Benchmarking the accuracy of structure-based binding affinity predictors on Spike-ACE2 Deep Mutational Interaction Set

Burcu Ozden[1,2], Eda Şamiloğlu[1,2], Mehmet Erguven[1], Can Yükrük[1], Mehdi Koşaca[1,2], Melis Oktayoğlu[1], Muratcan Menteş[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

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. To this end, two deep mutational scanning studies traced the impact of all possible mutations across Receptor Binding Domain (RBD) of Spike and catalytic domain of human ACE2. By concentrating on the interface mutations of these experimental data, we benchmarked six commonly used structure-based binding affinity predictors (FoldX, EvoEF1, MutaBind2, SSIPe, HADDOCK, and UEP). These predictors were selected based on their user-friendliness, accessibility, and speed. As a result of our benchmarking efforts, we observed that none of the methods could generate a meaningful correlation with the experimental binding data. The best correlation is achieved by FoldX (R = -0.51). Also, when we simplified the prediction problem to a binary classification, i.e., whether a mutation is enriching or depleting the binding, we showed that the highest accuracy is achieved by FoldX with 64% success rate. Surprisingly, on this set, simple energetic scoring functions performed significantly better than the ones using extra evolutionary-based terms, as in Mutabind and SSIPe. These observations suggest plenty of room to improve the conventional affinity predictors for guessing the variant-induced binding profile changes of a host-pathogen system, such as the Spike-ACE2. To aid such improvements we provide our benchmarking data at https://github.com/CSB-KaracaLab/RBD-ACE2-MutBench with the option to visualize our mutant models at https://rbd-ace2-mutbench.github.io/

Key words: binding affinity prediction, deep mutagenesis, SARS-CoV-2, ACE2, RBD

37 **INTRODUCTION**

38 At the beginning of 21st century, the emergence of Severe Acute Respiratory
39 Syndrome Coronavirus (SARS-CoV)[1] and Middle East Respiratory Syndrome Coronavirus[2]
40 led to serious public health concerns. Evolving from these viruses, during late 2019, a new
41 SARS virus, SARS-CoV-2, caused the most severe pandemic of the 21st century[3]. SARS-CoV-
42 2 infection is initiated upon having its Spike protein interacting with the host Angiotensin
43 Converting 2 (ACE2) enzyme[4]. The widespread infection of SARS-CoV-2 compared to its
44 predecessors was linked to higher binding affinity of Spike to ACE2[5]. Relatedly, alpha, beta,
45 gamma, eta, iota, kappa, lambda, mu, and omicron SARS-CoV-2 variants were shown to have
46 at least one mutation across the Spike-ACE2 interface[6]. This realization placed the
47 characterization of interfacial Spike-ACE2 mutations at the center of COVID-19-related
48 research. Within this context, in 2020, two deep mutational scanning (DMS) studies explored
49 how Spike/ACE2 variants impact Spike-ACE2 interactions[7,8]. In these DMS studies, the
50 residues on the Receptor Binding Domain (RBD) of Spike and the catalytic domain of human
51 ACE2 were mutated into other 19 amino acid possibilities, followed by tracing of new RBD-
52 ACE2 binding profiles.

53 In parallel to these experimental efforts, a handful of structure- and force field-based
54 computational studies employed a comprehensive investigation of variation across the RBD-
55 ACE2 interface (Table 1). Three such studies utilized two fast and user-friendly tools, FoldX
56 and HADDOCK [9–11]. Blanco *et al.* used FoldX with the inclusion of water molecules
57 (FoldXwater) to trace the binding enhancing RBD and ACE2 mutations [9]. From their 21
58 binding enhancing mutation predictions, nine of them were confirmed as affinity enhancing
59 by the DMS set. Rodrigues *et al.* investigated the impact of ACE2 orthologs on their RBD

60 binding with HADDOCK, where they proposed five significantly affinity improving ACE2
61 mutations[10]. Among these, D30E and A387E were shown to be affinity enhancing in the
62 ACE2 DMS set. Complementary to this study, Sorokina *et al*. performed computational alanine
63 scanning on ACE2 with HADDOCK[11]. Here, three out of five mutations (N49A, R393A, and
64 P389A) were classified as binding enriching both by the computational predictions and the
65 DMS set. Two other studies made use of elaborate simulation techniques, mainly molecular
66 dynamics simulations. Laurini *et al.* performed molecular mechanics/Poisson−Boltzmann
67 alanine scanning and molecular dynamics simulations to find structurally and energetically
68 critical RBD-ACE2 residues[12]. They proposed eight hotspot positions on RBD and ACE2,
69 where three of them were characterized as affinity enhancing in the DMS sets. Gheeraert *et*
70 *al.* performed 1 µs long molecular dynamics simulations of five RBD variants (alpha, beta,
71 gamma, delta, and epsilon) in complex with ACE2[13]. They found that L452R, T478K (delta
72 variant) and N501Y (alpha, gamma variants) cause drastic structural changes across the RBD-
73 ACE2 interface. These mutations were classified as binding enriching in the RBD DMS set.

74 **Table 1. Affinity impacting RBD and ACE2 variant/hotspot predictions.** The predictions agreeing
75 with the experimental are underlined and shown in bold.

Work carried out by	**Important RBD residues/mutations**	**Important ACE2 residues/ mutations**	**Approach**
Blanco *et al.*[9]	V445M/R/W, **Q493F/L/M/Y**, **Q498F**/L/M/**Y**, T500K,	G326E	Variant modeling and interface score calculation with FoldXwater

	N501A/C/L/S/**T**, **V503R**/W/Y		
Rodrigues *et al.*[10]		Q24E, **D30E**, H34Y, L79H, **A387E**	Variant modeling and interface score calculation with HADDOCK
Sorokina *et al.*[11]		**N49A**, **R393A**, M383A, **P389A**, G354A	Alanine scanning and interface score calculation with HADDOCK
Laurini *et al.*[12]	**Q498**, T500, R403	D38, **K31**, E37, K353, **Y41**	Interaction and energy analysis through molecular dynamics and MM/PB alanine scanning
Gheeraert *et al.*[13]	**L452R**, **T478K**, **N501Y**		Interaction analysis through MD simulations

