ORIGINAL RESEARCH ARTICLE

Globasci

Bond strength between receptor binding domain of spike protein and human angiotensin converting enzyme-2 using machine learning

Abdulmateen Adebiyi[1], Puja Adhikari[2], Praveen Rao[1], Wai-Yim Ching[2*]

[1]Department of Electrical Engineering and Computer Science, University of Missouri-Columbia, Columbia, MO 65212, USA.

[2]Department of Physics and Astronomy, University of Missouri-Kansas City, Kansas City, MO 64110, USA.

*Correspondence to: Wai-Yim Ching, Department of Physics and Astronomy, University of Missouri-Kansas City, Kansas City, MO 64110, USA; E-mail: chingw@umkc.edu; ORCID ID: 0000-0001-7738-8822.

How to cite: Adebiyi A., Adhikari P., Rao P. and Ching W.Y. Bond strength between receptor binding domain of spike protein and human angiotensin converting enzyme-2 using machine learning. *BME Horizon*, 2024; vol2(1). DOI: https://doi.org/10.37155/2972-449X-vol2(1)-110.

Abstract: The spike protein (S-protein) of SARS-CoV-2 plays an important role in binding, fusion, and host entry. In this study, we have predicted interatomic bond strength between receptor binding domain (RBD) and angiotensin converting enzyme-2 (ACE2) using machine learning (ML), that matches with expensive *ab initio* calculation result. We collected bond order result from *ab initio* calculations. We selected a total of 18 variables such as bond type, bond length, elements and their coordinates, and others, to train ML models. We then trained five well-known regression models, namely, Decision Tree regression, KNN Regression, XGBoost, Lasso Regression, and Ridge Regression. We tested these models on two different datasets, namely, Wild type (WT) and Omicron variant (OV). In the first setting, we used 90% of each dataset for training and 10% for testing to predict the bond order. XGBoost model outperformed all the other models in the prediction of the WT dataset. It achieved an R2 Score of 0.997. XGBoost also outperformed all the other models with an R2 score of 0.9998 in the prediction of the OV dataset. In the second setting, we trained all the models on the WT (or OV) dataset and predicted the bond order on the OV (or WT) dataset. Interestingly, Decision Tree outperformed all the other models in both cases. It achieved an R2 score of 0.997.

Keywords: Machine learning; Spike protein; RBD-ACE2 interface; Interatomic bonding; *Ab initio* calculations; XGBoost; Decision Trees; Linear regression

1. Introduction

The COVID-19 pandemic started in November 2019, taking millions of lives globally. The severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) has several variants of concerns (VOC) such as Alpha [1], Beta [2], Delta [3], Gamma [4], and Omicron [5] and variants of interest (VOI) such as Eta [6], Iota [7], Kappa [8], Lambda [9], and Mu [10]. These VOC and VOI have shown the nature of rapidly mutating SARS-CoV-2. With overwhelming effort of the scientific community, the development of vaccines has saved billions of lives. In addition to ongoing research in medicine and biology, scientists from various discipline have collaborated in an effort to enhance prepardness of such a situation in the future.

SARS-CoV-2 is composed of four proteins: spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins (see **Figure 1(a)**). Among these four proteins, spike protein (S-protein) plays an important role and initiates the infection by binding with human angiotensin converting enzyme 2 (ACE2). S-protein has two subunits S1 and S2. S1 consists of signal peptide (SP), n-terminal domain (NTD), receptor binding domain (RBD), subdomain 1 (SD1), and subdomain 2 (SD2). Similary, S2 consists of fusion peptide (FP), heptad repeat 1 (HR1), central helix (CH), connecting domain (CD), heptad repeat 2 (HR2), transmembrane domain (TM), and cytoplasmic tail (CT). Among these domains of S-protein, RBD binds with ACE2 (see **Figure 1(b)**).
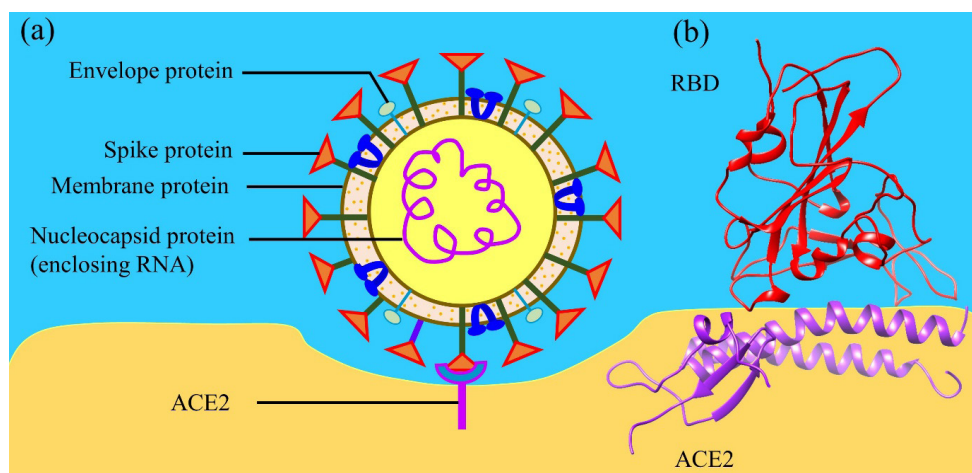

Figure 1. (a) Structure of SARS-CoV-2, highlighting its four key proteins and their interaction with angiotensin converting enzyme 2 (ACE2). **(b)** Ribbon structure of interface between receptor binding domain (RBD) and ACE2.

