Benchmarking the Widely Used Structure-based Binding Affinity Predictors on the Spike-ACE2 Deep Mutational Interaction Set

Burcu Ozden[1,2*], Eda Şamiloğlu[1,2*], Mehdi Koşaca[1,2], Melis Oktayoğlu[1], Can Yükrük[1], Mehmet Erguven[1], Nazmiye Arslan[1], Gökhan Karakülah[1,2], Ayşe Berçin Barlas[1,2], Büşra Savaş[1,2], Ezgi Karaca[1,2,+]

[1]Izmir Biomedicine and Genome Center, Dokuz Eylul University Health Campus, Balçova, Izmir 35330, Turkey

[2]Izmir International Biomedicine and Genome Institute, Dokuz Eylül University, Izmir 35340, Turkey

[*]Equal contribution

[+]Correspondence: ezgi.karaca@ibg.edu.tr

ABSTRACT

Since the start of COVID-19 pandemic, a huge effort has been devoted to understanding the Spike(SARS-CoV-2)-ACE2 recognition mechanism. As prominent examples, two deep mutational scanning studies traced the impact of all possible mutations/variants across the Spike-ACE2 interface. Expanding on this, we benchmark four widely used structure-based binding affinity predictors (FoldX, EvoEF1, MutaBind2, SSIPe) and two naïve predictors (HADDOCK, UEP) on the variant Spike-ACE2 deep mutational interaction set. Among these approaches, FoldX ranks first with a 64% success rate, followed by EvoEF1 with a 57% accuracy. Upon performing residue-based analyses, we reveal algorithmic biases, especially in ranking mutations with increasing/decreasing hydrophobicity/volume. We also show that the approaches using evolutionary-based terms in their scoring functions misclassify most mutations as binding depleting. These observations suggest plenty of room to improve the conventional affinity predictors for guessing the variant-induced binding profile changes of Spike-ACE2. To aid the improvement of the available approaches we provide our benchmarking data at https://github.com/CSB-KaracaLab/RBD-ACE2-MutBench

Word count: 155

Key words: binding affinity prediction, deep mutagenesis, FoldX, EvoEF1, SARS-CoV-2, point mutation, ACE2

INTRODUCTION

At the beginning of the 21st century, the Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV)[1] and the Middle East Respiratory Syndrome Coronavirus[2] caused serious public health concerns. During late 2019, a new SARS virus, SARS-CoV-2, has led to the most severe pandemic of the 21st century[3]. Since then, an enormous amount of effort has been devoted to dissecting the SARS-CoV-2 infection cycle. SARS-CoV-2 infection is initiated upon having its Spike protein interacting with the host Angiotensin Converting 2 (ACE2) enzyme[4]. The widespread infection of SARS-CoV-2 has been linked to higher binding affinity of Spike to ACE2[5]. Alpha, beta, gamma, eta, iota, kappa, lambda, mu, omicron variants have shown to have at least one mutation across Spike-ACE2 interface[6]. This realization has placed the characterization of interfacial Spike-ACE2 mutations at the center of COVID-19-related research. Within this context, in 2020, two deep mutational scanning (DMS) studies explored how Spike/ACE2 variants impact Spike-ACE2 interactions[7,8]. In these DMS studies, the residues on the Receptor Binding Domain (RBD) of Spike and the catalytic domain of human ACE2 were mutated to the other 19 amino acid possibilities, followed by tracing of the new RBD-ACE2 binding profiles.

