Method

Data Collection

The 73 species ACE2 sequences for constructing predictive models and evaluation were collected from published articles [1,2] and unpublished data. Eleven sequences from these 73 data were randomly selected as test dataset for model evaluation and were not involved in model training.

The sequences of mammalian ACE2 for prediction were downloaded as of 22 September 2020 with a total of 294 ACE2 sequences of mammalian species from 23 orders were gathered. We performed multiple sequence alignment on collection of 294 sequences with human ACE2 sequence, using software CLUSTAL (version: 2.1, parameter "complete multiple alignment "),[3] in which sequences with more than 10 consecutive amino acid missing in the head 100 sites were excluded from the subsequent analysis, resulting in 272 ACE2 sequences (204 unique species).

Model Construction and Evaluation

We selected key amino acid sites and used the log2 enrichment ratios values from Chan *et al*. to label the amino acids for the each ACE2 sequence[4] , with 20, 24 and 117 sites selected from Liu *et al.*,[1] Wang *et al.*[2]and Chan *et al.*,[4] respectively. The sequences screened for these three sites were divided into a training set and a test set based on 8:2 and used for training and testing of the model. As for prediction models, we used five different methods to learn three different collection of sites, including SVM, Decision Tree, Random Forest, AdaBoost and Gradient Boosting, resulting in 15 models of input data/methods. After hundreds of epochs of training, random combinations of the 15 models were evaluated based on precision (Precision = TP/(TP+FP) , where TP: True Positive, FP: False Positive). We selected six model combinations for ACE2 sequences prediction in the subsequent analysis, and set the prediction score (Prediction Score = Pn/Mn), where Pn indicates the number of one sequence was predicted to have binding ability and Mn is the total number of models used for prediction. The threshold value for the prediction score was set to 0.5, *i.e.* a prediction score ≥ 0.5 was considered to have the ability to binding with SARS-CoV-2. The 272 sequences were also screened for sites for binding ability prediction.

Model construction and prediction were carried out based on the scikit-learn module (version: 0.22.2) in the Python3. The functions used for model training were “svm” ,

"DecisionTreeClassifier", "RandomForestClassifier", "AdaBoostClassifier" and "GradientBoostingClassifier". The parameters used for SVM are: gamma = 'scale', class_weight = {0:2}, for decision tree classifier are default parameters, for random forest classifier are: n_estimators = 600, oob_score = True, n_jobs = -1, class_weight = {0:2}, for Ada boost classifier are: base_estimator=DecisionTreeClassifier(max_depth=2), n_estimators=500, and for gradient boosting classifier are: n_estimators = 100, learning_rate = 1.0, max_depth = 1, random_state = 0.

ACE2 sequence acquisition and Gene cloning

Twelve bat orthologs were randomly selected from the test sets. The full-length coding sequences (accession numbers are shown in Supplementary Table S2) of these orthologs were synthesized and cloned into the pEGFP-N1 vector for flow cytometry (FACS). The extracellular domain of these ACE2 orthologs were fused with the Fc domain of mouse IgG (mFc) and cloned into the pCAGGS expression vector for SPR.

Protein Expression and Purification

The SARS-CoV-2 RBD and SARS-CoV-2 NTD proteins used for flow cytometry and SPR were expressed and purified from the supernatants of HEK293F cells culture as described in our previous work.[5] Proteins were stored in a PBS buffer (1.8 mM KH2PO4, 10 mM Na2HPO4 (pH 7.4), 137 mM NaCl, 2.7 mM KCl) buffer. The indicated pCAGGS plasmids were transiently transfected into HEK293T cells (ATCC CRL-3216). Supernatant containing mFc-tagged ACE2 proteins were collected and concentrated at 48h post-transfection.

Flow cytometry analysis

To test the binding between each of the 12 ACE2s and SARS-CoV-2 RBD, the 12 bat ACE2s fused with eGFP were expressed on the cell surface by transfecting each of the 12 pEGFP-N1-ACE2s plasmids into BHK21 cells (ATCC, ATCC CCL-10) using PEI (Alfa). Cell culture was replaced with fresh media (DMEM with 10% FBS, Gibco) 4-6 h post-transfection. After 48 h, cells were collected and resuspended in PBS. Then, 2 × 105 cells were incubated with the histidine tagged test proteins (SARS-CoV-2 RBD, SARS-CoV-2 NTD) at a concentration of 10 μg/mL at 37°C for 30 min. Cells were then washed three times in PBS and stained with anti-His/APC antibodies (1:500, Miltenyi Biotec, AB_2751870) for 30 min at 37°C. FACS data were acquired on a BD FACSCanto

(BD Biosciences, Franklin Lakes, NJ) and analyzed using FlowJo V10 software (TreeStar Inc., Ashland, OR), with results shown in Figure S1.