There has been a lot of research including experimental[11-13] and computational [14-18] study focused on RBD-ACE2 interaction. In the computational research, most of the calculations are performed utilizing molecular dynamics (MD) [15, 19-21], with a few employing *ab initio* method. *Ab initio* calculations are known for their accuracy however, they are difficult and expensive especially for such large and complex biomolecules. In our past studies, we have conducted *ab initio* calculations for around 4999 atoms [18], which is very large but not sufficient to calculate biomolecules with several thousands of atoms. Hence, we have utilized *ab initio* methods on such biomolecules using the *divide-and-conquer* strategy. Using this strategy, we divide the complex biomolecules in small sections and calculate their properties such as electronic structure, interatomic bonding, and partial

charge. Among these properties interatomic bonding is a crucial one as it reflects the types of bonds involved, and their corresponding strength. Investigating the interatomic bonding within the RBD-ACE2 interface identifies key interacting amino acids. This information can be used to target specific amino acids. Given that RBD-ACE2 is the initial point of contact between the SARS-CoV-2 and human cells, the strategic targeting of these interacting amino acids can disrupt their interaction and stop virus attack. However, *ab initio* calculations can be costly. Therefore, finding ways to achieve comparable accuracy at a reduced expense is essential, and this is where machine learning (ML) steps in.

ML represents a sub-field within artificial intelligence (AI), employing algorithms to discern significant patterns and correlations within complex datasets,

facilitating the prediction of specific properties [22]. Using ML, computers can excel in several tasks that humans can perform using their knowledge and wisdom. ML delves into how computer system mimic human learning processes and explores methods for self-improvement to acquire new knowledge [23]. Widely utilized across diverse domains such as healthcare, finance, marketing, and telecommunications, ML models have become indispensable in today's real-world applications [23].

ML models have also been applied in physics and chemistry. For instance, Hansen *et al*. applied ML techniques to predict molecular properties [24]. The precise and efficient prediction of such properties is crucial for rational computer design in chemical and pharmaceutical industries. Du *et al*. proposed a method for learning the relativistic continuum mass table by using the kernel ridge regression [25]. They used it to learn the nuclear mass table obtained by the relativistic continuum Hartree-Boogoliubov theory. Recently, Adhikari *et al*. suggested the use of ML techniques for predicting potential unknown COVID-19 mutations [26].

In this study, we employed interatomic bonding result obtained through *ab initio* calculations and tested ML to predict these already obtained bonding outcomes. Our objective is to replicate results obtained via *ab initio* methods, with the eventual aim of overcoming the cost limitations associated with such calculations.

2. Method

2.1 Modelling RBD-ACE2 Interface
The interface structures were extracted from the PDB ID 6M0J [13] for the Wild type (WT) and PDB ID 7WBP [27] for the Omicron variant (OV). Amino acids within the sequence S19-I88 and G319-T365 were incorporated from ACE2, while the sequence T333-G526 was included from the RBD. The entire model consists of 311 amino acids. Hydrogen atoms were added using the Leap module with ff14SB force field in the AMBER package [28] resulting into a total of 4817 and 4873 atoms for WT and OV, respectively. The RBD-ACE2 of the OV contains 15 mutations.

2.2 Ab initio Packages
For the *ab initio* calculations, we have used two density functional theory (DFT) packages—Vienna *ab initio* simulation package (VASP) [29] and orthogonalized linear combination of atomic orbital (OLCAO) [30]. VASP was used for interface model optimization. In VASP, we used projector augmented wave (PAW) [31,32] method with Perdew-Burke-Ernzerhof (PBE) [31] exchange correlation functional within the generalized gradient approximation (GGA). We used the energy cut off 500 eV with electronic convergence of 10^{-4} eV, force convergence for ionic relaxation to -10^{-2} eV, and a single kpoint.

The optimized structure from VASP is used as an input for the OLCAO package. The combination of VASP with OLCAO is very effective for complex biomolecules [14, 16, 18, 30, 33, 34]. OLCAO uses atomic orbitals for basis function expansion. It is used to calculate the interatomic bonding in terms of bond order (BO). BO determines the strength of the bond. OLCAO uses Mulliken's population analysis to calculate BO. BO is the overlap population $\rho_{\alpha\beta}$ between pair of atoms (α, β) defined as:

$$\rho_{\alpha\beta} = \sum_{n,occ} \sum_{i,j} C_{i\alpha}^{*n} C_{j\beta}^{n} S_{i\alpha,j\beta} \quad \# \qquad (1)$$

where $S_{i\alpha,j\beta}$ are the overlap integrals between the ith orbital in α_{th} atom and the orbital in atom, and are the eigen vector coefficients of the band, j_{th} orbital in the atom. For the *ab initio* calculations using OLCAO method, the total wall clock time and CPU time utilized for RBD-ACE2 OV are approximately 617.89 hours and 617.75 hours, respectively, resulting in a CPU efficiency of 99.98%. For RBD-ACE2 WT, the total wall clock time and CPU time are approximately 619.07 hours, and 617.16 hours respectively, with a CPU efficiency of 99.69%. The dataset for WT and OV obtained from OLCAO were further used as described in section 2.3 below.

