Stability and expression of SARS-CoV-2 spike-protein mutations

Kristoffer T. Bæk¹, Rukmankesh Mehra², and Kasper P. Kepp¹*

¹ DTU Chemistry, Technical University of Denmark, Building 206, 2800, Kongens Lyngby, Denmark. E-mail: * kj@kemi.dtu.dk ORCID: 0000-0002-6754-7348.

² Department of Chemistry, Indian Institute of Technology Bhilai, Sejbahar, Raipur-492015, Chhattisgarh, India. E-mail: rukmankesh@iitbhidai.ac.in

* Correspondence: E-mail: kj@kemi.dtu.dk Tel: +45 45252409

ORCID
Kristoffer T. Bæk: 0000-0003-3243-3962
Rukmankesh Mehra: 0000-0001-6010-1514
Kasper P. Kepp: 0000-0002-6754-7348

Abstract
Protein fold stability likely plays a role in SARS-CoV-2 S-protein evolution, together with ACE2 binding and antibody evasion. While few thermodynamic stability data are available for S-protein mutants, many systematic experimental data exist for their expression. In this paper, we explore whether such expression levels relate to the thermodynamic stability of the mutants. We studied mutation-induced SARS-CoV-2 S-protein fold stability, as computed by three very distinct methods and eight different protein structures to account for method- and structure-dependencies. For all methods and structures used (24 comparisons), computed stability changes correlate significantly (99% confidence level) with experimental yeast expression from the literature, such that higher expression is associated with relatively higher fold stability. Also significant, albeit weaker, correlations were seen for ACE2 binding. The effect of thermodynamic fold stability may be direct or a correlate of amino acid or site properties, notably the solvent exposure of the site. Correlation between computed stability and experimental expression and ACE2 binding suggests that functional properties of the SARS-CoV-2 S-protein mutant space are largely determined by a few simple features, due to underlying correlations. Our study lends promise to the development of computational tools that may ideally aid in understanding and predicting SARS-CoV-2 S-protein evolution.

Keywords: SARS-CoV-2; covid-19; spike protein; mutations; protein stability
Introduction

The pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)[1–3] has led to intensive research into its spike-protein (S-protein, Figure 1a), whose evolution may lead to changes in its surface that reduce the efficiency of vaccine-induced antibodies.[4–6] The S-protein is a glycosylated homo-trimer on the surface of coronaviruses responsible for their characteristic shape.[7, 8] During entry of the virus particle into human host cells, the S-protein binds to the cell-surface receptor angiotensin-converting enzyme 2 (ACE2), Figure 1b.[9–11]

Due to the importance of S-protein epitopes for immune recognition, presentation of the S-protein is the rationale behind a range of important vaccines.[12, 13] The presence of prominent antibodies in a population may lead to selection pressure to change the S-protein surface to evade the antibodies learned from vaccination or infection with earlier variants.[14] Such antigenic drift may lead to variants capable of escaping vaccine-induced immunity, a problem that is likely to persist for many years. The omicron variant has been a hallmark example of such evolution, leading to a very large number of breakthrough infections in late 2021 and early 2022.[15]

Understanding such evolution effects requires protein structural data.[16–19] Since function is structure-dependent, the structure is the platform on which the amino acid evolution occurs, with evolution rates typically depending on the structural context of the amino acid site.[20–22] The recent technical breakthroughs in cryo-electron (cryo-EM) microscopy of large molecules such as proteins[23–27] are an excellent example of basic science providing essential innovation enormous utility, enabling publication of hundreds of SARS-CoV-2 S-protein structures during the last two years, both of the protein alone in the apo-state (apo-S-protein), in various conformation states, and in complex with a large range of antibodies and ACE2 at resolutions typically at 2-4 Å.[28]

Among the selection pressures acting on a protein beyond its direct function is the need to maintain the overall fold stability and translational fidelity, and thus evolution of new mutations often occurs with a tradeoff not to undermine these properties.[17, 29–32] Before fusion with a host cell, the S-protein is in a metastable conformation state under selection to evade antibodies and enhance ACE2 binding, which requires appropriate conversion to a conformation state with the receptor binding domain (RBD) in an up-wards conformation state.[33–37]
Figure 1: Structure of the SARS-CoV-2 S-protein, studied in this work. (a) Representative structure of the apo-S-protein state (PDB: 7DF3) with the studied mutated sites of the RBD shown in red color. (b) Representative cryo-EM structure of the S-protein in complex with ACE2, (PDB: 7DF4). (c) Overlay of the eight apo-S-protein structures used in this study to account for structural heterogeneity when computing mutation effects.

Understanding the importance of fold stability-function tradeoffs within the S-protein could be important for estimating the tolerability of new mutations and the likelihood and fitness of new emerging variants, which will remain an important topic even in the post-pandemic period. To understand this, we can compute the stability effects of the many mutations possible in the RBD. Many programs exist based on machine-learning or energy-based force-fields that can compute changes in protein stability upon mutation, with variable accuracy and issues relating to systematic errors (biases) and results being dependent on input structures used. The few experimental S-protein stability data from time-consuming differential scanning calorimetry or denaturation experiments prevent statistically meaningful analysis. Instead, we hypothesized that expression levels, which are available for very many SARS-CoV-2 S-protein mutations, may be a proxy of fold stability, as argued previously and explored further below.
Methods