In addition to these experimental efforts, several *in silico* studies investigate the impact of the variation on RBD-ACE2 interface (Table 1). As an example, Laurini *et al.* performed molecular mechanics/Poisson–Boltzmann surface area, computational alanine scanning mutagenesis, and interaction entropy calculations to find the critical RBD-ACE2 interactions[9]. As a result, they propose several hot spot residues on RBD and ACE2, where Q498F/H/W/Y on RBD and K31F/W/Y, Y41R on ACE2 are reported as affinity enhancers in the DMS sets. In parallel, Blanco *et al.* used FoldX with the inclusion of water molecules (FoldXwater) to trace the binding enhancing RBD and ACE2 mutations[10]. From their predictions, in the DMS sets, Q493F/L/M/Y, Q498F/Y, N501T, and V503R RBD mutations come up as affinity enhancing. On the ACE2 side, Rodrigues *et al.* investigated the impact of ACE2 orthologs on their RBD binding with HADDOCK, where they propose affinity improving ACE2 mutations[11]. Among these, D30E and A387E are profiled as affinity enhancing in the ACE2 DMS set. Complementary to this study, Sorokina *et al.* performed computational alanine scanning on ACE2 with HADDOCK[12]. Here, N49A, R393A, and P389A are classified as binding enriching in the ACE2 DMS set. Finally, Gheeraert *et al.* performed molecular dynamics simulations of five RBD variants (alpha, beta, gamma, delta, and epsilon) in complex with ACE2[13]. They find that mutations on the delta variant cause drastic changes across the RBD and ACE2 interface. These mutations are also classified as binding enriching in the RBD DMS dataset.

Table 1. Important RBD and ACE2 variants/hotspots according to the recent *in silico* studies. The predictions overlapping with the RBD and ACE2 DMS sets are underlined and highlighted in bold.

Work carried out by	Important RBD residues/mutations	Important ACE2 residues/ mutations
Laurini *et al.*[9]	**Q498**, T500, R403	D38, **K31**, E37, K353, **Y41**

Blanco *et al.*[10]	V445M/R/W, **Q493F/L/M/Y**, **Q498F**/L/M/**Y**, T500K, **N501**A/C/L/S/**T**, **V503R**/W/Y	G326E
Rodrigues *et al.*[11]		Q24E, **D30E**, H34Y, L79H, **A387E**
Sorokina *et al.*[12]		**N49A**, **R393A**, M383A, **P389A**, G354A
Gheeraert *et al.*[13]	**L452R**, **T478K**	

The prediction mismatches outlined in Table 1 portrays a certain level of misprediction for each prediction method, calling for a proper benchmarking of the variant-based affinity predictions. Expanding on this call, we benchmark four widely used structure-based binding affinity predictors (FoldX, EvoEF1, MutaBind2, SSIPe) and two naïve predictors (HADDOCK, UEP) on the variant Spike-ACE2 DMS sets[14–20]. Among these tools, FoldX and EvoEF1 use inter- and intra-molecular energies derived from empirical force field terms. Mutabind and SSPIe utilize FoldX and EvoEF1, respectively, to model the mutations that are scored with extra evolutionary-related terms. HADDOCK uses intermolecular van der Waals, electrostatics, and empirical desolvation terms, while UEP is based on statistically determined intermolecular contact potentials. As an outcome of our benchmarking efforts, we present the grounds of each method's success/failure. To further aid the improvement of the field, all our benchmarking files are deposited at https://github.com/CSB-KaracaLab/RBD-ACE2-MutBench with the visualization option at https://rbd-ace2-mutbench.github.io/[12].

RESULTS and DISCUSSION

Most of the enriching mutations tends to decrease the polarity of the interface

In ACE2 and RBD DMS sets, less than 15% of the mutations are located at the RBD-ACE2 interface[7,8]. From these interfacial mutations, we selected an equal number of *enriching* and *depleting* mutations to create an unbiased experimental set. As a result, we isolated 84 RBD (42 enriching, 42 depleting) and 179 ACE2 mutations (89 enriching, 90 depleting cases) (Figure 1, Table S1). In our final set, the most frequently appearing enriching mutation positions on RBD emanate from (in the decreasing frequency order) Q493, S477, F490, N501, V503, E484, Q498. Among these, Q493R, S477N, E484A are observed in omicron; E484K in beta, gamma, eta, iota, mu; E484Q in kappa; F490S in lambda, and N501Y in alpha, beta, gamma, mu, omicron variants[6]. On the ACE2 surface, the top enriching mutations come from T27, Q42, S19, and L79 positions (Figure 1A, Figure S1). All these residues, except S19, are reported as species-associated variations[21]. While appearing less frequently as binding enhancers, K31, E35, M82, and Y83 are earlier listed as critical residues for RBD-ACE2 interactions (Figure 1A, Figure S1)[9,10,22]. On the RBD and ACE2 surfaces, most of the enriching mutation positions come from polar residues, where the most impactful changes are observed for polar to non-polar mutations, especially for Q493 on RBD and T27, Q42 on ACE2 (Figure S2).