As presented in Table 1, accurate reproduction of RBD-ACE2 DMS profiles can be tricky even when elaborate simulation techniques are used. So, if such time-intensive simulation approaches are facing challenges in back-calculating the impact of RBD-ACE2 interface variation, how far are the fast prediction tools that were heavily used in the early months of pandemic, such as FoldX and HADDOCK, from accurately predicting the impact of RBD-ACE2 variations? To answer this question, we benchmarked six fast affinity prediction tools, i.e.,

83 FoldX, HADDOCK, EvoEF1, MutaBind2, SSIPe, and UEP [14–20] against the Spike-ACE2
84 interface DMS set. These predictors were selected based on their user-friendliness and speed,
85 since we wanted to put an emphasis on the accessibility of these tools to the researchers who
86 may not have programming experience or enough computing resources. Among these tools,
87 FoldX and EvoEF1 use intra- and inter-molecular energies derived from empirical force field
88 terms. HADDOCK scores complexes by combining intermolecular van der Waals,
89 electrostatics, and empirical desolvation terms. Mutabind and SSPIe utilize FoldX and EvoEF1,
90 respectively, to model the mutations. Both methods have their own scoring functions to
91 consider evolutionary-based information too. UEP scores mutations based on statistically
92 determined intermolecular contact potentials. HADDOCK, MutaBind2, and SSIPe can be run
93 through a web service. FoldX can be called over a GUI through a YASARA plugin. EvoEF1 and
94 UEP are available as stand-alone packages. Our benchmarking files can be accessed at
95 https://github.com/CSB-KaracaLab/RBD-ACE2-MutBench with the option to visualize our
96 mutant models at https://rbd-ace2-mutbench.github.io/

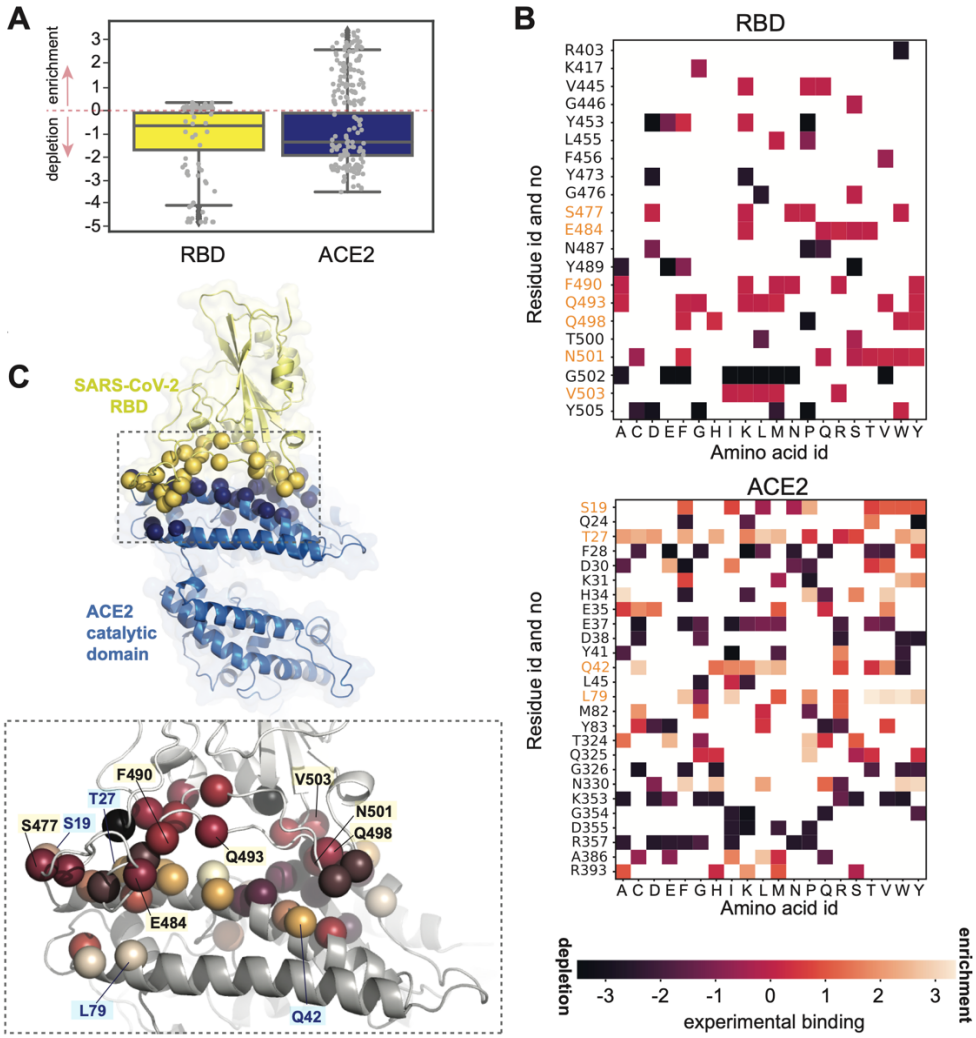
97 **RESULTS and DISCUSSION**

98 **Benchmark Compilation**

99 The deep mutational scanning (DMS) experiments performed by Chan *et al.* and Starr
100 *et al.* [7,8] scan the impact of all possible amino acid variations imposed on the RBD domain
101 of Spike and the catalytic domain of ACE2 on RBD-ACE2 binding. These studies classified
102 mutations as binding enriching or depleting when compared to the wild type interaction.
103 While compiling our benchmark set, our aim was (i) to select the DMS subset reporting on
104 the variations across RBD-ACE2 interface since the selected methods are tuned to predict the
105 impact of interface mutations, (ii) to include an equal number of binding enriching and
106 depleting cases to obtain a balanced benchmark set.