2.3 Dataset
The BO result obtained from OLCAO calculation served as the source data. The BO prediction was conducted on both interfaces: RBD-ACE2 WT and RBD-ACE2 OV datasets. Note that the RDB-ACE2 WT and RDB-ACE2 OV datasets contained 25,356 rows and 25, 753 rows, respectively. Each dataset had 21 variables. The datasets contained the X, Y and Z coordinates of the bond between the Atom 1 and Atom 2. The variables in the datasets are briefly explained in the table below. We used these datasets to predict the bond strength also known as the BO.

Table 1. Brief description of the variables

| #	Variable	Brief Description
1	Bond	Shows the bond
2	BL	Shows the bond length
3	Ele1	Element Participating in the bond
4	X1	Position Coordinates
5	Y1	Position Coordinates
6	Z1	Position Coordinates
7	Ele2	Element Participating in the bond
8	X2	Position Coordinates
9	Y2	Position Coordinates
10	Z2	Position Coordinates
11	PDBele1	Identifier for PDB
12	AA1	Amino acid participating in the bond
13	SeqNo1	Sequence number of the amino acid participating in the bond
14	ChainN1	Chain of the amino acids
15	PDBele2	Identifier for PDB
16	AA2	Amino acid participating in the bond
17	SeqNo2	Sequence number of the amino acid participating in the bond
18	ChainN2	Chain of the amino acids

2.4 Data Pre-processing

In the original dataset, the total number of variables were 21. We dropped Atom1 and Atom2 as they are both identifiers of the OLCAO package. Eighteen of them were used for training the models, and BO was the target variable for prediction. We used the variables bond, bond length, Ele1, x1, y1, z1, PDBele1, AA1, SeqNo1, ChainN1, Ele2, x2, y2, z2, PDBele2, AA2, SeqNo2 and ChainN2 as input features for our regression models. Label encoding was used to convert the categorical variables to numerical ones before passing them to the ML models. Label encoding is an efficient way to transform categorical data to numerical ones. It is simple, efficient and helps with memory efficiency.

2.5 Machine Learning

In our work, five well-known ML models were used for the BO prediction. The models were XGBoost Regression, K-Nearest Neighbor (KNN) Regression, Decision Trees Regression, Lasso Regression, and Ridge Regression. The ML models ran on average for 4 hours using grid-search hyperparameter tuning and 10-fold cross validation. These ML models have performed well on different datasets and are commonly used in research projects. XGBoost has outperformed so many models in literature via extreme gradient boosting [35]. Linear regression is used for its simplicity before using other complex models. Decision Trees are often used for many problems because of its interpretability. It can learn complex relationship between features. Next, we briefly describe each model and its unique characteristics for learning on data.
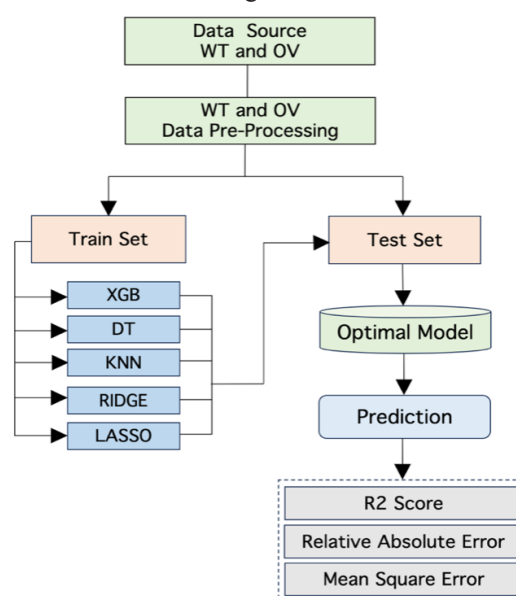
Figure 2. Overall approach for BO prediction

Lasso Regression: It is a type of linear regression that combines both linear regression and lasso hyperparameter tuning. It helps to pick the best regression model by discarding less relevant features[36]. The predicted BO is calculated as the product of the input features with the learned coefficients, and then adding the regularization term to penalize the absolute values of the coefficient.

Ridge Regression: It is a type of linear regression that uses Ridge Optimization and includes a regularization term to the regression function. The term penalizes the sum of the squared coefficients [37]. The predicted BO is calculated by multiplying the input features with the learned coefficients and then adding the regularization term that penalizes the squared magnitudes of the coefficients.

Decision Tree Regression: This regression type uses a tree-like model of decisions to predict the target value. The tree structure has the root node, decision nodes, and leaf nodes. The root node is the initial node of the decision tree. The predicted BO value is computed as the average numerical value for the training data that is stored in the selected leaf node. The optimal structure for the tree is learned during the training process.

