSS2 exhibited the highest cleavage score at the R815 site, while for furin, the optimal sites are R685 and also R815. Interestingly, other members of the PCSK family have the top site at R78, while R685 is only the second best. This observation could possibly explain why other members of the PCSK family beyond furin were not reported to activate the S-glycoprotein.

Influence of SARS-CoV-2 mutations on cleavage sites

We further analyzed whether any known SARS-CoV-2 mutations influenced the considered cleavage sites. Our analysis of mutations in positions around the cleavage sites was limited to three amino acids from both sides of the cleavage sites, in accordance with the size of the PWM matrix. We found that out of the eight aforementioned positions – 78, 346, 466, 634, 685, 790, 815, 847 – only R346 was substituted by lysine and tyrosine in different SARS-CoV-2 variants. These independent substitutions significantly reduce CSs at this position (Fig. S9, Fig. S10) and may indicate a selection toward the degradation of this presumable cleavage site.

We also paid special attention to the well-known mutations P681H/R, which are five amino acids away from the S1/S2 cleavage site. To obtain computational estimates of the influence of these mutations on S1/S2 cleavage efficiency, we constructed a new, larger matrix of furin sequence specificity (PWM) by considering positions within 5 residues around the cleavage site. We applied a new furin PWM to the S-glycoprotein's S1/S2 site of SARS-CoV-2 variants and found that Alpha and Omicron variants with the P681H mutation showed higher cleavage scores than the original Wuhan variant. Additionally, Kappa and Delta variants, carrying the P681R mutations, demonstrate even higher cleavage score values (Fig. 5b). We have also calculated a complete landscape of furin cleavage scores at the S1/S2 site relative to all possible substitutions at the 681 position (Fig. 5c). Interestingly, R and H substitutions were found to be associated with the highest and second-highest cleavage scores, respectively. Proline, the wild-type

amino acid at position 681 in the S glycoprotein of the Wuhan strain, obtained the fourth-highest cleavage score, following serine. These results align with recent experiments [48–50].

To prevent biases in our structural estimation of susceptibility to proteolysis at studied cleavage sites, associated with mutations in different SARS-CoV-2 variants and those introduced experimentally to enhance protein crystallizability, we included five additional structures related to various SARS-CoV-2 variants (Alpha, Beta, Delta, Kappa, Omicron) in the analysis for each of the three S-glycoprotein conformations. We found that our structural estimates are sufficiently stable for all mentioned cleavage sites (Figure S11).

Discussion

In this study, we presented a genome-wide bioinformatics analysis of capabilities of human proteases to proteolytically process the spike glycoprotein of SARS-CoV-2. This analysis was conducted using protease sequence specificity models, which were built based on known protease substrates, along with the method for estimating the structural susceptibility to proteolysis developed previously by our group. We validated our approach on well-studied S1/S2 and S2' cleavage sites of the S glycoprotein: our method identified furin and TMPRSS2 as the proteases with the highest cleavage scores for S1/S2 and S2' sites, respectively, and estimated that these sites are accessible for proteolysis in a structural sense. These results strongly align with the current knowledge of SARS-CoV-2 biology. It is worth noting that our structural method can consider the 3D structures of proteins with attached glycans. Although structures with glycans are underrepresented in the PDB due to challenges in producing and crystallizing glycosylated proteins [51], they can serve as input for our method. This can be important, as it is known that 40% of the surface of the spike protein is shielded by glycans [52,53].

By applying this approach to every peptide bond of the S glycoprotein for all 169 human proteases, we thoroughly explored the proteolytic capacity of the subset of human proteolytic enzymes. For the known S1/S2 and S2' cleavage sites, we found that, in addition to furin and TMPRSS2, there are other members of four families of serine proteases – PCSK, TTSP, kallikreins and coagulation factors – which demonstrate theoretical potential for cleaving S1/S2 and S2' sites based on their sequence specificity, colocalization, and expression in tissues associated with SARS-CoV-2 spread in the human body.

