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

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], Burcu Ozden[1,2*], Ezgi Karaca[1,2,3*]

[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

[3]Lead contact: ezgi.karaca@ibg.edu.tr

[*]Corresponding authors: Ezgi Karaca, Burcu Ozden

ABSTRACT

Since the start of COVID-19 pandemic, a huge effort has been devoted to understanding the Spike-ACE2 recognition mechanism. To this end, two recent deep mutational scanning studies traced the impact of all possible mutations/variants across the Spike-ACE2 interface. Expanding on these studies, we benchmarked 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 ranked first with a 64% success rate, followed by EvoEF1 with a 57% accuracy. Upon performing residue-based analyses, we revealed critical algorithmic biases, especially in ranking mutations with increasing/decreasing hydrophobicity/volume. We also showed that the approaches using evolutionary-based terms in their affinity predictions classify most mutations as affinity depleting. These observations suggest plenty of room to improve the conventional affinity predictors for predicting the binding affinity change of the viral host-pathogen system SARS-CoV-2-ACE2. To aid the improvement of the available approaches, we provide our mutant models, together with our benchmarking data at https://github.com/CSB-KaracaLab/ace2-rbd-point-mutation-benchmark

Word count: 166

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

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 increasing public health concerns. Following these, the emergence of a new SARS virus 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 earlier linked to the increased electrostatic forces across the Spike-ACE2 interface (5). Relatedly, alpha, beta, gamma, eta, iota, kappa, lambda, mu, omicron variants have shown to have at least one mutation across Spike-ACE2 interface (specifically, K417N/T, G446S, S477N, E484K/Q/A, F490S, Q493R, G496S, Q498R, N501Y, and Y505H variants) (6). This realization has placed the characterization of interfacial Spike-ACE2 mutations at the center of COVID-19-related research (7–12). Within this context, in 2020, two deep mutational scanning studies explored how Spike/ACE2 variants impact Spike-ACE2 interactions (7, 8). In these 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 the tracing of new RBD-ACE2 binding profiles.

In addition to these experimental efforts, four important *in silico* studies investigated the impact of the variation across RBD-ACE2 interface (9–12). Among these, 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 proposed Q498, T500, R403 RBD residues and D38, K31, E37, K353, Y41 ACE2 residues as hot spots. Q498F/H/W/Y on RBD and K31F/W/Y, Y41R on ACE2 were also reported as affinity enhancers in the deep mutational scanning datasets (7, 8). In parallel, Blanco *et al.* used FoldX with the inclusion of water molecules (FoldXwater) to trace the integral RBD and ACE2 variants for binding (10). They found V445M/R/W, Q493F/L/M/Y, Q498F/L/M/Y, T500K, N501A/C/L/S/T, V503R/W/Y as RBD binding enhancers and G326E as ACE2 binding enhancer. When these mutations were compared with those profiled in the deep mutational datasets, Q493F/L/M/Y, Q498F/Y, N501T, and V503R RBD mutations came up as affinity enhancing, and G326E ACE2 mutation as affinity depleting (7, 8). On the ACE2 side, Rodrigues *et al.* investigated the impact of ACE2 orthologs on their RBD binding with HADDOCK (11). Their work proposed Q24E, D30E, H34Y, L79H, A387E ACE2 mutations as affinity enhancing. Among these, D30E and A387E were also classified as affinity enhancing in the ACE2 deep mutational dataset (8). Finally, Sorokina *et al.* performed computational alanine scanning on ACE2 with HADDOCK (12). When their alanine scanning dataset is ranked according to their HADDOCK scores, N49A, R393A, M383A, P389A, and G354A came out as the most impactful mutations. N49A, R393A, and P389A were also classified as enriching in the ACE2 deep mutational dataset (8).

These timely *in silico* computational studies shed an important light on the critical RBD-ACE2 interactions. However, as they have never been benchmarked on a complete experimental set, their accuracy could never be properly explored. Expanding on this, we benchmarked the prediction capacity of the widely used structure-based binding affinity predictors (FoldX, EvoEF1, SSIPe, MutaBind2 (as main), and HADDOCK, UEP (naïve) (13–18)) on ACE2 and RBD deep mutational scanning datasets. We analyzed the performance of

these predictors from several aspects to reveal the grounds of their successes/failures. All our results, mutant models together, with our benchmarking analyses are provided at https://github.com/CSB-KaracaLab/ace2-rbd-point-mutation-benchmark .