KNN Regression: It is a supervised learning regression model in which the model predicts a continuous target variable based on the average or weighted average of the values of its k nearest neighbours. The predicted BO is computed by estimating the weighted average of the target values of the k-nearest neighbors.

XGBoost: It is a machine learning models that combines decision trees, ensemble learning, and gradient boosting. It is scalable and can be used to train on large datasets. The predicted BO is the sum of individual tree predictions multiplied by the learning rate, which is a hyperparameter used during training. Hyperparameters are the type of parameters that are used to control and manage an ML model during training.

In our experiments, we used the grid search approach to identify the best hyperparameters for the models. We also performed a 10-fold cross validation on our datasets. Cross validation is a technique that is used to evaluate a model by dividing the dataset into two segments [38, 39]. The first segment is used to train the model, and the other segment is used to test the model performance[39]. This process is repeated a number of times so that every data sample is used in the test set leading to a more robust evaluation of the model.

2.6 Bond Order Regression

Before the regression process, we first pre-processed our datasets so that it can be easily passed to the ML models. We converted the data that are not in numerical form to numerical form using label encoding. We then split our datasets into the training set and the testing set. We used four different settings in our work:

A. Train on 90% of the RBD-ACE2 WT dataset and evaluate the models on the remaining 10% of the dataset to predict the BO.

B. Train on 90% of the RBD-ACE2 OV dataset and evaluate the models on the remaining 10% of the dataset to predict the BO.

C. Train on the entire RBD-ACE2 WT dataset and evaluate the models on the entire RBD-ACE2 OV dataset to predict the BO.

D. Train on the entire RBD-ACE2 OV dataset and evaluate the models on the entire RBD-ACE2 WT dataset to predict the BO.

3. Result

In this section, we present the performance of the aforementioned ML models on the RBD-ACE2 WT and RBD-ACE2 OV datasets. We implemented all the models in Python using Scikit Learn [40], Numpy [41], Matplotlib, and XGBoost libraries [42]. We used the best model after the hyperparameter tuning for each approach for the BO prediction.

3.1 Performance Metrics

Next, we will discuss the different performance metrics that were used in this work to evaluate our models. Our regression models aimed to predict the BO in the chosen datasets. We evaluated the models based on the R2 Score, Root Mean Square Error (RMSE), and the Relative Absolute Error metrics (RAE), which are explained below.

R2: This metric is referred to as the goodness of fit or the coefficient of determination. It shows how the regression line approximates the actual data. The values are between 0 to 1 with 1 being the value when the model fits the data perfectly.

RAE: The RAE is the ratio of the mean error

(residuals) to the errors by the naïve model.

RMSE: This is calculated by computing the square root of the Mean Squared Error (MSE). MSE is the average of the squared differences between the predicted and expected target values. If all the predicted BO values are the same as the true BO values, RSME will be 0. RMSE is a good metric for regression analysis because it penalizes large error. The RMSE measures the standard deviation of the errors when the regression model makes a prediction.

Table 2. Performance Metrics of the ML models for Setting (A) (best value is shown in **bold**)

Models	R2	RAE	RMSE
XGBoost	**0.997**	0.029	**0.008**
Decision Trees	0.996	**0.025**	0.009
KNN	0.320	0.800	0.130
Ridge Regression	0.830	0.392	0.065
Lasso Regression	0.442	0.696	0.118

Table 3. Performance Metrics of the ML models for Setting (B) (best value is shown in **bold**)

Models	R2	RAE	RMSE
XGBoost	**0.998**	0.028	**0.006**
Decision Trees	0.996	**0.025**	0.009
KNN	0.339	0.726	0.126
Ridge Regression	0.826	0.405	0.065
Lasso Regression	0.444	0.712	0.116

First, we present the results for Settings (A) and (B). **Table 2** shows the performance metrics of the various ML models for Setting (A). XGBoost outperformed all the other models. It achieved an R2 score of 0.997, RAE of 0.029, and RMSE of 0.008. **Table 3** shows the performance metrics of the various ML models for Setting (B). Once again, XGBoost outperformed other models and achieved an R2 score of 0.998, RAE of 0.028, and RMSE of 0.006.

Next, we present the results for Settings (C) and (D). **Table 4** and **Table 5** show the performance results of our various ML models for Setting (C) and Setting (D), respectively. In both cases, Decision Trees outperformed other models.