Surface plasmon resonance (SPR) analysis

We tested the binding affinities between the mFc-tagged ACE2s and SARS-CoV-2 RBD or SARS-CoV RBD proteins by SPR using a BIAcore 8K (GE Healthcare) carried out at 25 °C in single-cycle mode. The PBST buffer (1.8 mM KH2PO4, 10 mM Na2HPO4 (pH 7.4), 137 mM NaCl, 2.7 mM KCl, and 0.05% (v/v) Tween 20) was used as the running buffer. The CM5 biosensor chip was first immobilized with anti-mIgG antibody (ZSGB-BIO, ZF-0513) as previously described.[1] The supernatants containing mFc-tagged ACE2s were injected and captured by the antibody immobilized on the CM5 chip at approximately 300-600 response units. The serially diluted SARS-CoV-2 RBD protein flowed over the chip surface, with another channel set as control. The chip was re-generated using pH 1.7 glycine after each reaction. The equilibrium dissociation constants (binding affinity, KD) for each pair of interaction were calculated with BIAcore_8K evaluation software (GE Healthcare) by fitting to a 1:1 Langmuir binding model. Data were analyzed using Origin 2018 (OriginLab).

Phylogenetic Tree

Phylogenetic tree was constructed by uploading the species names from 272 sequences into NCBI Taxonomy Common Tree (https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/). The visualization of the phylogenetic tree was based on iTol.[6]

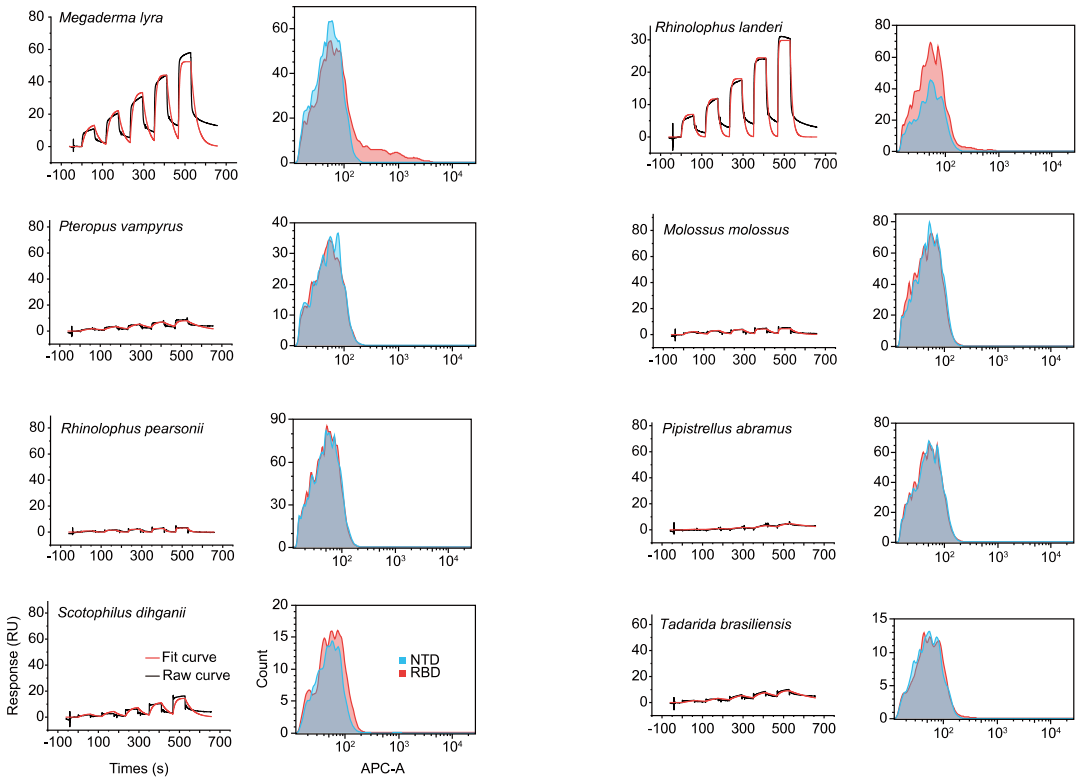

FIGURE S1. SPR and flow cytometry validation for multiple species' ACE2.

References

[1] Wang Q, Zhang Y, Wu L, *et al.* Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. Cell. 2020 May 14;181(4):894-904.e9.

[2] Liu Y, Hu G, Wang Y, *et al.* Functional and genetic analysis of viral receptor ACE2 orthologs reveals a broad potential host range of SARS-CoV-2. Proc Natl Acad Sci U S A. 2021 Mar 23;118(12):e2025373118.

[3] Higgins DG, Sharp PM. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene. 1988 Dec 15;73(1):237-44.

[4] Chan KK, Dorosky D, Sharma P, *et al.* Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2. Science. 2020 Sep 4;369(6508):1261-1265.

[5] Niu S, Wang J, Bai B, *et al.* Molecular basis of cross-species ACE2 interactions with SARS-CoV-2-like viruses of pangolin origin. EMBO J. 2021 Aug 16;40(16):e107786.

[6] Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021 Jul 2;49(W1):W293-W296.