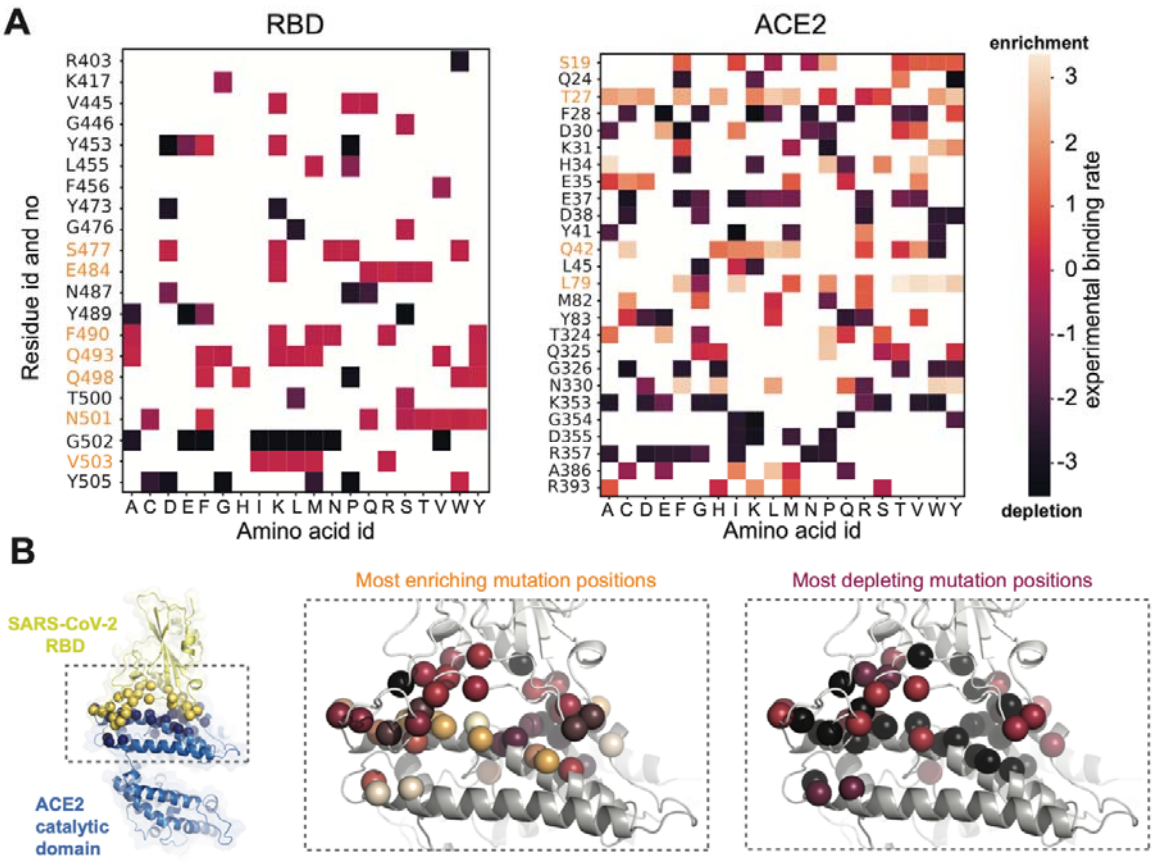

Figure 1. (A) The experimental enrichment and depletion binding rates on the RBD (left) and ACE2 (right). The values > 0 correspond to enhancing positions (light orange), where the values <0 are the

depleting ones (dark purple). The top enriching positions are highlighted in orange. **(B) The interactions between SARS-CoV-2 RBD (yellow) and ACE2 catalytic domain (blue)** (pdb id: 6M0J[20], in cartoon illustrated with PyMOL[23]. The interface residues investigated in this study are shown in spheres and colored according to the color scale given in panel A.