Table 4. Performance Metrics of the ML models for Setting (C) (best value is shown in **bold**)

Models	R2	RAE	RMSE
XGBoost	0.995	0.049	0.011
Decision Tress	**0.997**	**0.028**	**0.008**
KNN	0.386	0.708	0.121
Ridge Regression	0.824	0.406	0.065
Lasso Regression	0.386	0.710	0.115

Table 5. Performance Metrics of the ML models for Setting (D) (best value is shown in **bold**)

Models	R2	RAE	RMSE
XGBoost	0.991	0.064	0.015
Decision Tress	**0.997**	**0.024**	**0.007**
KNN	0.385	0.702	0.121
Ridge Regression	0.822	0.410	0.065
Lasso Regression	0.445	0.712	0.115

We additionally plotted the R2 values for the various models in Settings (A) and (B). These are illustrated in **Figures 3 and 4**, where the x-axis represents the actual BO and the y-axis denotes the predicted BO. For models that achieved high R2 score, we observed that the predictions were clustered around

the 45-degree line (shown as a dotted line). Decision Trees and XGBoost models performed very well. For models with low R2 score, we observed that the predicted values were scattered differently. Clearly, the linear regression models were unable to capture the underlying distribution of the BO values. KNN regression also performed poorly on the tested datasets. Similar trends were observed for Settings (C) and (D) where Decision Trees and XGBoost achieved high R2 scores. In the interest of space, the plots are shown in the supplementary information as **Figure S1** and **S2**, respectively.
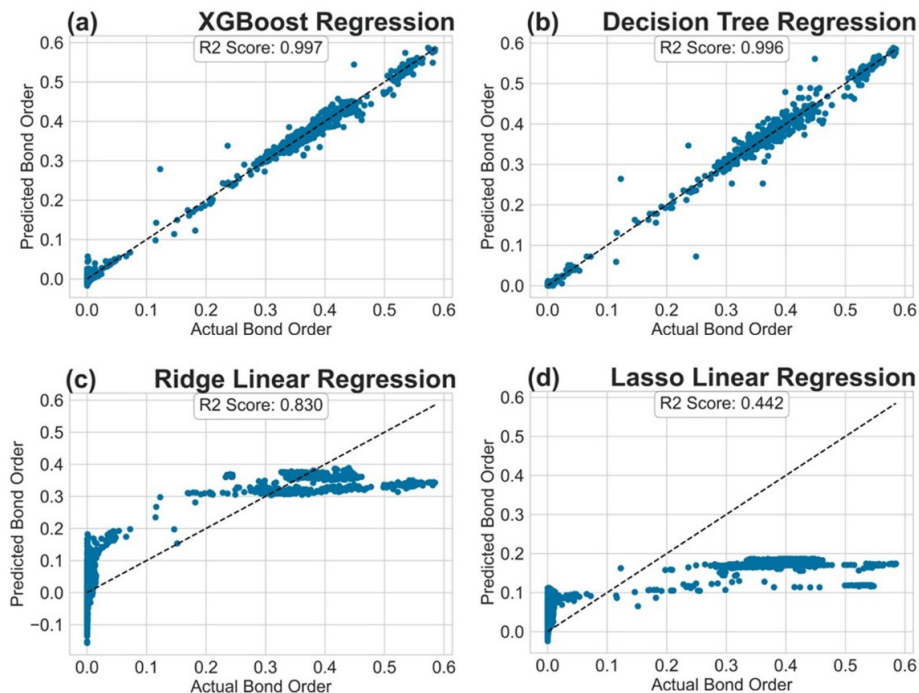
Figure 3. R2 plot for the tested regression models in Setting (A).
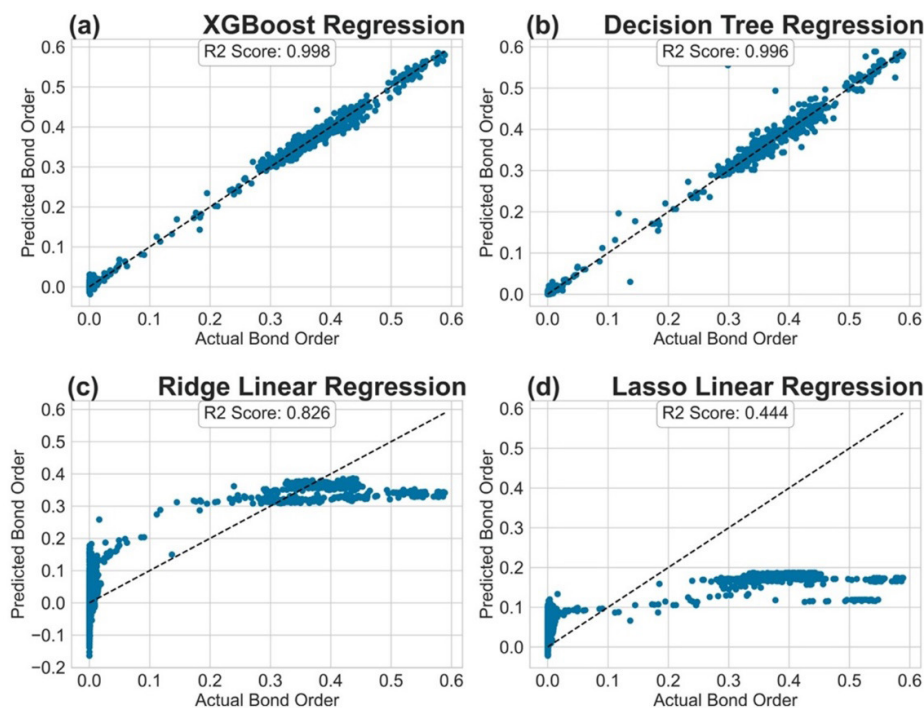
Figure 4. R2 Plot for the tested regression models in Setting (B).