Data for expression and ACE2 binding

Data for the effect of S-protein RBD single-point mutations on yeast expression and ACE2-binding were published by Starr et al.[53] These mutation sites are marked in red color in Figure 1a. The effect of mutations on expression is reported as the difference in log-mean fluorescence intensity (MFI) relative to wild-type (ΔlogMFI = logMFIvariant − logMFIwild-type), such that a positive value indicates higher RBD expression.[53] The effect of mutations on ACE2-binding (Figure 1b) is calculated from the apparent dissociation constants (KD, app) and shown as the difference in log10(KD,app) relative to wild-type (Δlog10(KD,app) = log10(KD,app)wild-type − log10(KD,app)variant), such that a positive value indicates higher variant ACE2 binding.[53] There were two independent measurements (one from each of two independent libraries) for each mutation, providing an important way to assess the quality and reproducibility of the individual data points.

Data curation

There is an overall good correlation between the results obtained from the two independent libraries (Figure S1a). We removed outliers by filtering out observations where the two replicates are different (for binding or expression), defined as residuals > 1, or where data for one replicate is missing. The rationale is that data points not similar in the two experimental replicates cannot be considered reproduced and may erroneously affect analysis. We furthermore removed data points with effects on binding in either replicate < -4.5 to avoid values near the detection limit (see Figure S1a). The correlation between the replicates after curation is shown in Figure S1b. Removing outliers and data points at the detection limit also excluded stop codon mutations (Figure S2). The described curation removed 685 of the original 4221 data points. Further analyses were performed using the remaining 3536 data points and the average of the binding and expression data from the two independent libraries, with detailed data collected in the supplementary file Table_S1.csv.

Structures and computer models used to compute stability effects

For each mutation in the dataset, the change in protein free energy of folding (ΔΔG, kcal/mol) was computed using three different methods: DeepDDG[54], mCSM Protein Stability[45]
(mCSM), and SimBa-IB[55]. DeepDDG and mCSM were accessed via their respective web servers (http://protein.org.cn/ddg.html, http://biosig.unimelb.edu.au/mcsm/stability), and Simba-IB was run from the command-line (http://github.com/kasperplaneta/SimBa2).

We have previously shown that analysis of functional properties can depend on input structure used,[56] and such heterogeneity is also seen in published cryo-EM structures of the S-protein.[28] Accordingly, we used eight experimental cryo-EM structures of the apo-S protein from the Protein Data Bank (PDB): 6VXX,[35] 6X6P,[57] 6X79,[58] 6Z97,[59] 6ZB4,[60] 7CAB,[61] 7DDD,[62] and 7DF3,[38] to account for such heterogeneity. These structures are shown in structural overlay in Figure 1c.

The relative solvent accessible surface area (RSA) of the mutated sites were calculated using SimBa-IB[55], which uses FreeSASA for this task.[63] Because the eight PDB structures represent homo-trimers, the ΔΔG and RSA values reported in this study are average ΔΔG and RSA values for the three chains (A, B, and C) of each structure.

It is important to note that the cryo-EM structures discussed may not reflect very precisely the real conformations of the S-protein at physiological temperature (37 °C): Cryo-EM structures are typically obtained with samples deposited with vitrified ice and rapidly cooled using a cryo-agent.[24, 25, 64] The freezing may remove some conformational dynamics.[65–69] Conformational changes of the S-protein may also be temperature-dependent.[70] In addition, the physiologically relevant state of the S-protein tends to be heavily glycosylated, which none of the experimental assays studying mutation impacts so far has addressed. Our goal is to identify whether experimental apo-S-protein data can be approximated by computed data, without translating these to the physiological significance of these data.
Results and discussion

Experimental and computed genotype-phenotype heat maps

Our main interest was to investigate whether experimentally measured expression and ACE2 binding of S-protein RBD mutants[53] can be related to the thermodynamic stability of the mutants. Amino acid substitutions often change the stability of a protein with a tendency for the distribution of such effects to be skewed towards destabilization.[71] Mutations in RBD affect expression of the S-protein and binding to ACE2 differently, but as discussed before[53] there is a relatively strong correlation (R=0.59) between the effect of mutations on expression and ACE2 binding (Figure S3a). This relationship is not intuitive but could relate to underlying effects such as ACE2 recognition and stability both relating to the mutating site's solvent accessibility, or correlations between codon use[72] which affects replication efficiency[73] and amino acid properties. Expression levels could also affect apparent binding constants even at the same specific ACE2 affinity, even if this is apparently accounted for, due to complex (e.g., tertiary) interactions. The large number of mutations having a small or "nearly neutral" impact, as noted by the authors[53] and reasonably expected, may also spuriously affect the relationships. However, if the data are grouped to adjust for the skewed distribution, the correlation between expression and binding becomes even larger and quite remarkable (R=0.88) (Figure S3b).

