

T.C.
DOKUZ EYLÜL UNIVERSITY
İZMİR INTERNATIONAL BIOMEDICINE AND GENOME
INSTITUTE

**BENCHMARKING THE BINDING AFFINITY
PREDICTORS OVER THE SPIKE:ACE2 DEEP
MUTAGENESIS DATASET**

EDA ŞAMİLOĞLU TENGRŞEK

DEPARTMENT OF MOLECULAR BIOLOGY AND
GENETICS

MASTER OF SCIENCE THESIS

İZMİR-2022

THESIS CODE: DEU.IBG.MSc-2019850031

T.C.
DOKUZ EYLÜL UNIVERSITY
IZMIR INTERNATIONAL BIOMEDICINE AND GENOME
INSTITUTE

**BENCHMARKING THE BINDING AFFINITY
PREDICTORS OVER THE SPIKE:ACE2 DEEP
MUTAGENESIS DATASET**

DEPARTMENT OF MOLECULAR BIOLOGY AND
GENETICS

MASTER OF SCIENCE THESIS

EDA ŞAMİLOĞLU TENGİRŞEK

Supervising Faculty Member: Asst. Prof. Ezgi KARACA EREK

THESIS CODE: DEU.IBG.MSc-2019850031

Dokuz Eylül Üniversitesi İzmir Uluslararası Biyotıp ve Genom Enstitüsü
Genom Bilimleri ve Moleküler Biyoteknoloji Anabilim Dalı,
Moleküler Biyoloji ve Genetik Yüksek Lisans programı öğrencisi
Eda Şamiloğlu Tengirşek
**'BAĞLANMA AFİNİTE TAHMİN ALGORİTMALARININ SPIKE:ACE2
DERİN MUTAGENEZ BAĞLANMA VERİLERİ ÜZERİNDE
DEĞERLENDİRİLMESİ'** konulu Yüksek Lisans tezini 10.05.2022
tarihinde başarılı olarak tamamlamıştır.

BAŞKAN

Dr. Öğr. Üyesi Ezgi KARACA EREK

Dokuz Eylül Üniversitesi
İzmir Biyotıp ve Genom Enstitüsü

ÜYE

Doç. Dr. Gökhan KARAKÜLAH
Dokuz Eylül Üniversitesi

ÜYE

Doç. Dr. Burcu KAPLAN TÜRKÖZ
Ege Üniversitesi

YEDEK ÜYE

Dr. Öğr. Üyesi Yavuz OKTAY
Dokuz Eylül Üniversitesi

YEDEK ÜYE

Dr. Öğr. Üyesi İsmail Hakkı AKGÜN
Ege Üniversitesi

TABLE OF CONTENTS.....	PAGE NUMBER
INDEX OF TABLES	iv
INDEX OF FIGURES	v
LIST OF ABBREVIATIONS	viii
ACKNOWLEDGEMENTS	ix
ABSTRACT	1
ÖZET	2
1. INTRODUCTION AND AIM	3
1.1. Statement and Importance of the Problem	3
1.2. Aim of the Study	3
1.3. Hypothesis of the Study.....	3
2. GENERAL INFORMATION	5
2.1. Coronaviruses Have Led Serious Pandemics in the 21st century.....	5
2.1.1 SARS-CoV-2: the agent of COVID-19 pandemic	7
2.2 The Structural Knowledge about SARS-CoV-2 proteins.....	8
2.2.1. Structure of SARS-CoV-2 Spike Protein	13
2.2.1.1. ACE2: the Receptor Protein of Spike.....	15
2.3. Close up into the SARS-CoV-2 RBD-ACE2 Interface	17
2.3.1 Mutations on the RBM Alter Binding Affinity of SARS-CoV-2 RBD with human ACE2	18
2.4 Deep Mutational Scanning Approach and Its Usage in the Case of COVID-19	20
2.4.1. Deep Mutational Scanning Experiment of ACE2.....	22
2.4.2. Deep Mutational Scanning Experiment of RBD	22
2.4.3. Studies Relying on Deep Mutational Scanning Dataset of ACE2 and RBD	22
2.5. Investigating the Effect of Variants on ACE2 and RBD Binding with Computational Tools	23
3. MATERIALS AND METHOD.....	26
3.1. Type of Study.....	26

3.2. Time and Location of Study	26
3.3. Population and Sample of Study.....	26
3.4. Materials of Study	26
3.5. Variable of Study	27
3.6. Tools for Data Collection	27
3.6.1. Web servers & resources used for data collection	27
3.6.2. Methodology	29
3.6.2.1 Obtaining the Experimentally Determined Structures of ACE2 and RBD Complex.....	30
3.6.2.2. Construction of the Benchmarking Dataset.....	30
3.6.2.3 Investigating Evolutionary Conservation of ACE2 and RBD Interface	32
3.6.2.4. Working Principles of Structure-Based Binding Affinity Predictors Used in This Study ..	33
3.6.2.4.1 HADDOCK-Based Prediction.....	34
3.6.2.4.2 UEP-Based Prediction.....	37
3.6.2.4.3 FoldX-Based Prediction	38
3.6.2.4.4 EvoEF1-Based Prediction.....	39
3.6.2.4.5. MutaBind2-Based Prediction	40
3.6.2.4.6. SSIPe-Based Prediction.....	41
3.6.2.4. Evaluating Prediction Performance of the Tools.....	41
3.6.2.5. Investigating Underlying Metrics of Prediction Performances	43
3.7. Study Plan and Calendar.....	45
3.8. Data Evaluation	46
3.9. Limitations of Study.....	46
3.10. Ethics Committee Approval	46
4. RESULTS	47
4.1. ACE2 and RBD Complex Has a Non-conserved Interface	47
4.2. Investigating Experimental Binding Value Distribution of Benchmarking Dataset	48
4.3. Comparing the Prediction Capacity of Each Tool.....	49
4.3.1 Examining Performances of Each Tool on UEP Dataset.....	52
4.4 Metric Analysis to Investigate Performance of Predictors	53
4.4.1. Examining Tool's Performances According to Quantitative Metrics: van der Waals Volume, Hydrophobicity, and Flexibility	54

4.4.2. Examining Tool's Performances According to Side Chain Properties of Amino Acids: Polar, Apolar, and Charge	59
5. DISCUSSION	61
6. CONCLUSION AND SUGGESTIONS.....	65
7. REFERENCES	67
APPENDIX A.....	78
APPENDIX B	87
APPENDIX C	87
APPENDIX D.....	89

INDEX OF TABLES

Table 1: Spike mutations of SARS-CoV-2 VOIs and VOCs are listed below. Deletional modifications are not included in the list since they are out of the scope of this study.	19
Table 2: Benchmarking dataset distribution is presented according to protein type (ACE2/RBD) and mutation effect (enriching/depleting).	31
Table 3: The UEP dataset distribution is presented according to protein type (ACE2/RBD) and mutation effect (enriching/depleting).	37
Table 4: Predicted binding affinity change scores of N501Y mutations and the success score of the predictors.	42
Table 5: The summary statistics of ConSurf scores of ACE2-RBD interface residues.	48
Table 6: Tools' success rates (%) were presented on the overall benchmarking dataset and its subsets. The highest and the lowest performances are colored red and green, respectively.	50
Table 7: Tools' success rates (%) were presented on the overall UEP dataset and its subsets. The highest and the lowest performances were colored red and green, respectively.	53
Table 8: The Δ Success rates of predictors were presented according to the metrics. The observed moderate biases were depicted in red	58

INDEX OF FIGURES

Figure 1: Coronaviruses have infected diverse species **A)** Coronaviruses caused three pandemics in the twenty-first century **B)** SARS-CoV, MERS-CoV, and SARS-CoV-2 are members of beta coronaviruses. Alpha and beta coronaviruses mainly infect mammals and humans. Gamma and delta coronaviruses infect birds and fish. Genres of coronaviruses are listed, and human pathogenic ones are indicated with asterisks.

.....6

Figure 2: The cryo-electron tomogram of the SARS-CoV-2 virus with spike and nucleocapsid proteins.....9

Figure 3: Genome organization and proteins of SARS-CoV-2. Non-structural proteins were depicted in green, structural proteins in orange and accessory proteins in blue.....11

Figure 4: Spike protein of SARS-CoV-2 mediates viral fusion. **A)** Cartoon representation showing viral membrane fusion. **B)** Structures of the homotrimeric transmembrane spike protein. **C)** After ACE2 binding, the spike is cleavage from S1/S2 cleavage site, and S2 undergoes postfusion conformation 14

Figure 5: ACE2 protein is a member of the renin-angiotensin hormone system in the human body. **A)** ACE2-Ang (1-7) axis is the antagonist of the ACE-Ang2 axis; these axes regulate homeostasis. **B)** Domain organization of human ACE2 complex with B⁰AT1. Spike uses ACE2 for attaching to the host cell 16

Figure 6: The interface of the RBD and human ACE2 complex. **A)** Contact residues were shown in the sticks and colored in yellow. **B)** Residues that mediate the formation of hydrogen bonds and salt bridges were labeled with residue names and depicted in yellow 18

Figure 7: Spike point mutations of SARS-CoV-2 variants (VOCs and VOIs) depicted on the structure in the sticks and colored in red. All point mutations are localized in RBD, specifically near RBM. 19

Figure 8: The schematic representation of the deep mutational scanning experiment workflow 21

Figure 9: The benchmarking datasets of ACE2 and RBD are provided as a heat map and colored with experimental binding values. The blue color corresponds to depleting mutations, and the red corresponds to enriching mutations. 32

Figure 10: The schematic representation of folder and file organization of this study. **A)** Tools were classified according to web-based or stand-alone. File and script organizations were provided. **B)** Predicted score files were deposited in the "input files" subfolder and used for preparing the benchmarking dataset in the desired format for analyses. The benchmarking and UEP datasets were located in the "output files" subfolder. Python notebooks and shell scripts were deposited in the "notebooks" and "shell scripts" subfolders, respectively. Variant models generated by tools were deposited in the "mutant models" folder. 34

Figure 11: The models of N501Y mutation. The side chain of wild-type amino acid Asparagine was depicted in sticks colored as white. Mutant amino acid Tyrosinewase was modeled by FoldX, FoldXwater, HADDOCK, EvoEF1, MutaBind2, and SSIPe; side chains of Asparagine were shown in sticks colored as yellow, orange, blue, green, red, and purple, respectively. 42

Figure 12: The mutation from Asparagine and Tyrosine at the 501st position was depicted, and corresponding wild-type and mutant metric values were presented 43

Figure 13: Schematic representation of amino acid side chain classes and labels of interchange of the categories.....45

Figure 14: The evolutionary conservation analyses of the ACE2-RBD complex. **A)** Conservation scores project onto the structure with a specific color code; green represents variable residues, and purple represents conserved residues. **B)** Conservation score distributions of ACE2 and RBD interface residues were presented. 47

Figure 15: The experimental binding values of the ACE2 and RBD benchmarking datasets were depicted 49

Figure 16: Performances of HADDOCK, FoldX, EvoEF1, MutaBind2, and SSIPe were presented with bar plots according to the overall dataset (gray), enrichment dataset (red), and depletion dataset (blue). The number of mutations for each dataset is depicted in the label. The significantly low performances were shown with an arrow. FoldXwater was discarded since it has a very similar performance to FoldX. 51

Figure 17: The success rates of tools were presented in terms of probability density over the volume, hydrophobicity, and flexibility metrics. The gray line represents the experimental dataset. The blue, yellow, orange, green, red, and purple

lines represent HADDOCK, FoldX, FoldXwater, EvoEF1, MutaBind2, and SSIPe, respectively.....56

Figure 18: Mutations were classified into five categories; no change, polarity gain, polarity loss, change gain, and charge loss on successfully predicted enrichment (light colors) and depletion (dark colors) datasets. Blue, yellow, orange, green, red, and purple represents the HADDOCK, FoldX, FoldXwater, EvoEF1, MutaBind2, and SSIPe, respectively.....60



LIST OF ABBREVIATIONS

Å	Angstrom
ACE2	Angiotensin Converting Enzyme-2
COVID-19	Coronavirus Disease 19
CT	Cytoplasmic Tail
DPP4	Dipeptidyl-peptidase 4
FP	Fusion Peptide
FACS	Fluorescence-Activated Cell Sorting
HADDOCK	High Ambiguity Driven biomolecular DOCKing
H	Hypothesis
IBG HPC	Izmir Biomedicine and Genome Center High-Performance Computing
Mpro	Main Protease
MM-PBSA	Molecular Mechanics Poisson-Boltzmann Surface Area
NTD	Amino Terminal Domain
OFRs	Open Reading Frames
PDB	Protein Data Bank
PDBePISA	Proteins, Interfaces, Structures, and Assemblies
RBD	Receptor Binding Domain
RCSB PDB	Research Collaboratory for Structural Bioinformatics PDB
S1	Subunit 1
S2	Subunit 2
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
TM	Transmembrane Domain
vdW	van der Waals
WHO	World Health Organization
VOC	Variants of Concern
VOI	Variants of Interest

ACKNOWLEDGEMENTS

Foremost, huge thanks to my thesis advisor, Dr. Ezgi Karaca for giving me the chance to do research in her group and for her patience and support during the time we had.

I would like to thank all my beautiful lab mates Burcu Özden, Büşra Savaş, Berçin Barlas, and Mehdi Koşaca for such a pleasant work environment, coffee and lunch breaks, and also their support.

I would like to thank the Health Institutes of Turkey (TUSEB) for financially supporting me during my master's studies.

I also thank Emre Tengirşek, Elif Arık, and Zeynep Güngör for their unconditional love and support. Life is sweeter and more cheerful with you.

Last but not least, I would like to thank all open access data servers, especially wwPDB. I am grateful for all the free online scientific events that I had a chance to attend during the Corona times to broaden my perspective. And also, I thank YouTube and SciHub for making my life easier.

BENCHMARKING THE BINDING AFFINITY PREDICTORS OVER THE SPIKE:ACE2 DEEP MUTAGENESIS DATASET

Eda Şamiloğlu Tengirşek, Izmir International Biomedicine and Genome Institute, Dokuz Eylul University Health Campus, Balçova, 35340, Izmir / Turkey

ABSTRACT

COVID-19 is caused by the SARS-CoV-2 virus. The SARS-CoV-2 virus binds to the host receptor protein ACE2 via its spike protein to initiate viral entrance and viral infection. Because of their critical roles, these proteins have been widely investigated by experimental and computational approaches. Relatedly, ACE2 and spike variant binding rates were obtained by the deep mutational scanning experiments. The deep mutational scanning method provides a high throughput binding dataset and thereby facilitates benchmarking of computational tools. By taking advantage of the deep mutagenesis datasets of ACE2 and spike, we aimed to investigate the performances in predicting the impact of interfacial mutations of FoldX, EvoEF1, MutaBind2, and SSIpe as main predictors in addition to HADDOCK and UEP as naïve predictors. In addition to their overall performance analysis, we performed metric-based analysis using volume, hydrophobicity, flexibility changes upon mutations, and change of side chain properties of mutations. As a result, we found that FoldX has the highest performance with a 64% success rate. Additionally, we highlighted that none of the predictors performed well on the binding increasing mutations. As a result, we revealed that (i) the top-ranking predictor, FoldX, predicts most of the hydrophobicity increasing mutations as affinity depleting, (ii) HADDOCK tends to tag mutations with increasing volume as affinity enhancing, (iii) Conservation-based tools, MutaBind2 and SSIpe predict most of the mutations as affinity depleting. All in all, we concluded that classical binding affinity predictors are not yet sufficient to predict binding affinity changes across the interface of the host-pathogen protein system SARS-CoV-2 spike -ACE2.

Keywords: COVID-19, SARS-CoV-2, spike, ACE2, Binding Affinity, FoldX

BAĞLANMA AFİNİTE TAHMİN ALGORİTMALARININ SPIKE:ACE2 DERİN MUTAGENEZ BAĞLANMA VERİLERİ ÜZERİNDE DEĞERLENDİRİLMESİ

Eda Şamiloğlu Tengirşek, İzmir Uluslararası Biyotıp ve Genom Enstitüsü,
Dokuz Eylül Üniversitesi, Sağlık Kampüsü, Balçova, 35340, İzmir / Türkiye

ÖZET

COVID-19, SARS-CoV-2 virüsünün neden olduğu son küresel salgındır. SARS-CoV-2 spike proteini aracılığıyla, viral girişi ve daha sonra viral enfeksiyonu başlatmak için partner proteini ACE2'ye bağlanır. ACE2 ve spike bu kritik görevleri nedeniyle deneysel ve hesaplamalı yaklaşımlarla geniş çapta araştırılmıştır. Ayrıca, derin mutasyonel tarama yöntemi kullanılarak ACE2 ve spike varyant bağlanma değerleriden deneysel olarak elde edilmiştir. Bu yöntem, yüksek ölçekli bağlama verisi sağlar ve hesaplamalı araçların performanslarını kıyaslama fırsatı sunar. Bu tezde, ACE2 ve spike'ın derin mutagenез veri setlerinden yararlanarak yapı tabanlı bağlanma afinitesi tahmin eden araçlardan FoldX, EvoEF1, MutaBind2, SSIPe'nin ve naif araç olarak HADDOCK ve UEP'nin performanslarını araştırdık. Ek olarak, mutasyonları hacim, hidrofobiklik, esneklik değişimleri ve mutasyonların yan zincir özelliklerinin değişimlerine göre inceleyerek metrik analizleri yaptık. Analizlerimiz sonucunda, FoldX'in %64 başarı oranı ile en yüksek performansa sahip olduğunu belirledik. Bunun yanı sıra araçlardan hiçbirinin bağlanma afinitesini arttıran mutasyonlar üzerinde iyi performans göstermediğini ortaya koyduk. Ayrıca metrik analizleri sonucunda araçların farklı seviyelerde eğilimleri olduğunu tespit ettik. En iyi performans gösteren araç, FoldX'in, hidrofobisitesi artan mutasyonların bağlanma afinitesi azalıyormuş gibi tahmin ettiğini, HADDOCK'un hacimce büyüyen mutasyonları bağlanma afinitesini arttırıyormuş gibi tahmin ettiğini bulduk. Evrimsel korunmuşluk temelli araçlardan MutaBind2 ve SSIPe'nin mutasyonları bağlanma afinitesini azaltıyormuş gibi tahmin ettiğini bulduk. Sonuç olarak, klasik bağlanma afinitesi tahmin araçlarının, SARS-CoV-2 spike ve ACE2 gibi konak-patojen protein sisteminin arayüzlerindeki bağlanma afinitesi değişikliklerini tahmin etmeye henüz uygun olmadığı kanısına vardık.

Anahtar Kelime: COVID-19, SARS-CoV-2, spike, ACE2, Bağlanma Afinitesi, FoldX

1. INTRODUCTION AND AIM

1.1. Statement and Importance of the Problem

COVID-19 is a global pandemic caused by the SARS-CoV-2 virus (Lu et al., 2020a) since late 2019, it remains to be a burden on the public health, and the global economy. The SARS-CoV-2 virus binds to the human angiotensin-converting enzyme 2 (ACE2) via its receptor-binding domain (RBD) of spike glycoprotein. This interaction promotes membrane fusion of the virus to the host cell and initiates viral infection (W. Li et al., 2003). Therefore, clarifying this critical interaction mechanism is highly important. To this end, deep mutational scanning experiments have been carried out for both RBD and ACE2 to reveal the contribution of a point mutations to their binding (Chan et al., 2020; Starr et al., 2020). Even though deep mutagenesis experiments are invaluable for deciphering the interaction mechanism of RBD and ACE2, they are time-consuming, costly, and manually intensive methods. On the other hand, computational approaches can predict the mutation effect on binding time and cost-effectively. Therefore, we investigated the capacity of several computational methods over the deep mutagenesis datasets of ACE2 and RBD in this thesis.

1.2. Aim of the Study

In this thesis, we aimed to reveal the performance of widely used structure-based binding affinity predictors (FoldX, EvoEF1, Mutabind2, and SSIPE as main and HADDOCK and UEP as naïve) on the deep mutagenesis point mutation interaction dataset of RBD and ACE2.

1.3. Hypothesis of the Study

H1: Structure-based binding affinity predictors could reliably predict binding affinity changes on the RBD and ACE2 complex caused by interface mutations.

H2: Differences in mutant and wild-type binding affinity values ($\Delta G_{\text{mutant}} - \Delta G_{\text{wild-type}}$) of RBD and ACE2 complex reflect the impact of the mutation on binding ($\Delta\Delta G$). So, the predicted binding score could be correlated with the experimental binding affinity of the variant complex obtained from the deep mutational scanning experiment.

H3: This system-specific benchmarking on RBD and ACE2 interface residues reflects the general performances of structure-based binding affinity prediction tools.



2. GENERAL INFORMATION

2.1. Coronaviruses Have Led Serious Pandemics in the 21st century

Coronaviruses are a member of the subfamily Coronavirinae in the Coronaviridae family (Chen, Liu, & Guo, 2020). Viruses of the Coronaviridae family generally have single-stranded, positive-sense RNA characterized by club-like spike proteins that envelope the virus's surface. Coronaviruses have a large ~30 Kb genome with the 5'-cap structure and the 3'-poly-A tail. This genomic arrangement allows coronaviruses to have a unique replication strategy, with positive-sense RNA acting as an mRNA for replicase protein translation. (Yang & Wang, 2020). The coronavirinae subfamily contains four coronavirus genres: alpha, beta, gamma, and delta. Mammals and humans are infected mainly by alpha coronaviruses and beta coronaviruses. Whereas gamma and delta coronaviruses infect birds and fish. (Figure 1.A). The first coronavirus was isolated in chickens in 1937. Twenty-three years later, the first human coronavirus was identified (Yang & Wang, 2020). Before 2019, six human-pathogenic coronaviruses have been identified: HCoV-229E, HCoV-OC43, HCoV-NL63, and HKU1 caused mild upper respiratory disease; SARS-CoV and MERS-CoV infect the lower lungs and cause a severe respiratory syndrome in human (Figure 1.A).

Coronaviruses have caused three highly pathogenic zoonotic diseases during the twenty-first century (Figure 1.B). Firstly, SARS-CoV emerged in the fall of 2002 in southern China and caused the Severe Acute Respiratory Syndrome (SARS), which spread to 29 countries with a fatality rate of ~10% (8422 confirmed cases and 916 deaths) (Cherry & Krogstad, 2004; Kuiken et al., 2003). SARS epidemic was stopped before it became a pandemic disease, thanks to a robust international response based on quarantine procedures. According to the studies on the zoonotic origin of SARS-CoV, the viral genome isolated from palm civet has 99.8% homologous to the human SARS CoV (Cherry & Krogstad, 2004). Additional studies suggest that it serves as an intermediate host between an unidentified animal reservoir and humans (Salata, Calistri, Parolin, & Palù, 2020). SARS-CoV uses angiotensin-converting enzyme 2 (ACE2) to attach host cells such as HCoV-NL63. (Fehr & Perlman, 2015).

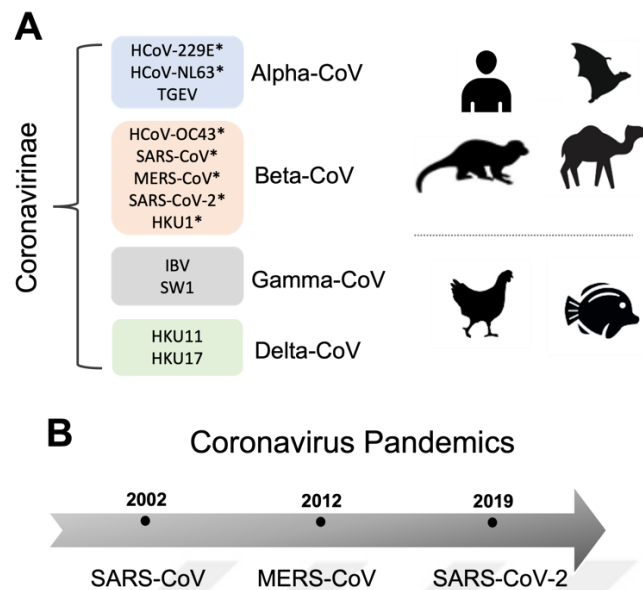


Figure 1: Coronaviruses have infected diverse species **A)** Coronaviruses caused three pandemics in the twenty-first century **B)** SARS-CoV, MERS-CoV, and SARS-CoV-2 are members of beta coronaviruses. Alpha and beta coronaviruses mainly infect mammals and humans. Gamma and delta coronaviruses infect birds and fish. Genres of coronaviruses are listed, and human pathogenic ones are indicated with asterisks.

After the SARS-CoV epidemic in 2002, a novel coronavirus MERS-CoV was identified in an older male in Saudi Arabia (Moh Zaki, Van Boheemen, Bestebroer, Osterhaus, & Fouchier, 2012). MERS-CoV led to an epidemic in the Middle East and spread to 26 intercontinental countries. This pandemic is known as the Middle East Respiratory Syndrome (MERS), characterized by high fever, cough, acute respiratory tract infection, expectorations, shortness of breath, and multiorgan dysfunction (Durai, Batool, Shah, & Choi, 2015). MERS has an increased fatal rate of ~35% (1,728 confirmed cases and 624 deaths) (De Wit, Van Doremalen, Falzarano, & Munster, 2016). According to several studies on MERS-CoV, dromedary camels were identified as the main reservoir for MERS-CoV (Chu et al., 2014; Reusken et al., 2013; W. et al., 2018). However, it is widely accepted that bats serve as a reservoir for SARS-CoV and MERS-CoV since bat coronaviruses' genomes have high sequence similarities with these viruses. (De Wit et al., 2016; Salata et al., 2020). MERS-CoV attaches to host

human cells via the dipeptidyl-peptidase 4 (DPP4) protein, a different receptor than SARS-CoV. (Stalin Raj et al., 2013).

2.1.1 SARS-CoV-2: the agent of COVID-19 pandemic

Since late 2019, Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has emerged and spread fast all over the world and caused Coronavirus disease 2019 (COVID-19) (Platto, Xue, & Carafoli, 2020). The outbreak's origin was defined as Wuhan Huanan Seafood Wholesale Market in China, as many of the first confirmed cases had direct or indirect contact with the market (Yang & Wang, 2020). Symptoms of SARS-CoV-2 infection are fever, cough, difficulties in breathing, and eventually severe pneumonia. SARS-CoV-2 is a member of beta coronavirus like the SARS-CoV and MERS-CoV. SARS-CoV-2 shows 79% sequence similarity with SARS-CoV, and they both use the same receptor, ACE2, to initiate viral entry (Lu et al., 2020b). Notably, a whole-genome comparison between SARS-like coronavirus from the bat (BatCoV RaTG13) and SARS-CoV-2 shows 96% similarity, suggesting that SARS-CoV-2 is mainly originated from bats (P. Zhou et al., 2020). Although the SARS-CoV and MERS-CoV affect thousands of people, SARS-CoV-2 has infected over 500 million people and caused the death of over six million people so far (WHO coronavirus (COVID-19) dashboard. Retrieved May 13, 2022, from <https://covid19.who.int>). So, SARS-CoV-2 has been posing a significant threat to public health and the global economy.

