Increased preference for lysine over arginine in spike proteins of SARS-CoV-2 BA.2.86 variant and its sublineages

Anže Božič*
Department of Theoretical Physics, Jožef Stefan Institute, Ljubljana, Slovenia

Rudolf Podgornik†
Department of Theoretical Physics, Jožef Stefan Institute, Ljubljana, Slovenia
School of Physical Sciences, University of Chinese Academy of Sciences, Beijing, China
Kavli Institute for Theoretical Sciences, University of Chinese Academy of Sciences, Beijing, China and Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou, China

The COVID-19 pandemic offered an unprecedented glimpse into the evolution of its causative virus, SARS-CoV-2. Since its outbreak in late 2019, the virus has been estimated to have explored all possible alternatives in terms of missense mutations for all sites of its polypeptide chain. Spike protein of the virus—responsible for target recognition, cellular entry, and endosomal escape—exhibits the largest sequence variation in particular. Here, we detail the evolutionary changes in ionizable amino acids which can acquire either positive or negative charge and which have been shown to have led to a significant increase in the overall charge on the spike protein during the initial period of the pandemic. While the previously observed charge trends have recently come to a halt, we show that mutations in ionizable amino acids continue to occur with the newly emerging variants. What is more, we show that in mutations of amino acids which can acquire positive charge, there is a particular preference for lysine residues over arginine residues, a trend which has gained in prominence with BA.2.86 and its sublineages, including the recently dominant JN.1 and KP.3 variants. The high lysine-to-arginine ratio in the spike proteins of BA.2.86 and its sublineages does not appear to directly influence viral fitness, including receptor binding, and the reasons behind it remain unclear, even though similar trends having been observed in several other systems as well.

INTRODUCTION

From the time of its outbreak in the late 2019, severe acute respiratory coronavirus 2 (SARS-CoV-2) has been undergoing numerous evolutionary changes which are reflected in changes in its rate of transmission, immune escape, and overall fitness [1]. The first significant change in the virus' adaptation to humans has been observed about a year into the COVID-19 pandemic with the emergence of the first variant of concern (VOC) termed Alpha; soon thereafter, further VOCs such as Beta, Delta, and Omicron emerged and became dominant by outcompeting previous variants [1]. This process has led to the fairly recent emergence of Omicron BA.2.86 with enhanced antibody evasion [2] and the current prevalence of some of its descendant lineages, including JN.1 and KP3. Of the latter two, JN.1 shows a considerably higher infectivity and immune evasion at the expense of reduced binding to ACE2 [3, 4]; KP.3, on the other hand, shows both strong immune evasion as well as increased receptor binding capability [5].

It has been estimated that the average neutral single-nucleotide mutation in the SARS-CoV-2 genome has occurred $\sim$ 15000 independent times since its emergence, implying that the virus has already explored all possible alternatives in terms of missense mutations [6]. This mutational spectrum is, however, highly uneven and shows preferences for certain mutations over others [7, 8], which can have important structural implications [9]. Spike protein of SARS-CoV-2—and its receptor-binding (RBD) and N-terminal (NTD) domains in particular—is the part of the virus that exhibits the greatest sequence variation [10]. Various specific mutations in the spike protein have been linked to increased transmissibility of the virus and its binding to the angiotensin-converting enzyme 2 (ACE2) and other receptors [11–14]. Some of these mutations create changes to and from different *ionizable amino acids*, that is, amino acids whose residues carry either a positive or negative charge [15–19]. Apart from the effect of individual mutations, the resulting changes in the *total charge* on the RBD have been shown to correlate with RBD–ACE2 affinity in the main VOCs of SARS-CoV-2 [20]. Moreover, in the early stages of the pandemic, Pawłowski [21, 22] observed a tendency of the mutations of different VOCs to increase the overall number of positively charged amino acids on the spike protein in general.