To assess the correlation between the effects of mutations on expression, ACE2 binding and protein stability, we computationally predicted the stability changes (ΔΔG) caused by the mutations in RBD using three state-of-the-art methods (DeepDDG, mCSM, and SimBa-IB) and compared with the experimentally observed changes in expression and binding. Figure 2a and Figure 2b show heat maps build from the experimental data by Starr et al.[53] after curation for non-reproducible replicates as described in Methods. These experimental heat maps are compared with the computational heat maps of ΔΔG derived in the present work, using DeepDDG (Figure 2c), mCSM (Figure 2d), and SimBa-IB (Figure 2e). SimBa produces more stabilizing trends overall, as it was developed to handle destabilization biases (the method performs similarly to other methods in benchmarks despite this feature)[55, 74].
Figure 2. Heat maps of mutation effects on RBD experimental expression and ACE2 binding (from Starr et al.[53]) compared with computed stability change (this work). (a) expression; (b) ACE2-binding affinity. Protein stability change computed with: (c) deepDDG; (d) mCSM (e) SimBa-IB. Rectangles are colored by mutational effect according to scale bars on the right. Black dots indicate the wild-type residues. White rectangles represent mutations for which there is no data in the curated dataset.
Figure 3. Effects of mutations mapped onto the SARS-CoV-2 RBD structure. Each residue is color-coded from cyan to red according to the average absolute effects of mutation. (a) RBD expression. (b) ACE2 binding. (c) Protein stability estimated by DeepDDG. (d) Protein stability estimated by mCSM, and (e) Protein stability estimated by SimBa-IB. Cyan indicates no effect, red indicates a strong effect. The color scales are relative within each panel. The ACE2 binding motif is shown on the right-hand side in each structure. (PDB: 6Z97, Chain A).

The heat maps in Figure 2 of experimental binding and expression do have some residual similarities, as also mentioned by Starr et al.[53] The computed stability effects provide estimates of the impact of all possible mutations in the RBD, and show some similarities to the expression data, notably with nearly neutral effects (grey) being common to both experimental and computed data in the N- and C-terminals and in some regions of the protein around 444-450 and a larger area around 470-488. In contrast, mutations in the region 388-396 have strong effects on protein stability but are nearly neutral with respect to expression and ACE2 binding. Overall, there is less similarity between ACE2 binding and the computed stability changes as perhaps expected.

RBD sites with non-neutral effects cluster in the structured regions

Figure 3 shows the effect of mutation at each site on expression, ACE2 binding and predicted protein stability mapped onto the RBD structure. The effect at each site is calculated as the average of the absolute effect (without sign) of all 19 mutations as a measure of the tolerance
to mutation at each site. Mutations that affect expression are mainly located in the core RBD subdomain, and in particular in the central beta sheet and its flanking alpha helices (Figure 3a). Mutations that affect binding are located in the ACE2 binding subdomain or, similarly to mutations that affect expression, in the central beta sheet, while large parts of the RBD domain are tolerant to mutations with regard to ACE2 binding (Figure 3b). Mutations that affect protein stability are mainly located in the structured core of the RBD subdomain, similarly to the effect of mutations on expression (Figure 3c-3e). When evaluating the computational methods in this way, their ability to identify the most tolerant sites for mutation (or equally, the sites more likely to be neutral from an evolutionary perspective) becomes more evident than in the heat maps of Figure 2, while the better agreement with expression remains clear.

**Computed and experimental S-protein mutant properties correlate significantly**

As discussed above, mutations in the RBD affect expression, ACE2 binding and protein stability differently, but with some overlap in site effects especially for protein stability and expression. To quantify this relationship, we plotted the predicted ΔΔG values for all mutations against the observed changes in RBD expression (Figure 4a) and ACE2 binding (Figure 4b), respectively, for each experimental structure used as input for the three methods (24 comparisons for all studied RBD mutations for expression and 24 comparisons for ACE2 binding). Correlations between RBD expression and ΔΔG for individual PDB structures are consistently observed, but their magnitude depends on the prediction method, with correlation coefficients ranging from 0.40 to 0.48 for DeepDDG, 0.27 to 0.34 for mCSM, and 0.12 to 0.22 for SimBa-IB (Figure 4a).

The observed ACE2 binding and the predicted ΔΔG also correlate, depending on the prediction method (Figure 4b), which is in line with the correlation between experimental measures of expression and ACE2 binding (Figure S3). However, the correlations are weaker than for expression with correlation coefficients ranging from 0.29 to 0.37 for DeepDDG, 0.13 to 0.26 for mCSM, and 0.09 to 0.15 for SimBa-IB. In all 48 comparisons of computed and experimental data in Figure 4, the correlations are statistically significant at the 99% confidence level (p-values of linear regression < 0.01).
Figure 4. Predicted change in protein stability as a function of mutations effect on RBD expression and ACE2 binding. Three different prediction methods (DeepDDG, mCSM, and SimBa-IB; indicated on the right) were used to predict the change in protein stability for each mutation using eight different experimental structures of the S-protein (indicated on top). Orange lines indicate the resulting linear regression, and the correlation coefficients (R) are shown. The p values for all correlations are <0.001. (a) ΔΔG plotted against RBD expression, and (b) ΔΔG plotted against ACE2 binding.
Figure 5. Average change in protein stability for the eight used structures vs. effect on RBD expression and ACE2 binding. Stability changes upon mutation for DeepDDG, mCSM, and SimBa-IB correlate with the experimental data at 99% significance (p-values of linear regression). Orange lines indicate the resulting linear regression.