Predicting enriching mutations are more difficult than depleting ones

Here, we benchmark the prediction capacity of four widely used structure-based binding affinity predictors (FoldX, EvoEF1, MutaBind2, SSIPe) and two naïve predictors (HADDOCK, UEP) on the variant Spike-ACE2 DMS sets[14–20] (Table S1, Figure 2A, Figure S3, Supplementary Text). On the DMS set, the overall success rates of predictors vary between 54% and 64%, where the top-ranking predictor is FoldX (Figure 2A). Surprisingly, FoldXwater ranks the second, implying that the inclusion of water effects does not improve the prediction accuracy. When we score the lowest ranking predictor, HADDOCK's models with FoldX, HADDOCK's success rate increases by 5% (from 54% from 59%) (Figure S4A).

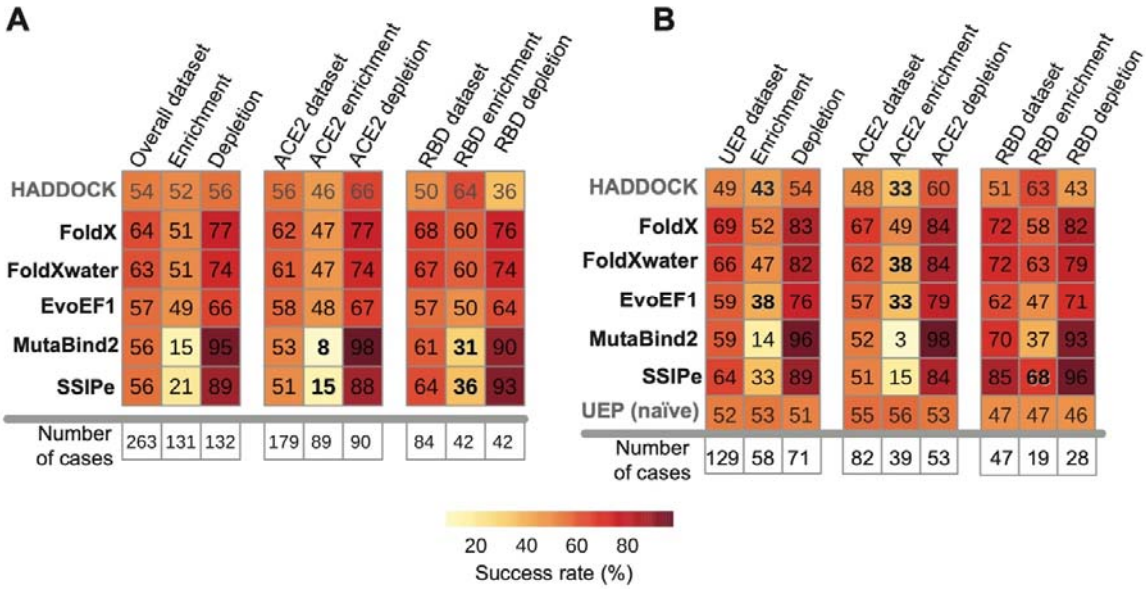

Figure 2. (A) The success rates on the curated DMS set. The naïve predictor HADDOCK success rates are shown in gray. The significantly low prediction rates are highlighted in bold. **(B) Success rates of predictors are calculated by using highly packed residues only.** If the success rate changes drastically according to overall dataset (left), it is represented in bold. The naïve predictors (HADDOCK and UEP) are highlighted in gray. The plots are prepared by using [24–27].

Seventy mutations are successfully predicted by all approaches (26.6% of all cases, composed of 31 enriching and 39 depleting cases (Table S3)). The approaches fail to classify 33 cases correctly (12.5% of the overall dataset), where most of them are enriching mutations (Figure 2A). When we analyze ACE2 and RBD subsets individually, better prediction rates for depleting mutations are consistently observed, highlighting that this outcome is independent of the depletion value ranges. Strikingly, MutaBind2 and SSIPe predicts most mutations as depleting, hinting at a problem in using evolutionary-based terms in scoring an host pathogen interaction like RBD-ACE2.