* anze.bozic@ijs.si
† podgornikrudolf@ucas.ac.cn

The early trend towards an *increase of the positive charge* on the spike protein has been confirmed by later studies on larger sets of SARS-CoV-2 lineages [23, 24], which have furthermore shown a plateauing of the total charge on the spike protein with the emergence of Omicron and subsequent variants [24–26].

While the changes in the total charge on the spike protein might have come to a halt, mutations that change the number and nature of ionizable amino acids continue to occur unabated; for instance, in the recently emerged BA.2.86, electrostatic changes have been shown to contribute to its immune evasion [27]. In this work, we show that despite the plateauing of the total charge, the BA.2.86 variant and its daughter lineages show an increased preference for lysine over arginine residues, despite both amino acid types carrying positive charge. We also demonstrate that these changes consistently occur despite the fact that mutations from one amino acid type to another typically do not have a significant effect on viral fitness, ACE2 binding affinity, or RBD expression.

RESULTS AND DISCUSSION

There are six ionizable amino acid types that typically carry charge [28]: tyrosine (TYR) as well as aspartic (ASP) and glutamic (GLU) acid acquire negative charge while arginine (ARG), lysine (LYS), and histidine (HIS) acquire positive charge (see also Methods). Missense mutations concerning ionizable amino acids in the spike protein can either *(i)* convert a neutral amino acid to a charged one or vice versa; *(ii)* convert a positively charged amino acid to a negatively charged one; or *(iii)* convert between different amino acid types that otherwise carry the same charge. We have previously identified several different clusters of SARS-CoV-2 lineages that differ in their patterns of change in the number of ionizable amino acids on the spike protein [24, 25]. The currently dominant set of lineages tracing their origin to BA.2.86 variant presents a very distinct cluster, characterized by an increase in the number of LYS residues and a decrease in the number of ARG and GLU residues (highlighted box in Fig. 1). The pattern of change in the number of ionizable amino acids in this cluster is very distinct from previous variants, including, for instance, XBB and other recombinant variants. It also shows large changes compared to the early VOCs such as Beta and Delta, where the increase towards overall positive charge on the spike protein has first been observed [21, 22].

While the initial evolution of the number of ionizable amino acids in the early SARS-CoV-2 lineages resulted in a significant increase in the total charge on the spike protein [21–24], this trend slowed down and all but disappeared with the emergence of Omicron and later lineages [24, 25]. As Fig. 1 shows, this does not mean that mutations in ionizable amino acids have not kept occurring with the evolution of the virus, in particular since individual mutations can still have significant influence on both the properties of the spike protein as well as the overall viral fitness [29, 30]. In fact, we observe a still ongoing trend that shows an increase in the number of LYS residues in the spike protein and a slight decrease in the number of ARG residues (Fig. 2), despite both amino acids carrying a positive charge at typical solution conditions. The increase