107 The DMS sets contain 988 interfacial mutations, measured over 26 RBD and 26 ACE2
108 residues (calculated by PDBePISA[21] on 6m0j PDB[22]). 13% of these 988 mutations were
109 profiled as binding enriching (42 for RBD and 89 for ACE2, Figure 1A) and the rest as binding
110 depleting. As shown in Figure 1A, the binding enriching RBD mutations span a narrow
111 enrichment range [0.01, 0.30], while for ACE2 this range increases to [0.03, 3.37]. We added
112 all enriching cases into our benchmark. To fairly represent the depleting cases, we selected
113 131 mutations sampling the whole depleting data spread (Figure 1A). We then analyzed the
114 individual binding profiles of selected mutations with heatmaps (Figure 1B). As can be
115 observed from these heatmaps, on the RBD side, several mutations on Q493, S477, F490,
116 N501, V503, E484, Q498 lead to better RBD-ACE2 binding (Figure 1B, Figure S1, Table S1).
117 Among these, Q493R and S477N were observed in omicron; E484K in beta, gamma, eta, iota,
118 mu; E484Q in kappa; N501Y in alpha, beta, gamma, mu, omicron variants[6]. On the ACE2
119 surface, the top enriching mutations came from T27, Q42, S19, and L79 positions (Figure 1C,
120 Figure S1). All these residues, except S19, were reported as species-associated variations[23].

121 While appearing less frequently as binding enhancers, K31, E35, M82, and Y83 were earlier
122 listed as critical residues for RBD-ACE2 interactions (Figure 1B-C and Figure S1)[9,12,22]. All
123 these residue positions are situated around the core (at the rim) of the RBD-ACE2 interface.
124 The top enriching mutations on the RBD side are Q493M, S477D, F490K, N501F, V503M,
125 E484R, Q498H, and on the ACE2 side are T27L, Q42C, S19P (Table S1). Further investigation
126 of these mutations did not lead to a generalized pattern for understanding RBD-ACE2
127 recognition.



128

129 **Figure 1. (A) RBD-ACE2 DMS benchmark set.** The experimental binding profile distributions of the
130 interfacial RBD (yellow, n=494) and ACE2 (blue, n=494) mutations are represented with box-and-
131 whisker plots. Values >0 indicate binding enriching and values <0 indicate binding depleting

132 mutations. 26.5% of this data set is selected as our benchmark set (n=262): n=84 for RBD (42
133 enriching, 42 depleting) and n=178 for ACE2 mutations (89 enriching, 89 depleting cases)),
134 represented as gray dots. This panel was generated in Python 3.8 by using pandas, Numpy and
135 Seaborn libraries[34–39]. **(B) The experimental binding enrichment and depletion values of our**
136 **benchmark set, RBD (top) and ACE2 (bottom).** The values > 0 correspond to binding enriching
137 positions (light orange), while the values <0 represent the depleting ones (dark purple). The
138 positions leading often to binding enriching mutations are highlighted in orange. **(C) The structural**
139 **depiction of enriching mutations.** The interface residues of RBD-ACE2 complex are shown in
140 spheres. The color code of the spheres follows the largest binding value measured for a given
141 residue, as shown in Figure 1B. The important positions are highlighted with labels (yellow: RBD,
142 blue: ACE2). The illustration is generated in PyMOL[24] using PDB 6M0J[22].
143

144 **Benchmarking the affinity predictors: seeking for a linear correlation**

145 For all the mutations in our benchmark set (n=262), we calculated the score change
146 imposed by the mutations with FoldX, FoldXwater, EvoEF1, MutaBind2, SSIPe, and HADDOCK.
147 [14–20] (Figure S2). Then, we investigated whether there is any meaningful correlation with
148 the calculated score changes and experimental binding enrichment/depletion values (Figure
149 2A). Here, the expected perfect correlation would be -1 as larger experimental values and
150 smaller binding scores point to improved binding. As a result, we observed insignificant linear
151 correlations for HADDOCK, EvoEF1, and SSIPe predictors. Mediocre correlations with R values
152 ranging from -0.51 to -0.45 were observed for FoldX and MutaBind2. The best correlation (the
153 highest R-value) was obtained by FoldX (R = -0.51). Interestingly, including the water effect
154 into FoldX predictions by using FoldXwater did not improve the accuracy of the original
155 approach (R = -0.45 vs. R = -0.51). Furthermore, the enhanced scoring function of MutaBind2
156 built upon FoldX did not improve the original FoldX scoring (R = -0.49 vs. R = -0.51). The same

157 was observed for SSIPe, since it was built upon EvoEF1's sampling (R = -0.28 vs. -0.27). As the
158 naïve predictor, we run UEP on a subset of our benchmark (n=129), as UEP is tuned to predict
159 mutation-induced changes of residues that are in contact with more than two atoms. This
160 effort also resulted in an insignificant correlation.

161 We further calculated the pairwise correlations of score changes predicted by each
162 algorithm (Figure 2B). This comparison revealed that HADDOCK reports the most
163 distinct scores compared the other algorithms. This is understandable, as all the other
164 approaches stem from a similar prediction mentality of FoldX (Materials and Methods). We
165 computed all-atom Root Mean Square Deviations (RMSDs) of each generated mutant model
166 in an all-to-all fashion to understand whether the distinct behavior of HADDOCK came from
167 differentially modeled side chain formations (Figure S3). This analysis demonstrated that
168 HADDOCK indeed generates the largest RMSD models compared to the models computed by
169 the other tools. Notably, MutaBind2 and FoldX conformers resulted in the second highest
170 RMSD cases, indicating that the further minimization steps used by MutaBind2 significantly
171 impacts the final conformation of the side chains. As expected, EvoEF1 and SSIPe mutant
172 models came out to be identical, since SSIPe utilizes EvoEF1 to structurally model the
173 mutations.