Figure 5 shows the aggregate data for all structures, to account for structural heterogeneity effects. The relationships observed in Figure 4 still hold true when averaging over all eight structures, i.e., our result is robust to structural heterogeneity. The effect of mutations on protein stability predicted using DeepDDG correlates better with both expression and ACE2 binding than using mCSM and especially SimBa-IB (Figure 4 and Figure 5). Judging from Figure 3c-e, the three methods agree to a large extent on how they predict the tolerability of a site to mutation, whereas they differ in how they predict the effect of individual mutations (Figure 2c-e), and this is most likely the cause of the differences in the resulting correlations.
Figure 6. Predicted change in protein stability as a function of mutations effect on RBD expression and binding to ACE2 using binned data. Three different prediction methods (DeepDDG, mCSM, and SimBa-IB) were used to predict the change in protein stability for each mutation using eight different experimental structures of Spike protein. The average change in stability for the eight structures is used in the calculation. (a) Binding and expression data grouped in 0.5-value bins (b) Binding and expression data grouped in 0.25-value bins.

Both the binding and expression effects are highly skewed, with an overrepresentation of data points between 0 and -1 (Figure S2). In order to adjust for this, we grouped the data into bins of 0.5-width and 0.25-width and calculated the mean expression and binding effect in each bin, and the mean predicted ΔΔG value for each bin. Figure 6 shows these data as averages of all PDBs (data for individual structures in Figures S4-S7). The correlation increases substantially upon binning, with each data point more well-determined, whereas the p-values decrease substantially due to the few aggregate data points after averaging. Remarkably, the computational data correlate extremely well with the binned experimental data, especially for DeepDDG, much more than normally seen.[74] Considering that the models were developed to predict fold stability effects, not expression (which also depends on effects at the nucleic acid level) and that we used the full S-protein structures whereas the experiments express RBD on the yeast surface with expected modifications, this result is very surprising. One interpretation of this result is that broader functional properties of the mutant space of the S-protein are in fact largely determined by a few simple features, due to underlying correlations.
**Figure 7.** Site-averaged impact on ACE2 binding and expression vs. stability changes. This figure shows experimental-computational relationships for site-averaged properties (all mutations with data for each site averaged) as an indicator of the site's tolerance to mutation.

To understand the correlations on a per-site basis, **Figure 7** shows the relationships between the experimental and computed data averaged over sites, as an indicator of the site's tolerance to mutations. As this removes some amino-acid specific variations between the three computer models used, the correlations now become more similar between the methods. In all cases, there is a significant correlation (99% confidence level, p-values of linear regression), such that sites more neutral to expression and ACE2 binding effects experimentally are also more neutral towards computed stability effects.

Mutations in residues that are buried in the core of a protein tend to have larger effect on protein stability, which is also the case for the RBD domain, where mutations in the core subdomain are less tolerated (**Figure 3c-3e**) than mutations near the surface. To quantify if surface exposure of residues in RBD correlated with the effect of the mutations on expression and binding, we plotted these variables against the RSA for each site, and we see a moderate correlation (R = 0.29 for binding, and R = 0.43 for expression) with a notable overall tolerance to mutation at sites with high solvent exposure (**Figure S8**). As expected, there is also a good correlation between surface exposure and predicted stability changes, with correlation coefficients ranging from 0.27 (for SimBa-IB) to 0.60 (for DeepDDG; **Figure S9**).
Concluding remarks and biological implications

The SARS-CoV-2 S-protein fusion with human ACE2 is a prerequisite for host cell entry.[9, 10] However, S-protein antibodies from previous infection or vaccines in the population bind the S-protein and neutralize some virus particles, thus reducing infectivity. During evolution of SARS-CoV-2, these two effects need to be balanced upon changes in the S-protein.[28] Most SARS-CoV-2 evolution until the omicron variant involved positive selection for better ACE2 binding[75], a feature maintained with omicron despite substantial evasion of existing antibodies via its many mutations in the S-protein.[76] In addition to these two, protein fold stability is commonly important during the evolution of new functionality, as it constrains the options available (function-stability tradeoffs)[29, 30, 77, 78] and has been speculated to play a role in SARS-CoV-2 S-protein evolution as well.[28] In order to understand and possibly predict SARS-CoV-2 evolution, an important health challenge,[28, 39] we evaluated such contributions of protein fold stability to the S-protein RBD mutation space.

Our work shows that computed protein stability effects correlate significantly for all 48 comparisons of data sets (8 structures, 3 methods, and two properties) with expression levels and to lesser extent ACE2 binding observed in experiments.[53] Thus, experimental mutant properties can partly be predicted, and the phenotypes may correlate, possibly due to underlying correlators between e.g. codon use, amino acid chemical properties, and site solvent exposure. Such correlations could impact the mutability of the SARS-CoV-2 S-protein and affect phenotype tradeoffs, and thus ultimately the virus evolution, given that S-protein fold stability in the prefusion state is an important property of SARS-CoV-2 mutants,[33] which suggests including protein stability into a fitness function of SARS-CoV-2.[28] To the extent that protein stability maintenance is an important selection pressure, it will affect the future SARS-CoV-2 S-protein evolution, including constraints on its antigenic drift.

Although single amino acid changes as analyzed above and in assays[79, 80] are likely to be not additive in real variants with multiple substitutions, due to amino acid correlation effects (within-protein epistasis), and possibly epistasis with other virus genes.[81–84] These epistasis effects haunt the protein evolution field and are not easily accounted for, but recent work suggests that substantial parts of the epistasis is already utilized.[85] Although this remains to be studied in future work, and does not include potential inter-gene epistasis such as e.g. processing of S-protein RNA by non-structural proteins during virus replication within the host cell. Still, the maintenance of S-protein fold stability in the lipid surface of the virion is likely to be an important constraint on ACE2 binding and antigenic drift making the heat maps studied here of interest both as a proxy of expression and S-protein stability but possibly also as a model contribution to the computational estimates of the overall fitness function of SARS-CoV-2.
Supporting Information

The supporting information pdf file contains additional data and analysis of relevance. The data set file Table_S1.csv contains the curated experimental data sets used for the study.