RESULTS and DISCUSSION

Our curated experimental set contains an equal amount of *enriching* and *depleting* interfacial RBD-ACE2 mutations, totaling 263 variants (Figure 1, Table S1). In the dataset, we have twice more ACE2 mutations (179) than the RBD ones (84) (Figure 1B, Materials and Methods). On the RBD surface, the most frequently appearing enriching mutations stem from (in the decreasing frequency order) Q493, S477, F490, N501, V503, E483, Q498 positions. Among these, Q493R, S477N, E484A were 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 came from positions T27, Q42, S19, and L79 (Figure 1B, Figure S1). All these residues, except S19, were reported as species-associated variations (19). While less frequently appearing as binding enhancers, E35, M82, Y83, and K31 were listed as critical residues for RBD-ACE2 interactions (Figure 1B, Figure S1) (9, 10, 20). On the RBD and ACE2 surfaces, most of the enriching mutation positions emanate from polar amino acids, where the most impactful changes were observed for polar to non-polar mutations, especially for Q493 on RBD and T27, Q42 on ACE2 (Figure S2).

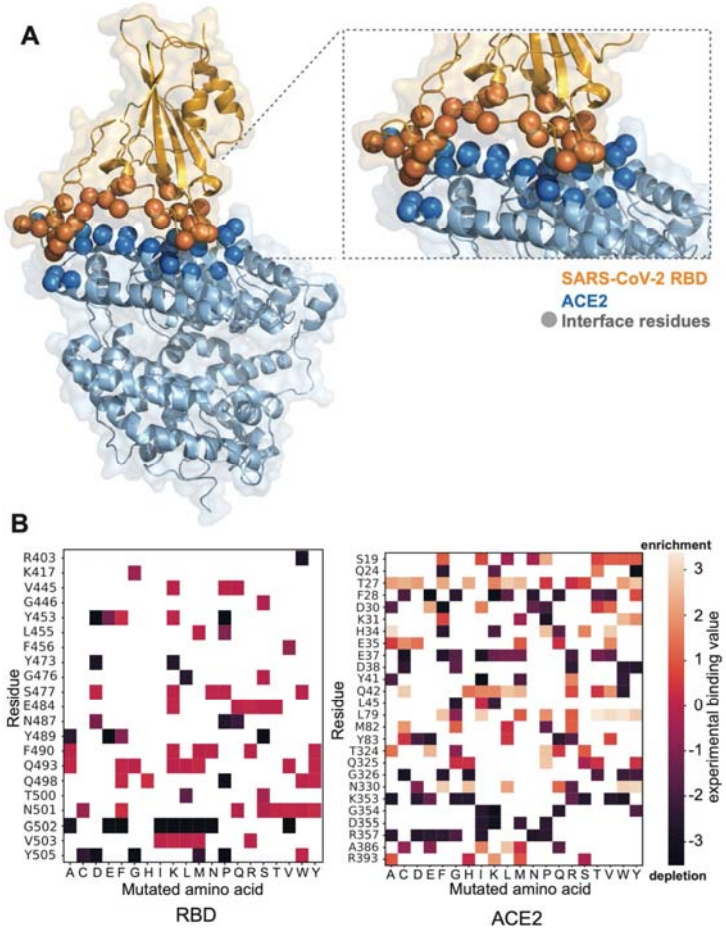

Figure 1. (A) The interactions between RBD (orange) and ACE2 (blue) (pdb id: 6M0J, (21), in cartoon illustrated with PyMOL(22)). The 52 residues investigated in this study are shown in spheres. **(B) The experimental enrichment and depletion values on the RBD (left) and ACE2 (right).** The values > 0

correspond to enhancing positions (light orange), where the values <0 are the depleting ones (dark purple).