174

175

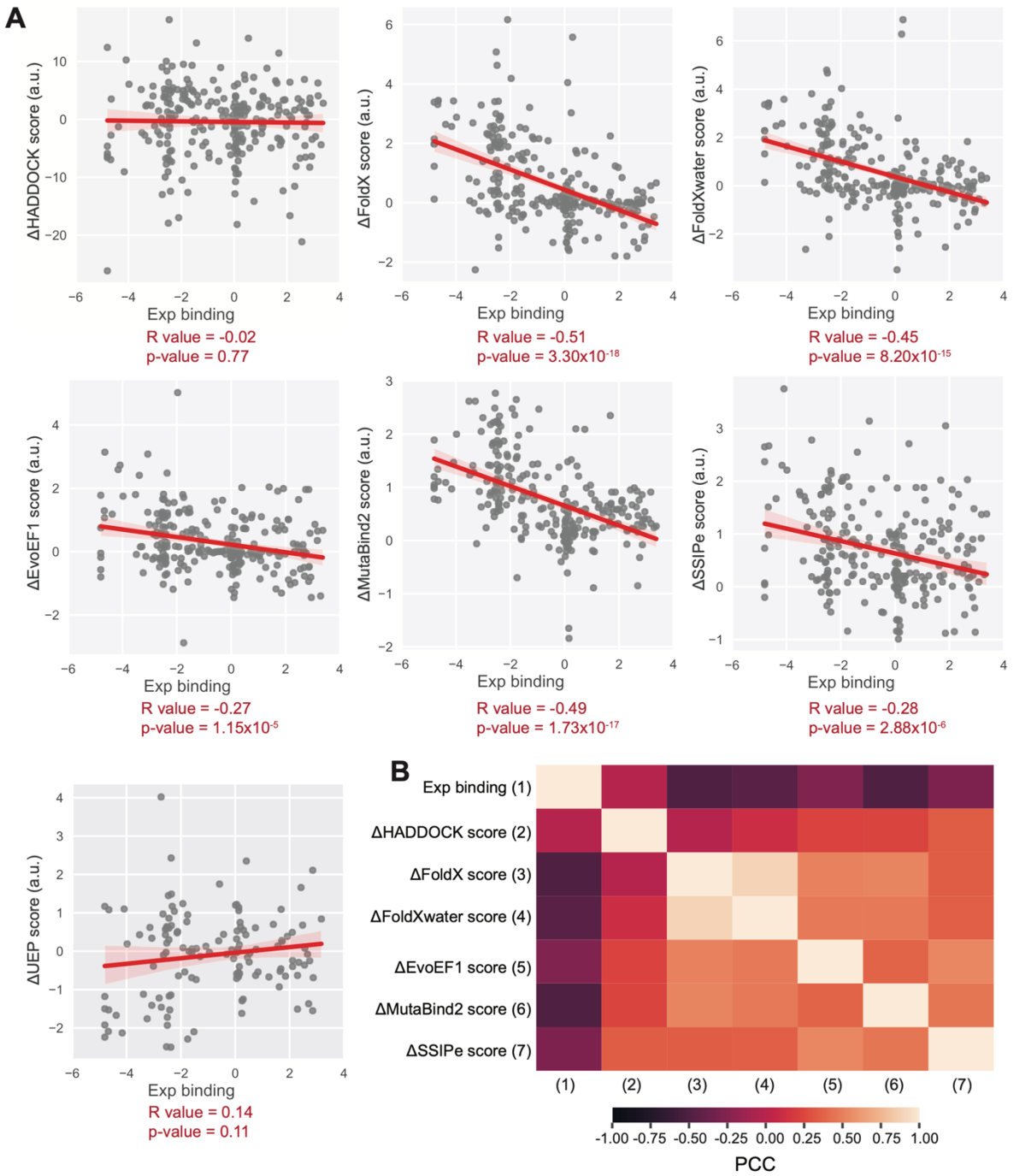


Figure 2. (A) Correlations between experimental DMS benchmark set and the binding affinity predictors. Data points for all scenarios are n=262, except UEP where the number of points is 129 (~50% of the benchmark set). R- and p-values were calculated by using statistics and scipy libraries of Python 3.8. **(B) The correlation heatmap of the computational and the experimental scores.** The correlation values are expressed in terms of Pearson Correlation Coefficients (PCC), ranging between

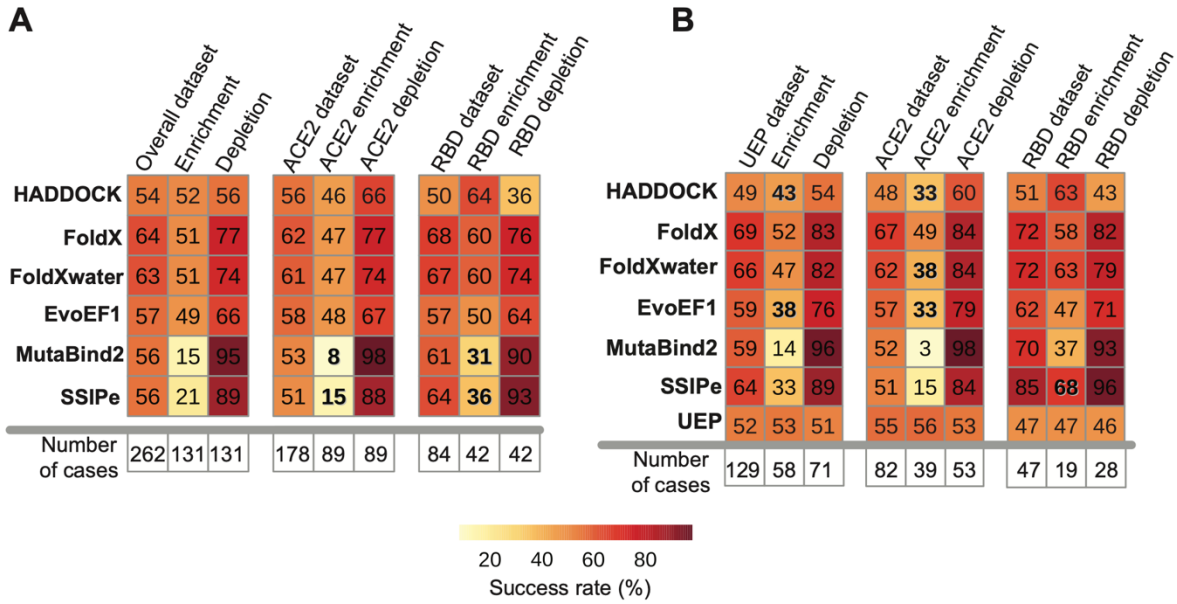
182 -1 and 1. Highly negatively (PCC = -1) and positively (PCC = 1) correlated predictors are colored with
183 dark purple and light orange, respectively.
184
185 **Benchmarking the affinity predictors: seeking for a binary classification**
186 Apart from analyzing the correlation between ΔScores and experimental binding
187 values, we also performed a binary assessment. In this assessment, the tools were tested
188 whether they can predict the direction of binding affinity change (enriching or depleting).
189 Accordingly, we counted a prediction as successful if the experimental and computational
190 data would agree whether a mutation is enriching or depleting. In this regard, the overall
191 prediction accuracy was calculated as the percentage of correct predictions (the success rate).
192 According to this binary assessment, the overall success rates of predictors varied between
193 54% and 64% (Figure 3A), where the top-ranking predictor again came out to be FoldX (64%)
194 (Figure 2A). FoldXwater ranked the second, once again implying that the inclusion of water
195 effects did not improve the prediction accuracy. When we analyzed ACE2 and RBD subsets
196 individually, better prediction rates for depleting mutations were consistently observed,
197 despite the narrow prediction range posed by the experimental data (Figure 3A). Strikingly,
198 MutaBind2 and SSIPe predicted most mutations as depleting, hinting at a problem in using
199 evolutionary-based terms in scoring the host-pathogen system RBD-ACE2. These
200 observations did not change, when we calculated the success rates only for the residues that
201 frequently lead to enriching mutations (the highlighted residues in Figure 1B-C). When we
202 further scored the lowest ranking predictor, HADDOCK's models with FoldX, HADDOCK's
203 success rate increased by 5% (from 54% from 59%) (Figure S4A). Normalizing HADDOCK
204 scores by the buried surface area (BSA) of the interface did not improve the success rates
205 (Figure S4A). On the UEP subset, the overall prediction performances vary within a broader