SARS-CoV-2 has been progressively evolving into new fast-spreading variants through evolutionary pressure, despite its having a low mutation-rated exonuclease (Lippi, Mattiuzzi, & Henry, 2022). SARS-CoV-2 variants are classified into Variants of Interests (VOIs) and Variants of Concerns (VOCs), depending on the extent of increased transmissibility, virulence, and decreased therapeutic response (Dubey, Choudhary, Kumar, & Tomar, 2021). According to the World Health Organization (WHO) dynamic nomenclature system of SARS-CoV-2 variants; Lambda and Mu are designated as VOIs, and Alpha, Beta, Gamma, Delta, and Omicron are defined as VOCs (Konings et al., 2021) (Tracking sars-COV-2 variants. World Health

Organization. Retrieved May 14, 2022, from <https://www.who.int/en/activities/tracking-SARS-CoV-2-variants>).

Accumulating mutations in the SARS-CoV-2 genome result in variants having varying replication dynamics, pathogenicity, and infectivity. Most importantly, spike protein mutations have been posing great interest as this protein mediates the viral entrance of the host cell and is mainly targeted by the humoral immune cells (P. Zhou et al., 2020). Among the SARS-CoV-2 variants, Omicron has heavily mutated spike protein with a mutation frequency of >1% (B. Xia et al., 2022). Omicron spike mutations are G339D, R346K, S371L, S373P, S375F, T376A, D405N, R408S, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, and Y505H. Due to these mutations, Omicron is associated with 2.4-fold higher transmissibility (Lippi et al., 2022) and 8.4 fold humoral immune escape potential (L. Zhang et al., 2022) compared to wild-type SARS-CoV-2.

2.2 The Structural Knowledge about SARS-CoV-2 proteins

SARS-CoV-2 genome contains fourteen functional open reading frames (ORFs) with two noncoding regions at both ends and multiple regions at the 3' end for structural, non-structural, and accessory proteins (Lu et al., 2020b). Following the viral infection, the ribosomal machinery synthesis proteins within the 3' end of the viral genome. SARS-CoV-2 has 29 proteins: four structural, six accessory, and sixteen non-structural (Bai, Zhong, & Gao, 2022). Structural proteins of coronavirus are spike, membrane, envelope, and nucleocapsid proteins that encapsulate the viral genome within a lipid bilayer membrane to form a virus envelope. Non-structural (Nsp 1-16) and accessor proteins assist viral RNA synthesis and evasion of cell defenses and the immune response of the host, respectively (Yang & Wang, 2020).

Due to the substantial impact of the COVID-19 pandemic on public health, structures of SARS-CoV-2 proteins were rapidly determined and released into the Protein Data Bank (PDB). The molecular architecture of SARS-CoV-2 was revealed using cryo-electron tomography with 12Å subtomogram averaging and deposited at

Electron Microscopy Data Bank with EMD-30430 ID (Yao et al., 2020). According to this tomogram, SARS-CoV-2 has an ellipsoidal shape with a diameter of 108 nm. SARS-CoV-2 has spike proteins with different conformations on the outer surface and RNA-associated nucleocapsid proteins inside (Figure 2). First, the SARS-CoV-2 main protease (ProM) structure was determined with X-ray crystallography and released in the PDB two weeks after the first viral genome was sequenced (J. Zhang & Chen, 2022). To date, 2124 SARS-CoV-2 structures have been deposited in the PDB archive (RCSB PDB. Retrieved May 14, 2022, from <https://bit.ly/PDB-SARS-CoV-2>). These structures are further used to reveal the structural and functional properties of the virus and provide critical information for drug development and vaccine design (J. Zhang & Chen, 2022).

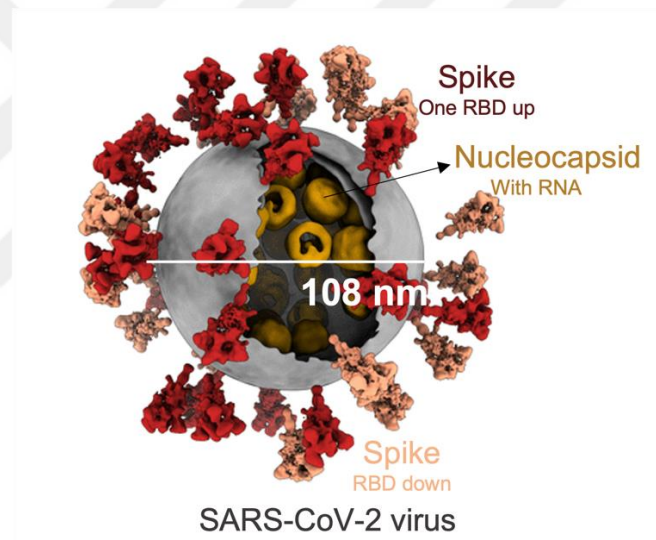


Figure 2: The cryo-electron tomogram of the SARS-CoV-2 virus with spike and nucleocapsid proteins.

More than half of the SARS-CoV-2 genome (~70%) consists of ORF1a and ORF1b within the 5'-region of the viral RNA genome. These regions of the viral genome encode sixteen non-structural proteins (Nsp1-16). Non-structural proteins conduct the enzymatic process in the cell, including viral genome replication, transcription, and translation. Nsp1 acts as a host shutoff factor by specifically binding to 40S host cell ribosome and suppressing host gene expression. The nsp1 of SARS-CoV and SARS-CoV-2 shares 84% sequence homology. This observation indicates that nsp1 is a

conserved protein. (Matthias et al., 2020). The crystal structure of nsp1 was determined and released to the PDB with an ID of 7K3N (Semper, Watanabe, & Savchenko, 2021). Nsp1 protein is also a target protein for therapeutics since blocking the interaction of nsp1 and ribosome would reinforce the immune response and impede viral survival (Arya et al., 2021). Nsp2 proteins are the most variable non-structural protein among the coronaviruses, and viral growth is slowed in the absence of this protein. This observation may explain the increased contagious between SARS-CoV and SARS-CoV-2 (Graham, Sims, Baric, & Denison, 2006). The structure of the nsp2 was determined using cryo-EM and released to the PDB with 7MSW ID (Gupta et al., 2021). Nsp3 protein is the largest non-structural protein with 213 kD molecular mass, and nsp3 of SARS-CoV and SARS-CoV-2 show 75% sequence similarity. Nsp3 interaction with other non-structural proteins promotes the construction of the replication-transcription complex (Bai et al., 2022). The structure of nsp3 was determined with the X-ray diffraction method, and the structure was deposited in the PDB with 6WEY ID (Frick, Viridi, Vuksanovic, Dahal, & Silvaggi, 2020). Nsp4 has a role in coronavirus replication by mediating the interaction with cellular membranes. There is no experimentally determined structure of nsp4 in PDB (Bai et al., 2022). Nsp5 is a homodimer protein that shows catalytic activity and is widely known as the main protease (Mpro). The Mpro cuts to ORF1a and ORF1b from eleven conserved cleavage sites. Due to this critical role, Mpro is one of the targets for drug development studies. Structure of nsp5 protein present in the PDB with 6LU7 ID (Jin et al., 2020). Nsp6 is the endoplasmic reticulum transmembrane protein that has a role in generating autophagosomes (Bai et al., 2022). There is no experimental structure for nsp6 protein. Nsp7, nsp8, nsp12, and nsp13 are present in the SARS-CoV-2 as a polymerase complex that catalyzes the synthesis of viral RNA. Nsp7 and nsp8 act as a cofactor of nsp12, which is highly conserved between SARS-CoV, and SARS-CoV-2 (with a ~96% sequence homology) (Dong et al., 2020). Nsp13 regulates helicase and ATP activity and is almost identical between SARS-CoV and SARS-CoV-2 (~99% sequence homology) (G. Yan et al., 2020). Nsp9 is a dimeric nucleic acid-binding protein encoded from ORF1a and has a role in RNA replication. This protein is highly conserved between SARS-CoV and SARS-CoV-2 with a 97% sequence identity (Arya et al., 2021). The crystal structure of the nsp9 protein is deposited in the PDB with

6WXD ID (Littler, Gully, Colson, & Rossjohn, 2020). Nsp10, nsp14, and nsp16 proteins connect with other non-structural proteins and have enzymatic roles in viral genome replication and immune evasion. These proteins are highly similar between SARS-CoV and SARS-CoV-2 (>%90 sequence homology) (Bai et al., 2022). Nsp10 and nsp16 complex structures and the structure of nsp14 were deposited to the PDB with 6WVN and 7N0C IDs, respectively (Chang et al., 2021; Monica et al., 2020). However, the function and structure of nsp11 are still unknown. Last but not least, nsp15, known as Uridylate-specific endoribonuclease, has a role in viral RNA degradation. The structure of nsp15 was determined and deposited in the PDB with 6WXC ID (Kim et al., 2021). All resolved non-structural proteins of SARS-CoV-2 were shown in Figure 3 (green).

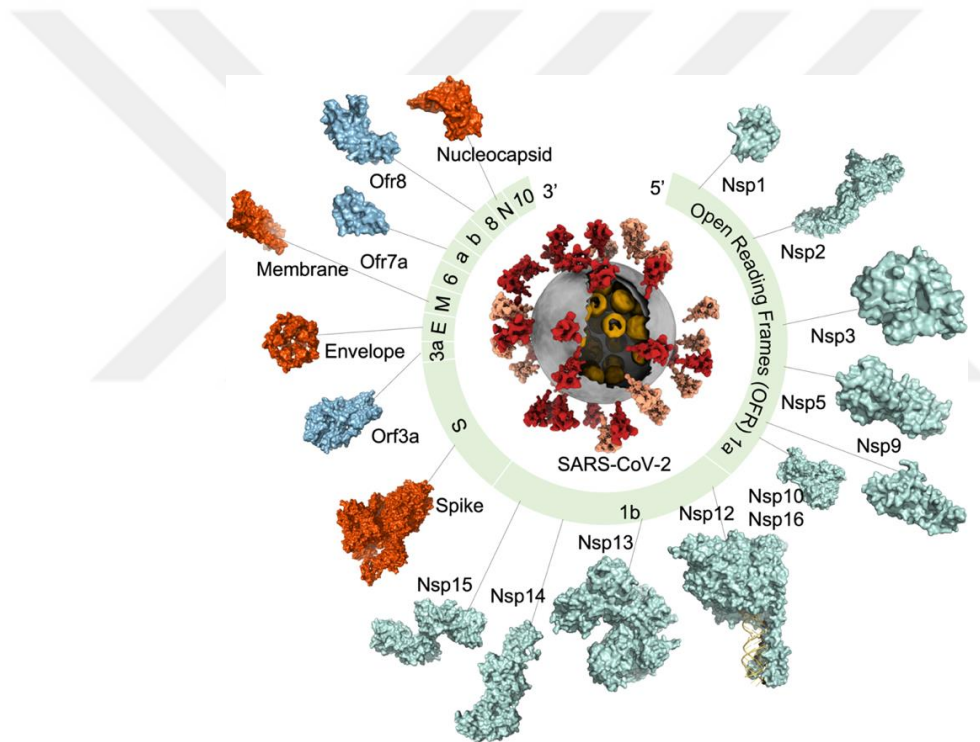


Figure 3: Genome organization and proteins of SARS-CoV-2. Non-structural proteins were depicted in green, structural proteins in orange and accessory proteins in blue.

Besides the non-structural proteins, SARS-CoV-2 has six accessory proteins that help to evade immune cells. These are named according to the ORFs in which they have been encoded: ORF3, ORF6, ORF7, ORF8, ORF9, and ORF10. ORF3 consists of ORF3a and ORF3b. ORF3a is an ion channel protein specifically related to

inflammasome activation and apoptosis. Also, this protein is highly conserved among the beta coronaviruses (Nieva, Madan, & Carrasco, 2012). The structure of ORF3a was determined using cryo-EM and deposited to the PDB with 6XDC ID (Kern et al., 2021). ORF3b is located in the nucleus and cytoplasm, including the mitochondria outer membrane; however, there is no structure since it is a disordered protein (Bai et al., 2022). ORF6 has a role in inhibiting cellular transmission, and its structure is not determined experimentally yet. ORF7 is an accessory protein that acts as an immunomodulating factor by interacting with immune cells (monocytes) (Hardenbrook & Zhang, 2022). Its experimental structure is present in the PDB with 7CI3 ID (Z. Zhou et al., 2021). ORF8 has an immunoglobulin-like fold such as ORF7 and its function as immune suppression in the infected cell. The structure of ORF8 was determined by Nelson et al. from the National Institutes of Health, National Institute Of Allergy and Infectious Diseases, and deposited to the PDB with 7JX6 ID. ORF9 protein has a role in the mitochondrial antiviral signaling pathway, and it modifies host mitochondrial activity through signal relaying (Hardenbrook & Zhang, 2022). Last but not least, ORF10 is a small peptide consisting of 38 amino acids, and there is no evidence of expression of it and no determined structure for ORF10 (Wu et al., 2020). All known structures of SARS-CoV-2 accessory proteins are depicted in Figure 3 (blue).

Aside from non-structural and accessory proteins, SARS-CoV-2 has four structural proteins comprising the virion. These proteins are required to produce the new viral particles and encode from the 3' region of the viral genome. Among structural proteins, the smallest one is envelope protein with 75 residues. The envelope is a cationic ion channel protein that modifies the membrane permeability of infected cells and is also crucial for virion integrity. Many envelope proteins are produced and located at the Golgi and Endoplasmic Reticulum during the infection cycle. However, only a tiny amount of them localize the viral membrane after the budding and virion particles are released (Hardenbrook & Zhang, 2022; Liao, Yuan, Torres, Tam, & Liu, 2006). SARS-CoV-2 envelope protein is determined using NMR and deposited to the PDB with 7K3G ID. Interestingly, even though SARS-CoV and SARS-CoV-2 proteins share high homology, the structures of SARS-CoV envelope protein (PDB ID: 5X29) and SARS-CoV-2 envelope protein are noticeably different from each other (Mandala et

al., 2020). The membrane protein is another structural protein of SARS-CoV-2, which regulates the assembly formation and determines the shape of the virion with envelope protein. In addition, nucleocapsid protein is another essential structural protein of SARS-CoV-2, which has a dimeric architecture that locates inside the virus. Nucleocapsid stabilizes the viral genome and acts as a chaperone protein for forming RNA conformation suitable for replication. Structure of the RNA binding domain and C-terminal dimerization domain of nucleocapsid protein are deposited in the PDB with 6VYO ID (Kang et al., 2020) and 6YUN ID (Zinzula et al., 2021), respectively. Spike is another structural protein of SARS-CoV-2, which has an essential function in attaching to the host cell, thereby drawing significant interest. In this study, we focused on the spike protein and its interaction to host receptor ACE2; therefore, spike protein is examined in detail in the next section. All probed structural proteins of SARS-CoV-2 were depicted in Figure 3 (red).

2.2.1. Structure of SARS-CoV-2 Spike Protein

Spike is a homotrimeric transmembrane glycoprotein embedded in the cell membrane, composed of S1 and S2 subunits with 1273 residues. The S1 subunit contains the extracellular part of the protein encompassing the N-terminal domain (NTD), the receptor binding domain (RBD), whereas the S2 subunit contains transmembrane intracellular components; fusion peptide (FP), transmembrane domain (TM), and cytoplasmic tail (CT) (Xia, 2021). The primary role of spike protein is to mediate attachment of the virus to the host receptor ACE2 for initiating membrane fusion and viral infection (Figure 4.A). The spike S1 subunit mediates the first step of viral attachment by binding to host receptor ACE2 with its RBD. The RBD shows two different conformations, up and down, caused by hinge-like movement (Harvey et al., 2021), and during up conformation, RBD can attach to its host receptor ACE2. The spike protein structure is determined before and after binding to ACE2. The whole structure of spike protein with one up-stated RBD presents at the PDB with 6VYB ID (Walls et al., 2020) (Figure 4.B). Additionally, the structure of the ACE2-RBD complex has been determined using X-ray crystallography and Cryo-EM, which are presented at the PDB with 6M0J and 6M17 IDs, respectively (Lan et al., 2020a; R. Yan et al.,

transmembrane spike protein. **C)** After ACE2 binding, the spike is cleavage from S1/S2 cleavage site, and S2 undergoes postfusion conformation.

Variants of the SARS-CoV-2 virus contain mutations in their spike proteins. The investigation of spike mutations is ongoing because of their critical role in host attaching and viral entrance. Furthermore, the RBD of the spike is a primary target for new antiviral strategies like antibody-mediated neutralization, vaccine, and therapeutic development. These strategies are based on blocking the interaction between RBD and ACE2 receptor protein. Thereby, the spike protein of the virus is surrounded by therapeutics, and RBD-ACE2 binding is prevented; thus, viral fusion is stopped (Du et al., 2009).

2.2.1.1. ACE2: the Receptor Protein of Spike

ACE2 is an 805 amino acids long, homodimer transmembrane protein that presents several types of capillary rich organs, including the lung, kidney, endothelium, but also intestine, heart, and brain (Roca-Ho, Riera, Palau, Pascual, & Soler, 2017; H. Zhang, Penninger, Li, Zhong, & Slutsky, 2020). ACE2 protein evolved from ACE protein by gene duplication, and these proteins share 42% sequence similarity (Donoghue et al., 2000). However, their physiological functions in the cell are different. ACE2 has a role in the renin-angiotensin system, a complex network that regulates homeostasis by maintaining vascular resistance and blood pressure in the human body. In the RAS network, ACE2 precisely converts AngII to Ang-(1–7) and increases vasodilation, whereas ACE alters AngI to AngII and induces vasoconstriction (Figure 5.A) (Medina-Enríquez et al., 2020; Samavati & Uhal, 2020).

Additionally, ACE2 is the chaperone protein of B⁰AT1; therefore, they present as a complex in the cell. The structure of full-length ACE2 with B⁰AT1 protein is determined using Cryo-EM, and the structure is released in the PDB with 6M1D and 6M18 IDs (R. Yan et al., 2020). ACE2 consists of four parts: peptidase domain, neck, linker, and transmembrane domain. The neck domain is responsible for the major dimerization interface, and the peptidase domain is responsible for the minor interface.

Major dimerization interface of ACE2 is stabilized with polar bonds. However, the minor interface consists of weak interactions, and two peptidase domains disappear from each other by 25Å while the ACE2 in the open conformation (Figure 5.B) (R. Yan et al., 2020).

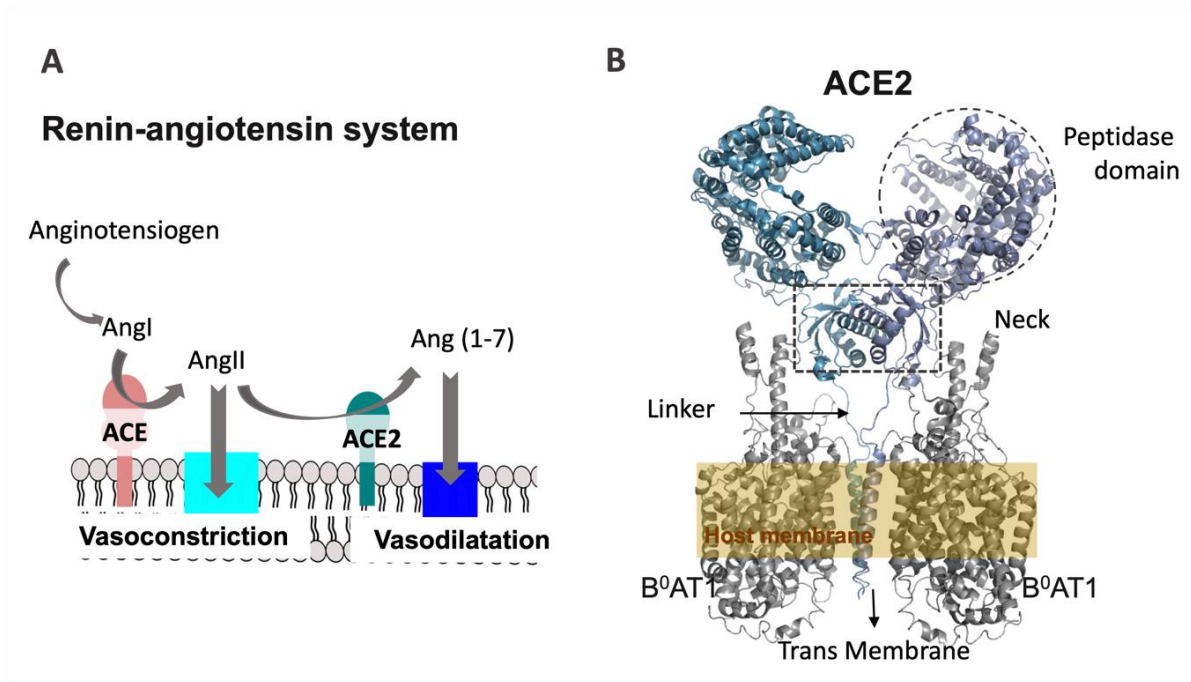


Figure 5: ACE2 protein is a member of the renin-angiotensin hormone system in the human body. **A)** ACE2-Ang (1-7) axis is the antagonist of the ACE-Ang2 axis; these axes regulate homeostasis. **B)** Domain organization of human ACE2 complex with B⁰AT1. Spike uses ACE2 for attaching to the host cell.

In addition to its homeostatic role in the cardiovascular system and chaperone-like function in the cell, SARS-CoV and SARS-CoV-2 use ACE2 as a functional receptor to enter the host cell (W. Li et al., 2003). Previous studies have shown that SARS-CoV2 infection causes downregulation of ACE2 and increases ACE2 levels in blood and urine. Additionally, downregulation of ACE2 causes an imbalance of the renin-angiotensin hormone system, promoting multiple organ injuries in the disease of SARS-CoV-2. Also, AngII may induce the viral progression causing organ injury by vasoconstriction, inflammation, and fibrosis. However pathogenicity of SARS-CoV-2 remains unclear, and further studies are needed (Braga et al., 2020; Furuhashi, Moniwa, Takizawa, Ura, & Shimamoto, 2020).

The N-terminal helix of the peptidase domain contains most of the ACE2 contact residues. SARS-CoV and SARS-CoV-2 bind to ACE2 from the same binding site (root mean square deviation of 1.2 Å) (Lan et al., 2020a; W. Li et al., 2005). Also, previous experimental and computational studies have shown that SARS-CoV-2 binds to ACE2 with a much higher affinity than SARS-CoV (Nguyen et al., 2020). SARS-CoV-2 and ACE2 complex has more engaged residues with a larger buried surface area than the SARS-CoV-ACE2 complex. This interface leads to more non-covalent contacts, such as van der Waals contacts and hydrogen bonds (Q. Wang et al., 2020). This increased binding affinity may interpret the enhanced infectivity of SARS-CoV-2.

2.3. Close up into the SARS-CoV-2 RBD-ACE2 Interface

As we stated previously, the interaction between RBD of spike and host receptor ACE2 is crucial for viral fusion. Therefore, the interaction between ACE2 and RBD was investigated at the atomic level. The RBD consists of five antiparallel beta-sheets (β 1-5), short alpha-helices, and connecting loops. The β 5 and β 6 extended insertion regions of RBD are called Receptor Binding Motif (RBM), and this motif has most of the contact residues with the ACE2 receptor protein (Figure 6.A). The peptidase domain of ACE2 consists of two lobes and a peptide-binding site between them. N-terminal helix at the small lobe of ACE2 accommodates the connecting residues of interaction. These residues involve 13 hydrogen bonds and two salt bridges across the hydrophilic interface. Hydrogen bonds appear within cut-off distance 4Å, between N487-Q24, K417-D30, Q493-E35, Y505-E37, Y449-D38, T500-Y41, N501-Y41, G446-Q42, Y449-Q42, Y489-Y83, N487-Y83, G502-K353, and Y505-R393 residues (Figure 6.B). All the RBD contact residues (except K417) are located on RBM. Two salt bridges appear between K417-D30 residues. Additionally, a cysteine pair at the 480 and 488 positions connects the loops, and stabilizing these flexible loops enhances binding to ACE2 (Lan et al., 2020a).

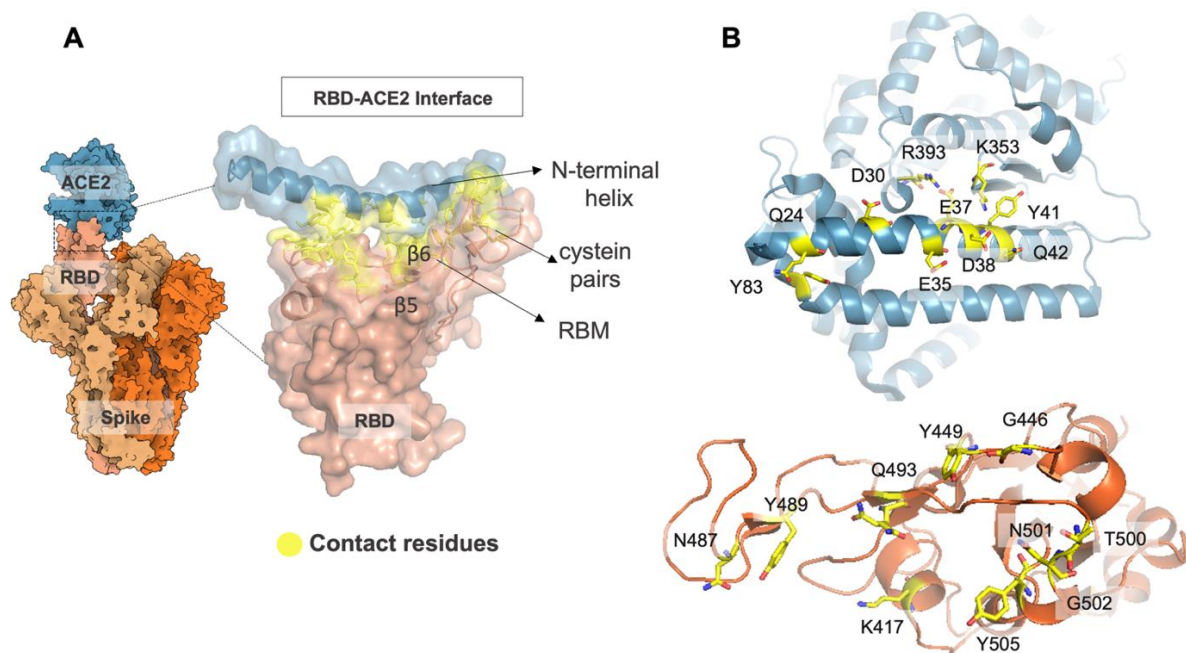


Figure 6: The interface of the RBD and human ACE2 complex. **A)** Contact residues were shown in sticks and colored in yellow. **B)** Residues that mediate the formation of hydrogen bonds and salt bridges were labeled with residue names and depicted in yellow.