Over this curated experimental set, we benchmarked the prediction capacity of four widely used binding affinity predictors, i.e., FoldX (and FoldXwater), EvoEF1, MutaBind2, and SSIPe (Table S1, Figure 2A, Figure S3). As naïve predictors, we used HADDOCK and UEP, as they contain simple physics-based and contact-based terms, respectively (Materials and Methods). For benchmarking of UEP, we used a subset of the curated experimental set, as UEP can calculate the binding affinity change of highly packed residues only (129/263 cases) (Table S2, Figure 2B). FoldXwater was benchmarked too since the inclusion of water into FoldX was shown to improve the binding affinity predictions (23). On the curated experimental set, the overall success rates of predictors vary between 54% and 64%. The top-ranking predictor is FoldX, followed by EvoEF1 (Figure 2A). We did not observe any significant performance improvement when we used FoldXwater. All predictors could predict the *depleting* mutations more accurately than the *enriching* ones. This fact is significantly pronounced in the case of MutaBind2 and SSIPe, as they classified most cases as depleting (Figure 2A). The same tendency was observed when we analyzed RBD and ACE2 subsets. So, irrespective of the source of mutation, the inclusion of evolutionary-based terms introduced a strong bias towards classifying the mutations as depleting. This should stem from having RBD bind to an evolutionary non-conserved ACE2 surface (Figure S4) (19). In general, 70 mutations could be successfully predicted by all approaches (26.6% of all cases, composed of 31 enriching and 39 depleting cases (Table S3)). All predictors failed to successfully classify 33 cases (12.5% of the overall dataset), where 79% of all-failed cases were composed of enriching mutations.

On the UEP subset, the overall prediction performances vary within a broader range, i.e., 49%-69% (Figure 2B). The top predictor becomes FoldX with a slightly reduced accuracy for FoldXwater (69% vs. 66%). The lowest ranking predictors are UEP and HADDOCK, with 52% and 49% success rates, respectively (Figure 2B). When we focused on the *enriching* and *depleting* mutations, the accuracy of EvoEF1 and HADDOCK on enriching cases decreased to 38% and 43%, respectively. Strikingly, on the enriching cases, all the predictors perform worse than the naïve predictor UEP. It also seems that predicting the packed enriching mutations for ACE2 was more challenging than the ones for RBD (Figure 2B).

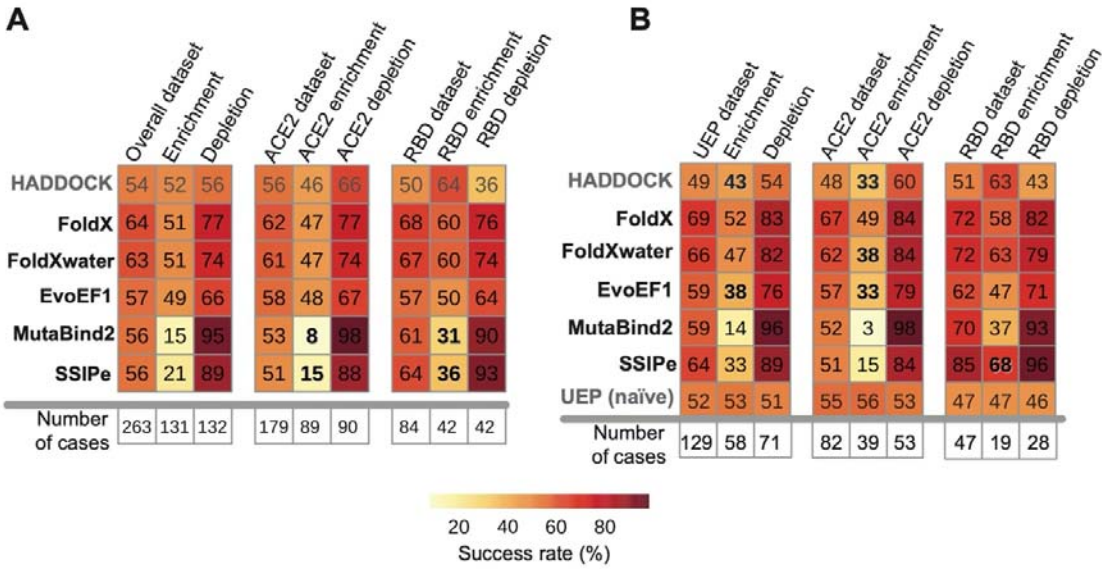

Figure 2. (A) The success rates on the curated experimental set. The naïve predictor HADDOCK success rates are shown in gray. The significantly low prediction rates are highlighted in bold. **(B) Success rates of predictors are calculated by using highly packed residues only.** If the success rate changes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