206 range, i.e., 49%-69% (Figure 3B), where the top two predictors became FoldX and FoldXwater

207 (69% vs. 66%) and the lowest performing predictors became UEP and HADDOCK (52% and

208 49%). On both datasets, all predictors had difficulties in predicting binding-enriching

209 mutations compared to the binding-depleting ones (Figure 3).



210

211 **Figure 3. (A) The success rates on the benchmark set (n=262). (B) Success rates of predictors are**

212 **calculated by using UEP's data set (n= 129).** In both panels, if the success rate changes drastically

213 compared to the overall dataset (first column on the left), it is shown in bold. The plots were

214 prepared by using R-Studio [25–28].

215

216 **Volume and hydrophobicity biases are the most obvious misprediction determinants**

217 To explore the prediction dependencies of each tool, we assessed their success rates

218 after classifying the benchmark cases according to volume, hydrophobicity, and flexibility

219 changes imposed by the mutations (Figure 4, Table S2, Figure S4B). The residue-based volume

220 values were taken from Zhi-hua *et al* [29], the hydrophobicity ones from Eisenberg *et al.* [30],

221 and the flexibility ones from Shapovalov and Dunbrack [31]. Per property, we calculated the

222 frequency distributions of the computed changes, given the enrichment or depletion status
223 of the original mutation (Figure 4A-C). To serve as a background, we calculated the frequency
224 distributions of the experimental data too (gray lines in Figure 4A-C). Here our assumption
225 was a discrepancy between the tails of the experimental data and the predictors' distributions
226 should point to an obvious bias for a given property. Accordingly, when we investigated the
227 volume change plots, we saw that all predictors had difficulties in predicting volume
228 decreasing mutations when the mutation induces an enrichment in binding (Figure 4A). In the
229 case of depleting mutations, only HADDOCK has an apparent bias toward classifying volume
230 increasing mutations wrongly. In the case of hydrophobicity change, for all predictors, there
231 is a slight tendency to classify hydrophobicity decreasing binding enriching mutations
232 wrongly. This is more pronounced for hydrophobicity decreasing and binding depleting
233 mutations for HADDOCK. We did not observe any bias for the predictors regarding the
234 changes in the side chain flexibilities.

235 To gauge whether the enriching and depletion mutation predictions were balanced
236 for a given property, we calculated the difference between the area under the curve of the
237 successfully predicted enriching and depletion mutations (ΔSuccess, Figure S4B). Here, our
238 assumption was, if the prediction is balanced, then the ΔSuccess should be close to zero. On
239 the other hand, the ΔSuccess values towards 100 and -100 should indicate extreme biases.
240 This analysis revealed that HADDOCK has moderate volume change bias with 34 ΔSuccess
241 score. Normalizing HADDOCK scores by the BSA of the interface doubled this bias leading to
242 78 ΔSuccess score. When HADDOCK models were scored with FoldX, the original 34 ΔSuccess
243 shifted to -31 ΔSuccess, resulting in a bias toward depleting mutations (Figure S4B). When we
244 considered the success rates based on hydrophobicity, we found out that FoldX tends to
245 predict depleting cases better with -31 ΔSuccess score. Inclusion of water in FoldX

246 (FoldXwater) increases this moderate bias (ΔSuccess score -36 vs -31). In this analysis too, we
247 could not find any relation between the flexibility change and success rate of predictors
248 (Figure 4C and Figure S4B). Finally, we found that ΔSuccess is extremely skewed for all metrics
249 for MutaBind2 and SSIPe, since they predict almost all mutations as depleting.

250 To present a complete analysis, we also investigated the impact of the
251 physicochemical property changes induced by each mutation on the prediction accuracy
252 (Figure 4D). The physicochemical change classes we considered were no change, polarity
253 loss/gain, and charge loss/gain. To serve as a background, we calculated the share of each
254 class within the original data set (gray bars in Figure 4D). As a result, we could not observe
255 any specific bias given a physicochemical property change. The success shares reflect the
256 general trends observed in successfully predicting binding enriching or depletion mutations,
257 as shown in Figure 4A.