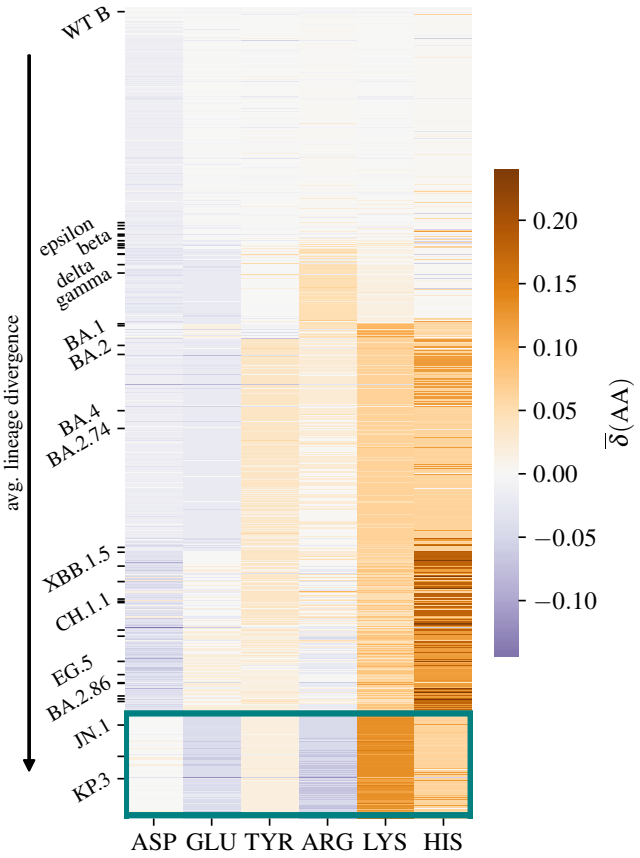

Figure 1. **Average change in the number of ionizable amino acids on the spike protein of SARS-CoV-2.** Shown is the average change $\overline{\delta}$(AA) in the number of negatively charged (ASP, GLU, and TYR) and positively charged (ARG, LYS, and HIS) amino acids on the spike protein of 2510 different SARS-CoV-2 lineages compared to WT. The lineages are ordered according to their (average) divergence from WT. Highlighted box includes the currently dominant set of lineages, including JN.1 and KP.3, all showing similar patterns of change in the number of ionizable amino acids.

in the number of LYS residues—from ~ 60 at the start of the COVID-19 pandemic to ~ 70 with the most recently emerged lineages—stands out in particular (Fig. 2b) and is also the main reason for the observed increase in the LYS/ARG ratio (Fig. 2c). While this ratio has first increased with the early Omicron lineages—where also the total charge on the spike protein reached its peak [24, 25]—we observe another pronounced increase with the most recently emerged lineages, including BA.2.86, JN.1, and KP.3 (reflected also in the highlighted box in Fig 1).

The increase in the LYS/ARG ratio is particularly interesting since it does not in any obvious way impact the overall positive charge on the spike protein. As Fig. 3a shows, the mutations to and from LYS and ARG residues most often occur in the RBD region of the spike protein, and several of these mutations are characteristic of BA.2.86 variants and its descendant lineages. The NTD region, on the other hand, exhibits a few ARG deletions, which is in line with observations that this region tends to acquire a more negative charge [31]. We can draw on previous studies [8, 32] to examine whether ARG↔LYS mutations in the spike protein of SARS-CoV-2 impact viral fitness [8], its binding affinity to ACE2, or the expression of RBD (a proxy for folding and stability) [32] (panels

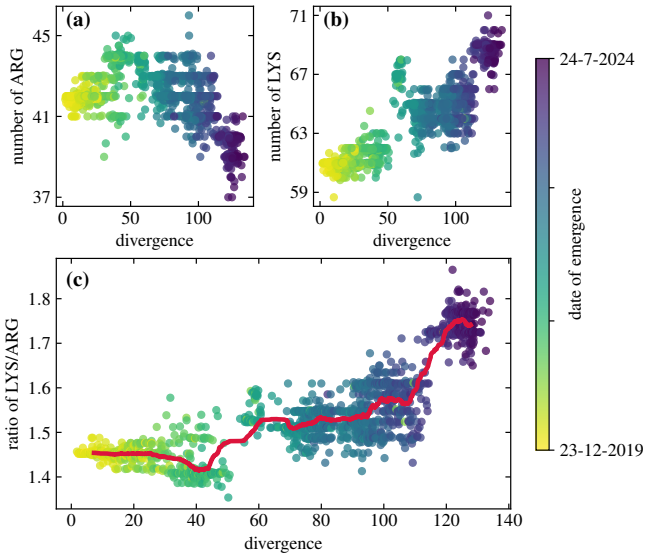

Figure 2. **Evolutionary change in the number of arginine and lysine residues on the spike protein of SARS-CoV-2.** Shown is the change in the number of **(a)** ARG and **(b)** LYS residues as well as **(c)** their ratio as a function of the (average) divergence of 2510 different SARS-CoV-2 lineages. Thick line in panel (c) shows a rolling average of the LYS/ARG ratio with lineage divergence (window size contains 100 lineages).