The capabilities of predictors were assessed by the properties of mutant and wild-type amino acids. Here, we considered Van der Waals volume, hydrophobicity, flexibility, and physicochemical property changes (Table S4, see Materials & Methods, Figure 3). As a result, we revealed a volume bias for HADDOCK, since it classified most of the volume-increasing mutations as affinity enhancing vice versa (Figure 3A). We also found that FoldX has hydrophobicity bias, as it accurately predicted enriching mutations when the mutation led to a decrease in hydrophobicity (Figure 3A). To quantify these biases, we calculated the difference between success rates of depleting and enhancing mutations (ΔSuccess, Figure S5). ΔSuccess varies between -100 and 100, where 0 means no bias and 100 and -100 mean extreme bias. Since MutaBind2 and SSIPe predict almost all mutations as depleting, their ΔSuccess became extremely skewed for all metrics. FoldX and HADDOCK turned out to have moderate biases for hydrophobicity and volume change metrics with 34 and -31 ΔSuccess scores. Interestingly, we could not any bias toward flexibility and physicochemical property change (Figures 3C-D and S5). Finally, we studied the similarities/differences of generated mutant models by calculating the all-atom Root Mean Square Deviations (RMSDs) over all conformers generated (Figure S6). This analysis showed that HADDOCK generates the most distinctive models compared to the others. 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 came 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.