Our analysis focused on cathepsin L revealed a potential cleavage site at position K790, which is located close to S2' and on the same loop. Subsequent structural examination showed the accessibility of this position for proteolysis and allowed us to suggest that cleavage of this site can induce structural changes similar to the cleavage of the S2' R815 position, which leads to the exposure of the fusion peptide and activation of fusion with the membrane. Interestingly, our proposed cleavage site differs from the cathepsin L cleavage sites previously suggested by both bioinformatics and experimental studies. However, the robust validation results of our approach on known cleavage sites have allowed us to designate the K790 site as a strong candidate for experimental validation.

In addition to analyzing established cleavage sites, such as S2' and S1/S2, for susceptibility to other proteases, and examining known implicated proteases like cathepsin L for their specific cleavage sites, we performed an unbiased search to identify additional potential cleavage sites of the S glycoprotein and the corresponding proteases capable of cleaving them. The search for additional cleavage sites was conducted in our study since we cannot exclude the possibility of unknown proteolytic processing of the spike glycoprotein in the SARS-CoV-2 life cycle. By analyzing both calculated sequence and structural proteolytic scores, we identified five positions on the S glycoprotein that likely correspond to previously unrecognized cleavage sites: R78, R346, R466, R634 and R847. These sites are located in various segments of the S glycoprotein including the N-terminal domain (NTD), receptor-binding domain (RBD), the loop connecting the S1 and S2 regions, and downstream of the fusion peptide. We have also examined the known and potential cleavage sites for preferences by furin, TMPRSS2, as well as other proteases from

the PCSK, TTSP, kallikreins, and coagulation factors families. Interestingly, while furin displays the highest cleavage scores at the S1/S2 and S2' sites, other members of the PCSK family demonstrate greater cleavage score values at position R78 than at the S1/S2 site. We speculate that this may be a contributing factor explaining why these proteases with similar specificity to furin are not linked to the pre-activation of the S glycoprotein at the S1/S2 site.

We have also examined whether mutations found in common SARS-CoV-2 variants are located near the considered cleavage sites and have the potential to impact their proteolytic efficiency. However, since we chose a sufficiently narrow width of the positional-weight matrix (PWM) involving three amino acids on each side of the cleavage, we were only able to estimate the impact of mutations located in the immediate vicinity of the cleavage sites. This limited size of the PWM was chosen to increase the number of proteases with modeled specificity, as a smaller PWM size requires fewer identified substrates to build the PWM. Finally, we identified only two independent mutations that influenced the aforementioned cleavage sites, and both of them led to the degradation of the presumable R346 cleavage site.

To the best of our knowledge, our study represents the first systematic analysis of human protease capabilities in the proteolytic processing of the SARS-CoV-2 spike glycoprotein. Our bioinformatics predictions provide direction for further experimental exploration and validation. The comprehensive understanding of the role of host factors, including human proteases, in SARS-CoV-2 infection, will significantly contribute to the development of new therapeutic agents, effective treatment strategies, and the prevention of future pandemics.

Acknowledgements

The reported study was funded by Russian Science Foundation, grant number 22-14-00132. We thank Irina Ponomareva for the design of the preprint's layout.