2.3.1 Mutations on the RBM Alter Binding Affinity of SARS-CoV-2 RBD with human ACE2

According to the previous studies, SARS-CoV and SARS-CoV-2 RBDs share sequence similarity of ~73%; however, at the RBM, the sequence homology decrease to ~50% (Wan, Shang, Graham, Baric, & Li, 2020). Also, SARS-CoV-2 binds to ACE2 with ~4-fold increased affinity than SARS-CoV (Daniel et al., 2020). This observation confirms that SARS-CoV-2 has been accumulating adaptive mutations in its RBD, specifically RBM. Adaptive mutations on RBM confer more connection to ACE2, thereby offering better ACE2-binding to SARS-CoV-2 than SARS-CoV, which may lead to increasing transmissibility to the virus (Shang et al., 2020). Furthermore, SARS-CoV-2 progressively spreads into new variants, and recent studies highlight that SARS-CoV-2 variants contain mutations mainly in their RBM (Han et al., 2021). According to that, we collected point mutation sets of the spike protein of SARS-CoV-

2 VOIs and VOCs from the websites of WHO (World Health Organization Tracking sars-COV-2 variants. Retrieved May 19, 2022, from <https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/>) and Pango (Pango Cov-lineages. Retrieved May 19, 2022, from <https://cov-lineages.org/constellations.html>) and listed in Table 1. Spike mutations are modeled on the structure. We realized that all these mutations are near RBM (Figure 7).

Table 1: Spike mutations of SARS-CoV-2 VOIs and VOCs are listed below. Deletional modifications are not included in the list since they are out of the scope of this study.

VOI or VOC	Variant name	Mutations
VOI	Lambda	L452Q, F490S
VOI	Mu	R346K, E484K, N501Y
VOC	Alpha	L452R, E484K, N501Y
VOC	Beta	E484K, N501Y
VOC	Gamma	K417T, E484K, N501Y
VOC	Delta	K417N, L452R, T478K, E484K
VOC	Omicron	G339D, R346K, S371L, S373P, S375F, T376A, D405N, R408S, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H

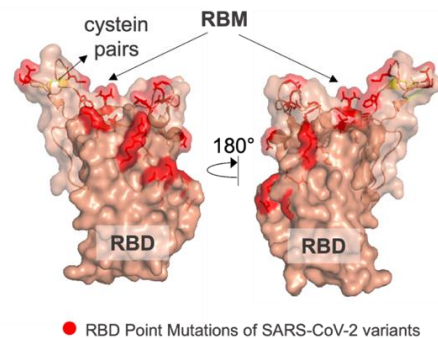


Figure 7: Spike point mutations of SARS-CoV-2 variants (VOCs and VOIs) depicted on the structure in sticks and colored in red. All point mutations are localized in RBD, specifically near RBM.

These mutations enhance ACE2 affinity by ~1.5-fold in the Alpha, Beta, Delta, Lambda, and Mu variants, 2-fold in the Gamma variant, and 2.4-fold in the Omicron variant. (Lippi et al., 2022). Moreover, SARS-CoV-2 variants cause a change in resistance of antibody responses and mediate escaping from humoral immunity (Dejnirattisai et al., 2021; Garcia-Beltran et al., 2021). According to the previous studies, E484K and N501Y/T mutations in the Beta, Gamma, and Omicron lineages; L452R mutations in the Alpha lineage are primarily responsible for enhanced immune evasion (Lucas et al., 2021).

As explained above, mutations in the SARS-CoV-2 spike and human ACE2 proteins alter the functional properties and binding affinities of spike-ACE2. Spike and ACE2 proteins were scanned with point mutations using the deep mutational scanning experiment to characterize the effect of mutation on the binding dynamics. Thereby the mutation effect on binding was revealed quantitatively (Chan et al., 2020; Starr et al., 2020).

2.4 Deep Mutational Scanning Approach and Its Usage in the Case of COVID-19

After the high-throughput sequencing revolution, characterizing the protein sequence-function relationship is of the utmost importance in biology. To this end, the high-throughput deep mutational scanning method has been developed to detect functional changes in protein variants (Araya & Fowler, 2011). The deep mutational scanning experiment combines molecular biology methods to produce large-scale functional scores such as expression, folding efficiency, stability, and activity profile of a protein within the cell. The deep mutagenesis experiment is the multiple stepped method generally described in 3 steps: creating a plasmid library of protein of interest, selecting the library, and deeply sequencing the library to obtain the functional scores (Fowler, Stephany, & Fields, 2014). The functional scores obtained from deep mutagenesis experiments are used for diverse applications, such as revealing functional activity landscapes of proteins (Sarkisyan et al., 2016) and enlightening disease-mutation relationships (Livesey & Marsh, 2020). Additionally, these functional

scores can be used for guiding protein engineering (R., John-Paul, E., Ryszard, & A., 2017), predicting unknown protein structures (Procko et al., 2013), and revealing mutational tolerance of protein (Canale, Cote-Hammarlof, Flynn, & Bolon, 2018).

In the case of COVID19, due to their critical roles in viral entrance SARS-CoV-2 RBD and human ACE2 proteins were separately subjected to the deep mutational scanning experiment by Starr et al. and Chan et al., respectively. These studies quantitatively presented mutation effects on RBD and ACE2 binding. In these studies, firstly, mutant plasmid libraries of ACE2 and RBD were prepared, and these plasmids were transfected into the cells. Following that, cells with variant protein on their membrane were incubated with its soluble receptor. Then cells were sorted using fluorescent activated cell sorting. Following that, deep sequencing was used to determine the binding affinity of the corresponding variant protein (Figure 8) (Chan et al., 2020; Starr et al., 2020).

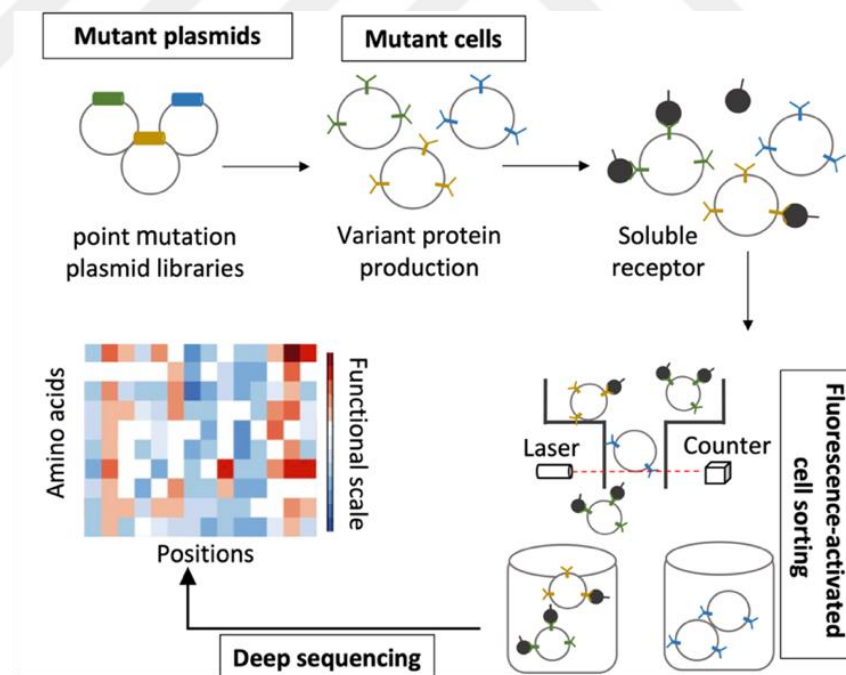


Figure 8: The schematic representation of the deep mutational scanning experiment workflow.

2.4.1. Deep Mutational Scanning Experiment of ACE2

For ACE2, 117 positions were scanned with point mutations by converting each residue to other possible 19 amino acids (117x19 mutations) (Chan et al., 2020). Each variant ACE2 protein is tagged with a c-myc epitope expressed in a human cell line, then folded and localized on the cell surface. Surface expression of ACE2 was detected using a c-myc epitope to avoid protein stability and localization errors. Cells that seamlessly have ACE2 incubated with RBD tagged with a super folded green fluorescent protein (sfGFP). In the following step, cells have ACE2-RBD complex on their surface were sorted by fluorescence-activated cell sorting (FACS). Then RNA transcripts were isolated from selected cells and sequenced with Illumina technology. The quantitative variant effect of each mutation finds by dividing the frequency of the variant in the sorted cell transcripts by the frequency of the naive plasmid library. This value is called the enrichment ratio, and it does not represent a direct measurement of the binding affinity change. However, enrichment ratio and binding affinity change are well correlated (Whitehead et al., 2012).

2.4.2. Deep Mutational Scanning Experiment of RBD

Starr et al. followed a similar procedure for RBD and scanned 201 positions (201x19 mutations) (Starr et al., 2020). This time yeast cell was used instead of the human cell because glycan can have a role in the interaction. For this, the first RBD was attached to the yeast cell using the yeast mating factor Aga2p. Following FACS selection, instead of Illumina sequencing in the ACE2 deep mutagenesis study, the variant effect was measured using barcode reading in varying soluble ACE2 concentrations to obtain a quantitative value of variant effect on binding.

2.4.3. Studies Relying on Deep Mutational Scanning Dataset of ACE2 and RBD

The deep mutational scanning dataset of ACE2 and RBD paves the way for developing neutralizing therapeutics for blocking ACE2 and RBD binding and deciphering the mutational tolerance of spike proteins of the new SARS-CoV-2 variant

(Chan et al., 2020; Starr et al., 2020). In a recent study, the deep mutagenesis dataset of RBD was used for developing a machine learning guided binding status predictor of RBD variants with a high performance (accuracy between 87% - 94%) (William and Reddy, 2021). In another study, an antibody-escape estimator was developed based on the RBD deep mutagenesis binding dataset to predict the antigenic impact of arbitrary combinations of mutations (Greaney, Starr, & Bloom, 2022). As an additional example, mutations were computationally built on RBD and ACE2 interface. Their binding energies were predicted using the MM/PBSA method, and the predicted binding energies were validated with the deep mutational scanning dataset (Laurini, Marson, Aulic, Fermeglia, & Pricl, 2021).

The deep mutational scanning method can require intensive time, cost, and labor. Also, there are several handicaps to this method, including data quality. For instance, we assume that, aside from intentional mutations, there are no extra mutations caused by biological complexity. However, data reliability is strived to increase by setting up the experiment with a control. Nonetheless, the high-throughput data of deep mutagenesis offers invaluable information to decipher the relation between genotype and phenotype of proteins in their native environment (Procko, 2020). In this thesis, we used deep mutational scanning datasets of variant ACE2 and RBD to benchmark the widely used structure-based binding affinity predictors. We aimed to suggest a prevailed predictor for investigating the mutation effect of RBD and ACE2 interaction in a time-efficient manner and without exhaustive effort.

2.5. Investigating the Effect of Variants on ACE2 and RBD Binding with Computational Tools

As it is previously described, ACE2 and RBD binding is crucial for attaching the virus to host cells, and this binding is directly related to virus infectivity. ACE2 and RBD binding dynamics are changing since the ACE2 has naturally-occurring variants in the mammalian populations, and RBD accumulates mutations through virus evolution. Binding dynamics of RBD and ACE2 were widely investigated with computational

methods in terms of variant infectivity, the zoonotic origin of the virus, animal tropism, and cross-species infection.

In the literature, various studies are based on structure-based binding affinity predictors. For instance, HADDOCK, a docking tool, was used for COVID-19 drug screening studies to identify new therapeutics from clinically tested drugs (HADDOCK Screen of 2000 Approved Drugs for Covid-19. Retrieved May 21, 2022, from <https://instruct-eric.eu/haddock-screen-of-2000-approved-drugs-for-covid19>). In another study, HADDOCK was used to reveal molecular features of ACE2 orthologs from 28 different species to predict infection susceptibility (Rodrigues et al., 2020). Additionally, in another study, HADDOCK was used to identify the effect of naturally-occurring mutations in human populations on RBD and ACE2 interaction (Sorokina et al., 2020). The widely used structure-based binding affinity predictor FoldX is used for COVID-19 related studies, particularly for predicting the stability of ACE2 and RBD interaction in different species (Blanco, Hernandez-Alias, Cianferoni, & Serrano, 2020; Hadi-Alijanvand & Rouhani, 2020). Apart from studies that rely on investigating variant ACE2 and RBD interaction dynamics, EvoEF1 was used for developing a peptide to block ACE2 and RBD interaction to prevent viral attaching (EvoEF: Energy potential for protein design. Retrieved May 21, 2022, from <https://zhanggroup.org/EvoEF/>). SSIPe was used to state binding affinity between hotspot residues of variant RBD and ACE2 (Saha, Ghosh, Sharma, & Nandi, 2021). Also, SSIPe was used to identify ACE2 residues that are important for viral recognition by determining binding affinity between primate ACE2 orthologs and RBD interaction. (Melin, Janiak, Marrone, Arora, & Higham, 2020). Last but not least, MutaBind2 was used for predicting the effect of Spike mutations on RBD and ACE2 interaction in terms of infectivity and antigenicity (Z. Li & Zhang, 2022; L. Zhang et al., 2021), and investigating hotspot residues of SARS-CoV and SARS-CoV-2 interfaces (Sobitan et al., 2021).

In addition to the probed structure-based predictors, the Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) method is used to predict the mutation's effect on ACE2 and RBD binding. MM-PBSA is an extensively used method for binding free energy prediction, generally based on molecular dynamic simulations and

computationally intensive than the probed classical binding affinity predictors (E. Wang et al., 2019). MM-PBSA was used to predict the ACE2 and RBD hotspot residues and variant contribution to binding, and results were validated with experimental binding values obtained from the deep mutagenesis experiment. These methods have quite high performance with a 92% success rate (Laurini et al., 2021; Laurini, Marson, Aulic, Fermeiglia, & Pricl, 2020)

On the other hand, the probed structure-based binding affinity predictors lack a comprehensive experimental validation even though some of the studies present limited proofs of performances with other tools (Z. Li & Zhang, 2022). Therefore, we benchmarked structure-based binding affinity predictors FoldX, EvoEF1, SSIPe, and MutaBind2 with the naïve predictors HADDOCK and UEP on the deep mutational scanning dataset of ACE2 and RBD. We revealed the performances of these tools and examined their performances from various perspectives to determine the reasons for their success or failure.

3. MATERIALS AND METHOD

3.1. Type of Study

This thesis composes of a theoretical study. We used web servers, binding energy prediction tools, molecular visualization software, Shell scripting for file manipulation, and Python for statistical analyses throughout this study.

3.2. Time and Location of Study

This thesis has been carried out at the Izmir Biomedicine and Genome Center and at the Izmir International Biomedicine and Genome Institute between February 2021 and May 2022.

3.3. Population and Sample of Study

In this thesis, we have investigated the performances of structure-based binding energy predictors in predicting the effect of interfacial mutations over the RBD and ACE2 deep mutational scanning dataset.

3.4. Materials of Study

In this thesis, we used HADDOCK, FoldX, EvoEF1, MutaBind2, SSIPe, and UEP to score point mutated variant models of RBD and ACE2. FoldX, EvoEF1, and UEP-related analyses were carried out in IBG-HPC. HADDOCK, MutaBind2, and SSIPe-related studies were conducted on their web servers.

An experimentally determined binding value dataset is required for benchmarking the performance of computational binding energy predictors. This dataset was taken from the previous experimental deep mutagenesis study of RBD and ACE2.

3.5. Variable of Study

The dependent variable of this study is experimentally determined RBD and ACE2 complex (PDB ID: 6M0J (Lan et al., 2020b)). All other generated RBD and ACE2 variant models are the independent variables of this study.

3.6. Tools for Data Collection

3.6.1. Web servers & resources used for data collection

We used the listed tools and websites to fulfill this study.

- **RCSB PDB:** RCSB PDB (Research Collaboratory for Structural Bioinformatics PDB) (<https://www.rcsb.org>) is the US data center of the global PDB archive where the 3D structures of biological macromolecules are deposited. We used this website to retrieve the crystal structure of the ACE2-RBD complex (PDB ID: 6M0J) (Lan et al., 2020b).
- **PDBePISA:** PDBePISA (Proteins, Interfaces, Structures, and Assemblies) (<https://www.ebi.ac.uk/pdbe/pisa/>) is a web tool for exploring macromolecular interfaces (Krissinel & Henrick, 2007).
- **ConSurf:** ConSurf is a web server (<https://consurf.tau.ac.il/>) predicting the evolutionary conservation of amino acid residues of a protein based on the sequence homology (Ashkenazy et al., 2016).
- **HADDOCK:** HADDOCK (High Ambiguity Driven protein-protein DOCKing) (<https://wenmr.science.uu.nl/haddock2.4/>) is an information-driven docking tool. We used this tool through its web server as a naïve predictor since its scoring function is based on simple force field terms (Honorato et al., 2021; Van Zundert et al., 2016a).

- **FoldX:** FoldX (<http://foldx.embl.de/>) is an empirical force field that calculates the binding value of a macromolecule or mutated version of it (Schymkowitz et al., 2005). We benchmarked FoldX on the deep mutagenesis dataset of RBD and ACE2.
- **EvoEF1:** EvoEF1 (EvoDesign Physics-based Energy Function) (<https://github.com/tommyhuangthu/EvoEF>) is a stand-alone empirical force field that calculates the binding value of a mutated protein complex used for specifically identifying hot spot mutations in the interaction interface (Pearce, Huang, Setiawan, & Zhang, 2019). We benchmarked EvoEF1 on the deep mutagenesis dataset of RBD and ACE2.
- **MutaBind2:** MutaBind2 (<https://lilab.jysw.suda.edu.cn/research/mutabind2/>) is an online force field that calculates the binding affinity change of a mutation. It is specifically used to determine the effects of mutation on protein-protein binding (N. Zhang et al., 2020). We included MutaBind2 in our benchmarking study as the main predictor.
- **SSIPE:** SSIPE (<https://zhanggroup.org/SSIPE/>) is an online force field that calculates the binding affinity change of a mutated protein complex (Huang, Zheng, Pearce, Zhang, & Zhang, 2020). We included SSIPE in our benchmarking study as the main predictor.
- **UEP:** UEP (<https://github.com/pepamengual/UEP>) is a stand-alone tool based on simple contact potentials, used to quickly estimate binding affinity change of point mutation on highly packed residues of protein-protein complexes (Amengual-Rigo, Fernández-Recio, & Guallar, 2021). We used UEP as a naïve predictor as its scoring function consists of simple contact terms.

- **PyMOL:** PyMOL (<https://pymol.org/2/>) is a molecular visualization tool. We used this tool to generate figures related to the interaction of RBD and ACE2.
- **Protein Imager:** Protein Imager (<https://3dproteinimaging.com/protein-imager/>) is an online molecular visualization tool for generating illustrations of diverse molecules (Tomasello, Armenia, & Molla, 2020). We used Protein Imager to create figures of this study.
- **IBG HPC:** IBG HPC (Izmir Biomedicine and Genome Center High-Performance Computing) is the local HPC service of our institute. We used this service for running UEP, EvoEF1, and FoldX, storing and manipulating our score and structure files.
- **PDB-tools:** PDB-tools (<http://www.bonvinlab.org/pdb-tools/>) is set of command-line tools to manipulate wwPDB file (Rodrigues, Teixeira, Trellet, & Bonvin, 2018). We used this tool for creating the HADDOCK input structure with single occupancy and a single temperature factor column.
- **Inkspace:** Inkspace (<https://inkscape.org/release/inkscape-1.2/>) is an open-source vector graphics software. We used the Inkscape to generate refined figures.

Throughout this thesis Python 3.8.5 (<https://www.python.org/downloads/release/python-385/>) and Shell scripting were used for data preparation and statistical analyses.

3.6.2. Methodology

3.6.2.1 Obtaining the Experimentally Determined Structures of ACE2 and RBD Complex

Due to the emergency of the COVID-19 situation, hundreds of ACE2-spike complex structures have been resolved using experimental structure determination methods, and these structures are deposited to PDB. To date, the structures of ACE2 orthologs from different mammals complex with the spike and structures of human ACE2 complexes with the spike from different SARS-CoV-2 variants were determined. Some of these structures have a full-length spike or ACE2; in our study, we required only RBD of the SARS-CoV-2 spike protein and peptidase domain of the human ACE2 protein. Therefore, we used the crystal structure of the SARS-CoV-2 RBD - human ACE2 peptidase domain complex (PDB ID: 6M0J) throughout the study. This structure was determined by X-ray diffraction and had 791 residues with a 2.45 Å resolution.

3.6.2.2. Construction of the Benchmarking Dataset

To create a benchmarking dataset for evaluating the performances of structure-based binding affinity predictors, we focused on interface residues of ACE2 and RBD. These residues contact non-covalent bonds and have an active role in binding by stabilizing protein-protein interactions. Therefore, mutations in these positions dramatically affect binding affinity between partner proteins. We determined interface residues of ACE2 and RBD using the crystal structure of the ACE2-RBD complex (PDB ID: 6M0J) by PDBePISA. PDBePISA applies crystal symmetry operation to input structure to obtain the most stable interface. However, we did not use this option and remained with input coordinates (x, y, z symmetry). PDBePISA suggested fifty-two interacted residues in the ACE2 and RBD complex. Twenty-six interface positions from ACE2: 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, and R393. Twenty-six interface positions from RBD: R403, K417, V445, G446, Y449, Y453, L455, F456, Y473, A475, G476, S477, E484, G485, F486, N487, Y489, F490, Q493, G496, Q498, T500, N501, G502, V503, and Y505.

Experimental binding values of mutations in the interface positions were selected from deep mutational scanning datasets of ACE2 and RBD. To avoid bias, we picked the same number of affinity enriching and decreasing mutations. We took all enriching interface mutations since the number of enriching cases less than depleting ones. Depleting cases were distributed over a broader range of binding values; therefore, we selected half of the depleting mutations from the highly depleting cases and the other half as randomly. This random selection was built in Python using the `default_rng` function from the NumPy package (with `seed=123`). Finally, we have 179 ACE2 mutations; 89 enriching and 90 depleting, and 84 RBD mutations; 42 enriching and 42 depleting (Table 2). So, the benchmarking dataset of this study includes 263 single mutations (APPENDIX A).

Table 2: Benchmarking dataset distribution is presented according to protein type (ACE2/RBD) and mutation effect (enriching/depleting).

	Total	Enriching	Depleting	
			Mostly	Random
ACE2	179	89	45	45
RBD	84	42	21	21
Overall	263	131	132	

Experimental binding values of ACE2 and RBD benchmarking datasets were provided in Figure 9. According to that, ACE2 has a diverse experimental value compared to the RBD dataset. RBD dataset is mainly composed of depleting and slightly enriching mutations.

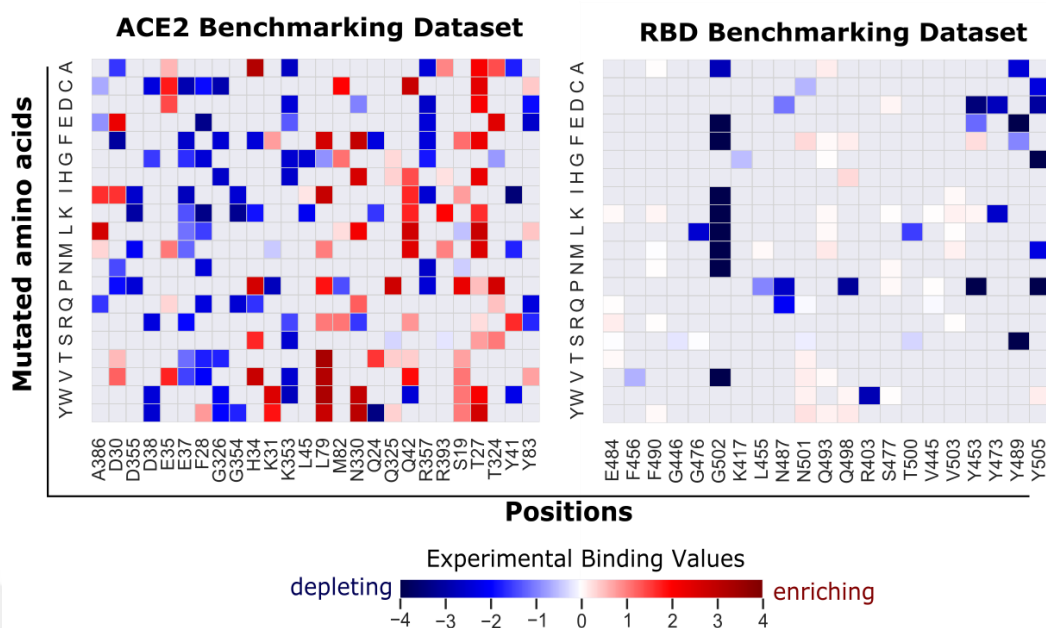


Figure 9: The benchmarking datasets of ACE2 and RBD are provided as a heat map and colored with experimental binding values. The blue color corresponds to depleting mutations, and the red corresponds to enriching mutations.

3.6.2.3 Investigating Evolutionary Conservation of ACE2 and RBD Interface

ConSurf is a widely used tool for mapping sequence homology information on the three-dimensional structure of protein complexes. ConSurf ranks residues from one to nine according to the evolutionary conservation; then, it colors residues using its conservation scales. To investigate the evolutionary conservation of ACE2 and RBD interface, we used ConSurf via its web server. We submitted two different runs for ACE2 and RBD using crystal structure of the ACE2-RBD complex (PDB ID: 6M0J). These runs took a couple of hours. Then we downloaded the output folders of ConSurf. We projected conservation scores onto the structure by opening the ConSurf output structure with PyMOL and running the consurf_new.py python code. Conservation scores of ACE2-RBD interface residues are presented in APPENDIX C.