(b)–(d) of Fig. 3). We see that changing any LYS residue to an ARG residue has little effect on any of the three measures. Conversely, this is not always the case when an ARG residue is changed to a LYS residue, as mutations of certain ARG residues lead to a decrease of all three quantities—viral fitness, ACE2 binding, and RBD expression. Since LYS→ARG mutations do not seem to influence the obvious markers of SARS-CoV-2 fitness, there should thus be another reason why mutations to LYS nonetheless seem to be preferred over mutations to ARG (Fig. 2). Deep mutational scanning by Taylor and Starr [32] identified potential mutations on the spike protein that would enhance RBD expression, where they observed that many stabilizing mutations *removed* positively charged residues at the ACE2 interface that have been introduced in variant evolution—in particular HIS and LYS. This contrasts with the observed preference for LYS over ARG residues in recent SARS-CoV-2 lineages (Fig. 2), especially since the RBD of BA.2.86 acquired two new HIS and four new LYS residues compared to BA.2 alone.

In general, studies have shown that viral proteins tend to contain more ARG than LYS [33] and that an infected cell absorbs far more ARG from the culture medium than an uninfected one [34]. This has led to the idea of diet restriction of ARG-rich foods during viral illnesses [35] and supplementation with LYS, its antagonist, which attenuates the growth-promoting effect of ARG [36]. There is a competitive antagonism between LYS and ARG, and if a cell gets saturated in one of the two, its absorption slows and leaves the other free to be more absorbed [37]. Some studies have shown positive results of ARG depletion on treating viral infections [35, 37], including SARS-CoV-2 [38]; however, a recent study by Rees *et al.* [39] suggests this might potentially exacerbate the effects of the infection by SARS-CoV-2 instead. Moreover, a study by Melano *et al.* [40] has demonstrated that it is in fact LYS that can attenuate SARS-CoV-2 infection, which appears to be in line with our observation that SARS-CoV-2 spike proteins of recent lineages prefer LYS over ARG (Fig. 2) and which stands in contrast to what is typically observed in other viruses [33].

The reasons behind the preference for LYS over ARG could be related to their physicochemical properties. While it is true that ARG is coded for by six codons and LYS by only two, codon usage bias in SARS-CoV-2 and other coronaviruses suggests that CpG codons are suppressed, with only two out of the six ARG codons being predominantly in use [41, 42]. At the same time, while the chemical reactivities of LYS and ARG side chains are similar, LYS residues are far more prone to post-translational modifications [43, 44]—for instance, LYS acetylation regulates the interaction between proteins and membranes [45]. In