Data availability statement

The full curated data used in this study are available as a csv file uploaded as supplementary information (Table_S1.csv), and scripts for analyzing these data are available at the web page https://github.com/ktbaek/Spike-ACE2.

Acknowledgements

The Danish Council for Independent Research (Grant case: 8022-00041B) is gratefully acknowledged for supporting this work.
References

1. Wu F, Zhao S, Yu B, et al (2020) A new coronavirus associated with human respiratory disease in China. Nature 579:265–269.
2. Zhu N, Zhang D, Wang W, et al (2020) A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 382:727–733.
3. Cevik M, Kuppalli K, Kindrachuk J, Peiris M (2020) Virology, transmission, and pathogenesis of SARS-CoV-2. bmj 371:
4. Dejnirattisai W, Zhou D, Ginn HM, et al (2021) The antigenic anatomy of SARS-CoV-2 receptor binding domain. Cell 184:2183-2200.e22. doi: 10.1016/j.cell.2021.02.032
5. Kemp SA, Collier DA, Datir RP, et al (2021) SARS-CoV-2 evolution during treatment of chronic infection. Nature 592:277–282. doi: 10.1038/s41586-021-03291-y
6. van Dorp L, Houldcroft CJ, Richard D, Balloux F (2021) COVID-19, the first pandemic in the post-genomic era. Curr Opin Virol 50:40–48.
7. Harvey WT, Carabelli AM, Jackson B, et al (2021) SARS-CoV-2 variants, spike mutations and immune escape. Nat Rev Microbiol 19:409–424.
8. Guruprasad L (2021) Human SARS CoV-2 spike protein mutations. Proteins Struct Funct Bioinforma 89:569–576. doi: https://doi.org/10.1002/prot.26042
9. Letko M, Marzi A, Munster V (2020) Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. Nat Microbiol 5:562–569.
10. Fehr AR, Perlman S (2015) Coronaviruses: an overview of their replication and pathogenesis. In: Coronaviruses Methods Protoc. Springer, New York, pp 1–23
11. Wang Q, Zhang Y, Wu L, et al (2020) Structural and functional basis of SARS-CoV-2 entry by using human ACE2. Cell 181:894–904.
12. Liu C, Zhou Q, Li Y, et al (2020) Research and development on therapeutic agents and vaccines for COVID-19 and related human coronavirus diseases. ACS Cent Sci 6, 3:315–331.
13. Forini G, Mantovani A (2021) COVID-19 vaccines: where we stand and challenges ahead. Cell Death Differ 28:626–639.
14. Yuan M, Huang D, Lee C-CD, et al (2021) Structural and functional ramifications of antigenic drift in recent SARS-CoV-2 variants. Science (80-) 373:818–823. doi: 10.1126/science.abh1139
15. Liu L, Iketani S, Guo Y, et al (2021) Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2. Nature 1–8.
16. Dasme P, Serohijos AWR, Kepp KP, Shakhnovich EI (2013) Positively Selected Sites in Cetacean Myoglobin Contribute to Protein Stability. PLoS Comput Biol 9:e1002929.
17. Liberles D a, Teichmann S a, Bahar I, et al (2012) The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci 21:769–785. doi: 10.1002/pro.2071
18. Bajaj M, Blundell T (1984) Evolution and the tertiary structure of proteins. Annu Rev Biophys Bioeng 13:453–492.
19. Wylie CS, Shakhnovich EI (2011) A biophysical protein folding model accounts for most mutational fitness effects in viruses. Proc Natl Acad Sci U S A 108:9916–9921. doi: 10.1073/pnas.1017572108
20. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341–352. doi: 10.1016/j.cell.2008.05.042
21. Topham CM, Srinivasan N, Blundell TL (1997) Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. Protein Eng 10:7–21.
22. Worth CL, Gong S, Blundell TL (2009) Structural and functional constraints in the evolution of protein families. Nat Rev Mol Cell Biol 10:709.
23. Blundell TL, Chaplin AK (2021) The resolution revolution in X-ray diffraction, Cryo-EM and other Technologies. Prog Biophys Mol Biol 160:2–4. doi: https://doi.org/10.1016/j.pbiomolbio.2021.01.003
24. Murata K, Wolf M (2018) Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. Biochim Biophys Acta - Gen Subj 1862:324–334. doi: 10.1016/j.bbagen.2017.07.020
25. Fernandez-Leiro R, Scheres SHW (2016) Unravelling biological macromolecules with cryo-electron microscopy. Nature 537:339–346. doi: 10.1038/nature19948
26. Danev R, Yanagisawa H, Kikkawa M (2019) Cryo-electron microscopy methodology: current aspects and future directions. Trends Biochem Sci 44:837–848.
27. Elmlund D, Elmlund H (2015) Cryogenic electron microscopy and single-particle analysis. Annu Rev Biochem 84:499–517.
28. Mehra R, Kepp KP (2022) Structure and Mutations of SARS-CoV-2 Spike Protein: A Focused Overview. ACS Infect Dis 8:29–58. doi: 10.1021/acsinfectdis.1c00433
29. Tokuriki N, Stricher F, Serrano L, Tawfik DS (2008) How protein stability and new functions trade off. PLoS Comput Biol 4:e1000002. doi: 10.1371/journal.pcbi.1000002
30. Goldstein RA (2011) The evolution and evolutionary consequences of marginal thermostability in proteins. Proteins 79:1396–1407. doi: 10.1002/prot.22964
31. Kepp KP (2020) Survival of the cheapest: how proteome cost minimization drives evolution. Q Rev Biophys 53:e7. doi: DOI: 10.1017/S0033583520000037
32. Bershtein S, Goldin K, Tawfik DS (2008) Intense neutral drifts yield robust and evolvable consensus proteins. J Mol Biol 379:1029–1044. doi: 10.1016/j.jmb.2008.04.024
33. Berger I, Schaffitzel C (2020) The SARS-CoV-2 spike protein: balancing stability and infectivity. Cell Res 30:1059–1060. doi: 10.1038/s41422-020-00430-4
34. Cai Y, Zhang J, Xiao T, et al (2020) Distinct conformational states of SARS-CoV-2 spike protein. Science (80-) 369:1586–1592.
35. Walls AC, Park Y-J, Tortorici MA, et al (2020) Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 181:281–292.
36. Henderson R, Edwards RJ, Mansouri K, et al (2020) Controlling the SARS-CoV-2 spike glycoprotein conformation. Nat Struct Mol Biol 27:925–933. doi: 10.1038/s41594-020-0479-4