3.6.2.4. Working Principles of Structure-Based Binding Affinity Predictors Used in This Study

In this thesis, we benchmarked four widely used structure-based binding affinity predictors: FoldX, EvoEF1, SSIPe, and MutaBind2. FoldX offers a water contribution option to involve water-mediated hydrogen bonds in calculating binding affinity. Therefore, we also used FoldX with a water option and named FoldXwater. We selected these tools according to their usage in the case of the COVID-19 situation (see Introduction 2.5) and their scoring functions. We have two tools that have physics-based terms in their scoring function: FoldX and EvoEF1. And we have two tools that consist of evolution-based terms: MutaBind2 and SSIPe. Also, we added HADDOCK and UEP to our benchmarking study as naïve predictors. HADDOCK is not a binding affinity prediction tool; however, we used its simple force field terms for predicting the mutation effect. Similarly, since the UEP has a simple contact term in its scoring function, we consider it a naïve predictor.

We produced the variant model and binding score for each mutation using each predictor throughout this study. FoldX, EvoEF1, and UEP were used via their stand-alone packages, and HADDOCK, MutaBind2, and SSIPe were used from their web servers. Thereby we have six models for each mutation, including models of HADDOCK, FoldX, FoldXwater, EvoEF1, MutaBind2, and SSIPe. Note that UEP does not output a variant structure. Between probed tools HADDOCK, FoldX, and EvoEF1 output predicted binding energy (ΔG) of the variant model, whereas UEP, MutaBind2, and SSIPe provide predicted affinity change ($\Delta\Delta G$) upon mutation. We calculated the predicted binding energy change for HADDOCK, FoldX, and EvoEF1 according to subtracting the predicted wild-type score from the predicted mutant score (Eq 1). These calculations were carried out using Python and shell scripting (Figure 10.A). The folder organization of the project is provided in Figure 10.B. All data relating to the project was released on GitHub (<https://github.com/EdaSamil/ace2-rbd-point-mutation-benchmark>).

$$\Delta\Delta G_{\text{predicted}} = \Delta G_{\text{mut}} - \Delta G_{\text{wt}} \quad (\text{Eq. 1})$$

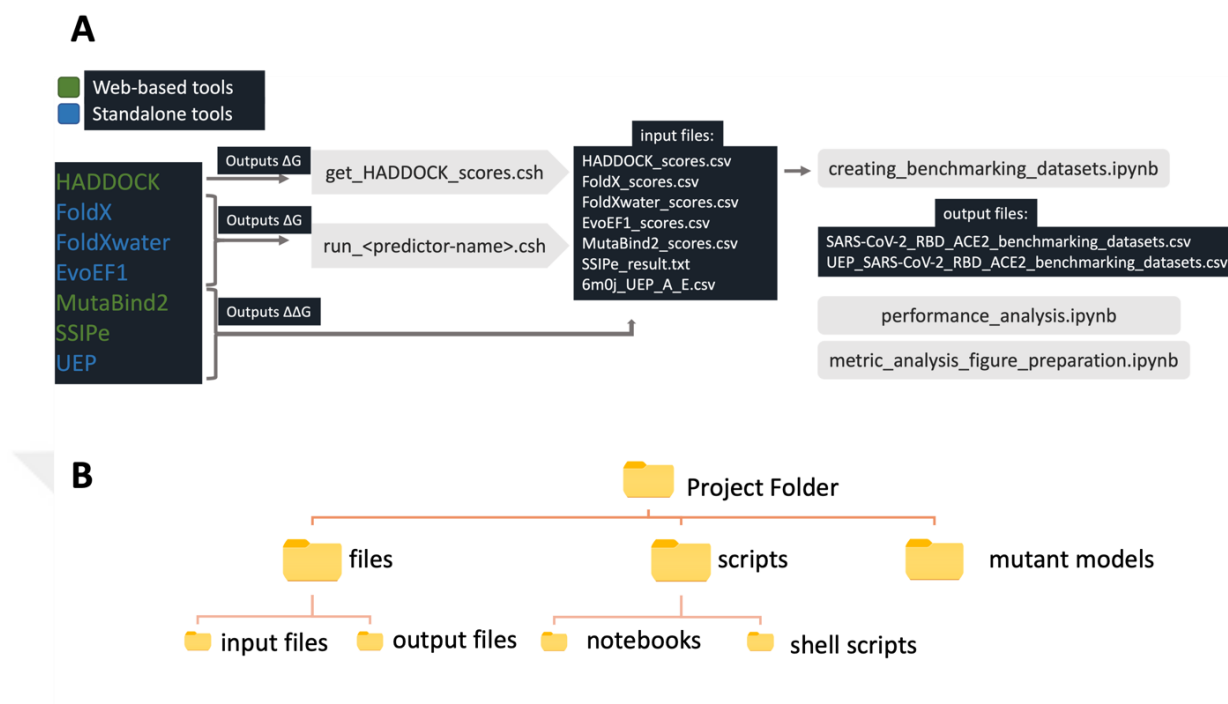


Figure 10: The schematic representation of folder and file organization of this study.

A) Tools were classified according to web-based or stand-alone. File and script organizations were provided. **B)** Predicted score files were deposited in the "input files" subfolder and used for preparing the benchmarking dataset in the desired format for analyses. The benchmarking and UEP datasets were located in the "output files" subfolder. Python notebooks and shell scripts were deposited in the "notebooks" and "shell scripts" subfolders, respectively. Variant models generated by tools were deposited in the "mutant models" folder.

3.6.2.4.1 HADDOCK-Based Prediction

HADDOCK is a widely used molecular docking tool that predicts the complex form of biomolecular systems. The working principles of HADDOCK (Van Zundert et al., 2016b) can be divided into three stages:

- *it0*, Rigid-body minimization: In this stage, partner proteins are randomly rotated around their center of mass while all geometrical parameters remain frozen (e.g., dihedral angles and bond length). Following this step,

partner proteins rotate and translate to obtain an energetically favorable pose of the complex. In this step, external information can be provided to guide the rigid-body docking.

- *it1*, Semi-flexible simulated annealing: In the second stage, generated models at the *it0* step are subjected to a three-step molecular dynamics-based refinement: *i*) orientation optimization of rigid-body partners, *ii*) semi-flexible simulation (flexibility is introduced only interface residues within a 5Å cutoff) for side chain conformation optimization, *iii*) flexibility is introduced to backbone and side chain of interface residues for finally optimize the model.
- *Water*, Refinement with explicit solvent: Selected models from the previous stage were subjected to short energy minimization in a solvent (TIP3 model water or DMSO environment) to improve the energetics of the interaction. Final models are generated after a short molecular dynamics simulation at 300K.

In our study, we used only the water refinement stage of the HADDOCK 2.4 webserver to calculate the binding energy of variant models (ΔG). HADDOCK uses a simple scoring function consisting of a linear combination of three physics-based energy terms to rank the generated models (Eq. 2).

$$E_{\text{HADDOCK}} = w \cdot E_{\text{vdW}} + w \cdot E_{\text{elec}} + w \cdot E_{\text{desolv}} \quad (\text{Eq. 2})$$

In order to model the variant ACE2-RBD complex, we used the following parameters:

1. *Distance restraints* section:

- Define center of mass restraints to enforce contact between the molecules → turn on

2. *Sampling parameters* section:

- Refine with short molecular dynamics in explicit solvent → turn on

3. *Advanced sampling parameters* section:

- Randomize starting orientations → turn off
- Perform initial rigid body minimization → turn off
- Allow translation in rigid body minimization → turn off
- Number of MD steps for rigid body high-temperature TAD → 0
- Number of MD steps during first rigid body cooling stage → 0
- Number of MD steps during second cooling stage with flexible side chains interface → 0
- Number of MD steps during third cooling stage with fully flexible interface → 0

HADDOCK does not implement mutation on the structure and does not work if the structure contains multiple conformations. Therefore, we created mutant structure files by hand. First, we cleaned water and ion molecules from the structure file using PyMOL. Then we used the pdb-tools to get rid of multiple conformations in the crystal structure of the RBD-ACE2 complex (PDB ID: 6M0J). For mutation generation, we deleted side chain atoms of wild-type residue, remained backbone atoms in the structure file using a text editor (e.g., BBEdit), and changed the wild-type amino acid name to the mutant amino acid. We used this manipulated structure file as an input of HADDOCK. HADDOCK notices missing side chain atoms and fills the missing atoms. After all, mutations were modeled and scored with the HADDOCK water refinement protocol within nearly five hours, and HADDOCK runs of each variant ACE2-RBD complex were downloaded. HADDOCK reports the predicted binding energy of the

model in the *structures_haddock-sorted.stat* file (can be found under the HADDOCK run folder: *structures/it1/water/*).

3.6.2.4.2 UEP-Based Prediction

UEP is a structure-based fast mutation effect classifier developed using the three-body contact potential of the Interactome3D data. Unlike common mutation effect predictors, UEP quickly scans the interface of complex structures within seconds without generating mutant models. UEP creates a contact matrix for the given protein complex by applying two restrictions: *i*) to define interface residues using a distance cutoff of 5Å, *ii*) selecting interface residues that are in contact with two or more residues of the partner protein. Due to UEP being able to work on highly packed interface residues, other positions can not be evaluated with UEP. After creating of contact matrix, UEP scores the mutation effect according to equation 3. Mutation and wild-type contacts are calculated by normalizing such interactions observed in the pre-trained UEP matrix with the expected number of interactions in the given UEP matrix. For calculating $\Delta\Delta G_{UEP}$, mutant contact divides into wild-type contacts. The log ratios can be represented as potential energies by applying the Boltzmann principle to this equation (Amengual-Rigo et al., 2021).

$$\Delta\Delta G_{UEP} = -\log((\text{Mutation contacts})/(\text{Wild-type contacts})) \text{ (Eq. 3)}$$

In our study, we used UEP as a stand-alone tool. UEP runs over the crystal structure of the ACE2-RBD complex (PDB ID: 6M0J), and result files are displayed in the same folder as the CSV file. Twenty-three positions have been selected as highly packed, and our benchmarking dataset narrowed down 129 mutations (Table 3). These mutations are indicated with aesthetics in Appendix A. Experimental binding value distribution of these mutations is presented in Appendix B.

Table 3: The UEP dataset distribution is presented according to protein type (ACE2/RBD) and mutation effect (enriching/depleting).

	Total	Enriching	Depleting
--	-------	-----------	-----------

			Mostly	Random
ACE2	82	39	26	17
RBD	47	19	16	12
Overall	129	58	71	

3.6.2.4.3 FoldX-Based Prediction

FoldX is a widely used structure-based binding affinity predictor that allows the creation of mutations in protein structures. We used the stand-alone Foldx 5.0 version to generate variant ACE2-RBD complexes and calculate the affinity change of mutations. For this, the mandatory individual_list.txt file is generated and mutations listed using a specific mutation format (e.g. [Wild-type amino acid][ChainID][residue position][mutated amino acid];, NE501Y;). FoldX provides a variety of commands to conduct different functions; we used RepairPDB, BuildModel, and AnalyseComplex commands in this study. RepairPDB was used to minimize the system's energy by repairing poor torsion angles and van der Waals' clashes in the structure. BuildModel was used to introduce mutations; optimize and output mutant and wild-type models for each mutation. AnalyseComplex was used to calculate the binding energy of generated wild-type and mutant models. Binding energy change ($\Delta\Delta G$) was calculated with corresponding wild-type binding energy for each mutation using equation 1. In order to calculate binding energy with the AnalyseComplex command, the scoring function of FoldX was used. FoldX scoring function consists of a linear combination of physics-based terms (Eq. 4).

$$E_{\text{FoldX}} = a*\Delta G_{\text{vdw}} + b*\Delta G_{\text{solvH}} + c*\Delta G_{\text{solvP}} + d*\Delta G_{\text{wb}} + e*\Delta G_{\text{hbond}} + f*\Delta G_{\text{el}} + g*\Delta G_{\text{kon}} + h*T \Delta S_{\text{mc}} + k*T \Delta S_{\text{sc}} + l*\Delta G_{\text{clash}} \quad (\text{Eq. 4})$$

In this equation, a to l are relative weights. ΔG_{vdw} is van der Waals energy, ΔG_{solvH} and ΔG_{solvP} are hydrophobic and polar group desolvation energies. In addition, those ΔG_{wb} is hydrogen bond energy from the water molecule; this term uses if the water molecule makes more than two interactions with groups of the protein. ΔG_{hbond} is hydrogen bond energy that is inferred from protein engineering studies. ΔG_{el} is

electrostatic energy calculated from the implementation of Coulomb's law. ΔG_{kon} is the electrostatic contribution of different polypeptides, which is a reasonable estimation of the complex's association rate (k_{on}). ΔS_{mc} and ΔS_{sc} are entropy costs for fixing a given conformation's backbone and side chain, respectively. Entropy cost for the backbone is obtained from a statistical examination of a given amino acid's phi–psi distribution as observed in a set of crystal structures, and entropy cost for the side chain is derived by scaling the calculated entropy parameters to the burial of the side chain. ΔG_{clash} is the energy of steric overlaps between atoms in the structure (Schymkowitz et al., 2005).

Additionally, FoldX offers the pdbWaters option, which contributes the water-mediated hydrogen bond energy to the calculation. We used the AnalyzeComplex command to define pdbWaters as "true". Generated scores and models of this analysis were saved in a separate folder named FoldXwater. All calculations were completed within several hours.

3.6.2.4.4 EvoEF1-Based Prediction

EvoEF1 is an energy function of the EvoDesign, a protein design tool. EvoEF1 consists of five physics-based terms representing the atomic interactions in a protein complex (Eq. 5). These terms are van der Waals energy (E_{vdw}), electrostatic energy (E_{elec}), hydrogen bond energy (E_{HB}), desolvation energy (E_{solv}), and reference energy (E_{ref}). For the E_{vdw} calculation, a modified version of the Lennard–Jones potential was used. The E_{elec} was used to calculate the electrostatic interactions of partially charged atoms within a distance cutoff of 6Å. The E_{HB} combines three energy terms: the distance of hydrogen-acceptor, the angle between the donor, hydrogen, and acceptor, and the angle between the hydrogen, acceptor, and base atoms. The E_{solv} term composes of desolvation energy of polar and nonpolar atoms with different weights. Lastly, the E_{ref} represents the approximate reference energy of unfolded state protein complex (Pearce et al., 2019).

$$E_{EvoEF1} = \sum [E_{vdw}(i,j) + E_{elec}(i,j) + E_{HB}(i,j) + E_{solv}(i,j)] - E_{ref} \quad (\text{Eq. 5})$$

In this study, we used a stand-alone version of EvoEF1 by running over the crystal structure of the ACE2-RBD complex (PDB ID: 6M0J). For using EvoEF1 mutations list in individual_list.txt file by following a specific mutation format (e.g. [Wild-type amino acid][ChainID][residue position][mutated amino acid];, NE501Y;). RepairStructure, BuildMutant, and ComputeBinding commands of EvoEF1 are used in this study. RepairStructure minimizes the system's energy by reducing steric clashes in the structure and repairing incomplete protein side chains. BuildModel was used to introduce mutations in the structure. ComputeBinding was used to calculate the binding energy of generated the protein-protein complex. Since the ComputeBinding command provides binding energy (ΔG) of the mutant model, binding energy change ($\Delta\Delta G$) upon mutation is calculated by subtracting wild-type binding energy from mutant (Eq. 1). All calculations were completed within several hours.

3.6.2.4.5. MutaBind2-Based Prediction

MutaBind2 is a structure-based binding affinity predictor that can implement mutation on the protein structure, generate variant models, and evaluate the mutation's effect. MutaBind2 follows FoldX's structure optimization and mutation implementation protocols for variant structure generation. Then 100-step energy minimization introduces backbone atoms in the variant and wild-type complex structure by the CHARMM force field. MutaBind2 consists of both physics and conservation-based terms built via the random forest method: $\Delta\Delta E_{vdw}$, $\Delta\Delta G_{solv}$, $\Delta\Delta G_{fold}$, SA_{part} , SA_{com} , CS , and N_{cont} . $\Delta\Delta E_{vdw}$ is the change of van der Waals interaction energy upon mutation, and $\Delta\Delta G_{solv}$ is the change of polar solvation energy upon mutation, which calculates using the PBEQ module of CHARMM program. $\Delta\Delta G_{fold}$ is the change of stability of the complex upon mutation calculated using FoldX software. SA_{part} and SA_{com} are solvent-accessible surface areas of the mutated residue in the wild-type unbound partner and complex structure, respectively. CS is the change of evolutionary conservation upon mutation calculated using the PROVEAN program. Lastly, N_{cont} is the number of interactive residues (within a distance cutoff of 10Å) between variant complex, performed only for the single mutations (N. Zhang et al., 2020). In this study, for using

MutaBind2, each mutation is introduced by hand, and output files are downloaded through its web server. MutaBind2 performs a run within less than an hour.

3.6.2.4.6. SSIPe-Based Prediction

SSIPe is a structure-based method to predict binding affinity change upon mutations. In the SSIPe approach, evolutionary conservation information of protein complex linearly combines with EvoEF1's physical energy function to estimate the mutation effect. Thereby, we describe the algorithm of SSIPe as two parts; $\Delta\Delta G_{SSIP}$ and $\Delta\Delta G_{EvoEF}$. $\Delta\Delta G_{SSIP}$ represents the conservation-related terms consisting of structure and sequence-based conservation profiles. For the structure-based profile, interface structural analogs determine by iAlign. For the sequence profile, sequence homologies calculate using STRING. Then, EvoEF1 was used for model generation and predicted energy change of EvoEF1 ($\Delta\Delta G_{EvoEF}$) linearly combined with $\Delta\Delta G_{SSIP}$ to obtain the final score (Huang, Zheng, Pearce, & Zhang, 2020). In this study, we used SSIPe from its web server. SSIPe runs over the crystal structure of ACE2-RBD complex (PDB ID: 6M0J), point mutations are presented as a list, and all results are downloaded directly from the website. All calculations were completed within several hours.

3.6.2.4. Evaluating Prediction Performance of the Tools

In order to determine the prediction performance of the predictors, we conducted a binary performance calculation. We classified mutations into two categories based on their effect on binding: enriching and depleting. Thermodynamically, a negative (<0) binding affinity means enriching mutation. However, for the experimental binding value from the deep mutagenesis experiment, a positive binding value (>0) represents enriching mutation and vice versa. We labeled each mutation as a success or failure by comparing the predicted and experimental categories. Thereby, we determined the performance of tools by dividing the number of accurate predictions by the total number of predictions. By multiplying 100, we obtained a percent success rate (Eq. 6).

$$\text{Success rate} = \text{Predictions}_{\text{Success}} / \text{Predictions}_{\text{all}} * 100 \quad (\text{Eq. 6})$$

For instance, the mutation from Asparagine and Tyrosine at the 501st position (N501Y) was modeled, and binding affinity change upon mutation was calculated with FoldX, EvoEF1, MutaBind2, and SSIPe (Figure 11). The experimental value of N501Y mutation is 0.24; it is an enriching mutation. Predicted binding affinity changes of N501Y are listed in Table 4. According to that, SSIPe and UEP predicted N501Y mutation as enriching; therefore, they were labeled as success, and other predictors were labeled as a failure. We performed this calculation for each mutation to determine the success rates of the tools. All analyses were performed using Python 3.8.5 with Pandas, Numpy, Seaborn, and Mathplotlib libraries.

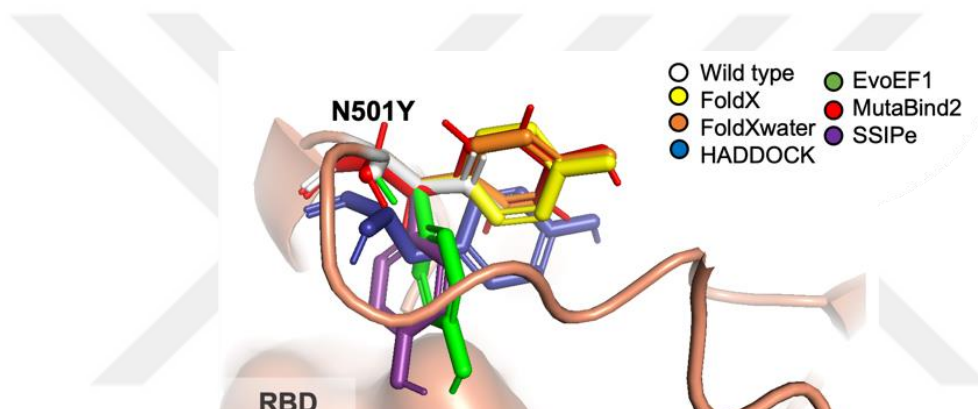


Figure 11: The models of N501Y mutation. The side chain of wild-type amino acid Asparagine was depicted in sticks colored as white. Mutant amino acid Tyrosinewase was modeled by FoldX, FoldXwater, HADDOCK, EvoEF1, MutaBind2, and SSIPe; side chains of Asparagine were shown in sticks colored as yellow, orange, blue, green, red, and purple, respectively.

Table 4: Predicted binding affinity change scores of N501Y mutations and the success score of the predictors.

	Experimental	HADDOCK	FoldX	FoldXwater	EvoEF1	MutaBind2	SSIPe	UEP
$\Delta\Delta G$ N501Y	0.24	1.90	3.03	6.28	0.32	0.65	-0.17	-1.62
Mutation effect	Enriching	Depleting	Depleting	Depleting	Depleting	Depleting	Enriching	Enriching

Success score	-	Failure	Failure	Failure	Failure	Failure	Success	Success
---------------	---	---------	---------	---------	---------	---------	---------	---------

3.6.2.5. Investigating Underlying Metrics of Prediction Performances

Mutations in the protein-protein interfaces cause physicochemical changes between interacted partners; thereby binding affinity of proteins has differed. We applied metric analysis to reveal the performance of predictors through physicochemical property changes. In this study, we used van der Waals volume (Lin, Long, Bo, Wang, & Wu, 2008), hydrophobicity (Eisenberg, Schwarz, Komaromy, & Wall, 1984), and flexibility (Shapovalov & Dunbrack, 2011) of amino acids as quantitative metrics to investigate tool performances. These metrics were selected to define physical and chemical changes upon mutation. Metric values of the related amino acid were obtained from the literature. These values are presented as a table in APPENDIX D. We depicted the mutation from Asparagine (N) to Tyrosine (Y) in the 501st position of the RBD structure in Figure 12.

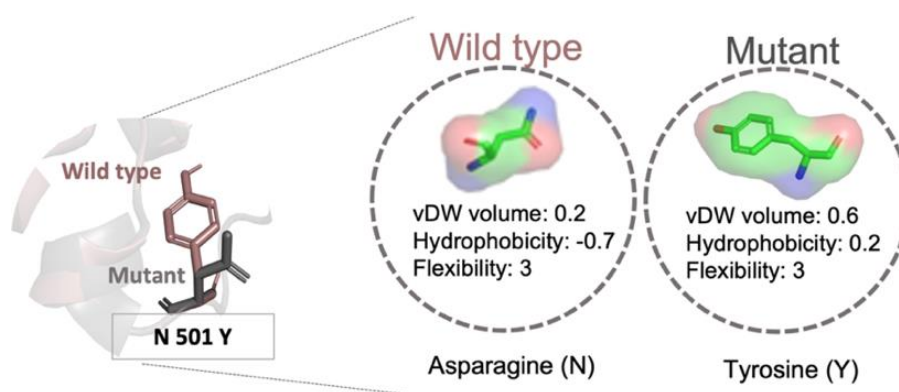


Figure 12: The mutation from Asparagine and Tyrosine at the 501st position was depicted, and corresponding wild-type and mutant metric values were presented.

To determine the physicochemical property change caused by mutation, we simply subtracted wild-type metric value from the mutant metric value (Eq. 7). For instance, the mutation from Asparagine to Tyrosine causes increasing in van der Waals volume from 0.2 to 0.6 and hydrophobicity from -0.7 to 0.2. The flexibility remains the same since the flexibility of Asparagine and Tyrosine are 3.

$$\text{Metric Change} = \text{Mutant}_{\text{metric-value}} - \text{Wild-type}_{\text{metric-value}} \quad (\text{Eq. 7})$$

- Volume change $\rightarrow 0.6 - 0.2 = 0.4$
- Hydrophobicity change $\rightarrow 0.2 - (-0.7) = 0.9$
- Flexibility change $\rightarrow 3 - 3 = 0$

We calculated the frequency distribution (probability density) of accurately predicted cases on the enriching and depleting datasets to investigate the relationship between performance and metrics. To investigate the density shift, we calculated the performance of tools where the metric value is positive (>0). Then we calculated a Δ Success rate by subtracting depleting and enriching performances to obtain predictor biases related to metrics. This rate changes between -100 and 100; at two extremes represents strong bias; positively and negatively correlated with the metric. For example, if the Δ Success rate of van der Waals volume is 99 for a given tool, this tool has a strong bias toward volume change; the tool predicts most volume increasing cases as enrichment.

In addition to these physicochemical properties, we used the side chain properties of amino acids for classifying mutations. The side chain of amino acids is generally classified into three categories: polar, apolar, and charged (Hagemans, van Belzen, Luengo, & Rüdiger, 2015).

- Polar amino acids: Asparagine, Glutamine, Serine, Threonine, Tyrosine.
- Nonpolar amino acids: Alanine, Glycine, Isoleucine, Leucine, Methionine, Phenylalanine, Proline, Tryptophan, Valine, Cysteine.
- Charged amino acids: Histidine, Glutamic Acid, Aspartic Acid, Arginine, Lysine.

According to this classification, if the wild-type and mutant amino acids are from the same class, we labeled the mutation "no change". If the wild-type and mutant amino acids are from different categories, the mutation is labeled as "charge gain", "charge

loss", "polarity gain", or "polarity loss" (Figure 13). For example, a mutation from Glycine to Lysine is labeled as charge gain since Glycine is a nonpolar amino acid and Lysine is a charged amino acid.