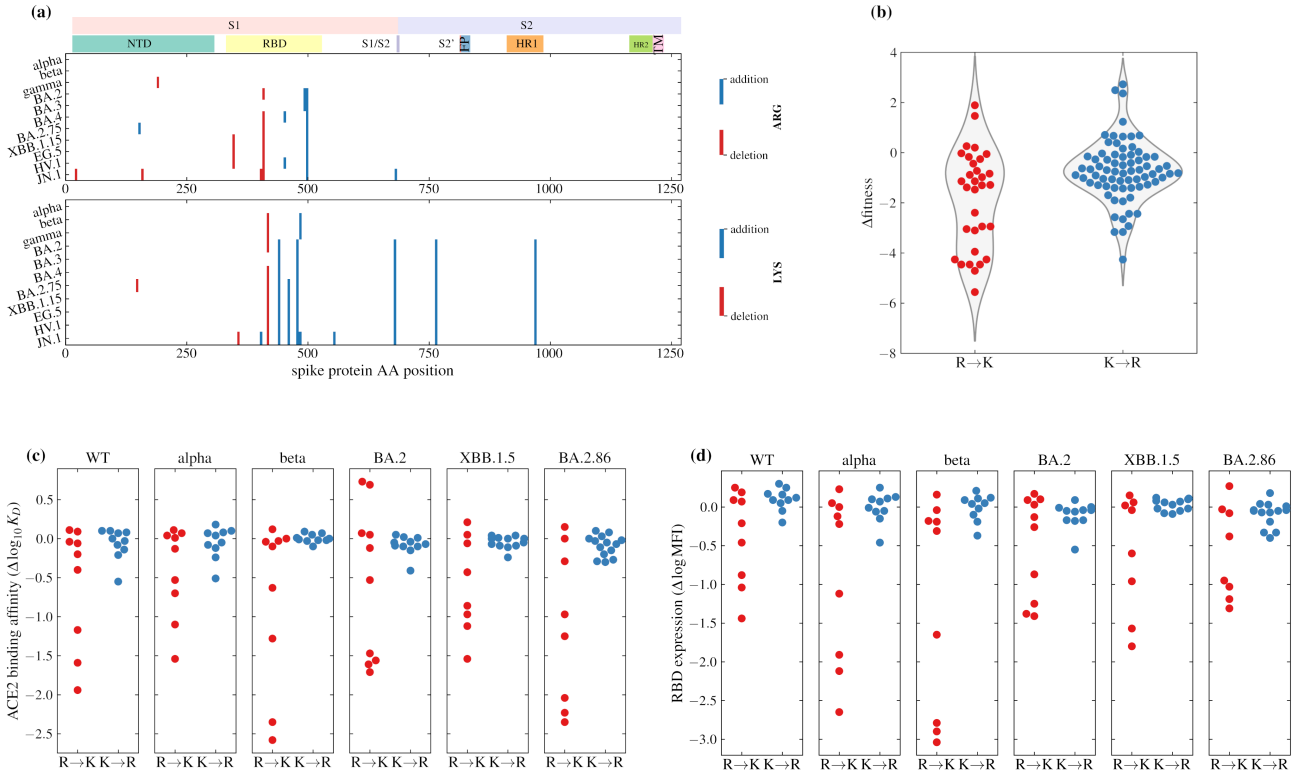

Figure 3. **Influence of arginine and lysine mutations on different aspects of SARS-CoV-2 fitness.** **(a)** Mutations in ARG and LYS residues in the spike proteins of select SARS-CoV-2 variants, showing changes to ("addition") or from ("deletion") the two residues. **(b)** Change in viral fitness upon a mutation in the spike protein from ARG to LYS (R→K) or LYS to ARG (K→R). Data from Ref. [8]. Deep mutation sequencing profiles of changes in **(c)** ACE2 binding affinity and **(d)** RBD expression upon mutations in the spike proteins of select SARS-CoV-2 variants. Data from Ref. [32].

the membrane, LYS residues readily deprotonate while ARG residues maintain their charge [46, 47] and can more easily disrupt and permeabilize lipid membranes; thus, preference for LYS over ARG could thus influence the electrostatic interaction of SARS-CoV-2 spike protein in specific environments [48–50]. Studies have also shown that preference for LYS over ARG can influence protein structural stability and folding [51–53], and a computational structural analysis of SARS-CoV-2 spike protein revealed that three specific ASP→LYS mutations in the central core region contribute to a preference for the alteration of the spike protein conformations [54]. Lastly, the ratio of ARG to LYS is also a known factor in determining protein solubility, with LYS being enriched relative to ARG in many of the more soluble proteins [55, 56].

CONCLUSION

Our study identified an increasing ratio of LYS to ARG residues in the spike protein of SARS-CoV-2 as a consequence of mutations that have been occurring since the early Omicron variants and which have seen a recent uptick with BA.2.86 and its daughter lineages. Combined with observations based on previous studies [8, 32], we showed that the choice of LYS over ARG in these mutations most often does not come with a significant benefit associated with viral fitness, binding with the ACE2 receptor, or the expression of RBD. While the reason behind the increased ratio thus remains an open question, we have outlined several potential mechanisms that could play a role.

