High-throughput molecular simulations of SARS-Cov-2 receptor binding domain mutants quantify correlations between dynamic fluctuations and protein expression.

V. Ovchinnikov,[1,a] and M. Karplus[1,2,b]
[1] Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, 02138
[2] Laboratoire de Chimie Biophysique, ISIS, Université de Strasbourg, 67000 Strasbourg, France
[a] ovchinnv@georgetown.edu
[b] marci@tammy.harvard.edu

SUPPORTING MATERIAL

Supporting results for modeling $lMFI$ and lK_D

#	c_P (p-value)	c_S (p-value)	c_P (p-value)	c_S (p-value)	c_P (p-value)	c_S (p-value)
	RMSF(COM)		RMSF(C_α)		RMSD(C_α)	
1	-0.36(7.8e-3)	-0.42(1.5e-3)	-0.35(9.9e-3)	-0.42(1.8e-3)	-0.35(8.7e-3)	-0.42(1.8e-3)
2	-0.22(1.7e-7)	-0.29(5.5e-12)	-0.22(1.5e-7)	-0.29(3.5e-12)	-0.19(5.1e-6)	-0.25(2.1e-9)
3	-0.30(2.4e-2)	-0.31(2.2e-2)	-0.30(2.6e-2)	-0.30(2.5e-2)	-0.23(9.2e-2)	-0.37(5.1e-3)
4	-0.34(2.1e-16)	-0.35(2.3e-17)	-0.34(9.5e-17)	-0.35(5.0e-18)	-0.32(1.2e-14)	-0.34(5.4e-16)

TABLE S1. Comparison of three simulation metrics chosen to model expression, as measured by the logarithm of the mean fluorescence intensity[1] ($lMFI$).

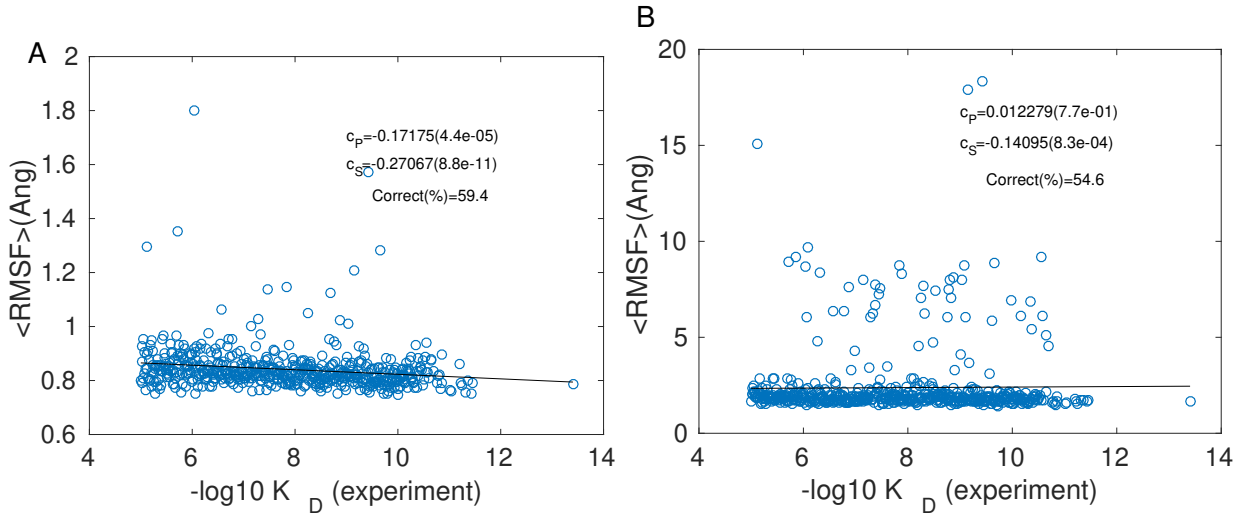

FIG. S1. Scatter plots of RMSFs from MD simulations *vs.* logarithm of experimental K_D (lK_D); A. The average RMSF was computed over RBD residues; B. The average RMSF was computed over RBD and ACE2 residues.