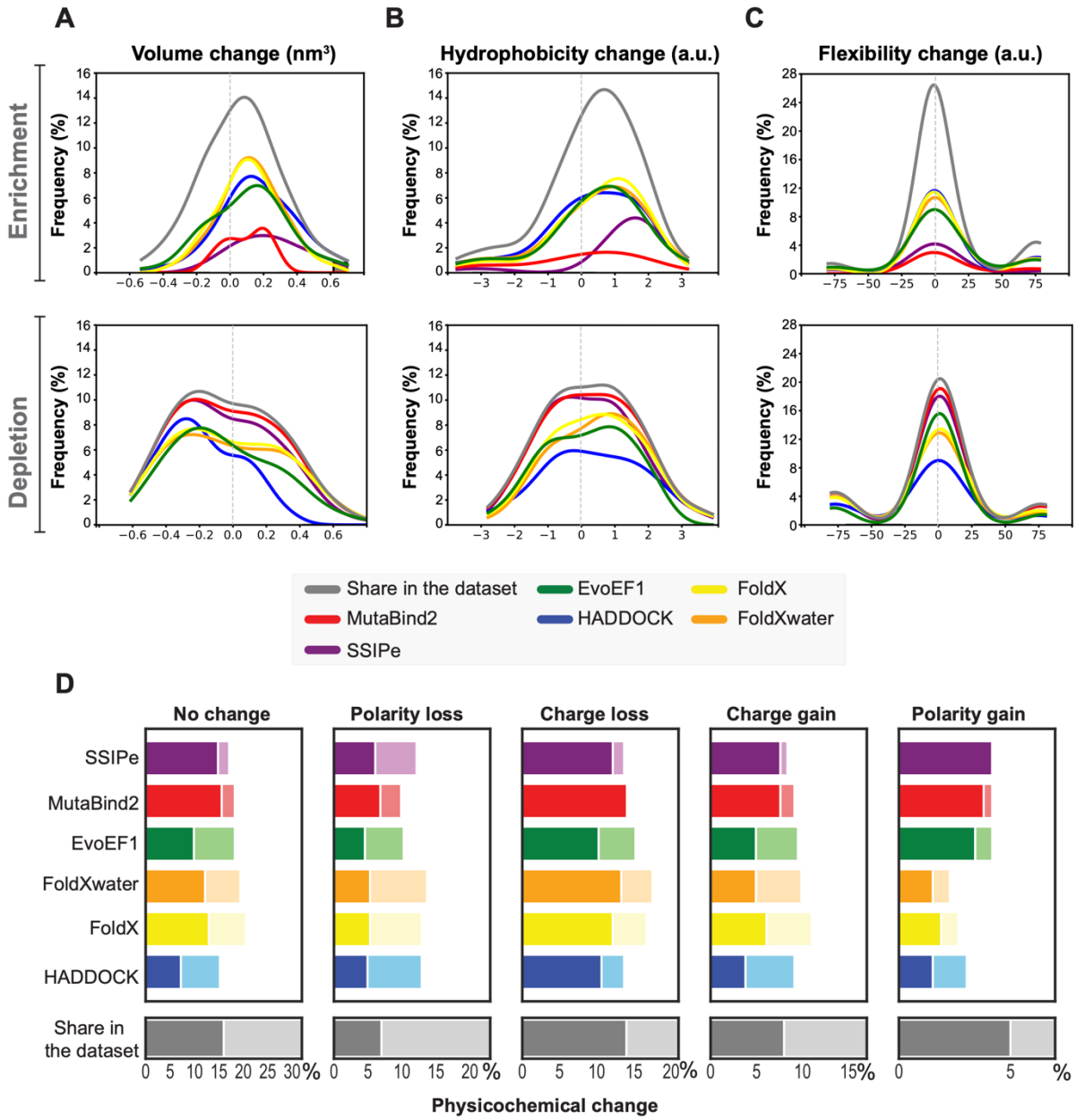


Figure 4. The effects of mutation-induced changes in the amino acid physical properties on the predictor's success rates according to (A) Volume change, **(B)** Hydrophobicity change, **(C)** Flexibility change. The original data set distribution for a given binding class is plotted in gray. The other predictors are colored as provided in the legend. Volume/hydrophobicity/flexibility-increasing mutations reside on the "positive side" of the plot (x-axis value >0), whereas volume/hydrophobicity/flexibility-decreasing mutations reside on the "negative side" of the pot (x-axis value <0) **(D)** The percentage of the correctly predicted cases, given the physicochemical change

266 induced upon mutation when polarity and charge states are considered. and light colors represent

267 successfully predicted depleting and enriching cases, respectively. The original data share for a given

268 class in the dataset is plotted in grey.

269

CONCLUSION

In the early months of SARS-CoV-2 pandemic, several fast and user-friendlt mutation modeling and scoring tools, such as FoldX and HADDOCK, were heavily used to predict the impact of Spike/ACE2 variations across the Spike-ACE2 interface. Expanding on these efforts, in this work, we benchmarked six fast and commonly used structure-based binding affinity predictors (FoldX, EvoEF1, MutaBind2, SSIPe, HADDOCK, and UEP) against the RBD-ACE2 DMS binding data. As a result, we observed that none of the predictors could produce a meaningful correlation with the experimental data (best correlation R-value -0.51 was obtained with FoldX). Even when a binary classification (binding enriching/depleting) was considered, the highest accuracy was obtained by FoldX with 64% success rate. Furthermore, all predictors had difficulties in predicting binding enriching mutations, especially the ones using conservation-based terms in their scoring. Finally, the most obvious biases in mispredictions were found to be toward volume and hydrophobicity changes, especially for HADDOCK and FoldX, respectively.

These observations suggest plenty of room to improve the conventional affinity predictors for guessing the variant-induced binding profile changes of host-pathogen systems, such as Spike-ACE2. To aid such improvements we provide our benchmarking data at https://github.com/CSB-KaracaLab/RBD-ACE2-MutBench with the option to visualize our mutant models at https://rbd-ace2-mutbench.github.io/ . We hope that our benchmarking study will guide the computational community for being prepared not only for combatting SARS-CoV-2-related health concerns but also other infectious diseases.

291 **MATERIALS and METHODS**

292 **Interfacial DMS value selection**

293 The original DMS sets contained 3,819 and 2,223 point mutations for RBD and ACE2,
294 respectively. From this set, we isolated the interfacial 988 RBD-ACE2 point mutations, for the
295 following residues (calculated over 6m0j[22] with PDBePISA[21] (Figure 1C)):
296 *Twenty-six RBD positions: R403, K417, V445, G446, Y449, Y453, L455, F456, Y473, A475, G476,*
297 *S477, Q484, G485, F486, N487, Y489, F490, Q493, G496, Q498, T500, N501, G502, V503, and*
298 *Y505.*
299 *Twenty-six ACE2 positions: S19, Q24, T27, F28, D30, K31, H34, E35, E37, D38, Y41, Q42, L45,*
300 *L79, M82, Y83, T324, Q325, G326, N330, K353, G354, D355, R357, A386, R393.*
301

302 **Structure-based binding affinity predictors**

303 Below are the brief scoring description of each predictor used. Except UEP, all
304 predictors explicitly model the mutation according to predictor's force field.
305 **FoldX:** The scoring function of FoldX is a linear sum of vdW energy (ΔG_{vdw}), hydrophobic
306 (ΔG_{solvH}) and polar group desolvation (ΔG_{solvP}) energies, hydrogen bond energy from water
307 molecules (ΔG_{wb}), hydrogen bond energy (ΔG_{hbond}), electrostatic energy (ΔG_{el}), the
308 electrostatic contribution of different polypeptides (ΔGk_{on}), entropic penalty for backbone
309 (ΔS_{mc}), entropic penalty of the side chain (ΔS_{sc}), and steric overlaps (ΔG_{clash}) (Eq. 1). FoldX also
310 has an option to include the contribution of water molecules to the binding affinity
311 (FoldXwater).