MATERIALS AND METHODS

Data collection

Collection of sequences from different SARS-CoV-2 lineages follows our approach described previously [24, 25]. In short, we use a list of SARS-CoV-2 Pango lineages from `CoV-Lineages.org` [57] to download SARS-CoV-2 genomic and protein data from NCBI Virus database [58] together with accompanying annotations. We define the date of emergence of a lineage as the earliest full record (i.e., year, month, and day) of isolate collection. Lineage divergence, defined through the number of mutations in the entire genome relative to the root of the phylogenetic tree (the start of the outbreak), is obtained from the global SARS-CoV-2 data available at `Nextstrain.org` [59], and only those entries with a genome coverage of > 99 % are selected. Finally, we retain the lineages whose downloaded fasta protein file is not empty, resulting in a total number of $N = 2510$ analyzed SARS-CoV-2 lineages. All the quantities used in our work are determined as averages over all spike protein sequences included under each lineage; we denote this averaging by $\overline{X}$.

Ionizable amino acids

Typically, six different amino acids can acquire charge [28]; aspartic (ASP) and glutamic (GLU) acid as well as tyrosine (TYR) can acquire negative charge while arginine (ARG), lysine (LYS), and histidine (HIS) can acquire positive charge. We neglect cysteine (CYS) as it is usually not considered to be an acid [28, 60]. We use Biopython [61] to parse the protein fasta files and count the number of ionizable amino acids on the spike proteins of different SARS-CoV-2 lineages. To compare the relative change in the number of ionizable amino acids on the spike proteins of different lineages to the WT version (lineage B), we define

$$\overline{\delta}(\mathrm{AA}) = \frac{\overline{N}(\mathrm{lineage})}{\overline{N}(\mathrm{WT\ B})} - 1. \tag{1}$$

ACKNOWLEDGMENTS

A.B. acknowledges support from Slovenian Research Agency (ARIS) under contract no. P1-0055. R.P. acknowledges support from National Natural Science Foundation of China (NSFC) [Key Project 12034019].

[1] A. M. Carabelli, T. P. Peacock, L. G. Thorne, W. T. Harvey, J. Hughes, S. J. Peacock, W. S. Barclay, T. I. De Silva, G. J. Towers, and D. L. Robertson, Nature Reviews Microbiology **21**, 162 (2023).
[2] H. Yang, H. Guo, A. Wang, L. Cao, Q. Fan, J. Jiang, M. Wang, L. Lin, X. Ge, H. Wang, et al., Nature Communications **15**, 7715 (2024).
[3] S. Yang, Y. Yu, Y. Xu, F. Jian, W. Song, A. Yisimayi, P. Wang, J. Wang, J. Liu, L. Yu, et al., The Lancet Infectious Diseases **24**, e70 (2024).
[4] Y. Kaku, K. Okumura, M. Padilla-Blanco, Y. Kosugi, K. Uriu, A. A. Hinay, L. Chen, A. Plianchaisuk, K. Kobiyama, K. J. Ishii, et al., The Lancet Infectious Diseases **24**, e82 (2024).
[5] F. Jian, J. Wang, A. Yisimayi, W. Song, Y. Xu, X. Chen, X. Niu, S. Yang, Y. Yu, P. Wang, et al., bioRxiv (2024).
[6] N. Balasco, G. Damaggio, L. Esposito, V. Colonna, and L. Vitagliano, International Journal of Biological Macromolecules **266**, 131054 (2024).
[7] J. D. Bloom, A. C. Beichman, R. A. Neher, and K. Harris, Molecular Biology and Evolution **40**, msad085 (2023).
[8] J. D. Bloom and R. A. Neher, Virus Evolution **9**, vead055 (2023).
[9] Q. Wang, Y. Guo, L. Liu, L. T. Schwanz, Z. Li, M. S. Nair, J. Ho, R. M. Zhang, S. Iketani, J. Yu, et al., Nature **624**, 639 (2023).
[10] W. T. Harvey, A. M. Carabelli, B. Jackson, R. K. Gupta, E. C. Thomson, E. M. Harrison, C. Ludden, R. Reeve, A. Rambaut, C. G. U. C.-U. Consortium, et al., Nat. Rev. Microbiol. **19**, 409 (2021).
[11] B. Jawad, P. Adhikari, R. Podgornik, and W.-Y. Ching, J. Phys. Chem. Lett. **13**, 3915 (2022).
[12] F. Obermeyer, M. Jankowiak, N. Barkas, S. F. Schaffner, J. D. Pyle, L. Yurkovetskiy, M. Bosso, D. J. Park, M. Babadi, B. L. MacInnis, et al., Science **376**, 1327 (2022).
[13] P. Sang, Y.-Q. Chen, M.-T. Liu, Y.-T. Wang, T. Yue, Y. Li, Y.-R. Yin, and L.-Q. Yang, Int. J. Mol. Sci. **23**, 14796 (2022).