37. Wrapp D, Wang N, Corbett KS, et al (2020) Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science (80-) 367:1260–1263. doi: 10.1126/science.abb2507

38. Xu C, Wang Y, Liu C, et al (2021) Conformational dynamics of SARS-CoV-2 trimeric spike glycoprotein in complex with receptor ACE2 revealed by cryo-EM. Sci Adv 7:eabe5575.

39. Maher MC, Bartha I, Weaver S, et al (2022) Predicting the mutational drivers of future SARS-CoV-2 variants of concern. Sci Transl Med 14:eabk3445. doi: 10.1126/scitranslmed.abk3445

40. Kulshreshtha S, Chaudhary V, Goswami GK, Mathur N (2016) Computational approaches for predicting mutant protein stability. J Comput Aided Mol Des 30:401–412.

41. Montanucci L, Savojardo C, Martelli PL, et al (2019) On the biases in predictions of protein stability changes upon variations: the INPS test case. Bioinformatics 35:2525–2527.

42. Pucci F, Bernaerts K V., Kwasigroch JM, Rooman M (2018) Quantification of biases in predictions of protein stability changes upon mutations. Bioinformatics 34:3659–3665. doi: 10.1093/bioinformatics/bty348

43. Potapov V, Cohen M, Schreiber G (2009) Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein Eng Des Sel 22:553–560. doi: 10.1093/protein/gzp030

44. Worth CL, Preissner R, Blundell TL (2011) SDM—a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Res 39:W215–W222. doi: 10.1093/nar/gkr363

45. Pires DEV, Ascher DB, Blundell TL (2014) MCSM: Predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30:335–342. doi: 10.1093/bioinformatics/bt491

46. Dehouck Y, Grosfils A, Folch B, et al (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics 25:2537–2543. doi: 10.1093/bioinformatics/btp445

47. Capriotti E, Fariselli P, Casadio R (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33:W306–W310.

48. Christensen NJ, Kepp KP (2012) Accurate stabilities of laccase mutants predicted with a modified FoldX protocol. J Chem Inf Model 52:3028–3042. doi: 10.1021/ci300398z

49. Kepp KP (2014) Computing stability effects of mutations in human superoxide dismutase 1. J Phys Chem B 118:1799–1812. doi: 10.1021/jp4119138

50. Kepp KP (2015) Towards a “Golden Standard” for computing globin stability: Stability and structure sensitivity of myoglobin mutants. Biochim Biophys Acta - Proteins Proteomics 1854:1239–1248. doi: 10.1016/j.bbapap.2015.06.002
51. Thiltgen G, Goldstein RA (2012) Assessing predictors of changes in protein stability upon mutation using self-consistency. PLoS One 7:e46084. doi: 10.1371/journal.pone.0046084

52. Caldararu O, Mehra R, Blundell TL, Kepp KP (2020) Systematic Investigation of the Data Set Dependency of Protein Stability Predictors. J Chem Inf Model 60:4772–4784. doi: 10.1021/acs.jcim.0c00591

53. Starr TN, Greaney AJ, Hilton SK, et al (2020) Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell 182:1295–1310.

54. Cao H, Wang J, He L, et al (2019) DeepDDG: predicting the stability change of protein point mutations using neural networks. J Chem Inf Model 59:1508–1514.