4. Conclusion

In this work, we explored the efficacy of well-known ML models in predicting BO values derived from ab initio calculations, which are computationally demanding to simulate on high performance computing platforms. Specifically, we employed XGBoost Regression, Decision Tree Regression, KNN Regression, Lasso Regression, and Ridge Regression for BO prediction. The XGBoost model outperformed all other models when the training and testing samples were from the same dataset (i.e., Settings (A) and (B)). However, Decision Trees outperformed all other models when the training and testing samples were from different datasets (i.e., Settings (C) and (D)). Linear regression models and KNN regression performed poorly in all settings. Our investigation shows that ML models such as Decision Trees and XGBoost can be used to accurately predict the BO values without expensive simulation-based computations.
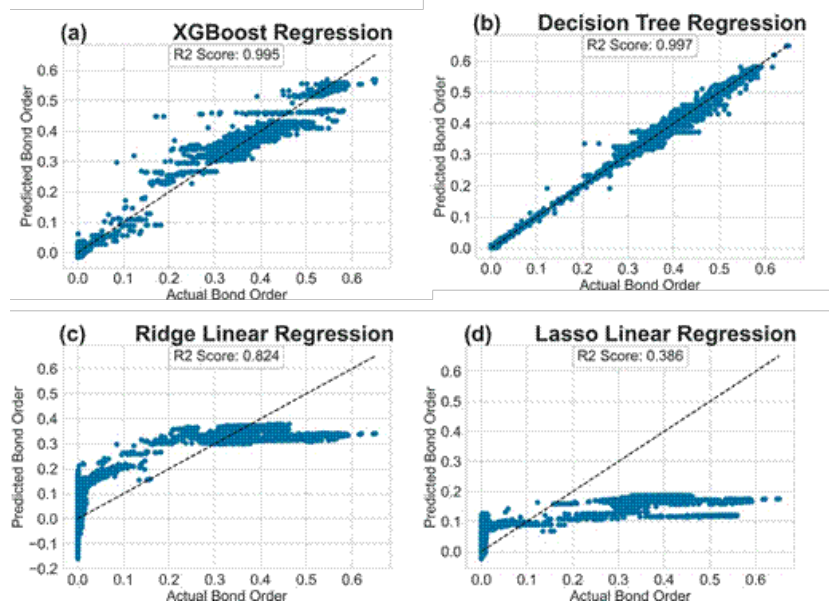
Supplementary Information

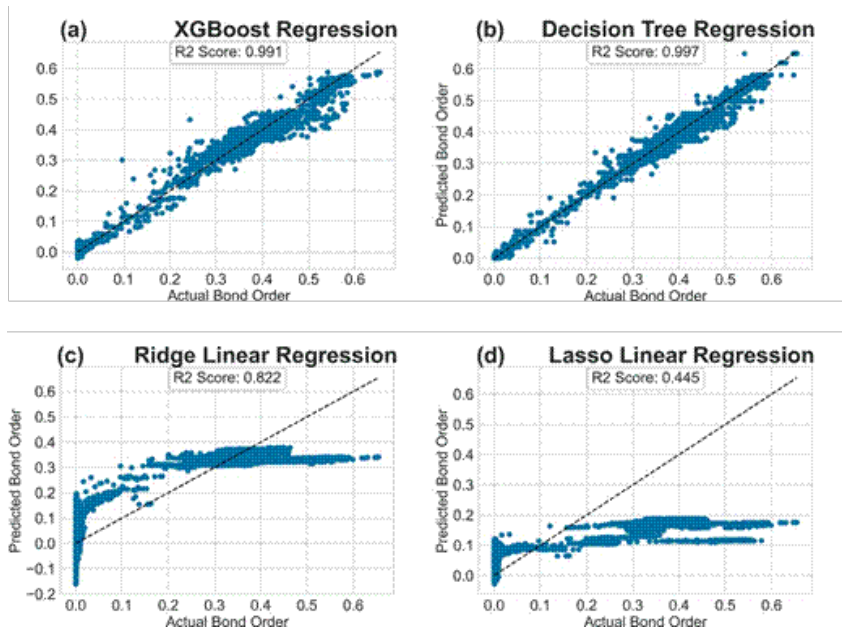**Figure S1:** R2 plot for the tested regression models in Setting (C).

Figure S2: R2 plot for the tested regression models in Setting (D).

Acknowledgements

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Rambaut, A., et al. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. 2020; Available from: https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563 (accessed on 20 January, 2022).

[2] Tegally, H., et al., Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *MedRxiv*, 2020, 1-19, https://doi.org/10.1101/2020.12.21.20248640.

[3] Singh, J., S.A. Rahman, N.Z. Ehtesham, S. Hira, and S.E. Hasnain, SARS-CoV-2 variants of concern are emerging in India. *Nature medicine*, 2021, 27, 1-3, https://doi.org/10.1038/s41591-021-01397-4.

[4] Faria, N.R., et al., Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. *Virological*, 2021, 372, 815-821.