For benchmarking UEP, we use a subset of curated DMS set, as UEP can calculate the binding affinity changes of highly packed residues only (50% cases) (Table S2, Figure 2B). On

the UEP subset, the overall prediction performances vary within a broader range, i.e., 49%-69% (Figure 2B), where again the top two predictors become FoldX and FoldXwater (69% vs. 66%). The naïve predictors are the lowest ranking ones with 52% and 49% success rates for UEP and HADDOCK, respectively (Figure 2B). Even so, we see that all the predictors perform worse than the naïve predictor UEP for the ACE2 enrichment category. So, predicting packed enriching ACE2 mutations turns out to be a major challenge (Figure 2B).

The volume and hydrophobicity changes impact the prediction accuracy

To explore the prediction dependency on the type of mutations, we assess the prediction accuracies according to Van der Waals volume, hydrophobicity, flexibility, and physicochemical changes (Table S4, see Materials & Methods, Figure 3). As a result, we observe that HADDOCK has a volume bias, as it classifies most of the volume-increasing mutations as affinity enhancing (Figure 3A). We also discover that FoldX has a hydrophobicity bias, as it accurately predicts enriching mutations when the mutation leads to a decrease in the hydrophobicity (Figure 3A). Interestingly, we do not observe any bias toward flexibility and physicochemical changes (Figures 3C-D and S4B).

To quantify these biases, we calculated the difference between success rates of depleting and enhancing mutations (ΔSuccess, Figure S4B). ΔSuccess varies between -100 and 100, where 0 means no bias and 100 and -100 mean extreme biases. Since MutaBind2 and SSIPe predict almost all mutations as depleting, their ΔSuccess is extremely skewed for all metrics. According to this metric as well, FoldX and HADDOCK show moderate biases for hydrophobicity and volume changes with 34 and -31 ΔSuccess scores. Normalizing HADDOCK scores by buried surface area (BSA) of the interface does not alleviate this dependency, instead it introduces a strong enrichment bias (Figure S4A). When HADDOCK models are scored with FoldX, a moderate flexibility bias with -30 ΔSuccess score is observed.

When we study the similarities/differences of generated mutant models by calculating the all-atom Root Mean Square Deviations (RMSDs), we reveal that HADDOCK generates the most distinctive models compared to the others (Figure S5). Further work is needed to understand the dependency of HADDOCK's performance on the mutation modeling. Since SSIPe utilizes EvoEF1 to structurally model the mutations, EvoEF1 and SSIPe mutant models come out identical. MutaBind2 utilizes FoldX to generate the structural variation. However, as MutaBind2 employs further minimization on the mutant model generated, their models diverge from each other.