Additional tests of regression model

In this section, we show how the ML model described here performs with different training/testing sets. As described in the main text (see Eq. (1)), the model is linear in amino acid coordinates, and therefore

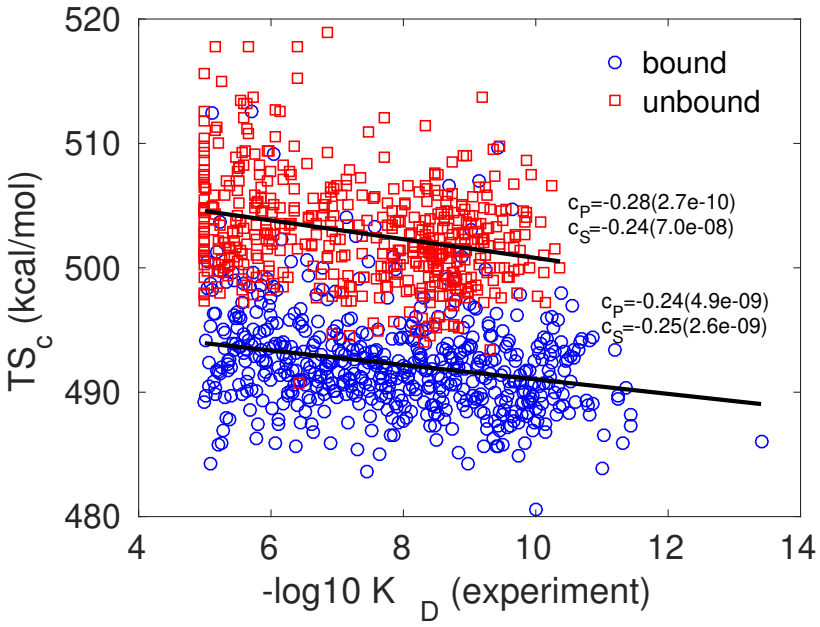

FIG. S2. Scatter plots of temperature-scaled classical entropy computed from quasiharmonic analysis[10,11] of MD simulation trajectories *vs.* negative lK_D; Linear fits are indicated as black lines. The correlation coefficients are given with the corresponding p-value in parentheses.

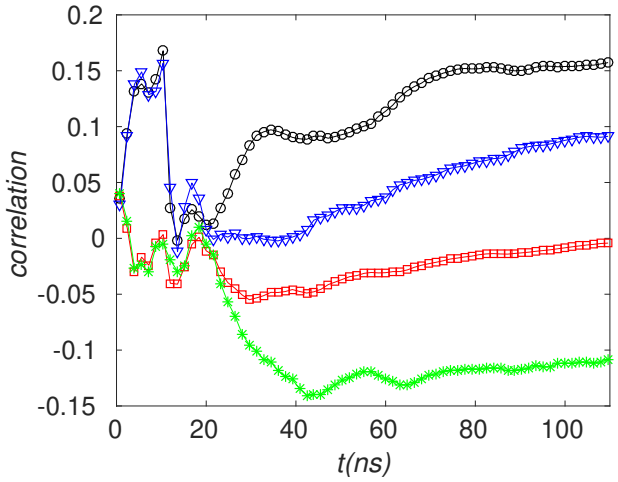

FIG. S3. Time-evolution of the Pearson and Spearman correlation coefficients between RMSDs and lK_D (red □ and green *, respectively), and the average corelation between the RMSDs of simulation replicates (black ○ and blue ▽ , respectively). The RMSDs were computed using all C_α RBD atoms.

does not capture inter-residue couplings (epistasis). Consequently, we expect model performance to degrade progressively for higher-order mutants. In Figs. S5 and S6, we show expression predictions when the model is trained on 1- and 2-residue mutations, respectively. Both figures show that model prediction decreases, as the composition of the test set changes from 1-residue to 3-residue (or greater) mutants. For example, for the model trained on 1-residue mutations, the Pearson coefficients c_P are 0.84, 0.74, and 0.62 for 1-, 2-, and ≥3-residue mutants. Similarly, if the model is trained on 2-residue mutants, the corresponding Pearson correlations are 0.83, 0.75, and 0.66, which also shows that the expected model performance degradation is