[14] W. Zhang, K. Shi, Q. Geng, G. Ye, H. Aihara, and F. Li, Proc. Natl. Acad. Sci. USA **119**, e2206509119 (2022).
[15] P. Adhikari, B. Jawad, R. Podgornik, and W.-Y. Ching, Int. J. Mol. Sci. **23**, 2870 (2022).
[16] P. Adhikari, B. Jawad, R. Podgornik, and W.-Y. Ching, Microorganisms **10**, 1999 (2022).
[17] C. Nie, A. K. Sahoo, R. R. Netz, A. Herrmann, M. Ballauff, and R. Haag, ChemBioChem **23**, e202100681 (2022).
[18] H. H. Gan, J. Zinno, F. Piano, and K. C. Gunsalus, Front. Virol. **2** (2022).
[19] S. H. Kim, F. L. Kearns, M. A. Rosenfeld, L. Votapka, L. Casalino, M. Papanikolas, R. E. Amaro, and R. Freeman, Cell Rep. Phys. Sci. **4** (2023).
[20] F. L. Barroso da Silva, C. C. Giron, and A. Laaksonen, J. Phys. Chem. B **126**, 6835 (2022).
[21] P. H. Pawłowski, Infect. Drug Resist. **14**, 5099 (2021).
[22] P. Pawłowski, Acta Biochim. Pol. **69**, 263 (2022).
[23] M. Cotten and M. V. Phan, iScience **26**, 106230 (2023).
[24] A. Božič and R. Podgornik, Virus evolution **9**, vead040 (2023).
[25] A. Božič and R. Podgornik, Bioinformatics Advances **4**, vbae053 (2024).
[26] F. Scarpa, E. Imperia, I. Azzena, M. Giovanetti, D. Benvenuto, C. Locci, M. Casu, P. L. Fiori, A. Maruotti, G. Ceccarelli, et al., The Journal of Infection **86**, 596 (2023).
[27] L. Li, K. Shi, Y. Gu, Z. Xu, C. Shu, D. Li, J. Sun, M. Cong, X. Li, X. Zhao, et al., Structure **32**, 1055 (2024).
[28] R. J. Nap, A. Božič, I. Szleifer, and R. Podgornik, Biophys. J. **107**, 1970 (2014).
[29] Z. Zhang, J. Zhang, and J. Wang, Frontiers in Public Health **10**, 952916 (2022).
[30] B. Dadonaite, J. Brown, T. E. McMahon, A. G. Farrell, M. D. Figgins, D. Asarnow, C. Stewart, J. Lee, J. Logue, T. Bedford, et al., Nature **631**, 617 (2024).
[31] R. J. Parsons and P. Acharya, Cell Rep. **42** (2023).
[32] A. L. Taylor and T. N. Starr, Virus Evolution **10**, veae067 (2024).
[33] M. D. Sanchez, A. C. Ochoa, and T. P. Foster, Antiviral research **132**, 13 (2016).
[34] S. Bol and E. M. Bunnik, BMC Veterinary Research **11**, 1 (2015).
[35] M. C. Pedrazini, E. F. Martinez, V. A. B. dos Santos, and F. C. Groppo, Future Journal of Pharmaceutical Sciences **10**, 1 (2024).
[36] D. J. Maggs, B. K. Collins, J. G. Thorne, and M. P. Nasisse, American Journal of Veterinary Research **61**, 1474 (2000).
[37] M. C. Pedrazini, M. H. da Silva, and F. C. Groppo, British Journal of Clinical Pharmacology **88**, 4708 (2022).
[38] J. M. Grimes, S. Khan, M. Badeaux, R. M. Rao, S. W. Rowlinson, and R. D. Carvajal, International Journal of Infectious Diseases **102**, 566 (2021).