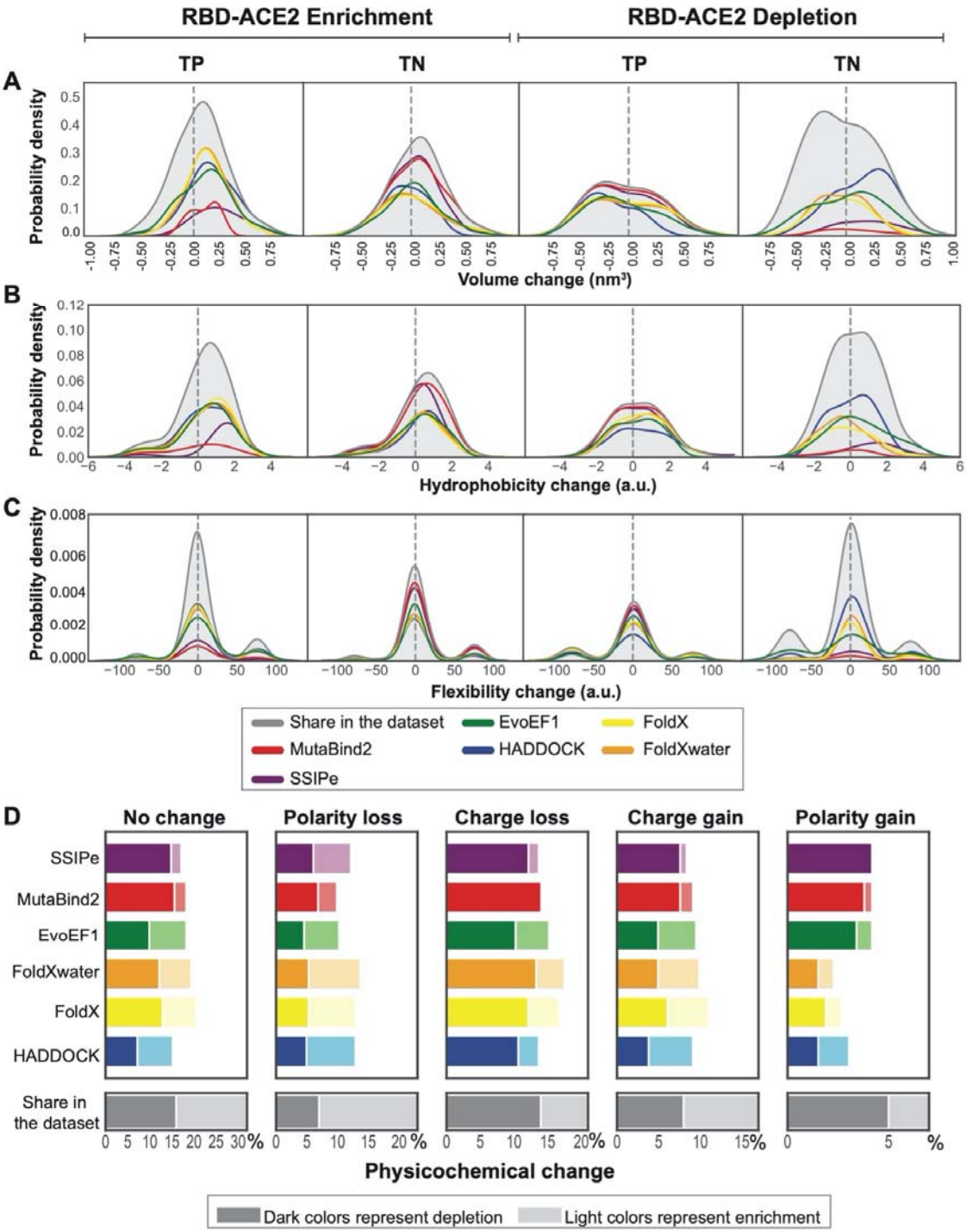

Figure 3. The effects of change in physical properties of amino acids upon a mutation on the success rate of predictors. (A) Volume change, **(B)** hydrophobicity change, **(C)** flexibility change, **(D)** physicochemical change in amino acids upon point mutations are depicted for the experimental dataset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

CONCLUSION

In this work, we benchmarked 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. As a result, we showed that these approaches have certain biases in predicting the impact of RBD-ACE2 variations. To be fair, we should note here that the experimental dataset used in this study should also contain a certain level of noise, which we could not incorporate into our performance assessment. For example, when we checked the known impactful RBD variants (K417N/T, G446S, S477N, E484K/Q/A, F490S, Q493R, G496S, Q498R, N501Y, and Y505H) in the deep mutational scanning dataset of Starr *et al.* we found that only three were classified as affinity enhancing (E484K, E484Q, N501Y). So, further benchmarking studies with deep mutational scanning datasets, particularly on host-pathogen systems are needed to reveal the affinity prediction performances of computational predictors. Our work provides an important step toward this direction.

MATERIALS and METHODS

Benchmark compilation

For benchmarking, we used 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 were mutated into other 19 amino acid possibilities. Among the probed mutations, the binding *enriching* ones correspond to 8.2% and 36% of RBD and ACE2 mutations, respectively. Less than 15% of these mutations (52 residues in total) were located across the RBD-ACE2 interface (RBD: 13.4%, ACE2: 11.1%). These interface positions were calculated over 6m0j (21) with PDBePISA (28) (Figure 1A):

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.

To create an unbiased set, we selected an equal number of *enriching* and *depleting* mutations from the interfacial RBD and ACE2 sets, respectively. As a result, we isolated 84 RBD (42 enriching, 42 depleting) and 179 ACE2 mutations (89 enriching, 90 depleting cases) (Figure 1B, Table S1). Half of the depleted interactions are highly depleting (50%), whereas the other half was randomly selected (by using the *default*_rng function of the NumPy package, with seed=123). The heatmaps in Figure 1B were generated with pandas, Numpy and seaborn libraries of Python 3.8 (29–34). For all the selected mutations, we made sure that we selected the constructs with high expression rates. Our final dataset contains 263 RBD and ACE2 variants (Table S1).

Structure-based binding affinity predictors

For our benchmarking efforts, we selected four widely used structure-based binding affinity predictors: FoldX (also FoldXwater), EvoEF1, Mutabind2, and SSIPe (13–16). All these predictors rely on physics-based terms, while Mutabind2 and SSIPe use additional evolutionary-based terms. Next to these, we also benchmarked HADDOCK (17, 35) and UEP (18) as naïve predictors, since their scoring functions are based on simple force field and contact terms, respectively.

FoldX: The scoring function of FoldX (13) is a linear sum of vdW energy (ΔG_{vdw}), hydrophobic (ΔG_{solvH}) and polar group desolvation (ΔG_{solvP}) energies, hydrogen bond energy from water molecules (ΔG_{wb}), hydrogen bond energy (ΔG_{hbond}), electrostatic energy (ΔG_{el}), the electrostatic contribution of different polypeptides (ΔGk_{on}), entropic penalty for backbone (ΔS_{mc}), entropic penalty of the side chain (ΔS_{sc}), and steric overlaps (ΔG_{clash}) (Eq. 1). FoldX also has an option to include the contribution of water molecules to the binding affinity (FoldXwater). Thus, we calculated the predicted binding affinities with and without water molecules.

$$E_{FoldX} = a*\Delta G_{vdw} + b*\Delta G_{solvH} + c*\Delta G_{solvP} + d*\Delta G_{wb} + e*\Delta G_{hbond} + f*\Delta G_{el} + g*\Delta Gk_{on} + h*T\Delta S_{mc} + k*T\Delta S_{sc} + l*\Delta G_{clash} \quad \text{(Eq. 1)}$$