[5] Kupferschmidt, K., New mutations raise specter of 'immune escape'. 2021, 371(6527), 329-330, https://doi.org/10.1126/science.371.6527.329.

[6] Ozer, E.A., L.M. Simons, O.M. Adewumi, A.A. Fowotade, E.C. Omoruyi, J.A. Adeniji, T.J. Dean, B.O. Taiwo, J.F. Hultquist, and R. Lorenzo-Redondo, High prevalence of SARS-CoV-2 B. 1.1. 7 (UK variant) and the novel B. 1.5. 2.5 lineage in Oyo State, Nigeria. *medRxiv*, 2021, 1-32, https://doi.org/10.1101/2021.04.09.21255206.

[7] Annavajhala, M.K., H. Mohri, J.E. Zucker, Z. Sheng, P. Wang, A. Gomez-Simmonds, D.D. Ho, and A.-C. Uhlemann, A novel SARS-CoV-2 variant of concern, B. 1.526, identified in New York. *medRxiv*, 2021, 1-28, https://doi.org/10.1101/2021.02.23.21252259.

[8] Liu, C., H.M. Ginn, W. Dejnirattisai, P. Supasa, B. Wang, A. Tuekprakhon, R. Nutalai, D. Zhou, A.J. Mentzer, and Y. Zhao, Reduced neutralization of SARS-CoV-2 B. 1.617 by vaccine and convalescent serum. *Cell*, 2021, 184(16), 4220-4236. e13, https://doi.org/10.1016/j.cell.2021.06.020.

[9] Kimura, I., et al., SARS-CoV-2 Lambda variant exhibits higher infectivity and immune resistance. *bioRxiv*, 2021, 1-31, https://doi.org/10.1101/2021.07.28.454085.

[10] Laiton-Donato, K., et al., Characterization of the emerging B. 1.621 variant of interest of SARS-CoV-2. *medRxiv*, 2021, 1-17, https://doi.org/10.1101/2021.05.08.21256619.

[11] Tai, W., L. He, X. Zhang, J. Pu, D. Voronin, S. Jiang, Y. Zhou, and L. Du, Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine. *Cellular & molecular immunology*, 2020, 17(6), 613-620, https://doi.org/10.1038/s41423-020-0400-4.

[12] Hanson, Q.M., K.M. Wilson, M. Shen, Z. Itkin, R.T. Eastman, P. Shinn, and M.D. Hall, Targeting ACE2–RBD interaction as a platform for COVID-19 therapeutics: Development and drug-repurposing screen of an AlphaLISA proximity assay. *ACS Pharmacology & Translational Science*, 2020, 3(6), 1352-1360, https://doi.org/10.1021/acsptsci.0c00161.

[13] Lan, J., J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan, Q. Zhang, X. Shi, Q. Wang, and L. Zhang, Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *nature*, 2020, 581(7807), 215-220, https://doi.org/10.1038/s41586-020-2180-5.

[14] Adhikari, P., B. Jawad, R. Podgornik, and W.-

Y. Ching, Mutations of Omicron variant at the interface of the receptor domain motif and human angiotensin-converting enzyme-2. *International journal of molecular sciences*, 2022, 23(5), 2870, https://doi.org/10.3390/ijms23052870.

[15] Jawad, B., P. Adhikari, R. Podgornik, and W.-Y. Ching, Binding interactions between receptor-binding domain of spike protein and human angiotensin converting enzyme-2 in omicron variant. *The journal of physical chemistry letters*, 2022, 13(17), 3915-3921, https://doi.org/10.1021/acs.jpclett.2c00423.

[16] Jawad, B., P. Adhikari, R. Podgornik, and W.-Y. Ching, Key interacting residues between RBD of SARS-CoV-2 and ACE2 receptor: Combination of molecular dynamic simulation and density functional calculation. *Journal of Chemical Information and Modeling*, 2021, https://doi.org/10.1021/acs.jcim.1c00560.

[17] Chen, C., V.S. Boorla, D. Banerjee, R. Chowdhury, V.S. Cavener, R.H. Nissly, A. Gontu, N.R. Boyle, K. Vandegrift, and M.S. Nair, Computational prediction of the effect of amino acid changes on the binding affinity between SARS-CoV-2 spike RBD and human ACE2. *Proceedings of the National Academy of Sciences*, 2021, 118(42), e2106480118, https://doi.org/10.1073/pnas.2106480118.

[18] Ching, W.-Y., P. Adhikari, B. Jawad, and R. Podgornik, Towards Quantum-Chemical Level Calculations of SARS-CoV-2 Spike Protein Variants of Concern by First Principles Density Functional Theory. *Biomedicines*, 2023, 11(2), 517, https://doi.org/10.3390/biomedicines11020517.

[19] Jawad, B., P. Adhikari, R. Podgornik, and W.-Y. Ching, Impact of BA.1, BA.2, and BA.4/BA.5 Omicron Mutations on Therapeutic Monoclonal Antibodies. *Submitted to CIBM*, 2023.