[39] C. A. Rees, C. A. Rostad, G. Mantus, E. J. Anderson, A. Chahroudi, P. Jaggi, J. Wrammert, J. B. Ochoa, A. Ochoa, R. K. Basu, et al., Proceedings of the National Academy of Sciences **118**, e2101708118 (2021).
[40] I. Melano, L.-L. Kuo, Y.-C. Lo, P.-W. Sung, N. Tien, and W.-C. Su, Viruses **13**, 1301 (2021).
[41] H. Gu, D. K. Chu, M. Peiris, and L. L. Poon, Virus Evolution **6**, veaa032 (2020).
[42] P. C. Woo, B. H. Wong, Y. Huang, S. K. Lau, and K.-Y. Yuen, Virology **369**, 431 (2007).
[43] C. Azevedo and A. Saiardi, Advances in biological regulation **60**, 144 (2016).
[44] A. P. Loboda, S. M. Soond, M. Piacentini, and N. A. Barlev, Cell Cycle **18**, 1995 (2019).
[45] A. K. Okada, K. Teranishi, M. R. Ambroso, J. M. Isas, E. Vazquez-Sarandeses, J.-Y. Lee, A. A. Melo, P. Pandey, D. Merken, L. Berndt, et al., Nature Communications **12**, 6466 (2021).
[46] L. Li, I. Vorobyov, and T. W. Allen, The journal of physical chemistry B **117**, 11906 (2013).
[47] C. T. Armstrong, P. E. Mason, J. R. Anderson, and C. E. Dempsey, Scientific reports **6**, 21759 (2016).
[48] L. Javidpour, A. Božič, A. Naji, and R. Podgornik, Soft Matter **17**, 4296 (2021).
[49] J. Bigay and B. Antonny, Developmental cell **23**, 886 (2012).
[50] M. Matveeva, M. Lefebvre, H. Chahinian, N. Yahi, and J. Fantini, Viruses **15**, 1854 (2023).
[51] J. Dinic and M. V. Tirrell, Biomacromolecules **25**, 2838 (2024).
[52] J. I. Austerberry, A. Thistlethwaite, K. Fisher, A. P. Golovanov, A. Pluen, R. Esfandiary, C. F. van der Walle, J. Warwicker, J. P. Derrick, and R. Curtis, Biochemistry **58**, 3413 (2019).
[53] S. Sokalingam, G. Raghunathan, N. Soundrarajan, and S.-G. Lee, PloS one **7**, e40410 (2012).
[54] J. C. Boer, Q. Pan, J. K. Holien, T.-B. Nguyen, D. B. Ascher, and M. Plebanski, Frontiers in immunology **13**, 954435 (2022).
[55] S. Sallah and J. Warwicker, bioRxiv (2024).
[56] J. Warwicker, S. Charonis, and R. A. Curtis, Molecular pharmaceutics **11**, 294 (2014).
[57] Á. O'Toole, V. Hill, O. G. Pybus, A. Watts, I. I. Bogoch, K. Khan, J. P. Messina, T. COVID, B. C. G. Network, H. Tegally, et al., Wellcome Open Res. **6** (2021).
[58] E. L. Hatcher, S. A. Zhdanov, Y. Bao, O. Blinkova, E. P. Nawrocki, Y. Ostapchuck, A. A. Schäffer, and J. R. Brister, Nucleic Acids Res. **45**, D482 (2017).
[59] J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher, Bioinform. **34**, 4121 (2018).
[60] A. Božič and R. Podgornik, Biophys. J. **113**, 1454 (2017).
[61] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, et al., Bioinformatics **25**, 1422 (2009).