EvoEF1: The scoring function of EvoEF1 (14) contains van der Waals (E_{vdw}), electrostatics (E_{elec}), hydrogen bond (E_{HB}), desolvation energies (E_{solv}), and the energy of a reference state (E_{ref}) (Eq. 2).

$$E_{EvoEF1} = [E_{vdw} + E_{elec} + E_{HB} + E_{solv}] - E_{ref} \quad \text{(Eq. 2)}$$

MutaBind2 and **SSIPe**: use FoldX and EvoEF1, respectively, to explicitly model the desired mutation. Next to these, for scoring, MutaBind2 utilizes other force field and contact-based terms, together with a metric measuring the evolutionary conservation of the mutated site (36). All these terms are incorporated into a single scoring function through a random forest algorithm. Like Mutabind2, SSIPe uses EvoEF1 energy terms and residue conservation-related terms, extracted from iAlign (37) and PSI-BLAST (38).

HADDOCK: The scoring function of HADDOCK is the sum of four terms, which are van der Waals (vdW), electrostatic, desolvation energy, and Buried Surface Area (Eq.3) (35, 39). As this scoring function was not developed to predict binding affinities, but to score docking results, we used it as a naïve predictor.

$$E_{HADDOCK} = w*EvdW + w*Eelec - w*BSA + w*Edesolv \quad \text{(Eq. 3)}$$

UEP: was used as the second naïve predictor. UEP predicts the impact of all possible interfacial mutations, when the position of interest has interactions with at least two other residues (highly packed residues within 5Å range) (18). Because of this, UEP could be run on a subset of our benchmark, specifically on 129 mutations (Table S2). The scoring function of UEP expands on the statistically determined intermolecular contact potentials (40).

To run FoldX and EvoEF1, we used their stand-alone packages (http://foldxsuite.crg.eu/products#foldx, https://github.com/tommyhuangthu/EvoEF). HADDOCK, MutaBind2, and SSIPe were run on their servers, as given in https://alcazar.science.uu.nl/services/HADDOCK2.2/, https://lilab.jysw.suda.edu.cn/research/mutabind2/ https://zhanggroup.org/SSIPe/ . MutaBind2, SSIPe, and UEP directly provide the binding affinity predictions, whereas for the rest we calculated the predicted binding affinity change according to Eq. 4. A mutation is evaluated as enriching if the predicted affinity change ($\Delta\Delta G_{predicted}$) is <0 and depleting, if ($\Delta\Delta$Gpredicted) is >0. Except UEP, all the selected tools provide the structural model of the mutation too, which were saved for further analysis.

$$\Delta\Delta G_{predicted} = \Delta G_{mut} - \Delta G_{wt} \quad \text{(Eq. 4)}$$

Performance Evaluation

The prediction performances were 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, based on the vdW volumes presented by Lin *et al.* (41). Amino acid hydrophobicities were taken from Eisenberg *et al.* (42). To measure flexibility change,

we used the flexibility scale presented by Shapovalov & Dunbrack (43). The physicochemical properties are considered as: polar amino acids - nN, 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 (44). Success rate and metric evaluations were performed in Python 3.8.5 with Pandas, Numpy, seaborn, and Matplotlib libraries (29–34, 45). For each category, the percentage of successfully predicted cases was calculated by dividing the number of accurate predictions to the total number of cases in that category, as in [Success rate = $Predictions_{good}/Predictions_{all}*100$]. The predictors results and their performance evaluations, together with our benchmark set are provided at https://github.com/CSB-KaracaLab/ace2-rbd-point-mutation-benchmark

DATA AVAILABILITY

All results including the codes and notebooks are deposited in Github (https://github.com/CSB-KaracaLab/ace2-rbd-point-mutation-benchmark) and the models and the scores can be visualized in https://nzmye.github.io/ace2-rbd-point-mutation-benchmark/visualize/#. We created the visualization page by using the open-source codes of Rodrigues *et al.* study (11).

SUPPLEMENTARY DATA

Supplementary Data are available.

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).

ACKNOWLEDGEMENTS

All the simulations and analyses were carried out in the HPC resources of Izmir Biomedicine and Genome Center.

CONFLICT OF INTEREST

The authors declare no competing interests.