312 $$E_{FoldX} = a*\Delta G_{vdw} + b*\Delta G_{solvH} + c*\Delta G_{solvP} + d*\Delta G_{wb} + e*\Delta G_{hbond} + f*\Delta G_{el} + g*\Delta Gk_{on} + h*T$$

313 $$\Delta S_{mc} + k*T\Delta S_{sc} + l*\Delta G_{clash} \text{ (Eq. 1)}$$

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

317 $$E_{EvoEF1} = [E_{vdw} + E_{elec} + E_{HB} + E_{solv}] - E_{ref} \quad \text{(Eq. 2)}$$

318 **MutaBind2** and **SSIPe**: use FoldX and EvoEF1, respectively, to explicitly model the desired
319 mutation. MutaBind2 further imposes relaxation and utilizes extra force field and contact-
320 based terms, together with a metric measuring the evolutionary conservation of the mutation
321 site. All these terms are incorporated into a random forest based scoring algorithm. SSIPe
322 uses EvoEF1 energy terms and residue conservation-related terms, extracted from iAlign[32]
323 and PSI-BLAST[33].

324 **HADDOCK:** In this work, we used HADDOCK water refinement to model the mutations. The
325 complexes then scored according to the sum of three terms, van der Waals, electrostatic, and
326 desolvation energy (Eq.3). [34]

327 $$E_{HADDOCK} = w*EvdW + w*Eelec + w*Edesolv \quad \text{(Eq. 3)}$$

328 **UEP**: UEP predicts the impact of all possible interfacial mutations, when the position of
329 interest has interactions with at least two other residues (within 5Å range). The scoring
330 function of UEP expands on the statistically determined intermolecular contact potentials.

331 To run FoldX and EvoEF1, we used their stand-alone packages
332 (http://foldxsuite.crg.eu/products#foldx, https://github.com/tommyhuangthu/EvoEF).
333 HADDOCK, MutaBind2, and SSIPe were run on their servers, as given in
334 https://milou.science.uu.nl/services/HADDOCK2.2/,
335 https://lilab.jysw.suda.edu.cn/research/mutabind2/ https://zhanggroup.org/SSIPe/.
336 MutaBind2, SSIPe, and UEP directly provide the binding affinity change predictions, whereas
337 for the rest we calculated the predicted binding affinity change according to Eq. 4:

338 $$\Delta\Delta G_{predicted} = \Delta G_{mut} - \Delta G_{wt} \text{ (Eq. 4)}$$

339 A mutation is evaluated as binding enriching if the predicted binding value change
340 ($\Delta Score_{predicted}$) is <0 and binding depleting, if ($\Delta Score_{predicted}$) is >0.

341 **Performance evaluation according to change in amino acid physical properties upon a**
342 **mutation**

343 The predictions were evaluated from the perspectives of volume, hydrophobicity,
344 flexibility, and physicochemical property change upon mutation ($\Delta Property_{change}$ =
345 $Property_{mutation}$ - $Property_{wildtype}$, Table S2). The physicochemical properties considered were:
346 polar amino acids - N, Q, S, T, Y; non-polar amino acids - A, G, I, L, M, F, P, W, V, C; charged
347 amino acids - H, E, D, R, K. Success rate and metric evaluations were performed in Python
348 3.8.5 with Pandas, Numpy, seaborn, and Matplotlib libraries [35–40]. For each category, the
349 percentage of successfully predicted cases were calculated by Eq .5.

350 $$\text{Success rate} = \text{Correct-Predictions}/\text{All-Predictions}*100 \text{ (Eq. 5)}$$

351

352

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 INFORMATION

Supplementary Information 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 and providing help in building our results visualization page.

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

B.O. performed statistical analysis and RMSD calculations, generated structural and statistical figures, edited and finalized all the figures, prepared the Github page, and wrote the manuscript. E.Ş., C.Y., and M.M. performed predictor runs. E.Ş. calculated the physical property and binary success rates, prepared the Github page, contributed to writing Introduction and Methods & Materials sections. M.E. generated heatmap tables for success rate results and wrote the manuscript. M.K. generated heatmaps for the RMSD calculations

and predictor scores. M.O. performed the most enriching mutation analysis. N. A. and G. K. prepared the visualization page on Github. A.B.B. and B.S. worked on the conceptualization of the benchmark set. E.K. conceptualized the study, overlooked the project, and wrote the manuscript.

CONFLICT OF INTEREST

The authors declare no competing interests.