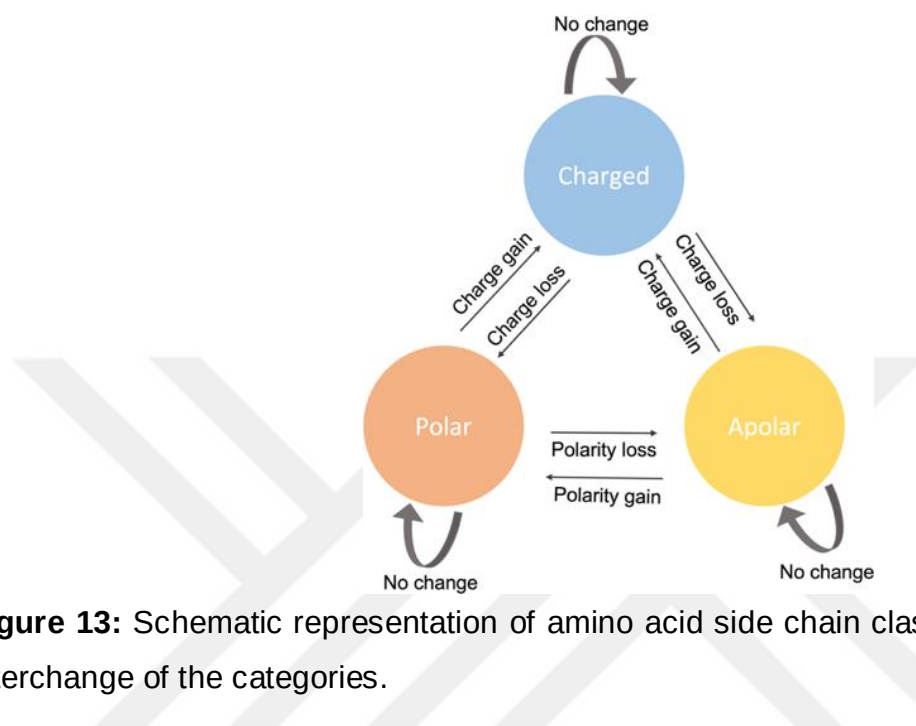


Figure 13: Schematic representation of amino acid side chain classes and labels of interchange of the categories.

3.7. Study Plan and Calendar

Mutation Selection of Benchmarking Dataset	February 2021 – March 2021
Selection of Binding Affinity Predictor	March 2021 – April 2021
Generation of Variant Models and Scores	May 2021 – August 2021
Performance Analyses of Predictors	September 2021 – November 2021
Metric Analyses of Predictors	December 2021 – January 2022
Writing the Thesis	February 2022 – May 2022

3.8. Data Evaluation

Crystal structure of the ACE2-RBD complex (PDB ID: 6M0J) (Lan et al., 2020b)) was evaluated by its resolution (in Å).

3.9. Limitations of Study

The deep mutational scanning experiment produces a noisy binding value since it is not a purified technique that performs on isolated systems.

We focused on the interaction interface of the protein complex, so we could not detect a wide range of conformational changes. We assume there is no allosteric effect in our mutation positions.

3.10. Ethics Committee Approval

The ethics committee has approved this thesis.

4. RESULTS

4.1. ACE2 and RBD Complex Has a Non-conserved Interface

We investigated the evolutionary conservation profile of ACE2 and RBD proteins using the ConSurf tool (Methods 3.6.2.3). ConSurf generates a conservation score between 1 to 9 for each residue in the protein complex. Additionally, ConSurf projects these scores onto the structure with a specific color code. In this analysis, we focused on the interface residues to identify the evolutionary conservation profile of the ACE2-RBD interface. The conservation coded structure of the ACE2-RBD complex is provided in Figure 14.A, and the conservation score distributions of ACE2 and RBD interface residues are presented in Figure 14.B. The conservation score mean of fifty-two ACE2-RBD interface residues is 2.3, the median is 1, and the upper quartile is 3 (Table 5). According to this analysis, we found that the ACE2-RBD has a non-conserved interface.

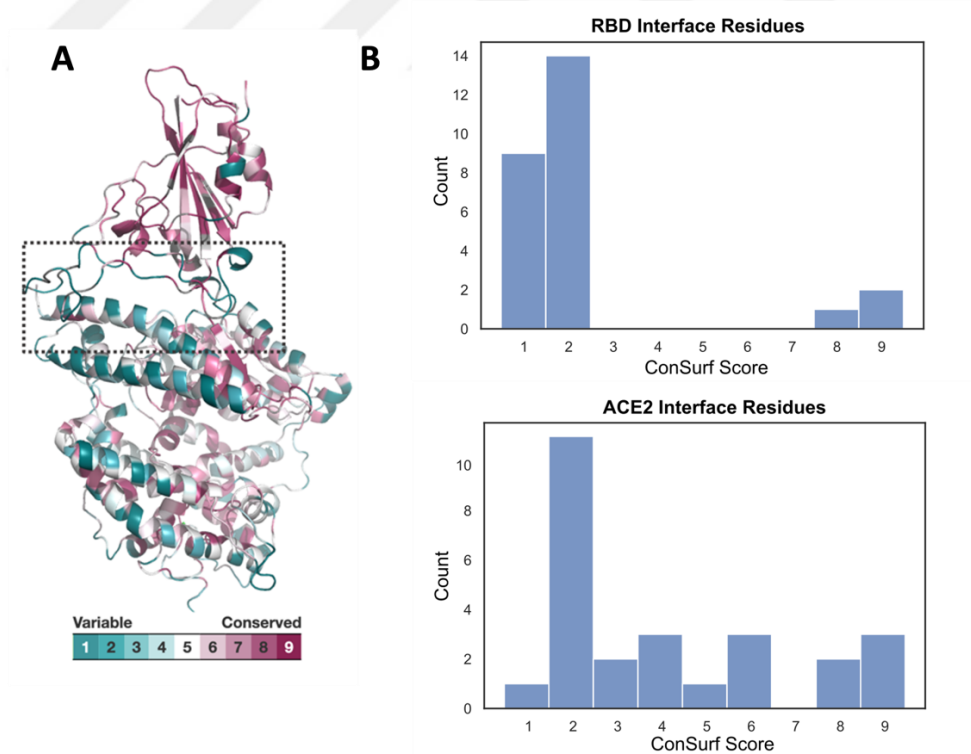


Figure 14: The evolutionary conservation analyses of the ACE2-RBD complex. **A)** Conservation scores project onto the structure with a specific color code; green

represents variable residues, and purple represents conserved residues. **B)** Conservation score distributions of ACE2 and RBD interface residues were presented.

Table 5: The summary statistics of ConSurf scores of ACE2-RBD interface residues.

Count (numbers of residue)	52	Lower Quartile	1.0
Mean	2.3	Median	1.0
Standard Deviation	2.7	Upper Quartile	3.0
Minimum	0.0	Maximum	9.0

4.2. Investigating Experimental Binding Value Distribution of Benchmarking Dataset

To benchmark structure-based binding affinity predictors, we selected interface mutations from the deep mutational interaction dataset of ACE2 and RBD. We have 263 mutations in our dataset, 179 from ACE2 and 84 from RBD. ACE2 and RBD datasets contain an equal amount of enriching and depleting mutations (Methods 3.6.2.2). Enriching mutations have a positive binding affinity, whereas depleting mutations are negative. We provided the experimental binding value distribution of the benchmarking dataset in Figure 15. We have seen that experimental binding values of ACE2 mutations change between -3.5 and 3.3; the lower quartile is -2.2, and the upper quartile is 1.6. Binding values of ACE2 have an almost symmetric distribution. However, binding values of RBD mutations stuck between 0.3 and -4.8, with the upper quartile of 0.06 and the median of 0. Especially enriching mutations (>0) overlapped within a constricted area. The RBD benchmarking dataset consists of slightly enriching mutations since the RBD experimental dataset does not have highly enriching interface mutations.

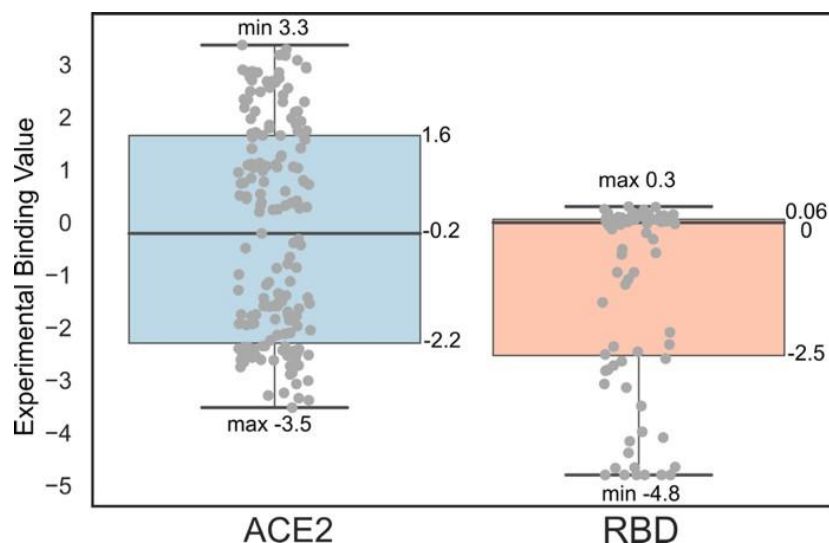


Figure 15: The experimental binding values of the ACE2 and RBD benchmarking datasets were depicted.

4.3. Comparing the Prediction Capacity of Each Tool

In this study, we elucidated the prediction performances of FoldX, FoldXwater, EvoEF1, MutaBind2, and SSIPe as main predictors and HADDOCK as a naïve predictor over the ACE2-RBD benchmarking dataset. This dataset consists of 263 single interface mutations (Table 1). As described in the Methods (Section 3.6.2.2), we performed a binary performance calculation by classifying mutations based on their effect on binding: enriching and depleting. Then, we compared the experimental and predicted classes for each mutation and labeled probed tools as failure or success (Methods 3.6.2.4). Thereby, we obtained performances of tools.

In line with the results of performance analyses, overall success rates ranged between 54% and 64%. (Table 6). The naïve predictor, HADDOCK, has the lowest prediction performance with a 54% success rate. FoldX has the highest success rate with 64%. EvoEF1 has the second-highest success rate, with 57%. Conservation-based tools, MutaBind2 and SSIPe, have a moderate performance with a 56% success rate. Additionally, performances of FoldX and FoldXwater are almost the same, with %64 and %63, respectively. The water contribution to the interface of the ACE2-RBD complex did not affect so much the performance of FoldX.

To understand the effect of protein type on binding affinity, we demonstrated the performance of tools on ACE2 and RBD datasets separately. For the ACE2 dataset, we discovered that performances varied between 51% and 62% (Table 6). Similarly, the RBD dataset's performances varied between 50% and 68%. FoldX is the highest-ranked predictor for each subset. Additionally, MutaBind2 and SSIPe have slightly higher success rates, and HADDOCK has slightly lower success rates on the RBD dataset than ACE2. As a result, we found that the protein type does not affect so much the performance of tools.

Table 6: Tools' success rates (%) were presented on the overall benchmarking dataset and its subsets. The highest and the lowest performances are colored red and green, respectively.

Benchmarking Dataset	Overall	ACE2	RBD	Enrichment	Depletion	ACE2 Enrichment	ACE2 Depletion	RBD Enrichment	RBD Depletion
HADDOCK	54	56	50	52	56	46	66	64	36
FoldX	64	62	68	51	77	47	77	60	76
FoldXwater	63	61	67	51	74	47	74	60	74
EvoEF1	57	58	57	49	66	48	67	50	64
MutaBind2	56	53	61	15	95	8	98	31	90
SSIPe	56	51	64	21	89	15	88	36	93

To understand the effect of binding affinity enhancer and decreaser mutations on the prediction performance of tools, we presented success rates on the enrichment and depletion datasets. Success rates over the enrichment dataset varied between 15% and 52%, whereas success rates on the depletion dataset varied between 56% and 95% (Table 6). Based on these results, there are great success rate differences between the lowest and highest success rates on the enrichment dataset. Predictors have high performances on the depletion dataset compared to the enrichment dataset. So, tools perform better on depleting mutations than on enriching mutations. Also, none of the predictors perform well on the enrichment dataset. In addition to these findings, it was discovered that HADDOCK, one of the naïve predictors of this study, has a similar performance on the enrichment (52%) and depletion (56%) datasets. To better represent, we provided the overall, enrichment and depletion performances in

Figure 16. Most interestingly, MutaBind2 and SSIPe, conservation-based tools, have the lowest performances on enrichment datasets and the highest performance on the depletion dataset. So, we observed a clear depletion bias for these tools since they classify non selectively most of the mutations as depleting of binding affinity. In the fact that physical energy-based terms in the scoring function of MutaBind2 and SSIPe are the same as FoldX and EvoEF1, respectively, we assume that the lower performance of these tools is due to conservation-related terms in their scoring functions (Methods 3.6.2.4.5 and 3.6.2.4.6).

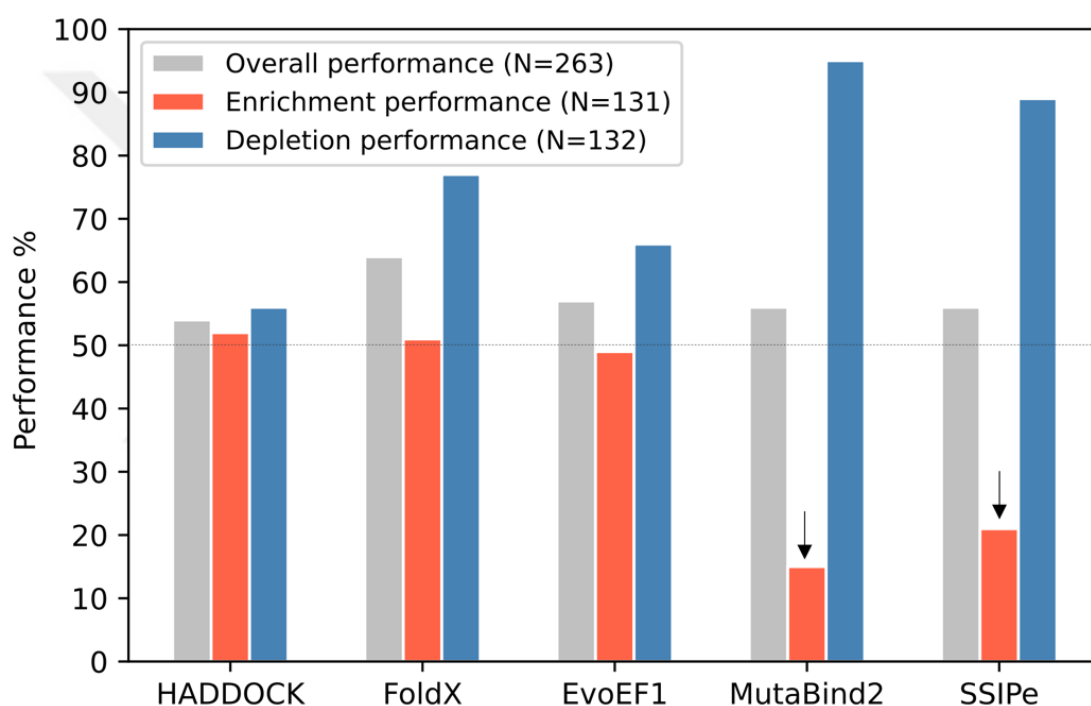


Figure 16: Performances of HADDOCK, FoldX, EvoEF1, MutaBind2, and SSIPe were presented with bar plots according to the overall dataset (gray), enrichment dataset (red), and depletion dataset (blue). The number of mutations for each dataset is depicted in the label. The significantly low performances were shown with an arrow. FoldXwater was discarded since it has a very similar performance to FoldX.

We divided enrichment and depletion datasets according to protein type to further enlighten the performance differences of tools. We revealed the performances on ACE2 enrichment, ACE2 depletion, RBD enrichment and RBD depletion subsets (Table 6). Tool performances change between 8%-48% and 31%-64% success rates

for ACE2 and RBD enrichment datasets, respectively. Performances on ACE2 and RBD depletion datasets varied between 66%-98% and 36%-93%, respectively. Similar to overall enrichment and depletion performances, predictors have a higher success rate on ACE2 and RBD depletion datasets than ACE2 and RBD enrichment datasets (except HADDOCK on RBD subsets). Additionally, when we compare the success rates of RBD enrichment and ACE2 enrichment subsets, we found that RBD enriching mutations remain more predictable than ACE2 enriching mutations for each tool.

4.3.1 Examining Performances of Each Tool on UEP Dataset

As stated in the Methods (session 3.6.2.3.2), UEP, one of the naïve predictors in this study, performs over highly packed residues only. Therefore, the UEP dataset narrowed from 263 to 129 mutations. Overall success rates of tools change between 49% and 69% over the UEP dataset (Table 7). FoldX is the highest-ranked tool, as with the benchmarking dataset. The naïve predictors of the study, HADDOCK and UEP, have similar success rates, 49% and 52%, respectively. In addition to the overall performances, we presented tools' performances on ACE2 and RBD subsets of the UEP dataset. According to this analysis, we found that the success rates of each tool on RBD are higher than ACE2, except UEP. Especially for MutaBind2 and SSIPE, the RBD mutations are more predictable than ACE2 mutations. As we expected from the naïve predictors, HADDOCK and UEP have similar performances on ACE2 and RBD subsets.

Table 7: Tools' success rates (%) were presented on the overall UEP dataset and its subsets. The highest and the lowest performances were colored red and green, respectively.

UEP dataset	Overall	ACE2	RBD	Enrichment	Depletion	ACE2 Enrichment	ACE2 Depletion	RBD Enrichment	RBD Depletion
HADDOCK	49	48	51	43	54	33	60	63	43
FoldX	69	67	72	52	83	49	84	58	82
FoldXwater	66	62	72	47	82	38	84	63	79
EvoEF1	59	57	62	38	76	33	79	47	71
MutaBind2	59	52	70	14	96	3	98	37	93
SSIpe	64	51	85	33	89	15	84	68	96
UEP	52	55	47	53	51	56	53	47	46

Additionally, we presented tools' performances on enriched and depleted datasets (Table 7). All main predictors, FoldX, EvoEF1, MutaBind2, and SSIpe, are good at predicting highly packed depleting mutations. The lowest and highest success rates on enrichment and depletion datasets are 14%-53% and 51%-96%, respectively. On the other hand, naïve predictors, HADDOCK and UEP have similar performances on enrichment and depletion datasets. Additionally, success rates of EvoEF1 on the enrichment dataset decreased from 49% to 38%, whereas depletion success rates increased from 66% to 76%. Surprisingly, the naïve predictor, UEP, has the highest performance on the enrichment dataset, which means that EvoEF1 particularly has poor performance on highly packed enriching mutations. MutaBind2 and SSIpe continue to bias toward depletion for highly packed mutations. However, SSIpe has a drastically high performance on the RBD enrichment dataset with a 68% success rate compared to the benchmarking dataset (Table 6, 36%).

4.4 Metric Analysis to Investigate Performance of Predictors

The previous results highlighted the prediction performance of predictors on the benchmarking and UEP datasets. Furthermore, we dissected datasets according to their protein (ACE2/RBD) and mutation effect type (enriching/depleting) to investigate the underlying parameters of predictor performances. Importantly, we observed most

of the predictors have a strong bias toward predicting mutations effect as depleting, with low performance on enrichment dataset.

In addition to that, the physicochemical properties of amino acids can be used for investigating the performance of tools. Once a mutation is created in the complex interface, certain metrics change at this location: for instance, van der Waals volume, side chain polarity, and charge. In this study, we selected various amino acid properties as metrics to represent mutations. Thereby we explored the performance of the predictors using metrics. For this, we chose van der Waals volume, hydrophobicity value, rotamer number (flexibility), and side chain property of amino acids to classify mutations (see Methods 3.6.2.5). These metrics represent the physical and chemical effects of mutations on the interface.

We evaluated van der Waals volume, hydrophobicity value, and flexibility since they are quantitative features. We also performed side chain property analysis, which is a qualitative feature. Amino acids are classified into polar, apolar, or charged; according to that, we classified mutation as no change, polarity loss/ gain, or charge loss/ gain (see Methods 3.6.2.5).

4.4.1. Examining Tool's Performances According to Quantitative Metrics: van der Waals Volume, Hydrophobicity, and Flexibility

Mutations cause physical and chemical changes in the interface ACE2 and RBD, directly affecting binding dynamics. For example, a mutation from Asparagine to Tyrosine increases van der Waals volume and hydrophobicity and does not change flexibility. For this mutation, the mutant amino acid Tyrosine is more hydrophobic and larger than wild-type amino acid Asparagine, and their rotamer numbers (flexibility) are identical. We performed metric analyses since the predictor performances might differ due to van der Waals, hydrophobicity, and flexibility change of mutations (Methods 3.6.2.5). To do so, we focused on only successfully predicted cases for each predictor, and we presented the distribution of predictor success rates over the metrics (Figure 17). In this way, we revealed the distinguished properties of successfully predicted

cases and explored biases related to the tools. According to our previous results, predictors have different prediction performances on enrichment and depletion datasets (see Result 4.3); therefore, we did the metric analyses on both datasets (Figure 17, left and right).

According to Figure 17, the volume change is presented on the x-axis. It takes values between -1 and 1: negative values represent volume decreasing cases, whereas positive ones represent volume increasing cases. According to this, the overall dataset was shown an almost normal distribution (gray line), slightly moved to the positive side, which means that we have slightly more volume-increasing mutations than volume-decreasing in the dataset.

For investigating the success rate of tools over hydrophobicity changes, we presented hydrophobicity changes of mutations on the x-axis and probability density on the y-axis. Hydrophobicity change takes a value between -6 and 6. The distribution of the experimental dataset is shown with a gray line. Positive and negative hydrophobicity change represents hydrophobicity increasing and decreasing mutations, respectively. According to Figure 17.B, most of the enriching mutations have positive hydrophobicity changes in the dataset. Also, depleting mutations are distributed near normally across the hydrophobicity changes.

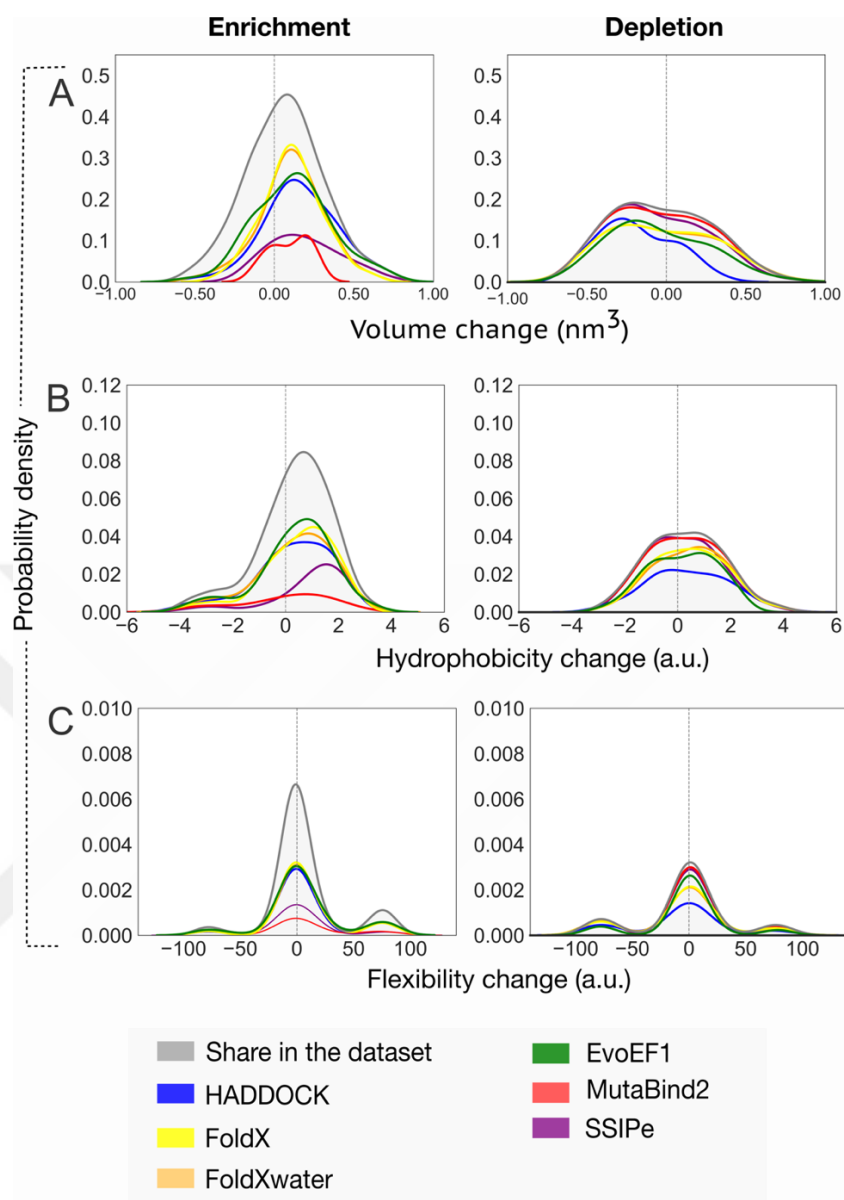


Figure 17: The success rates of tools were presented in terms of probability density over the volume, hydrophobicity, and flexibility metrics. The gray line represents the experimental dataset. The blue, yellow, orange, green, red, and purple lines represent HADDOCK, FoldX, FoldXwater, EvoEF1, MutaBind2, and SSIPe, respectively.

In Figure 17.C, we presented the frequency of accurately predicted mutations by each tool over the flexibility change of the mutations. The distribution of the experimental dataset is shown with a gray line. The flexibility of amino acids takes a value of 1, 2, 3, 9, 27, or 81. Therefore the distribution of mutations over flexibility

change has two small peaks at the negative and positive extremes and one big peak in the middle.

All in all, we observed that FoldX and FoldXwater have the highest peaks on the enrichment dataset for all metrics (Figure 17, left), as an expected result since the FoldX is the highest-ranked predictor and FoldX and FoldXwater have similar performances (see Result section 4.2). Additionally, we observed that MutaBind2 (red) and SSIPe (purple) have the lowest peak on the enrichment dataset for all metrics (Figure 17, left). MutaBind2 and SSIPe have the highest peaks on the depletion dataset for each metric because they have a high success rate on the depletion dataset. This result is expected, as conservation-based predictors are not good at predicting enriching mutations and have a depletion bias.

In the light of these observations, peak height (y-axis) gives us the frequency of accurately predicted mutations, and peak shift along the x-axis shows how the predictor performance changes as the metric value increases and decreases. Therefore, we decided to compare peak shifts of enrichment and depletion datasets to explore the biases of tools related to the metric. For this, we focused on the positive side of the graphics (metric increasing), and we found the under the curve ratios there. By taking the differences between the success ratios of enrichment and depletion datasets, we obtained a value that represents the magnitude of bias. We named this value as Δ Success rate (see Method 3.6.2.5). This rate changes between -100 and 100; at two extremes are represent strong biases. For example, in Figure 17.A, if we compare the distribution of the blue line in enrichment (left) and depletion (right) panels, we observe behaviors of HADDOCK depending on the volume change. It is evident that HADDOCK predicts volume-increasing cases as enrichment (peak shifts to the positive side on the enrichment panel) and predicts volume-decreasing cases as depletion (peak shifts to the negative side on the depletion panel). Calculating the Δ Success rate makes it easier to observe these kinds of behaviors (biases) of tools.