55. Baek KT, Kepp KP Data set and fitting dependencies when estimating protein mutant stability: Toward simple, balanced, and interpretable models. J Comput Chem. doi: https://doi.org/10.1002/jcc.26810

56. Caldararu O, Blundell TL, Kepp KP (2021) A base measure of precision for protein stability predictors: structural sensitivity. BMC Bioinformatics 22:88. doi: 10.1016/s12859-021-04030-w

57. Herrera NG, Morano NC, Celikgil A, et al (2021) Characterization of the SARS-CoV-2 S Protein: Biophysical, Biochemical, Structural, and Antigenic Analysis. ACS Omega 6:85–102. doi: 10.1021/acsomega.0c03512

58. McCallum M, Walls AC, Bowen JE, et al (2020) Structure-guided covalent stabilization of coronavirus spike glycoprotein trimers in the closed conformation. Nat Struct Mol Biol 27:942–949. doi: 10.1038/s41594-020-0483-8

59. Huo J, Zhao Y, Ren J, et al (2020) Neutralization of SARS-CoV-2 by Destruction of the Prefusion Spike. Cell Host Microbe 28:445-454.e6. doi: 10.1016/j.chom.2020.06.010

60. Toelzer C, Gupta K, Yadav SKN, et al (2020) Free fatty acid binding pocket in the locked structure of SARS-CoV-2 spike protein. Science (80-) 370:725–730. doi: 10.1126/science.abd3255

61. Lv Z, Deng Y-Q, Ye Q, et al (2020) Structural basis for neutralization of SARS-CoV-2 and SARS-CoV by a potent therapeutic antibody. Science (80-) 369:1505–1509. doi: 10.1126/science.abc5881

62. Zhang C, Wang Y, Zhu Y, et al (2021) Development and structural basis of a two-MAb cocktail for treating SARS-CoV-2 infections. Nat Commun 12:1–16.

63. Mitternacht S (2016) FreeSASA: An open source C library for solvent accessible surface area calculations. F1000Research 5:189.

64. Meents A, Gutmann S, Wagner A, Schulze-Briese C (2010) Origin and temperature dependence of radiation damage in biological samples at cryogenic temperatures. Proc Natl Acad Sci 107:1094–1099.

65. Mehra R, Dehury B, Kepp KP (2020) Cryo-temperature effects on membrane protein structure and dynamics. Phys Chem Chem Phys 22:5427–5438.

66. Linden AH, Franks WT, Akbey Ü, et al (2011) Cryogenic temperature effects and resolution upon slow cooling of protein preparations in solid state NMR. J Biomol
NMR 51:283–292. doi: 10.1007/s10858-011-9535-z

67. Tilton Jr RF, Dewan JC, Petsko GA (1992) Effects of temperature on protein structure and dynamics: X-ray crystallographic studies of the protein ribonuclease-A at nine different temperatures from 98 to 320K. Biochemistry 31:2469–2481.

68. Doster W, Bachleitner A, Dunau R, et al (1986) Thermal properties of water in myoglobin crystals and solutions at subzero temperatures. Biophys J 50:213–219. doi: 10.1016/S0006-3495(86)83455-5

69. Dunlop KV, Irvin RT, Hazes B (2005) Pros and cons of cryocrystallography: should we also collect a room-temperature data set? Acta Crystallogr Sect D Biol Crystallogr 61:80–87.

70. Edwards RJ, Mansouri K, Stalls V, et al (2021) Cold sensitivity of the SARS-CoV-2 spike ectodomain. Nat Struct Mol Biol 28:128–131.

71. Tokuriki N, Stricher F, Schymkowitz J, et al (2007) The stability effects of protein mutations appear to be universally distributed. J Mol Biol 369:1318–1332. doi: 10.1016/j.jmb.2007.03.069

72. Kandeel M, Ibrahim A, Fayez M, Al-Nazawi M (2020) From SARS and MERS CoVs to SARS-CoV-2: Moving toward more biased codon usage in viral structural and nonstructural genes. J Med Virol 92:660–666. doi: https://doi.org/10.1002/jmv.25754

73. Hanson G, Coller J (2018) Codon optimality, bias and usage in translation and mRNA decay. Nat Rev Mol cell Biol 19:20–30.

74. Pucci F, Schwersensky M, Rooman M (2022) Artificial intelligence challenges for predicting the impact of mutations on protein stability. Curr Opin Struct Biol 72:161–168. doi: https://doi.org/10.1016/j.sbi.2021.11.001

75. Ramanathan M, Ferguson ID, Miao W, Khavari PA (2021) SARS-CoV-2 B.1.1.7 and B. 1.351 Spike variants bind human ACE2 with increased affinity. Lancet Infect Dis 21:1070.

76. Mannar D, Saville JW, Zhu X, et al (2022) SARS-CoV-2 Omicron variant: Antibody evasion and cryo-EM structure of spike protein\&x2013;ACE2 complex. Science (80-) 375:760–764. doi: 10.1126/science.abn7760

77. Zeldovich KB, Chen P, Shakhnovich EI (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. Proc Natl Acad Sci U S A 104:16152–16157. doi: 10.1073/pnas.0705366104

78. Dasmeh P, Serohijos AWR, Kepp KP, Shakhnovich EI (2014) The Influence of Selection for Protein Stability on dN/dS Estimations. Genome Biol Evol 6:2956–2967. doi: 10.1093/gbe/evu223

79. Greaney AJ, Starr TN, Gilchuk P, et al (2021) Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. Cell Host Microbe 29:44–57.

80. Starr TN, Greaney AJ, Dingens AS, Bloom JD (2021) Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016. Cell Reports Med 2:100255.

81. Reetz MT (2013) The importance of additive and non-additive mutational effects in protein engineering. Angew Chemie Int Ed 52:2658–2666.
82. Phillips PC (2008) Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet 9:855–867.

83. Breen MS, Kemena C, Vlasov PK, et al (2012) Epistasis as the primary factor in molecular evolution. Nature 490:535–538.