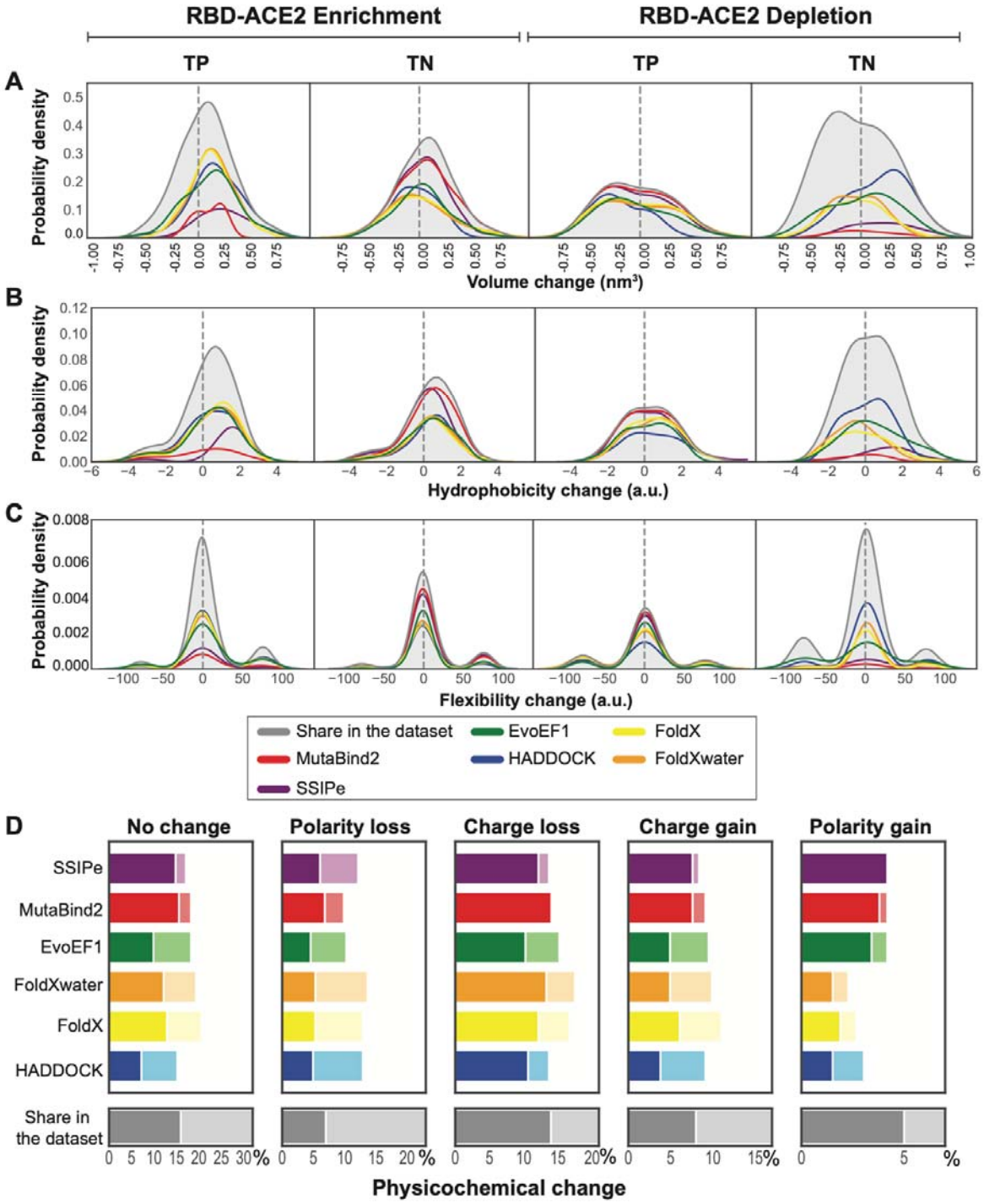

Figure 3. The effects of change in physical properties of amino acids upon a mutation on the success rate of predictors. (A) Volume change, **(B)** hydrophobicity change, **(C)** flexibility change, **(D)** physicochemical change in amino acids upon point mutations are depicted for the experimental dataset (gray), EvoEF1 (green), FoldX (yellow), MutaBind2 (red), HADDOCK (blue), FoldXwater (orange) and SSIP (purple). Dark and light colors represent depleting and enriching cases, respectively.

CONCLUSION

We present the first study benchmarking the commonly used structure-based binding affinity predictors in gauging the impact of Spike-ACE2 variations. As a result, we show that FoldX is the best performing method with moderate accuracy (64%). We also demonstrate that all predictors have difficulties in predicting binding-enhancing mutations. Some methods have biases towards mutations with increasing/decreasing hydrophobicity/volume. These imply that we should use these methods cautiously for drawing general conclusions in the absence of experimental data. However, we should also keep in mind that the DMS set we used contains a certain noise level that could not be reflected in our accuracy calculations. For this, further DMS-based benchmarking studies should be carried out. Finally, we hope that our work will aid the computational community for being prepared not only for combatting SARS-CoV-2-related health concerns, as well as other related infectious diseases.