We showed the Δ Success rates of predictors for each metric in Table 10. For volume change, HADDOCK has a moderate bias with 34%. HADDOCK predicts most

of the volume-increasing cases as binding enrichment. Volume bias of HADDOCK might stem from its scoring function since the HADDOCK scoring function consists of simple physical energy terms, and one of them is van der Waals Volume energy (Eq. 2). Next to HADDOCK, SSIPe and Mutabind2 have a very negative Δ Success rate of volume metric. We thought this was caused by the depletion bias of predictors, which we observed in the performance analysis (see Result 4.3).

According to the Δ Success rates of predictors for hydrophobicity change (Table 10), FoldX has a moderate bias with -31%. The negative Δ Success rate represents the reverse relationship between metric and binding affinity strength (enriching or depleting). As a result, FoldX predicted most of the hydrophobicity increasing cases as depletion of binding affinity. Not surprisingly, conservation-based tools, MutaBind2 and SSIPe, have very negative Δ Success rates, as with volume change; however, this is caused by the strong depletion bias of tools (see Result 4.3).

We could not find any moderate or strong bias toward flexibility (Table 10). Similar to our previous observations MutaBind2 and SSIPe showed very negative Δ Success rates; however, this is not linked to the flexibility since caused by their depletion bias (see Result 4.3).

Table 8: The Δ Success rates of predictors were presented according to the metrics. The observed moderate biases were depicted in red.

Metric	Predictor Names	Enrichment Success (%)	Depletion Success (%)	Δ Success Rate (%)
van der Waals Volume	HADDOCK	66	32	34
	FoldX	61	77	-16
	EvoEF1	54	58	-4
	MutaBind2	18	95	-77
	SSIPe	29	83	-54
Hydrophobicity	HADDOCK	48	58	-10
	FoldX	52	83	-31

	EvoEF1	48	67	-19
	MutaBind2	14	96	-82
	SSIpe	30	85	-55
Flexibility	HADDOCK	56	46	10
	FoldX	52	73	-21
	EvoEF1	48	67	-19
	MutaBind2	13	94	-81
	SSIpe	22	89	-67

4.4.2. Examining Tool's Performances According to Side Chain Properties of Amino Acids: Polar, Apolar, and Charge

Next to quantitative metrics such as van der Waals volume, hydrophobicity, and flexibility, we have classified amino acids according to their side chain properties as they are widely used to investigate the nature of amino acids. Amino acid side chains are generally categorized as polar, nonpolar, and charged. Mutation from one amino acid class to another can be labeled as charge gain/loss, polarity gain/loss, or no change (see Method 3.6.2.5). We tagged all mutations in the benchmarking dataset with these labels. As previously stated, examining success rates on enrichment and depletion datasets allows us to understand the performance of predictors better. Therefore, we separately presented the distribution of no change, polarity gain, polarity loss, charge gain, and charge loss mutations on successfully predicted enrichment and depletion datasets in Figure 18. So, we observed that most of the mutations fall into the "no change" category, followed by polarity loss and charge loss. We have a tiny amount of polarity gain mutations in the dataset (Figure 18 gray bars). Interestingly, we could not discover any bias related to side chain property change.

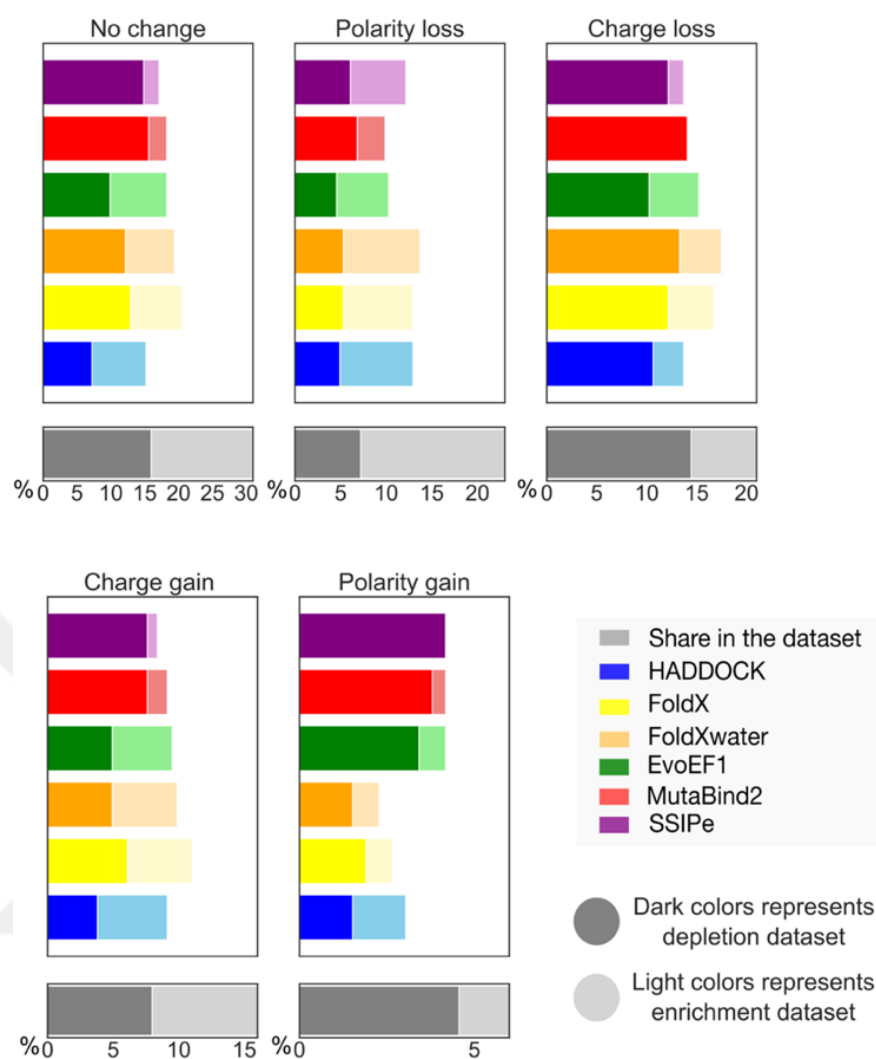


Figure 18: Mutations were classified into five categories; no change, polarity gain, polarity loss, change gain, and charge loss on successfully predicted enrichment (light colors) and depletion (dark colors) datasets. Blue, yellow, orange, green, red, and purple represents the HADDOCK, FoldX, FoldXwater, EvoEF1, MutaBind2, and SSIpe, respectively.

5. **DISCUSSION**

SARS-CoV-2 is a zoonotic virus that led to the COVID-19 pandemic. SARS-CoV-2 spike transmembrane protein binds to human ACE2 and then initiates viral entrance to the host cell. Therefore, this interaction is crucial and worth investigating in detail. Additionally, the binding dynamic of ACE2 and RBD changes since mammals have different ACE2 orthologs, and spike accumulates adaptive mutations across its receptor-binding domain. To quantify binding affinity changes upon ACE2 and RBD mutations, these proteins were scanned by point mutations using the deep mutagenesis experiment. Besides the experimental efforts, binding affinity changes upon mutations were investigated using computational approaches that provide cost and time efficiency. For instance, structure-based binding affinity predictors were used to estimate cross-species transmission and new emerging variants and develop new binders preventing SARS-CoV-2 infection.

However, the prediction capacity of structure-based binding affinity predictors was not validated with an experimental dataset. In this study, we aimed to identify the prediction capacity of the probed predictors. By taking advantage of the ACE2 and RBD deep mutagenesis datasets, we benchmarked widely used structure-based binding affinity predictors that have already been used in the case of COVID-19. FoldX, EvoEF1, MutaBind2, and SSIPe were evaluated as main predictors, and HADDOCK and UEP were used as naïve predictors. Also, we selected predictors that are based on various scoring functions; FoldX and EvoEF1 have physical energy-based terms, and MutaBind2 and SSIPe have both physical energy and conservation-based term.

To construct benchmarking dataset for this study, first we found contact residues of ACE2 and RBD. Because contact residues have a direct effect on binding since they mediate non-covalent interactions to the partner protein. Then we picked 263 interfacial mutations that on contact residues from deep mutagenesis dataset of ACE2 and RBD. While creating benchmarking dataset we selected an equal amount of enriching and depleting cases.

Performance analyses over benchmarking dataset revealed that FoldX is the highest-ranked predictor. FoldX's high performance might stem from its complex physics-based scoring function that includes entropy-related terms. Additionally, FoldX performance was presented with a water option (FoldXwater), and we found that contributing water to the ACE2-RBD interface does not affect FoldX's performance.

To dissect the effect of mutations, we investigated the performances of predictors on enrichment and depletion datasets of ACE2 and RBD. Through these analyses, we found that predictor performances are directly related to the effect of the mutations. Predictors have higher performances on depletion datasets than enrichment. We found that predicting ACE2-enriching mutations are more challenging than RBD- enriching mutations. Also, none of the predictors adequately predict binding affinity enhancer mutations. However, predicting enriching mutations is crucial in terms of estimating new emerging variants, predicting cross-species transmission, and developing neutralizer therapeutics to block ACE2 and RBD interaction.

Naive predictor performances confer us to evaluate performances of structure-based binding affinity predictors in a basic condition since their scoring function composes of simple terms. Performance analyses show that HADDOCK and UEP generally have similar and low performances. However, even though UEP is a naïve predictor, it has the highest performance, with a 53% success rate over the enrichment dataset. This also implies that enriching mutations is challenging for structure-based binding affinity predictors; current modeling and scoring protocols lack capturing the enrichment effect of mutations.

Interestingly MutaBind2 and SSIPe tend to predict most of the mutations as affinity depleting. This observation suggests that MutaBind2 and SSIPe are not sensitive to estimating ACE2 and RBD interfacial mutations' effect on binding. We assume that the depletion bias of MutaBind2 and SSIPe originates from their evolutionary conservation terms, considering that their physics-based terms are inherited from FoldX and EvoEF1, respectively. ACE2 and RBD have a non-conserved

interface that may have affected the prediction performances of conservation-based tools.

Further, we investigated metrics that affect the prediction capacity of tools; van der Waals volume, hydrophobicity flexibility, and side chain properties of the amino acid were selected as metrics. We have found that HADDOCK has a moderate bias; it classifies most of the enriching mutations as affinity enhancing and vice versa. We also identified that FoldX has a moderate bias for hydrophobicity. FoldX tends to predict hydrophobicity, decreasing mutations as affinity enhancers. We could not find any bias toward side chain properties and flexibility.

Regarding performance on naturally occurring SARS-CoV-2 mutations, we have only four cases in our dataset: G446S, S477N, E484K, and N501Y. So, we do not have a robust idea if FoldX (the high-ranked predictor) is good at predicting real cases or not. However, FoldX only predicts the E484K mutation accurately.

Since performances of these tools are evaluated on different datasets, we do not compare the literature performances of these tools. However, Amengual-Rigo *et al.* presented FoldX, EvoEF1, and UEP performances on the same dataset (N= 1251). They showed that Pearson Correlation Coefficients of predicted $\Delta\Delta G$ and experimental $\Delta\Delta G$ are 0.43, 0.33, and 0.22 for FoldX, EvoEF1, and UEP, respectively. According to this study, besides generally performances are not high, FoldX is the high ranked predictor, same with our results.

Previous studies showed that molecular dynamics-based methods (e.g., MM/PBSA) have higher performance in predicting mutation effect than classical binding affinity predictors (Laurini et al., 2021). Molecular dynamics-based methods are more time-consuming and computationally demanding than classical binding affinity predictors. However, intensive calculation protocols rely on side chain conformation sampling with some backbone flexibility providing high prediction capacity to these methods. This implies that classical binding affinity estimators may not capture the mutation effect due to their short molecular dynamics simulations.

Another critical point is that the deep mutagenesis experiment contains a certain level of noise since this experiment is based on cell-based assays, not based on isolated systems. In the deep mutagenesis experiment, enrichment values obtained using deep sequencing may not adequately reflect binding affinities. For instance, only eight mutations were classified as affinity enriching between 24 impactful mutations of SARS-CoV-2 variants (Table 1). We could not include noise of experimental method in our performance analyses. However, this experiment provides a high-throughput dataset for benchmarking computational tools or developing new methods based on a deep mutagenesis dataset, which is difficult to obtain with biophysics-based approaches.

All in all, we identified that none of the classical binding affinity predictors are reliable in predicting binding affinity changes of ACE2-RBD interfacial mutations. Even the highest-ranked tool, FoldX, has hydrophobicity bias and a poor success rate on the enrichment dataset. For subsequent pandemics, it is vital to estimate the mutation effect on binding quickly; therefore, stating the performance of structure-based classical affinity predictors on deep mutational scanning datasets is essential to reveal the advantages, approximations, and limitations of these tools. This study is an important example of predicting the mutation effect of host-pathogen systems over a deep mutagenesis dataset. Lastly, we note that this study provides an essential step toward future benchmarking studies that rely on deep mutagenesis datasets.

6. CONCLUSION AND SUGGESTIONS

- In this study, capacities in predicting the impact of mutations of HADDOCK, UEP, FoldX, EvoEF1, MutaBind2, and SSIPe were revealed on the deep mutagenesis dataset of ACE2 and RBD. None of the tools have acceptable performance on both enriching and depleting mutations. Therefore these classical binding affinity predictors remain unreliable in predicting mutation effect.
- According to performance analyses, we have seen that viral host-pathogen system, such as SARS-CoV-2 RBD and human ACE2, requires more advanced computational tools to predict binding affinity change upon mutations.
- Conservation-based tools, MutaBind2 and SSIPe, have a clear bias in predicting most ACE2-RBD interfacial mutations as affinity depleting. These tools may not accurately classify interfacial mutation effects on binding if the interface is non-conserved.
- We have seen that classical binding energy predictors are not fit to predict mutations' effect on the binding of SARS-CoV-2 RBD and ACE2 system. There is still plenty of work to be done to improve current tools to make them reliable for predicting binding affinity changes upon mutation. Recent studies showed deep learning applications on high-throughput deep mutagenesis datasets enhance the reliability of tools. So, deep mutagenesis datasets can be used to benchmark computational tools and develop new methods to predict mutation effects.
- We suggest that low performances of classical binding affinity predictors should be considered while using these tools for variant effect prediction and developing a therapeutic binder for blocking RBD-ACE2 interaction.

These studies might require more advanced and costly computational methods for reliable results.

- Our study is an important example of benchmarking the performance of computational tools in predicting mutation effects using a deep mutational scanning dataset. This performance calculation methodology can be applied to any tool over any deep mutagenesis binding affinity dataset.



7. REFERENCES

- Amengual-Rigo, P., Fernández-Recio, J., & Guallar, V. (2021). UEP: an open-source and fast classifier for predicting the impact of mutations in protein–protein complexes. *Bioinformatics*, 37(3), 334–341. <https://doi.org/10.1093/BIOINFORMATICS/BTAA708>
- Arya, R., Kumari, S., Pandey, B., Mistry, H., Bihani, S. C., Das, A., ... Kumar, M. (2021). Structural insights into SARS-CoV-2 proteins. *Journal of Molecular Biology*, 433(2), 166725. <https://doi.org/10.1016/j.jmb.2020.11.024>
- Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., & Ben-Tal, N. (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research*, 44(W1), W344–W350. <https://doi.org/10.1093/nar/gkw408>
- Bai, C., Zhong, Q., & Gao, G. F. (2022). Overview of SARS-CoV-2 genome-encoded proteins. 65(2), 280–294.
- Blanco, J. D., Hernandez-Alias, X., Cianferoni, D., & 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(12). <https://doi.org/10.1371/journal.pcbi.1008450>
- Braga, C. L., Silva-Aguiar, R. P., Battaglini, D., Peruchetti, D. B., Robba, C., Pelosi, P., ... Silva, P. L. (2020). The renin–angiotensin–aldosterone system: Role in pathogenesis and potential therapeutic target in COVID-19. *Pharmacology Research & Perspectives*, 8(4), e00623. <https://doi.org/https://doi.org/10.1002/prp2.623>
- Canale, A. S., Cote-Hammarlof, P. A., Flynn, J. M., & Bolon, D. N. A. (2018). Evolutionary mechanisms studied through protein fitness landscapes. *Current Opinion in Structural Biology*, 48, 141–148. <https://doi.org/https://doi.org/10.1016/j.sbi.2018.01.001>
- Chan, K. K., Dorosky, D., Sharma, P., Abbasi, S. A., Dye, J. M., Kranz, D. M., ... Procko, E. (2020). Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2. *Science*, 369(6508), 1261–1265. https://doi.org/10.1126/SCIENCE.ABC0870/SUPPL_FILE/ABC0870_DATA_FILE_S1.XLSX
- Chang, L., Wei, S., T., B. S., G., S. D., Bin, L., & Yang, Y. (2021). Structural basis of mismatch recognition by a SARS-CoV-2 proofreading enzyme. *Science*, 373(6559), 1142–1146. <https://doi.org/10.1126/science.abi9310>
- Chen, Y., Liu, Q., & Guo, D. (2020). Emerging coronaviruses: Genome structure, replication, and pathogenesis. *Journal of Medical Virology*, 92(4), 418–423. <https://doi.org/https://doi.org/10.1002/jmv.25681>

- Cherry, J. D., & Krogstad, P. (2004). SARS: The First Pandemic of the 21st Century. *Pediatric Research*, 56(1), 1–5. <https://doi.org/10.1203/01.PDR.0000129184.87042.FC>
- Chu, D. K. W., Poon, L. L. M., Gomaa, M. M., Shehata, M. M., Perera, R. A. P. M., Abu Zeid, D., ... Kayali, G. (2014). MERS coronaviruses in dromedary camels, Egypt. *Emerging Infectious Diseases*, 20(6), 1049–1053. <https://doi.org/10.3201/eid2006.140299>
- Daniel, W., Nianshuang, W., S., C. K., A., G. J., Ching-Lin, H., Olubukola, A., ... S., M. J. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*, 367(6483), 1260–1263. <https://doi.org/10.1126/science.abb2507>
- De Wit, E., Van Doremalen, N., Falzarano, D., & Munster, V. J. (2016). SARS and MERS: Recent insights into emerging coronaviruses. *Nature Reviews Microbiology*, Vol. 14, pp. 523–534. <https://doi.org/10.1038/nrmicro.2016.81>
- Dejnirattisai, W., Zhou, D., Supasa, P., Liu, C., Mentzer, A. J., Ginn, H. M., ... Screaton, G. R. (2021). Antibody evasion by the P.1 strain of SARS-CoV-2. *Cell*, 184(11), 2939–2954.e9. <https://doi.org/10.1016/j.cell.2021.03.055>
- Dong, S., Sun, J., Mao, Z., Wang, L., Lu, Y.-L., & Li, J. (2020). A guideline for homology modeling of the proteins from newly discovered betacoronavirus, 2019 novel coronavirus (2019-nCoV). *Journal of Medical Virology*, 92(9), 1542–1548. <https://doi.org/https://doi.org/10.1002/jmv.25768>
- Donoghue, M., Hsieh, F., Baronas, E., Godbout, K., Gosselin, M., Stagliano, N., ... Acton, S. (2000). A novel angiotensin-converting enzyme-related carboxypeptidase (ACE2) converts angiotensin I to angiotensin 1-9. *Circulation Research*, 87(5). <https://doi.org/10.1161/01.res.87.5.e1>
- Du, L., He, Y., Zhou, Y., Liu, S., Zheng, B.-J., & Jiang, S. (2009). The spike protein of SARS-CoV — a target for vaccine and therapeutic development. *Nature Reviews Microbiology*, 7(3), 226–236. <https://doi.org/10.1038/nrmicro2090>
- Dubey, A., Choudhary, S., Kumar, P., & Tomar, S. (2021). Emerging SARS-CoV-2 Variants: Genetic Variability and Clinical Implications. *Current Microbiology*, 79(1), 20. <https://doi.org/10.1007/s00284-021-02724-1>
- Durai, P., Batool, M., Shah, M., & Choi, S. (2015). *Middle East respiratory syndrome coronavirus: transmission, virology and therapeutic targeting to aid in outbreak control*. <https://doi.org/10.1038/emm.2015.76>
- Eisenberg, D., Schwarz, E., Komaromy, M., & Wall, R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *Journal of Molecular Biology*, 179(1), 125–142. [https://doi.org/10.1016/0022-2836\(84\)90309-7](https://doi.org/10.1016/0022-2836(84)90309-7)
- Fehr, A. R., & Perlman, S. (2015). Coronaviruses: An overview of their replication and

- pathogenesis. In *Coronaviruses: Methods and Protocols* (Vol. 1282, pp. 1–23). https://doi.org/10.1007/978-1-4939-2438-7_1
- Fowler, D. M., Stephany, J. J., & Fields, S. (2014). Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nature Protocols*, 9(9), 2267–2284. <https://doi.org/10.1038/nprot.2014.153>
- Furuhashi, M., Moniwa, N., Takizawa, H., Ura, N., & Shimamoto, K. (2020). Potential differential effects of renin-angiotensin system inhibitors on SARS-CoV-2 infection and lung injury in COVID-19. *Hypertension Research*, 43(8), 837–840. <https://doi.org/10.1038/s41440-020-0478-1>
- Garcia-Beltran, W. F., Lam, E. C., St. Denis, K., Nitido, A. D., Garcia, Z. H., Hauser, B. M., ... Balazs, A. B. (2021). Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell*, 184(9), 2372–2383.e9. <https://doi.org/10.1016/j.cell.2021.03.013>
- Graham, R. L., Sims, A. C., Baric, R. S., & Denison, M. R. (2006). *The NSP2 Proteins of Mouse Hepatitis Virus and Sars Coronavirus are Dispensable for Viral Replication* BT - *The Nidoviruses* (S. Perlman & K. V Holmes, Eds.). Boston, MA: Springer US.
- Greaney, A. J., Starr, T. N., & Bloom, J. D. (2022). An antibody-escape estimator for mutations to the SARS-CoV-2 receptor-binding domain. *Virus Evolution*, 8(1), veac021. <https://doi.org/10.1093/ve/veac021>
- Gupta, M., Azumaya, C. M., Moritz, M., Pourmal, S., Diallo, A., Merz, G. E., ... Verba, K. A. (2021). CryoEM and AI reveal a structure of SARS-CoV-2 Nsp2, a multifunctional protein involved in key host processes. *BioRxiv: The Preprint Server for Biology*, 2021.05.10.443524. <https://doi.org/10.1101/2021.05.10.443524>
- Hadi-Alijanvand, H., & Rouhani, M. (2020). Studying the Effects of ACE2 Mutations on the Stability, Dynamics, and Dissociation Process of SARS-CoV-2 S1/hACE2 Complexes. *Journal of Proteome Research*, 19(11), 4609–4623. <https://doi.org/10.1021/acs.jproteome.0c00348>
- Hagemans, D., van Belzen, I. A. E. M., Luengo, T. M., & Rüdiger, S. G. D. (2015). A script to highlight hydrophobicity and charge on protein surfaces. *Frontiers in Molecular Biosciences*, 2(OCT), 1–11. <https://doi.org/10.3389/fmolb.2015.00056>
- Han, P., Su, C., Zhang, Y., Bai, C., Zheng, A., Qiao, C., ... Qi, J. (2021). Molecular insights into receptor binding of recent emerging SARS-CoV-2 variants. *Nature Communications* 2021 12:1, 12(1), 1–9. <https://doi.org/10.1038/s41467-021-26401-w>
- Hardenbrook, N. J., & Zhang, P. (2022). A structural view of the SARS-CoV-2 virus and its assembly. *Current Opinion in Virology*, 52, 123–134. <https://doi.org/https://doi.org/10.1016/j.coviro.2021.11.011>

- Honorato, R. V., Koukos, P. I., Jiménez-García, B., Tsaregorodtsev, A., Verlato, M., Giachetti, A., ... Bonvin, A. M. J. J. (2021). *Structural Biology in the Clouds: The WeNMR-EOSC Ecosystem*. <https://doi.org/10.3389/fmolb.2021.729513>
- Huang, X., Zheng, W., Pearce, R., & 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(8), 2429–2437. <https://doi.org/10.1093/bioinformatics/btz926>
- Huang, X., Zheng, W., Pearce, R., Zhang, Y., & 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(8), 2429–2437. <https://doi.org/10.1093/bioinformatics/btz926>
- Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., ... Yang, H. (2020). Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature*, 582(7811), 289–293. <https://doi.org/10.1038/s41586-020-2223-y>
- Kang, S., Yang, M., Hong, Z., Zhang, L., Huang, Z., Chen, X., ... Chen, S. (2020). Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. *BioRxiv*, 2020.03.06.977876. <https://doi.org/10.1101/2020.03.06.977876>
- Kern, D. M., Sorum, B., Mali, S. S., Hoel, C. M., Sridharan, S., Remis, J. P., ... Brohawn, S. G. (2021). Cryo-EM structure of the SARS-CoV-2 3a ion channel in lipid nanodiscs. *BioRxiv*, 2020.06.17.156554. <https://doi.org/10.1101/2020.06.17.156554>
- Kim, Y., Wower, J., Maltseva, N., Chang, C., Jedrzejczak, R., Wilamowski, M., ... Joachimiak, A. (2021). Tipiracil binds to uridine site and inhibits Nsp15 endoribonuclease NendoU from SARS-CoV-2. *Communications Biology*, 4(1), 193. <https://doi.org/10.1038/s42003-021-01735-9>
- Konings, F., Perkins, M. D., Kuhn, J. H., Pallen, M. J., Alm, E. J., Archer, B. N., ... Van Kerkhove, M. D. (2021). SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse. *Nature Microbiology*, 6(7), 821–823. <https://doi.org/10.1038/s41564-021-00932-w>
- Krissinel, E., & Henrick, K. (2007). Inference of Macromolecular Assemblies from Crystalline State. *Journal of Molecular Biology*, 372(3), 774–797. <https://doi.org/10.1016/J.JMB.2007.05.022>
- Kuiken, T., Fouchier, R. A. M., Schutten, M., Rimmelzwaan, G. F., Van Amerongen, G., Van Riel, D., ... Osterhaus, A. D. M. E. (2003). Newly discovered coronavirus as the primary cause of severe acute respiratory syndrome. *Lancet*, 362(9380), 263–270. [https://doi.org/10.1016/S0140-6736\(03\)13967-0](https://doi.org/10.1016/S0140-6736(03)13967-0)
- Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., ... Wang, X. (2020a). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor.