84. Hopf TA, Ingraham JB, Poelwijk FJ, et al (2017) Mutation effects predicted from sequence co-variation. Nat Biotechnol 35:128–135. doi: 10.1038/nbt.3769

85. Rochman ND, Faure G, Wolf YI, et al (2022) Epistasis at the SARS-CoV-2 Receptor-Binding Domain Interface and the Propitiously Boring Implications for Vaccine Escape. MBio in press:e00135-22. doi: 10.1128/mbio.00135-22
SUPPLEMENTARY INFORMATION

Stability and expression of SARS-CoV-2 Spike-Protein Mutations

Kristoffer T. Bæk¹, Rukmankesh Mehra², and Kasper P. Kepp¹*

¹ DTU Chemistry, Technical University of Denmark, Building 206, 2800, Kongens Lyngby, Denmark. E-mail: * kpi@kemi.dtu.dk ORCID: 0000-0002-6754-7348.

² Department of Chemistry, Indian Institute of Technology Bhilai, Sejbahar, Raipur-492015, Chhattisgarh, India. E-mail: rukmankesh@iitbilai.ac.in

* Correspondence: E-mail: kpi@kemi.dtu.dk Tel: +45 45252409

ORCID

Kristoffer T. Bæk: 0000-0003-3243-3962

Rukmankesh Mehra: 0000-0001-6010-1514

Kasper P. Kepp: 0000-0002-6754-7348
Figure S1. Correlation in effects on RBD expression and binding to ACE2 by two independent mutant libraries. (a) The original dataset from Starr et al.\(^1\) containing 4221 data points. (b) The dataset was curated by removing outliers (absolute residuals > 1), removing data points with binding less than -4.5 in one or both libraries, and removing stop codon mutations, resulting in a dataset of 3536 data points. The orange lines are the linear regression lines for the scatter plots.
Figure S2. Distribution of mutations by effects on RBD expression and binding to ACE2.

The distribution of mutations is shown for (a) the original dataset, and (b) the curated dataset. Color indicates the type of mutation. The average of the two libraries is used.
Figure S3. Relationship between mutation effect on RBD expression and binding to ACE2. (a) Each point represents one mutation’s effect on expression and binding. (b) Data points were binned according to the effect on expression in 0.25-wide bins, and points indicate the mean binding effect for each bin as a function of the middle expression effect in each bin. Vertical bars show the standard deviation in each bin. Orange line indicates the linear regression, and R is the correlation coefficient.
Figure S4. Predicted change in protein stability as a function of mutations effect on RBD expression – binned data (wide bins). Three different prediction methods (DeepDDG, mCSM, and SimBa-IB) were used to predict the change in protein stability for each mutation using eight different experimental structures of S-protein. Data points were binned according to effect on expression in 0.5-wide bins, and points show the mean ΔΔG for each bin as a function of middle expression effect in each bin. Vertical bars show the standard deviation in each bin. Orange lines indicate the resulting linear regression.
Figure S5. Predicted change in protein stability as a function of mutations effect on RBD expression – binned data (narrow bins). Three different prediction methods (DeepDDG, mCSM, and SimBa-IB) were used to predict the change in protein stability for each mutation using eight different experimental structures of S-protein. Data points were binned according to effect on expression in 0.25-wide bins, and points show the mean ΔΔG for each bin as a function of middle expression effect in each bin. Vertical bars show the standard deviation in each bin. Orange lines indicate the resulting linear regression.
Figure S6. Predicted change in protein stability as a function of mutations effect on binding to ACE2 – binned data (wide bins). Three different prediction methods (DeepDDG, mCSM, and SimBa-IB) were used to predict the change in protein stability for each mutation using eight different experimental structures of S-protein. Data points were binned according to effect on binding in 0.5-effect bins, and points show the mean ΔΔG for each bin as a function of middle binding effect in each bin. Vertical bars show the standard deviation in each bin. Orange lines indicate the resulting linear regression.
Figure S7. Predicted change in protein stability as a function of mutations effect on binding to ACE2 – binned data (narrow bins). Three different prediction methods (DeepDDG, mCSM, and SimBa-IB) were used to predict the change in protein stability for each mutation using eight different experimental structures of S-protein. Data points were binned according to effect on binding in 0.25-wide bins and points show the mean ΔΔG for each bin as a function of middle binding effect in each bin. Vertical bars show the standard deviation in each bin. Orange lines indicate the resulting linear regression.
Figure S8. Relationship between RSA and the mutations’ effect on RBD expression and binding to ACE2. FreeSASA was used to calculate RSA for each wild-type residue using eight different experimental structures of S-protein. The average RSA value for the eight structures is used. (a) Each point represents one mutation. (b) Data points were binned according to RSA values in bins of 0.25 width, and points show the mean effect on binding or expression for each bin as a function of the middle RSA value in each bin. Vertical bars show the standard deviation in each bin. Orange lines indicate the resulting linear regression.
Figure S9: Change in stability as a function of relative solvent accessibility (RSA). Three different prediction methods (DeepDDG, mCSM, and SimBa-IB) were used to predict the change in protein stability for each mutation using the average of eight different experimental structures of S-protein. Each point represents one mutation.
References

(1) Starr, T. N.; Greaney, A. J.; Hilton, S. K.; Ellis, D.; Crawford, K. H. D.; Dingens, A. S.; Navarro, M. J.; Bowen, J. E.; Tortorici, M. A.; Walls, A. C. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. Cell 2020, 182 (5), 1295–1310.