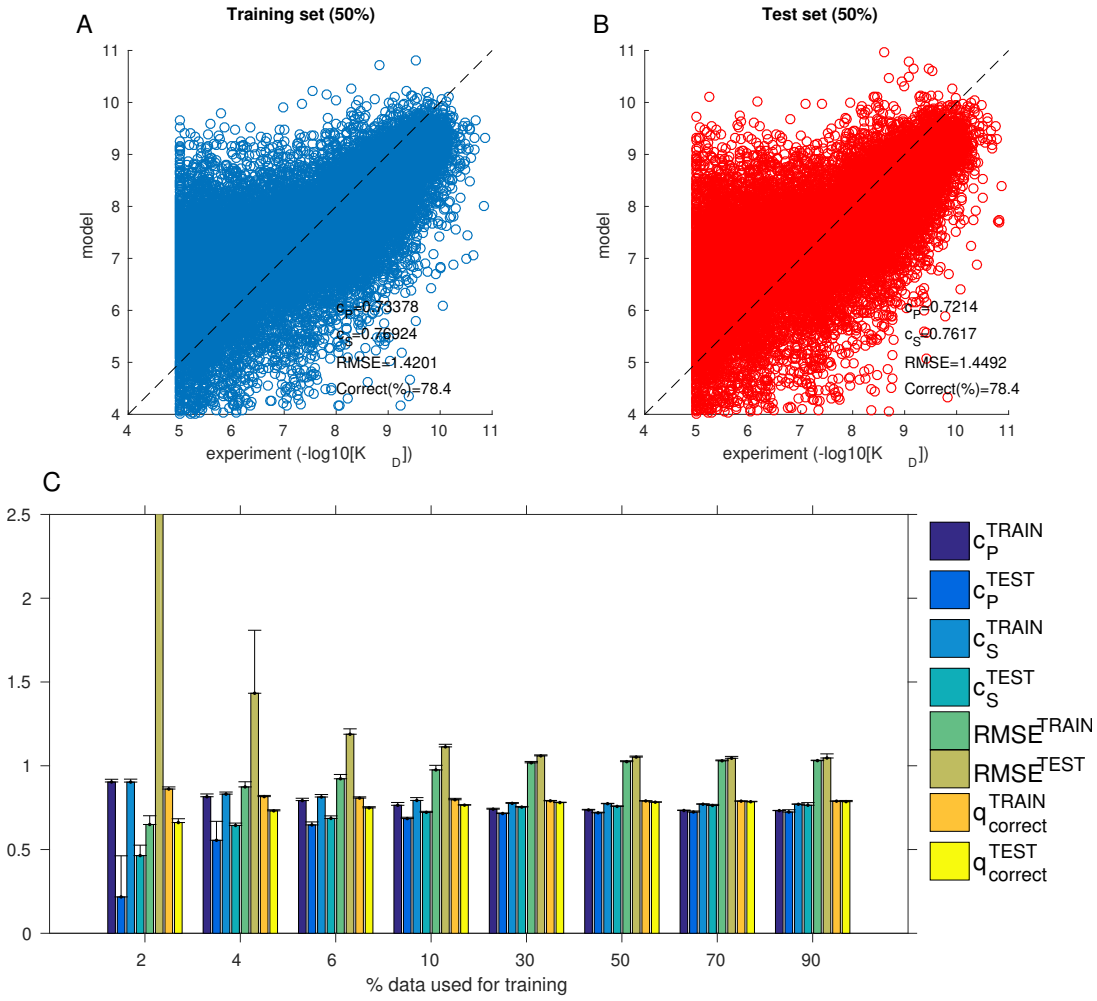

FIG. S4. Results of the linear regression model to predict lK_D expression. A & B: scatter plot of model predictions *vs.* experiment; A: training set; B: testing set; 50% of the data were assigned to each set. C. Model performance metrics as a function of the percent data used for training; for each percent value, the remaining data were used for testing; $q_{correct}$ represents the fraction of correct predictions of pairwise rank (see main text).

essentially independent of the training set.