[20] Srivastava, N., P. Garg, P. Srivastava, and P.K. Seth, A molecular dynamics simulation study of the ACE2 receptor with screened natural inhibitors to identify novel drug candidate against COVID-19. *PeerJ*, 2021, 9, e11171, https://doi.org/10.7717/peerj.11171.

[21] Celik, I., A. Khan, F.M. Dwivany, Fatimawali, D.-Q. Wei, and T.E. Tallei, Computational prediction of the effect of mutations in the receptor-binding domain on the interaction between SARS-CoV-2 and human ACE2. *Molecular Diversity*, 2022, 26(6), 3309-3324, https://doi.org/10.1007/s11030-022-10392-x.

[22] Pattern recognition and machine learning, ed. C.M. Bishop. 2006, New York: Springer New York. 645-678.

[23] Wang, H., C. Ma, and L. Zhou. A brief review of machine learning and its application. in *2009 international conference on information engineering and computer science*. 2009. IEEE.

[24] Hansen, K., F. Biegler, R. Ramakrishnan, W. Pronobis, O.A. Von Lilienfeld, K.-R. Muller, and A. Tkatchenko, Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *The journal of physical chemistry letters*, 2015, 6(12), 2326-2331, https://doi.org/10.1021/acs.jpclett.5b00831.

[25] Du, X.-K., P. Guo, X.-H. Wu, and S.-Q. Zhang, Examination of machine learning for assessing physical effects: Learning the relativistic continuum mass table with kernel ridge regression. *Chinese Physics C*, 2023, 47(7), 074108, https://doi.org/10.1088/1674-1137/acc791.

[26] Adhikari, P., B. Jawad, P. Rao, R. Podgornik, and W.-Y. Ching, Delta variant with P681R critical mutation revealed by ultra-large atomic-scale ab initio simulation: Implications for the fundamentals of biomolecular interactions. *Viruses*, 2022, 14(3), 465, https://doi.org/10.3390/v14030465.

[27] Han, P., L. Li, S. Liu, Q. Wang, D. Zhang, Z. Xu, P. Han, X. Li, Q. Peng, and C. Su, Receptor binding and complex structures of human ACE2 to spike RBD from omicron and delta SARS-CoV-2. *Cell*, 2022, 185(4), 630-640. e10, https://doi.org/10.1016/j.cell.2022.01.001.

[28] Pearlman, D.A., D.A. Case, J.W. Caldwell, W.S. Ross, T.E. Cheatham III, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman, AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*,

1995, 91(1-3), 1-41, https://doi.org/10.1016/0010-4655(95)00041-D.

[29] VASP - Vienna Ab initio Simulation Package. Available from: https://www.vasp.at/ (accessed on June 1, 2023).

[30] Ching, W.-Y. and P. Rulis, Electronic Structure Methods for Complex Materials: The orthogonalized linear combination of atomic orbitals. 2012, UK, London: OUP Oxford.

[31] Perdew, J.P., K. Burke, and M. Ernzerhof, Generalized gradient approximation made simple. *Physical review letters*, 1996, 77(18), 3865, https://doi.org/10.1103/PhysRevLett.77.3865.

[32] Kresse, G. and D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical review b*, 1999, 59(3), 1758, https://doi.org/10.1103/PhysRevB.59.1758.

[33] Adhikari, P. and W.-Y. Ching, Amino acid interacting network in the receptor-binding domain of SARS-CoV-2 spike protein. *RSC Advances* 2020, 10, 39831-39841, https://doi.org/10.1039/d0ra08222h.

[34] Adhikari, P., N. Li, M. Shin, N.F. Steinmetz, R. Twarock, R. Podgornik, and W.-Y. Ching, Intra- and intermolecular atomic-scale interactions in the receptor binding domain of SARS-CoV-2 spike protein: implication for ACE2 receptor binding. *Physical Chemistry Chemical Physics*, 2020, 22(33), 18272-18283, https://doi.org/10.1039/D0CP03145C.

[35] Wang, R., L. Wang, J. Zhang, M. He, and J. Xu, XGBoost machine learning algorism performed better than regression models in predicting mortality of moderate-to-severe traumatic brain injury. *World Neurosurgery*, 2022, 163, e617-e622, https://doi.org/10.1016/j.wneu.2022.04.044.

[36] Tibshirani, R., Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1996, 58(1), 267-288, https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[37] Hoerl, A.E. and R.W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 1970, 12(1), 55-67, https://doi.org/10.1080/00401706.1970.10488634.

[38] Abu-Mostafa, Y.S., M. Magdon-Ismail, and H.-T. Lin, Learning from data. Vol. 4. 2012, New York: AMLBook

[39] Refaeilzadeh, P., L. Tang, and H. Liu, Cross-validation. *Encyclopedia of database systems*, 2009, 532-538, https://doi.org/10.1007/978-0-387-39940-9_565.

[40] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 2011, 12, 2825-2830.

[41] Harris, C.R., K.J. Millman, S.J. Van Der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, and N.J. Smith, Array programming with NumPy. *Nature*, 2020, 585(7825), 357-362, https://doi.org/10.1038/s41586-020-2649-2.

[42] Chen, T. and C. Guestrin. Xgboost: A scalable tree boosting system. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.