MATERIALS and METHODS

Benchmark compilation

For benchmarking, we use deep mutational scanning coupled with interaction profiling data sets for Spike-RBD[8] and human ACE2[7]. In these sets, 201 residues of Spike-RBD and 117 residues of human ACE2 are mutated into other 19 amino acid possibilities. The RBD-ACE2 interface positions are calculated over 6m0j[22] with PDBePISA[29](Figure 1B):

Twenty-six interfacial Spike-RBD positions: R403, K417, V445, G446, Y449, Y453, L455, F456, Y473, A475, G476, S477, Q484, G485, F486, N487, Y489, F490, Q493, G496, Q498, T500, N501, G502, V503, and Y505.
Twenty-six interfacial ACE2 positions: S19, Q24, T27, F28, D30, K31, H34, E35, E37, D38, Y41, Q42, L45, L79, M82, Y83, T324, Q325, G326, N330, K353, G354, D355, R357, A386, R393.

Since most of the depleting cases are close to the neutral point, we selected half of the depleted interactions among the highly depleting ones, whereas the other half was randomly selected (by using the *default*_rng function of the NumPy package, with seed=123). The heatmaps in Figure 1A are generated with pandas, Numpy and Seaborn libraries of Python 3.8[30–35]. All the selected mutations show high expression rates.

Performance Evaluation

The predictions are evaluated from the perspectives of volume, hydrophobicity, flexibility, and physicochemical property change upon mutation ($\Delta Property_{change}$ = $Property_{mutation}$ - $Property_{wildtype}$, Table S4). Volume change is the Van der Waals (vdW) volume change[36]. Amino acid hydrophobicities are taken from Eisenberg *et al.*[37]. To measure flexibility change, we use the flexibility scale presented by Shapovalov & Dunbrack[38]. The physicochemical properties are considered as: polar amino acids - N, Q, S, T, Y; non-polar amino acids - A, G, I, L, M, F, P, W, V, C; charged amino acids - H, E, D, R, K. Success rate and metric evaluations are performed in Python 3.8.5 with Pandas, Numpy, seaborn, and Matplotlib libraries [30–35]. For each category, the percentage of successfully predicted cases are calculated by [Success rate = Correct-Predictions/All-Predictions*100].

DATA AVAILABILITY

All results including the codes and notebooks are deposited in Github (https://github.com/CSB-KaracaLab/RBD-ACE2-MutBench) and the models and the scores can be visualized at https://rbd-ace2-mutbench.github.io/ .

SUPPLEMENTARY DATA

Supplementary Data are submitted with the manuscript.

ACKNOWLEDGEMENTS

All the simulations and analyses were carried out in the HPC resources of Izmir Biomedicine and Genome Center. The authors would like to thank João P. G. L. M. Rodrigues for the critical reading of our manuscript.

FUNDING

This work was supported by EMBO Installation Grant (no. 4421), Young Scientist Award granted by the Turkish Science Academy, and TÜSEB Research Grant (no. 3933).

AUTHOR CONTRIBUTIONS

E. K. conceptualized the study and wrote the manuscript. B. O. and E. Ş. performed data analysis, prepared Github page and figures, and wrote the manuscript. M. K., M. O. and M. E. performed analysis and prepared figures. A.B.B., B.S., and C.Y. performed data analysis. N. A. and G. K. prepared our visualization page on Github.

CONFLICT OF INTEREST

The authors declare no competing interests.