As discussed in the main text, more sophisticated models can be parametrized. Perhaps the simplest generalization would be a model that is linear in residue couplings, which could be implemented as a weighted product (or sum) of residue coordinates. Alternatively, a more involved interpolation scheme could be used, such as gaussian process regression.[45] However, these models are beyond the present scope, as our main goal here was a qualitative comparison between simple MD-based and ML-based predictions, with ML giving superior results, albeit with preexisting training data.

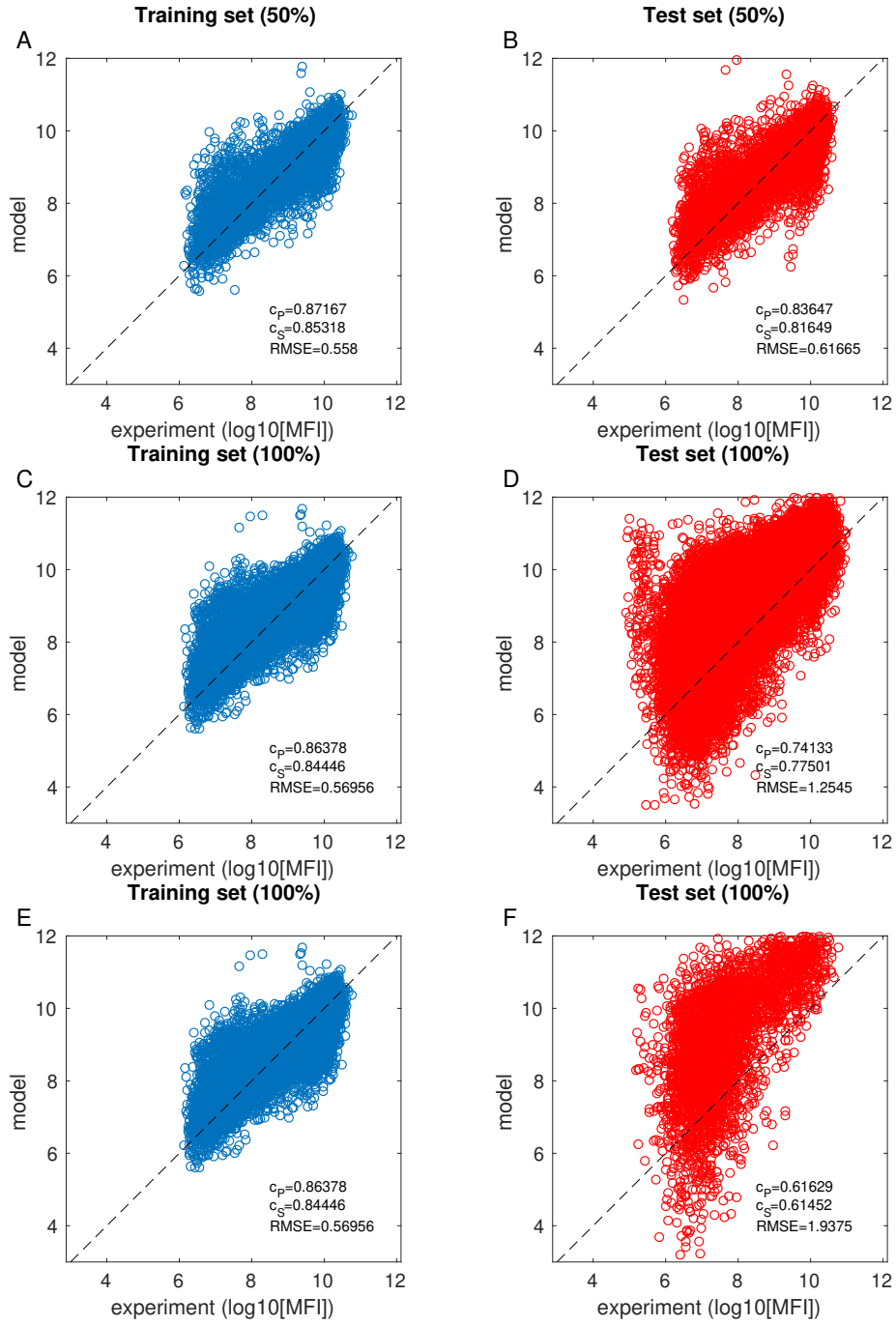

FIG. S5. Scatter plots of the linear regression model *vs.* experiment to predict $lMFI$ expression. Training and testing set are on the left and right, respectively. The model was trained on single-residue mutants only, and tested on single (A & B), double (C & D), or triple or higher (E & F) residue mutants. In A & B, randomly-chosen 50% of the single-residue mutant data was used for training, and the remainder, for testing. In C–F, all of the available data (subject to the above mutational constraints) was used for training and testing.