- Nature*, 581(7807), 215–220. <https://doi.org/10.1038/s41586-020-2180-5>
- Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., ... Wang, X. (2020b). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*, 581(7807), 215–220. <https://doi.org/10.1038/s41586-020-2180-5>
- Laurini, E., Marson, D., Aulic, S., Fermeglia, A., & Pricl, S. (2021). Computational Mutagenesis at the SARS-CoV-2 Spike Protein/Angiotensin-Converting Enzyme 2 Binding Interface: Comparison with Experimental Evidence. *ACS Nano*, 15(4), 6929–6948. <https://doi.org/10.1021/acsnano.0c10833>
- Laurini, E., Marson, D., Aulic, S., Fermeglia, M., & Pricl, S. (2020). Computational Alanine Scanning and Structural Analysis of the SARS-CoV-2 Spike Protein/Angiotensin-Converting Enzyme 2 Complex. *ACS Nano*, 14(9), 11821–11830. <https://doi.org/10.1021/acsnano.0c04674>
- Li, F., Li, W., Farzan, M., & Harrison, S. C. (2005). Structural biology: Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science*, 309(5742), 1864–1868. <https://doi.org/10.1126/science.1116480>
- Li, W., Moore, M. J., Vasilieva, N., Sui, J., Wong, S. K., Berne, M. A., ... Farzan, M. (2003). Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature*, 426(6965), 450–454. <https://doi.org/10.1038/nature02145>
- Li, W., Zhang, C., Sui, J., Kuhn, J. H., Moore, M. J., Luo, S., ... Farzan, M. (2005). Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. *EMBO Journal*, 24(8), 1634–1643. <https://doi.org/10.1038/sj.emboj.7600640>
- Li, Z., & Zhang, J. Z. H. (2022). Mutational Effect of Some Major COVID-19 Variants on Binding of the S Protein to ACE2. *Biomolecules*, Vol. 12. <https://doi.org/10.3390/biom12040572>
- Liao, Y., Yuan, Q., Torres, J., Tam, J. P., & Liu, D. X. (2006). Biochemical and functional characterization of the membrane association and membrane permeabilizing activity of the severe acute respiratory syndrome coronavirus envelope protein. *Virology*, 349(2), 264–275. <https://doi.org/10.1016/j.virol.2006.01.028>
- Lin, Z. hua, Long, H. xia, Bo, Z., Wang, Y. qiang, & Wu, Y. zhang. (2008). New descriptors of amino acids and their application to peptide QSAR study. *Peptides*, 29(10), 1798–1805. <https://doi.org/10.1016/J.PEPTIDES.2008.06.004>
- Lippi, G., Mattiuzzi, C., & Henry, B. M. (2022). Updated picture of SARS-CoV-2 variants and mutations. *Diagnosis*, 9(1), 11–17. <https://doi.org/doi:10.1515/dx-2021-0149>
- Littler, D. R., Gully, B. S., Colson, R. N., & Rossjohn, J. (2020). Crystal Structure of the SARS-CoV-2 Non-structural Protein 9, Nsp9. *IScience*, 23(7). <https://doi.org/10.1016/j.isci.2020.101258>

- Livesey, B. J., & Marsh, J. A. (2020). Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Molecular Systems Biology*, 16(7), e9380. <https://doi.org/10.15252/msb.20199380>
- Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., ... Tan, W. (2020a). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet*, 395(10224), 565–574. [https://doi.org/10.1016/S0140-6736\(20\)30251-8](https://doi.org/10.1016/S0140-6736(20)30251-8)
- Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., ... Tan, W. (2020b). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet*, 395(10224), 565–574. [https://doi.org/10.1016/S0140-6736\(20\)30251-8](https://doi.org/10.1016/S0140-6736(20)30251-8)
- Lucas, C., Vogels, C. B. F., Yildirim, I., Rothman, J. E., Lu, P., Monteiro, V., ... Initiative, Y. S.-C.-2 G. S. (2021). Impact of circulating SARS-CoV-2 variants on mRNA vaccine-induced immunity. *Nature*, 600(7889), 523–529. <https://doi.org/10.1038/s41586-021-04085-y>
- Mandala, V. S., McKay, M. J., Shcherbakov, A. A., Dregni, A. J., Kolocouris, A., & Hong, M. (2020). Structure and drug binding of the SARS-CoV-2 envelope protein transmembrane domain in lipid bilayers. *Nature Structural & Molecular Biology*, 27(12), 1202–1208. <https://doi.org/10.1038/s41594-020-00536-8>
- Matthias, T., Robert, B., Michael, A., Lennart, K., Timo, D., Maximilian, H., ... Roland, B. (2020). Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2. *Science*, 369(6508), 1249–1255. <https://doi.org/10.1126/science.abc8665>
- Medina-Enríquez, M. M., Lopez-León, S., Carlos-Escalante, J. A., Aponte-Torres, Z., Cuapio, A., & Wegman-Ostrosky, T. (2020). ACE2: the molecular doorway to SARS-CoV-2. *Cell & Bioscience*, 10(1), 148. <https://doi.org/10.1186/s13578-020-00519-8>
- Melin, A. D., Janiak, M. C., Marrone, F., Arora, P. S., & Higham, J. P. (2020). Comparative ACE2 variation and primate COVID-19 risk. *Communications Biology*, 3(1), 641. <https://doi.org/10.1038/s42003-020-01370-w>
- Moh Zaki, A., Van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D. M. E., & Fouchier, R. A. M. (2012). Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. *NEJM.Org. N Engl J Med*, 367, 1814–1834. <https://doi.org/10.1056/NEJMoa1211721>
- Monica, R.-L., George, M., Ludmilla, S., L., I. N., Olga, K., Joseph, B., & F., S. K. J. (2020). High-resolution structures of the SARS-CoV-2 2'-O-methyltransferase reveal strategies for structure-based inhibitor design. *Science Signaling*, 13(651), eabe1202. <https://doi.org/10.1126/scisignal.abe1202>
- Nguyen, H. L., Lan, P. D., Thai, N. Q., Nissley, D. A., O'Brien, E. P., & Li, M. S. (2020).

- Does SARS-CoV-2 Bind to Human ACE2 More Strongly Than Does SARS-CoV? *The Journal of Physical Chemistry B*, 124(34), 7336–7347. <https://doi.org/10.1021/acs.jpcb.0c04511>
- Nieva, J. L., Madan, V., & Carrasco, L. (2012). Viroporins: structure and biological functions. *Nature Reviews Microbiology*, 10(8), 563–574. <https://doi.org/10.1038/nrmicro2820>
- Örd, M., Faustova, I., & Loog, M. (2020). The sequence at Spike S1/S2 site enables cleavage by furin and phospho-regulation in SARS-CoV2 but not in SARS-CoV1 or MERS-CoV. *Scientific Reports*, 10(1), 16944. <https://doi.org/10.1038/s41598-020-74101-0>
- Pearce, R., Huang, X., Setiawan, D., & 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(13), 2467–2476. <https://doi.org/10.1016/J.JMB.2019.02.028>
- Platto, S., Xue, T., & Carafoli, E. (2020). COVID19: an announced pandemic. *Cell Death & Disease*, 11(9), 799. <https://doi.org/10.1038/s41419-020-02995-9>
- Procko, E., Hedman, R., Hamilton, K., Seetharaman, J., Fleishman, S. J., Su, M., ... Baker, D. (2013). Computational Design of a Protein-Based Enzyme Inhibitor. *Journal of Molecular Biology*, 425(18), 3563–3575. <https://doi.org/https://doi.org/10.1016/j.jmb.2013.06.035>
- R., K. J., John-Paul, B., E., W. E., Ryszard, M., & A., W. T. (2017). Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proceedings of the National Academy of Sciences*, 114(9), 2265–2270. <https://doi.org/10.1073/pnas.1614437114>
- Reusken, C. B. E. M., Haagmans, B. L., Müller, M. A., Gutierrez, C., Godeke, G. J., Meyer, B., ... Koopmans, M. P. G. (2013). Middle East respiratory syndrome coronavirus neutralising serum antibodies in dromedary camels: A comparative serological study. *The Lancet Infectious Diseases*, 13(10), 859–866. [https://doi.org/10.1016/S1473-3099\(13\)70164-6](https://doi.org/10.1016/S1473-3099(13)70164-6)
- Roca-Ho, H., Riera, M., Palau, V., Pascual, J., & Soler, M. J. (2017). Characterization of ACE and ACE2 Expression within Different Organs of the NOD Mouse. *International Journal of Molecular Sciences*, Vol. 18. <https://doi.org/10.3390/ijms18030563>
- Rodrigues, J. P. G. L. M., Barrera-Vilarmau, S., M. C. Teixeira, J., Sorokina, M., Seckel, E., Kastitis, P. L., & Levitt, M. (2020). Insights on cross-species transmission of SARS-CoV-2 from structural modeling. *PLOS Computational Biology*, 16(12), e1008449. Retrieved from <https://doi.org/10.1371/journal.pcbi.1008449>
- Rodrigues, J. P. G. L. M., Teixeira, J. M. C., Trellet, M., & Bonvin, A. M. J. J. (2018). Pdb-tools: A swiss army knife for molecular structures [version 1; referees: 2

- approved]. *F1000Research*, 7, 1–9.
<https://doi.org/10.12688/f1000research.17456.1>
- Saha, I., Ghosh, N., Sharma, N., & Nandi, S. (2021). Hotspot Mutations in SARS-CoV-2 . *Frontiers in Genetics* , Vol. 12. Retrieved from <https://www.frontiersin.org/article/10.3389/fgene.2021.753440>
- Salata, C., Calistri, A., Parolin, C., & Palù, G. (2020). Coronaviruses: A paradigm of new emerging zoonotic diseases. *Pathogens and Disease*, Vol. 77, p. 6. <https://doi.org/10.1093/femspd/ftaa006>
- Samavati, L., & Uhal, B. D. (2020). ACE2, Much More Than Just a Receptor for SARS-COV-2. *Frontiers in Cellular and Infection Microbiology*, 10, 317. <https://doi.org/10.3389/FCIMB.2020.00317/BIBTEX>
- Sarkisyan, K. S., Bolotin, D. A., Meer, M. V, Usmanova, D. R., Mishin, A. S., Sharonov, G. V, ... Kondrashov, F. A. (2016). Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603), 397–401. <https://doi.org/10.1038/nature17995>
- Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Research*, 33(suppl_2), W382–W388. <https://doi.org/10.1093/nar/gki387>
- Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., ... Li, F. (2020). Structural basis of receptor recognition by SARS-CoV-2. *Nature*, 581(7807), 221–224. <https://doi.org/10.1038/s41586-020-2179-y>
- Shapovalov, M. V., & Dunbrack, R. L. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, 19(6), 844–858. <https://doi.org/10.1016/J.STR.2011.03.019/ATTACHMENT/5A0063F1-B3D4-4E72-9AF3-17B9BE652037/MMC2.AVI>
- Sobitan, A., Mahase, V., Rhoades, R., Williams, D., Liu, D., Xie, Y., ... Teng, S. (2021). Computational Saturation Mutagenesis of SARS-CoV-1 Spike Glycoprotein: Stability, Binding Affinity, and Comparison With SARS-CoV-2 . *Frontiers in Molecular Biosciences* , Vol. 8. Retrieved from <https://www.frontiersin.org/article/10.3389/fmolb.2021.784303>
- Sorokina, M., M. C. Teixeira, J., Barrera-Vilarmau, S., Paschke, R., Papatotiriou, I., Rodrigues, J. P. G. L. M., & Kastiris, P. L. (2020). Structural models of human ACE2 variants with SARS-CoV-2 Spike protein for structure-based drug design. *Scientific Data*, 7(1), 309. <https://doi.org/10.1038/s41597-020-00652-6>
- Stalin Raj, V., Mou, H., Smits, S. L., W Dekkers, D. H., Müller, M. A., Dijkman, R., ... Haagmans, B. L. (2013). *Dipeptidyl peptidase 4 is a functional receptor for the emerging human coronavirus-EMC*. <https://doi.org/10.1038/nature12005>
- Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H. D., Dingens, A. S.,

- ... Bloom, J. D. (2020). Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*, 182(5), 1295-1310.e20. <https://doi.org/10.1016/J.CELL.2020.08.012/ATTACHMENT/FE8D0017-8DE9-419E-A27E-E3764D796022/MMC3.ZIP>
- Tomasello, G., Armenia, I., & Molla, G. (2020). The Protein Imager: a full-featured online molecular viewer interface with server-side HQ-rendering capabilities. *Bioinformatics*, 36(9), 2909–2911. <https://doi.org/10.1093/bioinformatics/btaa009>
- Van Zundert, G. C. P., Rodrigues, J. P. G. L. M., Trellet, M., Schmitz, C., Kastitis, P. L., Karaca, E., ... Bonvin, A. M. J. J. (2016a). The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology*, 428(4), 720–725. <https://doi.org/10.1016/J.JMB.2015.09.014>
- Van Zundert, G. C. P., Rodrigues, J. P. G. L. M., Trellet, M., Schmitz, C., Kastitis, P. L., Karaca, E., ... Bonvin, A. M. J. J. (2016b). The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology*, 428(4), 720–725. <https://doi.org/10.1016/j.jmb.2015.09.014>
- W., C. D. K., Y., H. K. P., M., P. R. A. P., Eve, M., Daniela, N., Jincun, Z., ... Malik, P. (2018). MERS coronaviruses from camels in Africa exhibit region-dependent genetic diversity. *Proceedings of the National Academy of Sciences*, 115(12), 3144–3149. <https://doi.org/10.1073/pnas.1718769115>
- Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., McGuire, A. T., & Veesler, D. (2020). Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell*, 181(2), 281-292.e6. <https://doi.org/10.1016/j.cell.2020.02.058>
- Wan, Y., Shang, J., Graham, R., Baric, R. S., & 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(7). <https://doi.org/10.1128/jvi.00127-20>
- Wang, E., Sun, H., Wang, J., Wang, Z., Liu, H., Zhang, J. Z. H., & Hou, T. (2019). End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chemical Reviews*, 119(16), 9478–9508. <https://doi.org/10.1021/acs.chemrev.9b00055>
- Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., ... Qi, J. (2020). Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. *Cell*, 181(4), 894-904.e9. <https://doi.org/10.1016/j.cell.2020.03.045>
- Whitehead, T. A., Chevalier, A., Song, Y., Dreyfus, C., Fleishman, S. J., De Mattos, C., ... Baker, D. (2012). Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nature Biotechnology*, 30(6), 543–548. <https://doi.org/10.1038/nbt.2214>
- William and Reddy, S. T. J. M. and W. C. R. and G. B. and E. R. A. and H. J. and F. L.

- and M. S. W. and Y. A. and K. (2021). Predictive Profiling of SARS-CoV-2 Variants by Deep Mutational Learning. *Cell*, 1(3394), 1–30. <https://doi.org/http://dx.doi.org/10.2139/ssrn.3982671>
- Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., ... Zhang, Y. Z. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798), 265–269. <https://doi.org/10.1038/s41586-020-2008-3>
- Xia, X. (2021). Domains and Functions of Spike Protein in SARS-Cov-2 in the Context of Vaccine Design. *Viruses* 2021, Vol. 13, Page 109, 13(1), 109. <https://doi.org/10.3390/V13010109>
- Yan, G., Liming, Y., Yucen, H., Fengjiang, L., Yao, Z., Lin, C., ... Zihe, R. (2020). Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science*, 368(6492), 779–782. <https://doi.org/10.1126/science.abb7498>
- Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., & Zhou, Q. (2020). Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science*, 367(6485), 1444–1448. https://doi.org/10.1126/SCIENCE.ABB2762/SUPPL_FILE/ABB2762S1.MP4
- Yang, P., & Wang, X. (2020). COVID-19: a new challenge for human beings. *Cellular & Molecular Immunology*, 17. <https://doi.org/10.1038/s41423-020-0407-x>
- Yao, H., Song, Y., Chen, Y., Wu, N., Xu, J., Sun, C., ... Li, S. (2020). Molecular Architecture of the SARS-CoV-2 Virus. *Cell*, 183(3), 730-738.e13. <https://doi.org/https://doi.org/10.1016/j.cell.2020.09.018>
- Zhang, H., Penninger, J. M., Li, Y., Zhong, N., & Slutsky, A. S. (2020). Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive Care Medicine*, 46(4), 586–590. <https://doi.org/10.1007/s00134-020-05985-9>
- Zhang, J., & Chen, B. (2022). Fighting SARS-CoV-2 with structural biology methods. *Nature Methods*, 19(4), 381–383. <https://doi.org/10.1038/s41592-022-01448-9>
- Zhang, J., Xiao, T., Cai, Y., & Chen, B. (2021). Structure of SARS-CoV-2 spike protein. *Current Opinion in Virology*, 50, 173–182. <https://doi.org/10.1016/j.coviro.2021.08.010>
- Zhang, L., Cui, Z., Li, Q., Wang, B., Yu, Y., Wu, J., ... Wang, Y. (2021). Ten emerging SARS-CoV-2 spike variants exhibit variable infectivity, animal tropism, and antibody neutralization. *Communications Biology*, 4(1), 1196. <https://doi.org/10.1038/s42003-021-02728-4>
- Zhang, L., Li, Q., Liang, Z., Li, T., Liu, S., Cui, Q., ... Wang, Y. (2022). The significant immune escape of pseudotyped SARS-CoV-2 variant Omicron. *Emerging Microbes & Infections*, 11(1), 1–5. <https://doi.org/10.1080/22221751.2021.2017757>

- Zhang, N., Chen, Y., Lu, H., Zhao, F., Alvarez, R. V., Goncarenco, A., ... Li, M. (2020). MutaBind2: Predicting the Impacts of Single and Multiple Mutations on Protein-Protein Interactions. *IScience*, 23(3). <https://doi.org/10.1016/J.ISCI.2020.100939>
- Zhou, P., Yang, X. Lou, Wang, X. G., Hu, B., Zhang, L., Zhang, W., ... Shi, Z. L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579(7798), 270–273. <https://doi.org/10.1038/s41586-020-2012-7>
- Zhou, Z., Huang, C., Zhou, Z., Huang, Z., Su, L., Kang, S., ... Chen, S. (2021). Structural insight reveals SARS-CoV-2 ORF7a as an immunomodulating factor for human CD14⁺ monocytes. *IScience*, 24(3). <https://doi.org/10.1016/j.isci.2021.102187>
- Zinzula, L., Basquin, J., Bohn, S., Beck, F., Klumpe, S., Pfeifer, G., ... Baumeister, W. (2021). High-resolution structure and biophysical characterization of the nucleocapsid phosphoprotein dimerization domain from the Covid-19 severe acute respiratory syndrome coronavirus 2. *Biochemical and Biophysical Research Communications*, 538, 54–62. <https://doi.org/https://doi.org/10.1016/j.bbrc.2020.09.131>

APPENDIX A.

List of mutations in benchmarking dataset with experimental and predicted binding values. UEP dataset was indicated with asteriks.

Case ID	Experimental Binding Value	Protein	HADDOCK $\Delta\Delta G$	FoldX $\Delta\Delta G$	FoldXwater $\Delta\Delta G$	EvoEF1 $\Delta\Delta G$	MutaBind2 $\Delta\Delta G$	SSIPe $\Delta\Delta G$
A386C	-0.66	ACE2	2.93	0.00	0.00	-0.04	0.32	0.19
A386E	-0.86	ACE2	-1.27	0.16	0.20	0.05	0.25	0.22
A386I	1.65	ACE2	-3.25	-0.22	-0.26	-0.01	0.28	0.20
A386L	2.68	ACE2	-1.30	0.12	-0.39	-0.12	0.30	0.16
A386M	0.32	ACE2	-1.02	-0.06	-0.08	-0.12	0.20	0.16
A386Q	-1.59	ACE2	-3.86	0.01	-0.08	0.01	0.02	0.21
D30A *	-1.59	ACE2	9.32	1.47	1.50	1.23	0.37	0.85
D30E *	2.29	ACE2	1.60	0.01	0.70	0.25	-0.07	0.71
D30F *	-3.07	ACE2	5.73	0.23	1.46	1.24	0.63	0.19
D30I *	1.62	ACE2	6.26	0.00	0.57	0.74	0.53	0.40
D30N *	-1.43	ACE2	13.20	-0.91	0.12	0.14	0.26	0.84
D30P *	-1.92	ACE2	9.51	-0.30	1.83	1.25	0.95	2.07
D30T *	0.54	ACE2	14.01	-0.67	0.30	0.72	0.38	0.26
D30V *	1.24	ACE2	8.82	-1.18	0.07	0.61	0.62	0.21
D38C *	-2.55	ACE2	8.36	1.92	1.71	0.52	2.77	0.47
D38G	-3.01	ACE2	-1.15	1.39	1.11	0.85	2.31	1.34
D38R *	-2.12	ACE2	1.94	1.23	1.03	0.78	1.62	0.30
D38W *	-2.44	ACE2	-1.21	3.42	3.26	0.32	2.20	1.88
D38Y *	-2.35	ACE2	0.17	1.86	2.64	1.22	1.02	0.57
D35I *	-1.54	ACE2	4.76	1.66	1.11	1.36	0.81	0.90
D35K *	-2.37	ACE2	-7.33	2.03	0.95	0.28	0.57	0.74
D35M *	-2.41	ACE2	-12.42	2.79	0.52	0.56	0.69	-0.28
D35P *	-2.53	ACE2	-4.17	1.00	0.95	1.09	0.80	0.02
E35A	0.59	ACE2	-4.94	-0.14	-0.21	-0.10	0.56	0.53
E35C	1.79	ACE2	-3.83	-0.16	-0.22	-0.14	0.59	0.23
E35D	1.41	ACE2	-0.27	0.03	0.06	-0.04	0.52	0.85

E35M	1.04	ACE2	-2.46	-0.11	-0.16	-0.13	0.17	-0.03
E35Q	0.32	ACE2	-3.42	-0.17	-0.26	-0.32	0.07	0.96
E35V	1.74	ACE2	-4.45	-0.23	-0.30	-0.14	0.45	0.09
E37C	-2.86	ACE2	1.46	2.33	2.65	0.55	1.11	0.55
E37F	-2.58	ACE2	-0.87	1.99	2.79	0.59	0.92	0.18
E37G	-1.64	ACE2	2.65	2.54	2.40	0.62	1.16	0.85
E37I	-2.75	ACE2	0.16	1.87	2.16	0.50	0.85	0.53
E37K	-1.35	ACE2	3.01	1.06	2.00	-0.31	0.83	1.12
E37L	-1.12	ACE2	-0.09	1.52	1.66	0.22	0.76	-0.30
E37M	-1.19	ACE2	-4.57	2.22	2.41	0.02	1.01	0.10
E37R	-1.95	ACE2	3.49	1.48	0.95	-0.60	0.79	0.62
E37T	-1.14	ACE2	4.70	2.15	1.64	0.52	0.92	0.40
E37V	-1.50	ACE2	-2.16	1.43	2.15	0.51	0.91	0.40
E484K	-1.93	ACE2	4.32	-0.18	-0.36	0.15	1.11	1.02
E484Q	-3.29	ACE2	2.50	-2.26	-2.63	0.16	2.62	1.63
E484R	-2.46	ACE2	0.25	-0.01	-0.12	0.15	1.30	1.30
E484S	-3.34	ACE2	2.32	-0.01	-0.18	0.15	1.28	1.71
E484T	-1.53	ACE2	-0.55	-0.41	-0.39	0.03	1.22	0.24
F28C	-2.43	ACE2	0.91	-1.16	-1.38	-0.40	1.33	1.46
F28E	-2.36	ACE2	0.99	-0.16	-0.17	0.15	1.39	1.91
F28G	-1.78	ACE2	5.20	-0.01	-0.12	0.56	1.54	1.03
F28K	-2.05	ACE2	1.74	-0.44	-0.13	0.17	1.22	0.89
F28L	0.80	ACE2	-3.81	-0.03	-0.18	-0.22	0.99	0.51
F28N	-2.89	ACE2	-2.11	-0.30	-0.32	-0.24	0.24	0.12
F28Q	-2.49	ACE2	-17.94	4.63	4.67	1.81	1.43	0.88
F28T	-2.61	ACE2	-9.97	2.59	2.56	-0.28	0.54	0.18
F28V	-1.75	ACE2	-6.00	-0.31	-0.87	-0.45	-0.70	0.08
F28Y	-2.10	ACE2	-16.99	6.17	0.50	1.55	1.58	0.88
F456V *	-1.97	ACE2	-10.49	4.19	4.03	5.02	1.35	2.00
F490A	-2.50	ACE2	-6.64	1.97	1.87	-1.03	1.16	0.25
F490K	-3.24	ACE2	-3.72	0.70	0.80	-0.67	-0.06	1.13
F490M	-2.50	ACE2	-9.05	-0.04	-0.08	1.15	0.54	1.85

F490N	-1.75	ACE2	-12.63	0.89	1.08	-2.88	1.58	-0.84
F490R	3.17	ACE2	6.28	0.57	0.43	0.35	0.91	0.92
F490Y	-2.43	ACE2	-5.55	-1.52	-2.10	-0.74	0.62	-0.11
G326C	-1.86	ACE2	4.23	-0.60	0.37	-0.38	0.21	1.16
G326F	2.66	ACE2	6.89	0.42	1.83	0.05	0.93	2.03
G326H	-1.60	ACE2	2.48	0.53	0.78	0.10	0.05	1.31
G326T	1.69	ACE2	11.44	0.36	0.62	-0.43	0.68	1.00
G326W	2.87	ACE2	3.67	0.35	0.67	-0.47	0.38	0.22
G326Y	0.75	ACE2	-14.16	-1.27	-1.42	-0.04	0.59	-0.35
G354I *	-0.43	ACE2	0.50	-1.34	-0.25	-0.72	0.50	-0.49
G354K *	-2.73	ACE2	9.04	2.57	2.57	0.62	2.18	1.74
G354Q *	2.56	ACE2	-21.15	-0.24	-0.61	0.10	0.99	-0.53
G354Y *	1.85	ACE2	-12.35	-1.79	-2.54	-1.31	0.30	-0.88
G446S	-2.66	ACE2	3.82	2.36	1.69	0.97	1.79	0.68
G476L	-2.44	ACE2	-1.59	2.22	2.39	1.06	1.60	1.03
G476S	-1.29	ACE2	-1.30	3.21	2.94	0.68	2.10	0.77
G502A *	-2.52	ACE2	1.29	2.47	1.72	1.66	1.07	0.91
G502E *	-2.57	ACE2	-7.75	3.41	3.24	0.26	0.96	0.46
G502F *	-1.44	ACE2	0.19	2.49	-0.32	0.71	0.15	0.80
G502I *	-2.53	ACE2	2.91	3.00	2.15	1.25	1.68	1.09
G502K *	-2.52	ACE2	5.15	5.08	4.79	0.78	2.24	0.18
G502L *	-2.72	ACE2	-12.99	1.75	1.45	1.20	1.41	-0.10
G502M *	-2.48	ACE2	3.49	0.44	0.42	0.15	0.92	0.26
G502N *	0.20	ACE2	-1.19	0.19	0.23	0.01	0.38	0.21
G502V *	-2.14	ACE2	5.14	0.22	0.00	-1.19	0.67	-0.10
H34A *	2.68	ACE2	-4.53	-0.28	-0.46	-0.50	0.16	0.76
H34F *	-0.87	ACE2	3.08	1.14	1.01	0.84	0.59	1.85
H34K *	2.90	ACE2	2.74	0.17	0.16	0.58	0.27	1.48
H34P *	1.03	ACE2	-0.18	-0.08	-0.09	-0.14	-0.19	0.96
H34Q *	1.86	ACE2	3.32	1.24	1.14	0.82	0.91	3.05
H34S *	1.06	ACE2	-1.50	0.47	0.53	-0.86	-0.89	1.88
H34V *	3.37	ACE2	2.78	0.67	0.31	0.67	0.27	1.38