References

1. Ramage, H.; Cherry, S. Virus-Host Interactions: From Unbiased Genetic Screens to Function. *Annu. Rev. Virol.* **2015**, *2,* 497–524, doi:10.1146/annurev-virology-100114-055238.
2. Li, G.; Hilgenfeld, R.; Whitley, R.; De Clercq, E. Therapeutic Strategies for COVID-19: Progress and Lessons Learned. *Nat. Rev. Drug Discov.* **2023**, *22*, 449–475, doi:10.1038/s41573-023-00672-y.
3. V'kovski, P.; Kratzel, A.; Steiner, S.; Stalder, H.; Thiel, V. Coronavirus Biology and Replication: Implications for SARS-CoV-2. *Nat. Rev. Microbiol.* **2021**, *19*, 155–170, doi:10.1038/s41579-020-00468-6.
4. Gorbalenya, A.E.; Baker, S.C.; Baric, R.S.; de Groot, R.J.; Drosten, C.; Gulyaeva, A.A.; Haagmans, B.L.; Lauber, C.; Leontovich, A.M.; Neuman, B.W.; et al. The Species Severe Acute Respiratory Syndrome-Related Coronavirus: Classifying 2019-NCoV and Naming It SARS-CoV-2. *Nat. Microbiol.* **2020**, *5*, 536–544, doi:10.1038/s41564-020-0695-z.
5. Baggen, J.; Vanstreels, E.; Jansen, S.; Daelemans, D. Cellular Host Factors for SARS-CoV-2 Infection. *Nat. Microbiol.* **2021**, *6*, 1219–1232, doi:10.1038/s41564-021-00958-0.
6. Takeda, M. Proteolytic Activation of SARS-CoV-2 Spike Protein. *Microbiol. Immunol.* **2022**, *66*, 15–23, doi:10.1111/1348-0421.12945.
7. Jackson, C.B.; Farzan, M.; Chen, B.; Choe, H. Mechanisms of SARS-CoV-2 Entry into Cells. *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 3–20, doi:10.1038/s41580-021-00418-x.
8. Walls, A.C.; Park, Y.J.; Tortorici, M.A.; Wall, A.; McGuire, A.T.; Veesler, D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **2020**, *181*, 281-292.e6, doi:10.1016/j.cell.2020.02.058.
9. Krammer, F. SARS-CoV-2 Vaccines in Development. *Nature* **2020**, *586*, 516–527, doi:10.1038/s41586-020-2798-3.
10. Zabiegala, A.; Kim, Y.; Chang, K.O. Roles of Host Proteases in the Entry of SARS-CoV-2. *Anim. Dis.* **2023**, *3*, 1–14, doi:10.1186/s44149-023-00075-x.
11. Benton, D.J.; Wrobel, A.G.; Xu, P.; Roustan, C.; Martin, S.R.; Rosenthal, P.B.; Skehel, J.J.; Gamblin, S.J. Receptor Binding and Priming of the Spike Protein of SARS-CoV-2 for Membrane Fusion. *Nature* **2020**, doi:10.1038/s41586-020-2772-0.