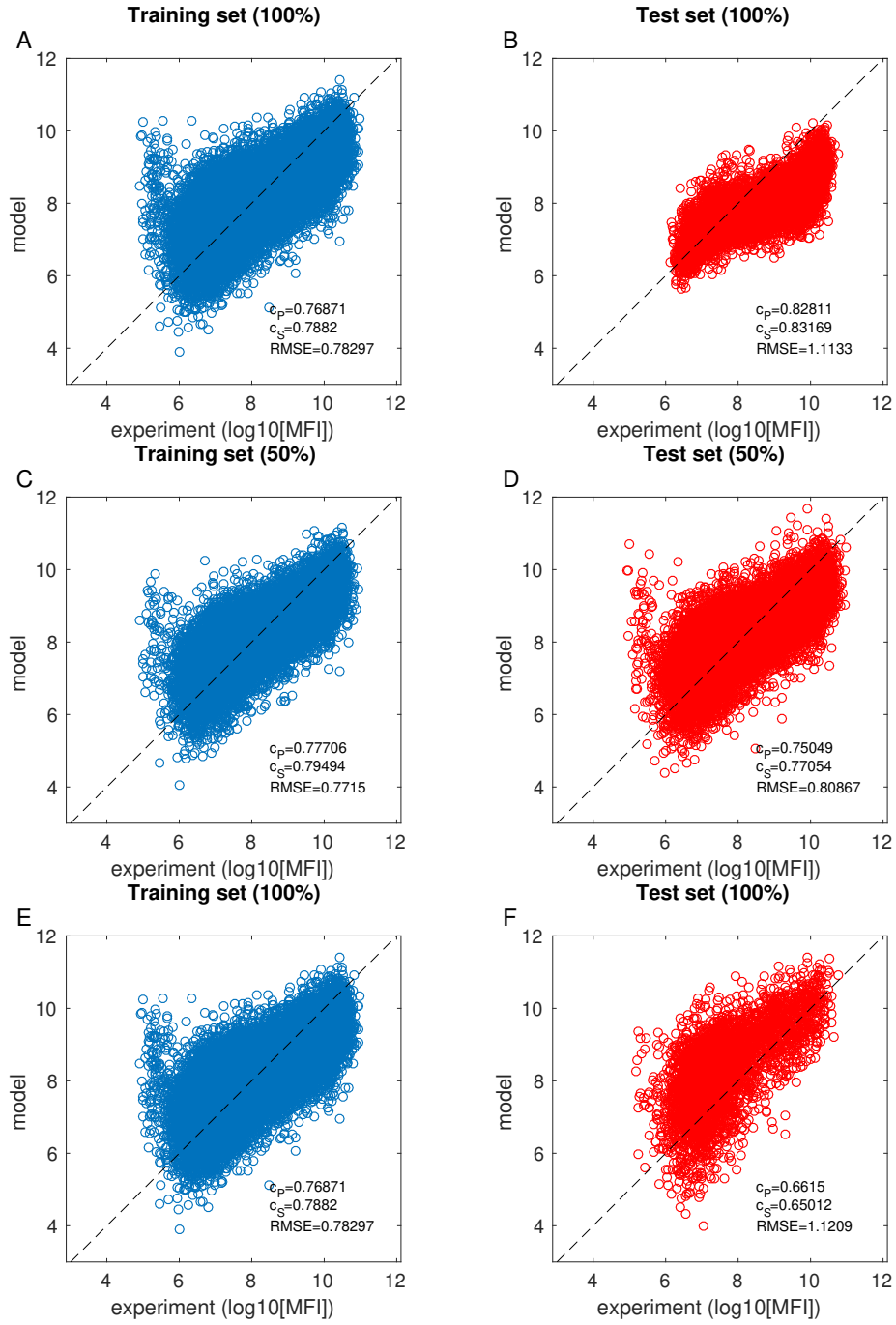

FIG. S6. Scatter plots of the linear regression model *vs.* experiment to predict $lMFI$ expression. Training and testing set are on the left and right, respectively. The model was trained on double-residue mutants only, and tested on single (A & B), double (C & D), or triple or higher (E & F) residue mutants. In C & D, randomly-chosen 50% of the single-residue mutant data was used for training, and the remainder, for testing. In A, B, E and F, all of the available data (subject to the above mutational constraints) was used for training and testing.

Root-mean-square fluctuations of residue centers-of-mass (RMSF-COM)

For each amino acid residue indexed by i composed of n_i atoms, we computed the coarse-grained (CG) coordinate triplet r_i^{CG} as

$$r_i^{CG} = \frac{\sum_{j=1}^{n_i} r_i^j m^j}{m_i^{CG}}, \tag{S1}$$

where

$$m_i^{CG} = \sum_{j=1}^{n_i} m_i^j, \tag{S2}$$

and r_i^j and m_i^j are the coordinate triplet and mass of atom j of residue i, respectively.

The CG coordinates were computed for each trajectory, with trajectory frames sampled in 400ps time increments. Next, the CG coordinates were shifted to their centers of mass, and rotated to achieve the best-fit superposition,[46] *i.e.*, for each trajectory frame k

$${}^k r_i^{CG} \mapsto {}^k r_i^{CG} - \frac{\sum_{i=1}^{N_{res}} {}^k r_j^{CG} m_j^{CG}}{\sum_{i=1}^{N_{res}} m_j^{CG}}, \tag{S3}$$