K31F *	3.18	ACE2	4.07	0.09	0.26	0.65	0.32	1.37
K31M *	3.29	ACE2	-4.39	-0.47	-0.70	-1.40	0.24	0.23
K31P *	2.86	ACE2	-8.26	-0.01	-0.15	-0.49	-0.08	0.90
K31W *	1.97	ACE2	1.24	1.19	0.86	0.69	0.32	1.09
K31Y *	1.11	ACE2	3.69	1.58	1.03	1.10	0.76	1.64
K353A *	0.27	ACE2	-0.91	0.23	0.38	-0.17	0.28	0.06
K353D *	-1.39	ACE2	4.13	1.69	1.59	0.60	1.25	2.55
K353E *	1.09	ACE2	3.98	0.46	-0.01	-0.29	-0.21	1.21
K353G *	-0.99	ACE2	2.31	0.10	-0.21	-0.15	1.44	0.80
K353H *	2.93	ACE2	-6.38	-1.28	-1.09	-0.81	0.41	-0.40
K353R *	2.64	ACE2	1.48	-0.05	0.18	0.22	0.63	0.63
K353S *	2.18	ACE2	0.20	-0.76	-0.54	-1.04	0.61	-0.85
K353V *	1.27	ACE2	3.25	-0.27	-0.22	-0.50	0.68	0.86
K353W *	-1.59	ACE2	-0.30	-0.49	-0.26	-0.57	0.47	0.49
K417G	2.95	ACE2	-3.45	-1.14	-0.97	-1.45	0.54	-0.86
L45G	3.08	ACE2	-2.99	-0.74	-0.54	0.86	0.40	0.09
L45I *	-2.37	ACE2	-5.37	2.92	2.47	-0.70	1.17	-0.48
L45K *	-1.59	ACE2	5.07	0.29	-0.62	-0.96	0.75	0.57
L79F	1.57	ACE2	6.66	-0.39	-0.77	-0.70	0.65	-0.24
L79G	-3.38	ACE2	-2.75	2.69	2.49	2.41	1.41	0.49
L79I	0.29	ACE2	2.39	-0.02	-0.14	-0.01	0.19	0.20
L79M	0.36	ACE2	2.67	0.01	0.01	-0.01	0.14	0.20
L79P	2.76	ACE2	0.75	-0.03	-0.03	-0.05	-0.32	0.18
L79R	-0.32	ACE2	1.47	-0.02	0.00	0.01	0.30	0.20
L79T	0.38	ACE2	0.66	-0.01	0.01	-0.01	0.11	0.20
L79V	0.37	ACE2	2.09	-0.97	0.01	-0.35	0.21	0.08
L79W	2.85	ACE2	2.82	-0.01	0.01	1.96	0.86	0.71
L79Y	1.40	ACE2	-4.93	0.03	-0.73	2.04	0.57	1.04
L455M *	1.71	ACE2	1.54	-0.15	-0.21	1.91	0.70	0.69
L455P	1.73	ACE2	-3.41	-0.05	0.02	0.66	0.91	1.12
M82C	2.74	ACE2	-0.53	0.01	-0.32	1.96	0.76	-0.03
M82G	2.48	ACE2	2.02	-0.03	-0.15	1.64	0.67	0.34

M82L	0.84	ACE2	0.07	0.07	-2.00	0.24	0.45	0.67
M82P	0.39	ACE2	7.66	-0.06	0.02	2.02	0.69	0.60
M82R	1.92	ACE2	3.37	-0.06	0.03	1.99	0.58	0.59
N330D	-2.44	ACE2	-7.10	-0.22	0.01	0.77	0.70	0.08
N330F	-2.41	ACE2	2.63	0.29	0.25	-0.21	1.83	0.59
N330H	-2.62	ACE2	4.49	0.80	0.54	-0.21	2.22	0.90
N330L	-2.42	ACE2	4.52	0.99	0.79	-0.14	1.80	0.77
N330Q	-2.40	ACE2	0.82	0.84	1.15	-0.35	2.34	-0.12
N330R	-1.84	ACE2	3.39	0.32	0.34	-0.21	2.36	0.59
N330W	-2.48	ACE2	6.38	0.37	0.45	-0.21	1.88	0.31
N330Y	-2.61	ACE2	3.88	0.41	0.54	-0.21	1.84	0.90
N487D *	-2.44	ACE2	5.04	0.67	0.55	-0.21	2.05	1.80
N487P *	0.95	ACE2	2.15	0.10	0.23	-0.08	1.04	0.74
N487Q *	0.24	ACE2	0.68	0.41	0.13	-0.08	1.52	0.74
N501C *	1.98	ACE2	-1.21	0.04	-0.02	-0.08	0.59	1.21
N501F *	0.98	ACE2	2.62	0.02	-0.09	-0.08	0.88	0.19
N501Q *	-0.21	ACE2	0.38	0.02	0.46	-0.08	1.38	1.04
N501S *	1.13	ACE2	-0.41	0.06	0.03	0.00	0.16	-0.13
N501T *	0.77	ACE2	-2.34	0.04	0.02	0.00	0.20	0.26
N501V *	-0.49	ACE2	1.58	0.04	0.04	0.00	0.17	-0.48
N501W *	-0.39	ACE2	3.59	0.07	0.04	0.00	0.09	0.85
N501Y *	2.33	ACE2	-2.97	0.01	-0.02	0.01	0.19	1.75
Q24F *	0.72	ACE2	4.87	0.00	0.00	0.00	0.10	0.13
Q24K *	1.09	ACE2	-1.22	0.04	0.03	0.00	0.17	0.13
Q24T *	1.07	ACE2	-1.44	0.07	0.03	0.00	0.16	-0.36
Q24Y *	1.05	ACE2	-1.07	0.06	0.04	0.00	0.16	-0.25
Q42C *	2.11	ACE2	6.42	0.85	1.05	1.01	0.52	1.51
Q42H *	2.42	ACE2	3.56	0.02	0.94	0.22	0.95	0.96
Q42I *	2.10	ACE2	0.69	-0.39	-0.52	1.02	0.73	1.84
Q42K *	2.29	ACE2	-11.56	-1.80	-1.68	-0.67	0.18	0.28
Q42L *	2.34	ACE2	-8.12	-0.23	0.45	-0.98	0.45	0.91
Q42M *	1.65	ACE2	0.74	0.24	0.25	-1.18	0.35	1.25

Q42R *	2.85	ACE2	0.70	-0.82	-0.61	-0.94	0.27	-0.20
Q42T *	2.55	ACE2	3.04	-1.32	-0.73	-0.55	0.21	0.42
Q42V *	0.51	ACE2	1.85	0.97	1.21	1.01	1.21	2.71
Q42W *	0.25	ACE2	-2.63	0.83	0.74	-0.64	0.66	1.15
Q325G	0.77	ACE2	-2.67	0.60	1.53	1.16	0.59	1.88
Q325H	2.11	ACE2	-16.64	-0.10	-0.10	-0.89	0.76	0.01
Q325P	2.70	ACE2	-6.54	-1.53	-1.51	-0.77	0.47	0.37
Q325S	1.40	ACE2	2.15	0.01	0.04	0.00	0.24	0.00
Q325T	2.49	ACE2	-6.73	-0.34	-1.12	0.15	0.23	0.26
Q325Y	-0.79	ACE2	6.77	0.06	0.02	0.00	0.61	0.00
Q493A *	2.69	ACE2	1.48	0.01	-0.07	0.00	-0.85	0.00
Q493F *	0.49	ACE2	-1.17	0.18	-0.81	0.74	0.24	0.47
Q493G *	1.05	ACE2	2.37	0.01	0.22	0.00	0.06	0.00
Q493K *	-1.78	ACE2	6.90	0.47	1.09	1.80	2.65	2.08
Q493L *	-3.52	ACE2	5.16	0.81	0.84	1.80	2.62	1.80
Q493M *	-1.75	ACE2	3.14	0.42	-0.02	1.71	1.76	1.51
Q493V *	1.68	ACE2	2.38	2.25	0.81	0.98	2.35	2.02
Q493Y *	-2.24	ACE2	0.65	2.07	0.60	0.83	1.87	0.87
Q498F *	0.41	ACE2	6.91	1.76	1.83	1.09	2.10	1.37
Q498H *	-1.93	ACE2	4.98	2.28	1.71	1.22	2.48	2.01
Q498P *	-2.57	ACE2	10.04	2.04	1.65	1.20	2.49	1.83
Q498W *	0.45	ACE2	1.44	1.67	1.32	0.73	1.40	0.51
Q498Y *	-2.35	ACE2	9.20	1.98	1.77	1.14	2.15	2.14
R357A	-1.70	ACE2	6.15	1.60	1.35	1.33	1.98	1.90
R357D	0.74	ACE2	7.27	1.75	1.83	1.07	2.09	1.23
R357E	0.06	RBD	3.92	-0.14	-0.16	-0.16	0.16	0.95
R357F	0.03	RBD	2.86	-0.08	-0.12	-0.04	0.11	1.00
R357G	0.15	RBD	1.62	-0.17	-0.20	0.12	0.13	0.74
R357I	-0.02	RBD	2.13	-0.09	-0.12	-0.19	-0.02	0.78
R357N	0.05	RBD	4.06	-0.08	-0.10	-0.19	0.04	0.05
R357P	-0.61	RBD	5.10	2.09	2.01	1.58	1.19	1.38
R393A	0.03	RBD	-0.57	0.01	0.00	0.01	0.34	0.21

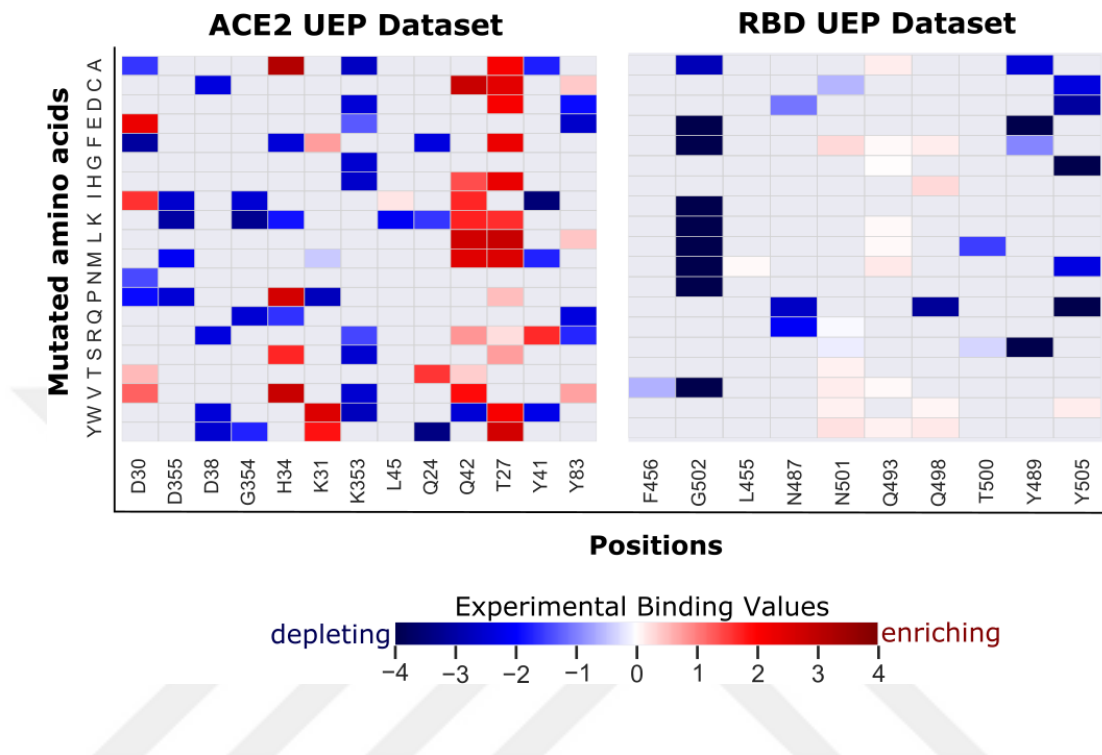
R393H	0.09	RBD	0.61	-0.01	-0.04	0.00	-0.02	0.00
R393K	-0.03	RBD	-2.56	0.01	0.00	0.00	0.09	0.00
R393M	0.02	RBD	-1.32	0.01	0.00	0.03	0.37	0.21
R393S	0.01	RBD	-4.96	-0.15	-0.12	0.27	0.27	0.30
R403W	0.06	RBD	-0.34	0.00	0.00	0.00	0.03	0.00
S19F	-0.20	RBD	-4.64	-0.11	-0.43	0.05	-0.18	0.99
S19I	-2.51	RBD	-2.88	1.44	0.91	-0.03	1.10	-0.36
S19L	-0.02	RBD	0.79	1.06	0.60	0.01	0.58	0.98
S19N	-2.80	RBD	-2.68	2.28	2.27	0.82	1.42	1.18
S19P	-4.67	RBD	-6.96	3.31	3.29	1.29	1.34	1.51
S19T	-4.80	RBD	-26.17	2.15	2.29	-0.57	1.05	0.04
S19V	-4.66	RBD	-2.73	3.42	3.40	1.07	0.76	0.98
S19W	-4.16	RBD	-9.05	2.57	2.58	2.59	0.97	2.30
S19Y	-4.80	RBD	-4.64	1.98	1.93	-0.18	0.98	-0.20
S477D	-4.80	RBD	-6.36	1.98	2.27	-0.80	1.20	0.07
S477K	-4.38	RBD	-1.27	2.64	2.48	1.18	1.09	1.67
S477N	-4.80	RBD	-6.64	3.39	3.36	0.75	1.01	0.73
S477P	-0.51	RBD	8.95	0.92	1.15	1.32	0.64	0.61
S477W	0.05	RBD	3.90	0.19	0.03	-0.59	0.20	0.84
T27A *	-0.95	RBD	-6.63	2.16	2.30	0.95	2.47	3.14
T27C *	-1.08	RBD	-3.98	-0.11	-0.89	0.05	1.22	0.61
T27D *	-2.64	RBD	5.25	2.19	2.12	-0.49	1.83	1.39
T27F *	-2.09	RBD	1.54	1.89	1.08	-0.19	1.39	0.76
T27H *	-0.58	RBD	-7.22	-0.11	-0.48	-0.01	-0.13	0.10
T27K *	0.29	RBD	-3.89	5.58	6.88	0.11	0.63	-0.26
T27L *	-0.06	RBD	-0.56	0.25	-0.30	-0.49	0.87	0.78
T27M *	-0.13	RBD	-6.11	0.21	0.40	0.41	-0.40	0.93
T27P *	0.10	RBD	1.18	-1.17	-1.93	0.01	-0.61	-0.16
T27R *	0.15	RBD	-3.01	-0.91	-1.34	0.01	-0.15	-0.03
T27S *	0.11	RBD	-18.15	4.05	2.80	-0.01	1.48	-0.85
T27W *	0.24	RBD	1.90	3.03	6.28	0.32	0.65	-0.17
T27Y *	0.13	RBD	6.13	0.24	0.24	0.75	0.54	0.44

T324A	0.06	RBD	-8.64	-1.33	-1.40	-0.44	-0.05	-0.63
T324E	0.01	RBD	5.74	0.36	0.38	0.77	0.98	0.45
T324G	0.05	RBD	-10.77	-0.41	-0.31	-0.73	0.12	0.47
T324P	0.05	RBD	-0.25	-0.89	-0.38	0.06	0.02	-0.84
T324Q	0.18	RBD	2.27	-1.08	-0.96	-1.06	-0.02	-0.74
T324S	0.05	RBD	-0.59	0.59	0.57	0.54	0.41	-0.05
T500L *	0.12	RBD	0.74	-1.54	-2.10	-1.30	-0.07	-0.99
T500S *	0.15	RBD	-7.18	-0.79	-1.56	-0.67	-1.65	-0.63
V445K	0.30	RBD	-5.97	-0.67	-0.36	-0.42	0.32	0.13
V445P	-3.14	RBD	2.41	-0.27	-0.23	-0.15	0.92	1.42
V445Q	0.07	RBD	-9.42	-1.35	-3.48	0.04	0.55	-0.62
V503I	0.16	RBD	-9.40	-1.61	-2.58	-0.64	-1.84	-0.86
V503K	-2.82	RBD	-4.20	-0.38	0.24	-0.18	0.68	0.14
V503L	0.09	RBD	1.34	-0.08	-0.10	0.20	0.20	0.28
V503M	0.03	RBD	-0.88	0.03	0.03	0.32	0.18	0.32
V503R	0.06	RBD	1.61	0.04	0.04	0.00	-0.11	0.00
Y41A *	0.06	RBD	-3.39	0.00	0.00	0.06	-0.20	0.23
Y41I *	0.02	RBD	-8.41	0.12	0.11	-0.19	0.16	0.14
Y41M *	-1.52	RBD	-8.07	-0.40	-0.65	1.24	0.47	0.59
Y41R *	-0.32	RBD	-1.21	0.21	0.11	1.62	0.39	2.06
Y41W *	0.02	RBD	-1.54	-0.27	-0.45	-0.09	-0.30	0.17
Y83C *	0.03	RBD	-12.87	0.04	0.05	0.03	-0.03	0.21
Y83D *	-0.04	RBD	-0.29	-0.24	-0.27	0.14	0.09	0.25
Y83E *	0.05	RBD	-4.52	-0.08	-0.08	-0.11	0.07	0.26
Y83L *	0.10	RBD	-5.96	-0.68	-0.77	-1.45	0.12	-0.09
Y83Q *	0.03	RBD	-2.04	-0.13	0.02	-0.04	0.07	0.28
Y83R *	0.10	RBD	-5.31	-0.01	-0.02	-0.40	0.08	0.15
Y83V *	0.01	RBD	-1.09	-0.24	-1.13	-1.19	0.11	-0.01
Y453D	-3.49	RBD	-2.31	0.07	0.06	-0.09	1.23	1.74
Y453E	-1.18	RBD	-1.80	-0.05	-0.08	-0.04	1.43	1.60
Y453F	0.25	RBD	0.99	-0.41	-0.31	-0.36	-0.19	0.68
Y453K	0.08	RBD	-4.07	0.52	0.45	-0.50	0.73	1.87

Y453P	-4.80	RBD	-5.90	0.12	0.14	-0.05	0.78	2.65
Y473D	-2.72	RBD	-0.65	0.41	0.28	0.20	0.90	1.75
Y473K	-2.59	RBD	-6.76	0.26	0.01	0.10	0.45	1.88
Y489A *	-2.46	RBD	17.21	1.86	1.70	1.64	2.23	2.02
Y489E *	-3.98	RBD	6.06	2.88	1.64	1.75	2.00	2.21
Y489F *	-0.95	RBD	-5.79	-0.35	-0.58	0.38	0.28	0.93
Y489S *	-4.80	RBD	12.42	1.23	1.32	1.78	1.89	2.37
Y505C *	-2.36	RBD	4.93	2.88	3.48	2.48	0.80	2.20
Y505D *	-3.07	RBD	4.03	3.29	3.81	3.08	2.05	2.78
Y505G *	-4.65	RBD	3.51	3.45	3.43	3.14	1.11	2.67
Y505M *	-2.32	RBD	5.45	2.52	2.54	1.53	0.93	1.38
Y505P *	-4.09	RBD	10.26	3.43	3.58	2.73	1.47	3.75
Y505W *	0.13	RBD	-5.56	0.70	-0.05	1.27	0.50	1.05

APPENDIX B.

Experimental binding values distribution of the UEP dataset was provided.



APPENDIX C.

The ConSurf values of ACE2-RBD interface residues were presented.

Protein	Interface Residue Position	ConSurf score
ACE2	S19	0
ACE2	Q24	1
ACE2	T27	1
ACE2	F28	7
ACE2	D30	1
ACE2	K31	1
ACE2	H34	1
ACE2	E35	2
ACE2	E37	3
ACE2	D38	1

ACE2	Y41	4
ACE2	Q42	1
ACE2	L45	2
ACE2	L79	1
ACE2	M82	1
ACE2	Y83	3
ACE2	T324	8
ACE2	Q325	1
ACE2	G326	5
ACE2	N330	5
ACE2	K353	1
ACE2	G354	3
ACE2	D355	9
ACE2	R357	9
ACE2	A386	5
ACE2	R393	7
RBD	R403	0
RBD	K417	8
RBD	V445	1
RBD	G446	0
RBD	Y449	1
RBD	Y453	9
RBD	L455	1
RBD	F456	1
RBD	Y473	0
RBD	A475	1
RBD	G476	1
RBD	S477	0
RBD	E484	1
RBD	G485	1
RBD	F486	1
RBD	N487	7
RBD	Y489	0
RBD	F490	1
RBD	Q493	1
RBD	G496	0
RBD	Q498	1

RBD	T500	0
RBD	N501	1
RBD	G502	0
RBD	V503	1
RBD	Y505	0

APPENDIX D.

List of van der Waals, hydrophobicity, and flexibility values of amino acids

Amino Acid	van Der Waals Volume (nm ³)	Hydrophobicity (a.u)	Flexibility (a.u)
A	0.05702	0.62	1
R	0.58946	-2.53	81
N	0.22972	-0.78	3
D	0.21051	-0.90	3
C	0.14907	0.29	3
Q	0.34861	-0.85	9
E	0.32837	-0.74	9
G	0.00279	0.48	1
H	0.37694	-0.40	3
I	0.37671	1.38	9
L	0.37876	1.06	9
K	0.45363	-1.50	81
M	0.38872	0.64	27
F	0.55298	1.19	3
P	0.2279	0.12	2
S	0.09204	-0.18	3
T	0.19341	-0.05	3
W	0.79351	0.81	3
Y	0.6115	0.26	3
V	0.25674	1.08	3



İZMİR BİYOTİP VE GENOM MERKEZİ
GİRİŞİMSEL OLMAYAN ARAŞTIRMALAR ETİK KURULU (İBG-GOEK)
KARARI

Toplantı Tarihi	: 26/07/2021	Toplantı Günü	: Pazartesi
Toplantı Sayısı	: 8	Toplantı Saati	: 10:30

Sayın Dr. Ezgi KARACA,

2021-019 Protokol No'lu; sorumlusu olduğunuz "Benchmarking the binding affinity predictors over the Spike:ACE2 deep mutagenesis data set" başlıklı araştırmanın uygulanmasında etik açıdan sakınca olmadığına oy birliği ile karar verilmiştir.

Bilgilerinizi ve gereğini rica ederiz.

Dr. Serap ERKEK ÖZHAN
Başkan V.

Pandemi süresince online ortamda gerçekleştirilen toplantımızda alınan kararlar tek imzalı olarak düzenlenmektedir. Pandemi sona erdikten sonra ıslak imzalı karar belgesi teslim edilecektir.



EDA ŞAMILOĞLU TENGİRŞEK

Personal Information

Contact Information

Contact Address

Telephone

Email

Web Page

04 February 2020 - Now (2 Year 4 Month)

Yüksek Lisans, Tezli Program, DOKUZ EYLÜL ÜNİVERSİTESİ, TÜRKİYE
İZMİR ULUSLARARASI BİYOTIP VE GENOM ENSTİTÜSÜ, MOLEKÜLER BİYOLOJİ VE
GENETİK (YL) (TEZLİ) (İNGİLİZCE)

Cumulative Grade Point Average: 3.93 / 4.0

Education Information

01 September 2014 - 17 June 2019 (4 Year 10 Month)

Lisans, Anadal/Normal Öğretim, İSTANBUL ÜNİVERSİTESİ, TÜRKİYE
FEN FAKÜLTESİ, MOLEKÜLER BİYOLOJİ VE GENETİK PR.

Diploma Number: 24598

Cumulative Grade Point Average: 3.49 / 4.0

İNGİLİZCE (Reading: İyi, Writing: İyi, Speaking: İyi)

Foreign Language Information

Scientific Technological Activity Fields

Keywords

Yapısal Biyoenformatik

Hesaplamalı Biyoloji

Biyoinformatik

Atomik ve Moleküler Etkileşimler

R&D Competency

Books

E. KARACA, M. İLTER, Ö. ŞENSOY, E. ŞAMILOĞLU TENGİRŞEK & N. B. İMAT, Protein Etkileşimlerinin DeneySEL ve Hesaplamalı Yaklaşımlarla Bütüncül Olarak Modellenmesi, S. E. ACUNER ZORLUYUSAL [Editors], Protein Yapısı, Mühendisliği, Etkileşimleri Dinamiği ve İlaç Tasarımındaki Yeri(203 - 221), ISBN: 49891, TÜRKİYE: Nobel Tıp, 31 December 2021, Kitapta Bölüm.

Articles

M. KOŞACA, E. Ş. TENGİRŞEK & E. KARACA EREK, Arayüz Mutasyonların Protein Etkileşimlerine Tesirini Tahmin Eden Algoritmalarla HADDOCKun Performansını Karşılaştırması, International Journal of Advances in Engineering and Pure Sciences, 2021, 2636-8277, 33, 4, 592--608.

Conference Papers

E. ŞAMİLOĞLU TENGİRŞEK, Benchmarking the Widely Used Structure-based Binding Affinity Predictors on the variant Spike-ACE2 Interaction Dataset, Sözlü Sunum, 5th International Symposium on Bioinformatics, 15 December 2021, 17 December 2021.

TÜBİTAK Scholarships and Fundings

Number of Panelist/Observer/Reporter Jobs

Refereeing/Panelist/Outer Consultant Count	ARDEB/BİDEB 0	TEYDEB 0	Total 0
Number of Observer/Consultant Jobs	ARDEB/BİDEB 0	TEYDEB 0	Total 0
Number of Reporter Jobs	ARDEB/BİDEB 0	TEYDEB 0	Total 0