12. Letko, M.; Marzi, A.; Munster, V. Functional Assessment of Cell Entry and Receptor Usage for SARS-CoV-2 and Other Lineage B Betacoronaviruses. *Nat. Microbiol.* **2020**, *5*, 562–569, doi:10.1038/s41564-020-0688-y.
13. Matsuyama, S.; Nao, N.; Shirato, K.; Kawase, M.; Saito, S.; Takayama, I.; Nagata, N.; Sekizuka, T.; Katoh, H.; Kato, F.; et al. Enhanced Isolation of SARS-CoV-2 by TMPRSS2- Expressing Cells. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 7001–7003, doi:10.1073/pnas.2002589117.
14. Shang, J.; Wan, Y.; Luo, C.; Ye, G.; Geng, Q.; Auerbach, A.; Li, F. Cell Entry Mechanisms of SARS-CoV-2. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, doi:10.1073/pnas.2003138117.
15. Callaway, E. The Coronavirus Is Mutating - Does It Matter? *Nature* **2020**, *585*, 174–177, doi:10.1038/d41586-020-02544-6.
16. Lubinski, B.; Whittaker, G.R. The SARS-CoV-2 Furin Cleavage Site : Natural. *The Lancet Microbe* **2023**, *5247*, 5247, doi:10.1016/S2666-5247(23)00144-1.
17. Whittaker, G.R. SARS-CoV-2 Spike and Its Adaptable Furin Cleavage Site. *The Lancet Microbe* **2021**, *2*, e488–e489, doi:10.1016/S2666-5247(21)00174-9.
18. Wu, Y.; Zhao, S. Furin Cleavage Sites Naturally Occur in Coronaviruses. *Stem Cell Res.* **2021**, *50*, doi:10.1016/j.scr.2020.102115.
19. Chan, Y.A.; Zhan, S.H. The Emergence of the Spike Furin Cleavage Site in SARS-CoV-2. *Mol. Biol. Evol.* **2021**, *39*, 1–7, doi:10.1093/molbev/msab327.
20. Whittaker, G.R.; Daniel, S.; Millet, J.K. Coronavirus Entry: How We Arrived at SARS-CoV-2. *Curr. Opin. Virol.* **2021**, *47*, 113–120, doi:10.1016/j.coviro.2021.02.006.
21. Burkard, C.; Verheije, M.H.; Wicht, O.; van Kasteren, S.I.; van Kuppeveld, F.J.; Haagmans, B.L.; Pelkmans, L.; Rottier, P.J.M.; Bosch, B.J.; de Haan, C.A.M. Coronavirus Cell Entry Occurs through the Endo-/Lysosomal Pathway in a Proteolysis-Dependent Manner. *PLoS Pathog.* **2014**, *10*, doi:10.1371/journal.ppat.1004502.
22. TZ., L.; H., Z.; H., L.; M., L.; R., Y.; J., P.; Q., X.; J., S.; B., L.; J., W.; et al. Identification of Common Deletions in the Spike Protein of SARS-CoV-2. *J. Virol.* **2020**, *94*.
23. Park, J.E.; Li, K.; Barlan, A.; Fehr, A.R.; Perlman, S.; McCray, P.B.; Gallagher, T. Proteolytic Processing of Middle East Respiratory Syndrome Coronavirus Spikes Expands Virus Tropism. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 12262–12267, doi:10.1073/pnas.1608147113.
24. Baggen, J.; Jacquemyn, M.; Persoons, L.; Vanstreels, E.; Pye, V.E.; Wrobel, A.G.; Calvaresi, V.; Martin, S.R.; Roustan, C.; Cronin, N.B.; et al. TMEM106B Is a Receptor Mediating ACE2-Independent SARS-CoV-2 Cell Entry. *Cell* **2023**, 3427–3442, doi:10.1016/j.cell.2023.06.005.
25. Meng, B.; Abdullahi, A.; Ferreira, I.A.T.M.; Goonawardane, N.; Saito, A.; Kimura, I.; Yamasoba, D.; Gerber, P.P.; Fatihi, S.; Rathore, S.; et al. Altered TMPRSS2 Usage by SARS-CoV-2 Omicron Impacts Infectivity and Fusogenicity. *Nature* **2022**, *603*, 706–714, doi:10.1038/s41586-022-04474-x.
26. Tang, T.; Jaimes, J.A.; Bidon, M.K.; Straus, M.R.; Daniel, S.; Whittaker, G.R. Proteolytic Activation of SARS-CoV-2 Spike at the S1/S2 Boundary: Potential Role of Proteases beyond Furin. *ACS Infect. Dis.* **2021**, *7*, 264–272, doi:10.1021/acsinfecdis.0c00701.
27. Kastenhuber, E.R.; Mercadante, M.; Nilsson-Payant, B.; Johnson, J.L.; Jaimes, J.A.; Muecksch, F.; Weisblum, Y.; Bram, Y.; Chandar, V.; Whittaker, G.R.; et al. Coagulation Factors Directly Cleave SARSCoV-2 Spike and Enhance Viral Entry. *Elife* **2022**, *11*, doi:10.7554/eLife.77444.
28. Benlarbi, M.; Laroche, G.; Fink, C.; Fu, K.; Mulloy, R.P.; Phan, A.; Ariana, A.; Stewart, C.M.; Prévost, J.; Beaudoin-Bussières, G.; et al. Identification and Differential Usage of a Host Metalloproteinase Entry Pathway by SARS-CoV-2 Delta and Omicron. *iScience* **2022**, *25*, doi:10.1016/j.isci.2022.105316.
29. UniProt Consortium, T. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2018**, *46*, 2699, doi:10.1093/nar/gky092.
30. Protein Data Bank: The Single Global Archive for 3D Macromolecular Structure Data. *Nucleic Acids Res.* **2019**, *47*, D520–D528, doi:10.1093/nar/gky949.
31. Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F.T.; De Beer, T.A.P.; Rempfer, C.; Bordoli, L.; et al. SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Res.* **2018**, *46*, W296–W303, doi:10.1093/nar/gky427.
32. Hodcroft, E.B. CoVariants: SARS-CoV-2 Mutations and Variants of Interest.
33. Binder, J.X.; Pletscher-Frankild, S.; Tsafou, K.; Stolte, C.; O'Donoghue, S.I.; Schneider, R.; Jensen, L.J. COMPARTMENTS: Unification and Visualization of Protein Subcellular Localization Evidence. *Database* **2014**, *2014*, 1–9, doi:10.1093/database/bau012.
34. Palasca, O.; Santos, A.; Stolte, C.; Gorodkin, J.; Jensen, L.J. TISSUES 2.0: An Integrative Web Resource on Mammalian Tissue Expression. *Database* **2018**, *2018*, 1–12, doi:10.1093/database/bay003.