followed by

$${}^k r_i^{CG} \mapsto^k A \, {}^k r_i^{CG}, \tag{S4}$$

where ${}^k A$ is the best-fit rotation matrix defined as the minimizer

$${}^k A = \arg \min_{B \in M_{3\times 3}} \| B \, {}^k r_i^{CG} - \langle \, {}^l r_i^{CG} \rangle_{1 \le l \le N_{fr}} \|, \tag{S5}$$

and angle brackets represent averages over the N_{fr} trajectory frames. Equations (S4) and (S5) are applied iteratively five times to achieve self-consistency.

The root-mean-square fluctuation (RMSF) of CG coordinate i is computed as

$$RMSF_i = \langle \| r_i^{CG} - \langle r_i^{CG} \rangle \|^2 \rangle^{1/2}, \tag{S6}$$

where the angle brackets represent trajectory averages, and $\|\cdot\|$ is the Euclidean norm. To obtain a scalar value for comparing to experimental $lMFI$'s or lK_D's, the RMSFs are averaged over the protein residues, as indicated in the main text.

To estimate the absolute entropy of the RBD mutants in bound and unbound states, we performed quasi-harmonic analysis in the CG coordinates of the RBD, as follows. The mass-weighted covariance matrix of CG coordinate displacements was computed as[10,11]

$$C_{ij} = (m_i^{CG} m_j^{CG})^{1/2} \langle (r_i^{CG} - \langle r_i^{CG} \rangle)(r_j^{CG} - \langle r_j^{CG} \rangle) \rangle, \tag{S7}$$

where the angle brackets repreent trajectory averages, as above.