384 **REFERENCES**

385 [1] LeDuc JW, Barry MA. SARS, the First Pandemic of the 21st Century. Emerg
386 Infect Dis 2004;10:e26. https://doi.org/10.3201/EID1011.040797_02.
387 [2] Durai P, Batool M, Shah M, Choi S. Middle East respiratory syndrome
388 coronavirus: transmission, virology and therapeutic targeting to aid in outbreak
389 control. Exp Mol Med 2015;47:e181. https://doi.org/10.1038/EMM.2015.76.
390 [3] Platto S, Xue T, Carafoli E. COVID19: an announced pandemic n.d.
391 https://doi.org/10.1038/s41419-020-02995-9.
392 [4] Li W, Moore MJ, Vasllieva N, Sui J, Wong SK, Berne MA, et al. Angiotensin-
393 converting enzyme 2 is a functional receptor for the SARS coronavirus. Nature
394 2003;426:450. https://doi.org/10.1038/NATURE02145.
395 [5] Ali A, Vijayan R. Dynamics of the ACE2–SARS-CoV-2/SARS-CoV spike protein
396 interface reveal unique mechanisms. Sci Rep 2020;10.
397 https://doi.org/10.1038/s41598-020-71188-3.
398 [6] CoVariants n.d. https://covariants.org/ (accessed June 27, 2022).
399 [7] Chan KK, Dorosky D, Sharma P, Abbasi SA, Dye JM, Kranz DM, et al. Engineering
400 human ACE2 to optimize binding to the spike protein of SARS coronavirus 2.
401 Science (1979) 2020;369. https://doi.org/10.1126/SCIENCE.ABC0870.
402 [8] Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, et al. Deep
403 Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals
404 Constraints on Folding and ACE2 Binding. Cell 2020;182:1295-1310.e20.
405 https://doi.org/10.1016/j.cell.2020.08.012.
406 [9] Delgado Blanco J, Hernandez-Alias X, Cianferoni D, Serrano L. In silico
407 mutagenesis of human ACE2 with S protein and translational efficiency explain
408 SARS-CoV-2 infectivity in different species. PLoS Comput Biol
409 2020;16:e1008450. https://doi.org/10.1371/journal.pcbi.1008450.
410 [10] Rodrigues JPGLM, Barrera-Vilarmau S, M. C. Teixeira J, Sorokina M, Seckel E,
411 Kastritis PL, et al. Insights on cross-species transmission of SARS-CoV-2 from
412 structural modeling. PLoS Comput Biol 2020;16:e1008449.
413 https://doi.org/10.1371/journal.pcbi.1008449.
414 [11] Sorokina M, M. C. Teixeira J, Barrera-Vilarmau S, Paschke R, Papasotiriou I,
415 Rodrigues JPGLM, et al. Structural models of human ACE2 variants with SARS-
416 CoV-2 Spike protein for structure-based drug design. Sci Data 2020;7:309.
417 https://doi.org/10.1038/s41597-020-00652-6.
418 [12] Laurini E, Marson D, Aulic S, Fermeglia M, Pricl S. Computational Alanine
419 Scanning and Structural Analysis of the SARS-CoV-2 Spike Protein/Angiotensin-
420 Converting Enzyme 2 Complex. ACS Nano 2020;14.
421 https://doi.org/10.1021/acsnano.0c04674.
422 [13] Gheeraert A, Vuillon L, Chaloin L, Moncorgé O, Very T, Perez S, et al. Singular
423 Interface Dynamics of the SARS-CoV-2 Delta Variant Explained with Contact
424 Perturbation Analysis. J Chem Inf Model 2022;62:3107–22.
425 https://doi.org/10.1021/ACS.JCIM.2C00350.
426 [14] Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web
427 server: An online force field. Nucleic Acids Res 2005;33.
428 https://doi.org/10.1093/nar/gki387.

[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 2019;431. https://doi.org/10.1016/j.jmb.2019.02.028.

[16] Zhang N, Chen Y, Lu H, Zhao F, Alvarez RV, Goncearenco A, et al. MutaBind2: Predicting the Impacts of Single and Multiple Mutations on Protein-Protein Interactions. IScience 2020;23. https://doi.org/10.1016/j.isci.2020.100939.

[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 2020;36. https://doi.org/10.1093/bioinformatics/btz926.

[18] Van Zundert GCP, Rodrigues JPGLM, Trellet M, Schmitz C, Kastritis PL, Karaca E, et al. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. J Mol Biol 2016. https://doi.org/10.1016/j.jmb.2015.09.014.

[19] Honorato R V., Koukos PI, Jiménez-García B, Tsaregorodtsev A, Verlato M, Giachetti A, et al. Structural Biology in the Clouds: The WeNMR-EOSC Ecosystem. Front Mol Biosci 2021;8. https://doi.org/10.3389/fmolb.2021.729513.

[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 2021;37. https://doi.org/10.1093/bioinformatics/btaa708.

[21] Krissinel E, Henrick K. Inference of Macromolecular Assemblies from Crystalline State. J Mol Biol 2007;372:774–97. https://doi.org/10.1016/j.jmb.2007.05.022.

[22] Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature 2020;581:215–20. https://doi.org/10.1038/s41586-020-2180-5.

[23] Wan Y, Shang J, Graham R, Baric RS, Li F. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. J Virol 2020;94. https://doi.org/10.1128/jvi.00127-20.

[24] Schrödinger LLC and DW. PyMOL n.d.

[25] Integrated Development Environment for R. RStudio n.d.

[26] A Language and Environment for Statistical Computing. R n.d.

[27] Kolde R. Pheatmap: Pretty Heatmaps 2019.

[28] Neuwirth E. RColorBrewer: ColorBrewer Palettes 2014.

[29] Lin Z, Long H, Bo Z, Wang Y, Wu Y. New descriptors of amino acids and their application to peptide QSAR study. Peptides (NY) 2008;29:1798–805. https://doi.org/10.1016/j.peptides.2008.06.004.

[30] Eisenberg D, Schwarz E, Komaromy M, Wall R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J Mol Biol 1984;179:125–42. https://doi.org/10.1016/0022-2836(84)90309-7.

[31] Shapovalov MV, Dunbrack RL. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. Structure 2011;19:844–58. https://doi.org/10.1016/j.str.2011.03.019.

[32] Gao M, Skolnick J. iAlign: a method for the structural comparison of protein–protein interfaces. Bioinformatics 2010;26:2259–65. https://doi.org/10.1093/bioinformatics/btq404.
[33] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 1997;25. https://doi.org/10.1093/nar/25.17.3389.
[34] Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. J Am Chem Soc 2003;125:1731–7. https://doi.org/10.1021/ja026939x.
[35] Van Rossum G, Drake FL, Harris CR, Millman KJ, van der Walt SJ, Gommers R, et al. Python 3 Reference Manual. vol. 585. 2009.
[36] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020;585. https://doi.org/10.1038/s41586-020-2649-2.
[37] Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, et al. Jupyter Notebooks—a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas - Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016, 2016. https://doi.org/10.3233/978-1-61499-649-1-87.
[38] Anaconda Software Distribution. Anaconda Documentation 2020.
[39] The pandas development team. Pandas-dev/pandas: Pandas. Zenodo 2020.
[40] Waskom M. seaborn: statistical data visualization. J Open Source Softw 2021;6. https://doi.org/10.21105/joss.03021.