35. Rawlings, N.D.; Barrett, A.J.; Thomas, P.D.; Huang, X.; Bateman, A.; Finn, R.D. The MEROPS Database of Proteolytic Enzymes, Their Substrates and Inhibitors in 2017 and a Comparison with Peptidases in the PANTHER Database. *Nucleic Acids Res*. **2018**, *46*, D624–D632, doi:10.1093/nar/gkx1134.

36. Wasserman, W.W.; Sandelin, A. Applied Bioinformatics for the Identification of Regulatory Elements. *Nat. Rev. Genet.* **2004**, *5*, 276–287, doi:10.1038/nrg1315.

37. Schechter, I.; Berger, A. On the Active Site of Proteases. 3. Mapping the Active Site of Papain; Specific Peptide Inhibitors of Papain. *Biochem. Biophys. Res. Commun.* **1968**, *32*, 898–902.

38. Matveev, E. V.; Safronov, V. V.; Ponomarev, G. V.; Kazanov, M.D. Predicting Structural Susceptibility of Proteins to Proteolytic Processing. *Int. J. Mol. Sci*. **2023**, *24*, 10761, doi:10.3390/ijms241310761.

39. Igarashi, Y.; Eroshkin, A.; Gramatikova, S.; Gramatikoff, K.; Zhang, Y.; Smith, J.W.; Osterman, A.L.; Godzik, A. CutDB: A Proteolytic Event Database. *Nucleic Acids Res*. **2007**, *35*, D546-9, doi:10.1093/nar/gkl813.

40. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res*. **2011**, *12*, 2825–2830.

41. Hoffmann, M.; Kleine-Weber, H.; Pöhlmann, S. A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. *Mol. Cell* **2020**, *78*, 779-784.e5, doi:10.1016/j.molcel.2020.04.022.

42. Bugge, T.H.; Antalis, T.M.; Wu, Q. Type II Transmembrane Serine Proteases. *J. Biol. Chem*. **2009**, *284*, 23177–23181, doi:10.1074/jbc.R109.021006.

43. Bosch, B.J.; Bartelink, W.; Rottier, P.J.M. Cathepsin L Functionally Cleaves the Severe Acute Respiratory Syndrome Coronavirus Class I Fusion Protein Upstream of Rather than Adjacent to the Fusion Peptide. *J. Virol*. **2008**, *82*, 8887–8890, doi:10.1128/jvi.00415-08.

44. Bollavaram, K.; Leeman, T.H.; Lee, M.W.; Kulkarni, A.; Upshaw, S.G.; Yang, J.; Song, H.; Platt, M.O. Multiple Sites on SARS-CoV-2 Spike Protein Are Susceptible to Proteolysis by Cathepsins B, K, L, S, and V. *Protein Sci*. **2021**, *30*, 1131–1143, doi:10.1002/pro.4073.

45. Zhao, M.M.; Zhu, Y.; Zhang, L.; Zhong, G.; Tai, L.; Liu, S.; Yin, G.; Lu, J.; He, Q.; Li, M.J.; et al. Novel Cleavage Sites Identified in SARS-CoV-2 Spike Protein Reveal Mechanism for Cathepsin L-Facilitated Viral Infection and Treatment Strategies. *Cell Discov.* **2022**, *8*, doi:10.1038/s41421-022-00419-w.