REFERENCE

1. Cherry,J.D. and Krogstad,P. (2004) SARS: The first pandemic of the 21st century. *Pediatric Research*, **56**, 1–5.
2. Durai,P., Batool,M., Shah,M. and Choi,S. (2015) Middle East respiratory syndrome coronavirus: transmission, virology and therapeutic targeting to aid in outbreak control. *Experimental & Molecular Medicine*, **47**, e181–e181.
3. Platto,S., Xue,T. and Carafoli,E. (2020) COVID19: an announced pandemic. *Cell Death & Disease*, **11**, 799.
4. Li,W., Moore,M.J., Vasilieva,N., Sui,J., Wong,S.K., Berne,M.A., Somasundaran,M., Sullivan,J.L., Luzuriaga,K., Greenough,T.C., *et al.* (2003) Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature*, **426**, 450–454.
5. Ali,A. and Vijayan,R. (2020) Dynamics of the ACE2–SARS-CoV-2/SARS-CoV spike protein interface reveal unique mechanisms. *Scientific Reports*, **10**, 14214.
6. Hodcroft,E.B. CoVariants: SARS-CoV-2 Mutations and Variants of Interest. https://covariants.org/.
7. Chan,K.K., Dorosky,D., Sharma,P., Abbasi,S.A., Dye,J.M., Kranz,D.M., Herbert,A.S. and Procko,E. (2020) Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2. *Science (1979)*, **369**, 1261–1265.
8. 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., *et al.* (2020) Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*, **182**, 1295-1310.e20.
9. Laurini,E., Marson,D., Aulic,S., Fermeglia,M. and Pricl,S. (2020) Computational Alanine Scanning and Structural Analysis of the SARS-CoV-2 Spike Protein/Angiotensin-Converting Enzyme 2 Complex. *ACS Nano*, **14**, 11821–11830.
10. Delgado Blanco,J., Hernandez-Alias,X., Cianferoni,D. and Serrano,L. (2020) In silico mutagenesis of human ACE2 with S protein and translational efficiency explain SARS-CoV-2 infectivity in different species. *PLOS Computational Biology*, **16**, e1008450.
11. Rodrigues,J.P.G.L.M., Barrera-Vilarmau,S., M. C. Teixeira,J., Sorokina,M., Seckel,E., Kastritis,P.L. and Levitt,M. (2020) Insights on cross-species transmission of SARS-CoV-2 from structural modeling. *PLOS Computational Biology*, **16**, e1008449.
12. Sorokina,M., M. C. Teixeira,J., Barrera-Vilarmau,S., Paschke,R., Papasotiriou,I., Rodrigues,J.P.G.L.M. and Kastritis,P.L. (2020) Structural models of human ACE2 variants with SARS-CoV-2 Spike protein for structure-based drug design. *Scientific Data*, **7**, 309.
13. Schymkowitz,J., Borg,J., Stricher,F., Nys,R., Rousseau,F. and Serrano,L. (2005) The FoldX web server: an online force field. *Nucleic Acids Research*, **33**, W382–W388.
14. Pearce,R., Huang,X., Setiawan,D. and Zhang,Y. (2019) EvoDesign: Designing Protein–Protein Binding Interactions Using Evolutionary Interface Profiles in Conjunction with an Optimized Physical Energy Function. *Journal of Molecular Biology*, **431**, 2467–2476.
15. Huang,X., Zheng,W., Pearce,R. and Zhang,Y. (2020) SSIPe: accurately estimating protein–protein binding affinity change upon mutations using evolutionary profiles in combination with an optimized physical energy function. *Bioinformatics*, **36**, 2429–2437.
16. Zhang,N., Chen,Y., Lu,H., Zhao,F., Alvarez,R.V., Goncearenco,A., Panchenko,A.R. and Li,M. (2020) MutaBind2: Predicting the Impacts of Single and Multiple Mutations on Protein-Protein Interactions. *iScience*, **23**, 100939.