The matrix C_{ij} was diagonalized in Matlab[38] using singular value decomposition routines to yield the diagonal eigenvalue matrix Λ (here, equivalent to the singular values) and eigenvectors (mass-weighted principal components) U,

$$C = U \Lambda U^{-1}. \tag{S8}$$

Six of the singular values were very close to zero, and the associated modes, which correspond to rigid-body motion, were discarded. To compute the quasiharmonic entropy, we added the contributions from the remaining eigenvalues λ_i, with $i = 7 \ldots 603$, using the classical harmonic oscillator entropy for each eigenvalue,

$$S_{class} = k_B \left(\sum_{i=7}^{603} 1 + \log \frac{\sqrt{k_B T \lambda_i}}{\hbar} \right), \tag{S9}$$

where T=300K, k_B is Boltzmann's constant, $\hbar$ is Planck's constant in radians, and log is the natural logarithm. The total number of quasi-harmonic modes was 603, corresponding to 3 times the number of CG residue beads, of which there were 201, spanning the residue range 331–531, as described in *Methods*. The coordinates of the ACE2 receptor were not included in the quasiharmonic analysis of the bound state, which neglects coupling between RBD and ACE2 motions.

SUPPORTING REFERENCES

45 C. E. Rasmussen and C. K. I. Williams, Gaussian Processes for Machine Learning (MIT Press, Cambridge, MA, 2006).

46 W. Kabsch, Acta Cryst. A32, 922 (1976).

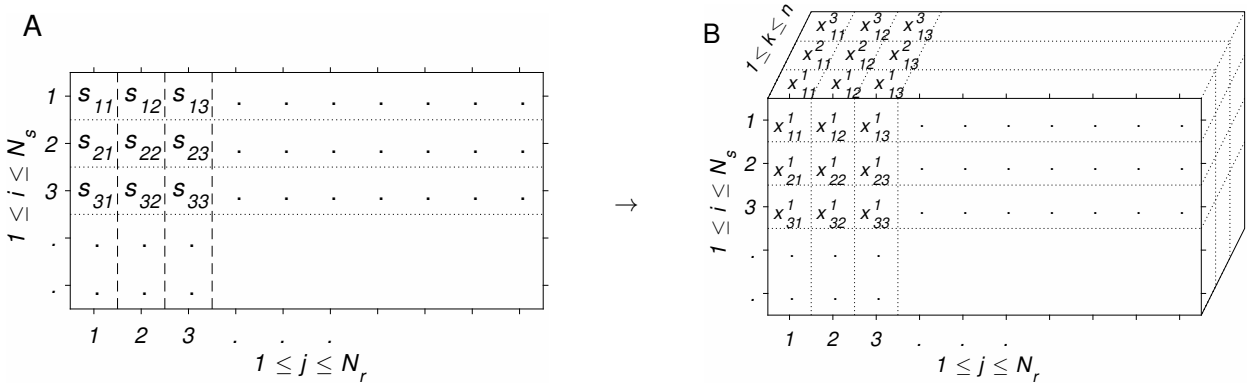

FIG. S7. Illustration of the amino acid encoding procedure.[40] A: Multiple alignment $\widehat{\boldsymbol{S}}$ of N_s sequences, each having N_r residues, including possible gaps, with $s_{ij} \in \{\mathrm{A}, \mathrm{C}, \mathrm{D}, \mathrm{E}, \mathrm{F}, \mathrm{G}, \mathrm{H}, \mathrm{I}, \mathrm{K}, \mathrm{L}, \mathrm{M}, \mathrm{N}, \mathrm{P}, \mathrm{Q}, \mathrm{R}, \mathrm{S}, \mathrm{T}, \mathrm{V}, \mathrm{W}, \mathrm{Y}, '-'\}$. B: Embedded alignment $\widehat{\boldsymbol{X}}$; each residue type in $\widehat{\boldsymbol{S}}$ shown in A is associated with a (n=3)-dimensional real-valued vector $\boldsymbol{x}_{ij}=\{x_{ij}^1, x_{ij}^2, x_{ij}^3\}$, which is interpreted as Cartesian coordinates of the residue.