REFERENCES

1. LeDuc, J. W. & Barry, M. A. SARS, the First Pandemic of the 21st Century. *Emerg. Infect. Dis.* **10**, e26 (2004).
2. Durai, P., Batool, M., Shah, M. & Choi, S. Middle East respiratory syndrome coronavirus: transmission, virology and therapeutic targeting to aid in outbreak control. *Exp. Mol. Med.* **47**, e181 (2015).
3. Platto, S., Xue, T. & Carafoli, E. COVID19: an announced pandemic. doi:10.1038/s41419-020-02995-9
4. Li, W. *et al.* Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* **426**, 450 (2003).
5. Ali, A. & Vijayan, R. Dynamics of the ACE2–SARS-CoV-2/SARS-CoV spike protein interface reveal unique mechanisms. *Sci. Rep.* **10**, (2020).
6. CoVariants. Available at: https://covariants.org/. (Accessed: 27th June 2022)
7. Chan, K. K. *et al.* Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2. *Science (80-.).* **369**, (2020).
8. Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295-1310.e20 (2020).
9. Laurini, E., Marson, D., Aulic, S., Fermeglia, M. & Pricl, S. Computational Alanine Scanning and Structural Analysis of the SARS-CoV-2 Spike Protein/Angiotensin-Converting Enzyme 2 Complex. *ACS Nano* **14**, (2020).
10. Blanco, J. D., Hernandez-Alias, X., Cianferoni, D. & Serrano, L. In silico mutagenesis of human ACE2 with S protein and translational efficiency explain SARS-CoV-2 infectivity in different species. *PLoS Comput. Biol.* **16**, (2020).
11. Rodrigues, J. P. G. L. M. *et al.* Insights on cross-species transmission of SARS-CoV-2 from structural modeling. *PLoS Comput. Biol.* **16**, (2020).
12. Sorokina, M. *et al.* Structural models of human ACE2 variants with SARS-CoV-2 Spike protein for structure-based drug design. *Sci. Data* **7**, 1–10 (2020).
13. Gheeraert, A. *et al.* Singular Interface Dynamics of the SARS-CoV-2 Delta Variant Explained with Contact Perturbation Analysis. *J. Chem. Inf. Model.* **62**, 3107–3122 (2022).
14. Schymkowitz, J. *et al.* The FoldX web server: An online force field. *Nucleic Acids Res.* **33**, (2005).
15. Pearce, R., Huang, X., Setiawan, D. & Zhang, Y. EvoDesign: Designing Protein–Protein Binding Interactions Using Evolutionary Interface Profiles in Conjunction with an Optimized Physical Energy Function. *J. Mol. Biol.* **431**, (2019).
16. Zhang, N. *et al.* MutaBind2: Predicting the Impacts of Single and Multiple Mutations on Protein-Protein Interactions. *iScience* **23**, (2020).
17. Huang, X., Zheng, W., Pearce, R., Zhang, Y. & Zhang, Y. SSIPe: Accurately estimating protein-protein binding affinity change upon mutations using evolutionary profiles in combination with an optimized physical energy function. *Bioinformatics* **36**, (2020).
18. Van Zundert, G. C. P. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* (2016). doi:10.1016/j.jmb.2015.09.014
19. Honorato, R. V. *et al.* Structural Biology in the Clouds: The WeNMR-EOSC Ecosystem. *Front. Mol. Biosci.* **8**, (2021).
20. Amengual-Rigo, P., Fernández-Recio, J. & Guallar, V. UEP: an open-source and fast

classifier for predicting the impact of mutations in protein-protein complexes. *Bioinformatics* **37**, (2021).

21. Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. *J. Virol.* **94**, (2020).
22. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, (2020).
23. Schrödinger, L. L. C. and D. W. PyMOL.
24. Integrated Development Environment for R. RStudio.
25. A Language and Environment for Statistical Computing. R.
26. Kolde, R. Pheatmap: Pretty Heatmaps. (2019).
27. Neuwirth, E. RColorBrewer: ColorBrewer Palettes. (2014).
28. Sorokina, M. *et al.* An Electrostatically-steered Conformational Selection Mechanism Promotes SARS-CoV-2 Spike Protein Variation. *J. Mol. Biol.* **434**, 167637 (2022).
29. Krissinel, E. & Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **372**, (2007).
30. Van Rossum, G. *et al. Python 3 Reference Manual*. *Nature* **585**, (2009).
31. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, (2020).
32. Kluyver, T. *et al.* Jupyter Notebooks—a publishing format for reproducible computational workflows. in *Positioning and Power in Academic Publishing: Players, Agents and Agendas - Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016* (2016). doi:10.3233/978-1-61499-649-1-87
33. Anaconda Software Distribution. *Anaconda Documentation* (2020).
34. The pandas development team. Pandas-dev/pandas: Pandas. *Zenodo* (2020).
35. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, (2021).
36. Lin, Z. hua, Long, H. xia, Bo, Z., Wang, Y. qiang & Wu, Y. zhang. New descriptors of amino acids and their application to peptide QSAR study. *Peptides* **29**, (2008).
37. Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* **179**, (1984).
38. Shapovalov, M. V. & Dunbrack, R. L. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **19**, (2011).