17. Dominguez,C., Boelens,R. and Bonvin,A.M.J.J. (2003) HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. *J Am Chem Soc*, **125**, 1731–1737.
18. Amengual-Rigo,P., Fernández-Recio,J. and Guallar,V. (2021) UEP: an open-source and fast classifier for predicting the impact of mutations in protein–protein complexes. *Bioinformatics*, **37**, 334–341.
19. Landau,M., Mayrose,I., Rosenberg,Y., Glaser,F., Martz,E., Pupko,T. and Ben-Tal,N. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Research*, **33**, W299–W302.
20. Wan,Y., Shang,J., Graham,R., Baric,R.S. and Li,F. (2020) Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. *Journal of Virology*, **94**.
21. Lan,J., Ge,J., Yu,J., Shan,S., Zhou,H., Fan,S., Zhang,Q., Shi,X., Wang,Q., Zhang,L., *et al.* (2020) Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*, **581**, 215–220.
22. Schrödinger,L.L.C. and DeLano,W. PyMOL.
23. Schymkowitz,J.W.H., Rousseau,F., Martins,I.C., Ferkinghoff-Borg,J., Stricher,F. and Serrano,L. (2005) Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proceedings of the National Academy of Sciences*, **102**, 10147–10152.
24. Neuwirth,E. (2014) RColorBrewer: ColorBrewer Palettes.
25. Kolde,R. (2019) heatmap: Pretty Heatmaps.
26. RStudio Team (2015) RStudio: Integrated Development Environment for R.
27. R Core Team (2019) R: A Language and Environment for Statistical Computing.
28. Krissinel,E. and Henrick,K. (2007) Inference of Macromolecular Assemblies from Crystalline State. *Journal of Molecular Biology*, **372**, 774–797.
29. van Rossum,G. and Drake,F.L. (2009) Python 3 Reference Manual.
30. Harris,C.R., Millman,K.J., der Walt,S.J. van, Gommers,R., Virtanen,P., David Cournapeau, Wieser,E., Taylor,J., Sebastian Berg, Smith,N.J., *et al.* (2020) Array programming with NumPy. *Nature*, **585**, 357–362.
31. Kluyver,T., Ragan-Kelley,B., Pérez,F., Granger,B., Bussonnier,M., Frederic,J., Kelley,K., Hamrick,J., Grout,J., Corlay,S., *et al.* (2016) Jupyter Notebooks - a publishing format for reproducible computational workflows. In Loizides,F., Scmidt,B. (eds), *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. IOS Press, Netherlands, pp. 87–90.
32. Anaconda Software Distribution (2020) *Anaconda Documentation*.
33. pandas development team,T. (2020) pandas-dev/pandas: Pandas. 10.5281/zenodo.3509134.
34. Waskom,M.L. (2021) seaborn: statistical data visualization. *Journal of Open Source Software*, **6**, 3021.
35. van Zundert,G.C.P., Rodrigues,J.P.G.L.M., Trellet,M., Schmitz,C., Kastritis,P.L., Karaca,E., Melquiond,A.S.J., van Dijk,M., de Vries,S.J. and Bonvin,A.M.J.J. (2016) The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology*, **428**, 720–725.
36. Choi,Y. and Chan,A.P. (2015) PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, **31**, 2745–2747.

37. Gao,M. and Skolnick,J. (2010) iAlign: a method for the structural comparison of protein–protein interfaces. *Bioinformatics*, **26**, 2259–2265.
38. Altschul,S. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.
39. Honorato,R. v., Koukos,P.I., Jiménez-García,B., Tsaregorodtsev,A., Verlato,M., Giachetti,A., Rosato,A. and Bonvin,A.M.J.J. (2021) Structural Biology in the Clouds: The WeNMR-EOSC Ecosystem. *Frontiers in Molecular Biosciences*, **8**.
40. Moal,I.H. and Fernandez-Recio,J. (2013) Intermolecular Contact Potentials for Protein–Protein Interactions Extracted from Binding Free Energy Changes upon Mutation. *Journal of Chemical Theory and Computation*, **9**, 3715–3727.
41. Lin,Z., Long,H., Bo,Z., Wang,Y. and Wu,Y. (2008) New descriptors of amino acids and their application to peptide QSAR study. *Peptides (N.Y.)*, **29**, 1798–1805.
42. Eisenberg,D., Schwarz,E., Komaromy,M. and Wall,R. (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *Journal of Molecular Biology*, **179**, 125–142.
43. Shapovalov,M.V. and Dunbrack,R.L. (2011) A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure*, **19**, 844–858.
44. Hagemans,D., van Belzen,I.A.E.M., Morán Luengo,T. and Rüdiger,S.G.D. (2015) A script to highlight hydrophobicity and charge on protein surfaces. *Frontiers in Molecular Biosciences*, **2**.
45. Hunter,J.D. (2007) Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, **9**, 90–95.