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

Victor Ovchinnikov[1, a)] and Martin Karplus[1, 2, b)]
[1)] *Harvard University, Department of Chemistry and Chemical Biology, Cambridge, MA, 02138*
[2)] *Laboratoire de Chimie Biophysique, ISIS, Université de Strasbourg, 67000 Strasbourg, France*

(Dated: 26 September 2024)

Prediction of protein fitness from computational modeling is an area of active research in rational protein design. Here, we investigated whether protein fluctuations computed from molecular dynamics simulations can be used to predict the expression levels of SARS-CoV-2 receptor binding domain (RBD) mutants determined in the deep mutational scanning experiment of Starr et al.[1] Specifically, we performed more than 0.7 milliseconds of molecular dynamics (MD) simulations of 557 mutant RBDs in triplicate to achieve statistical significance under various simulation conditions. Our results show modest but significant anticorrelation in the range [-0.4,-0.3] between expression and RBD protein flexibility. A simple linear regression machine learning model achieved correlation coefficients in the range [0.7,0.8], thus outperforming MD-based models, but required about 25 mutations at each residue position for training.

Keywords: molecular dynamics, deep mutational scanning, coronavirus, linear regression

INTRODUCTION The COVID pandemic of 2019 resulted in the development of effective vaccines for the SARS-CoV-2 coronavirus. However, random mutations in the coronavirus spike protein, high viral transmission rates, and evolutionary pressure exerted by antibodies continue to cause the emergence of escape variants, for which the antibodies induced by the vaccines have reduced affinity.[2] Because the descendants of the original COVID-19 strain are likely to become endemic,[3] there is a need to develop advanced coronavirus vaccines capable of eliciting immunity to different strains, ideally including some that are yet to emerge. To address this challenge using rational antigen design, a detailed understanding of the determinants of coronavirus infectivity is needed. For rational vaccine design, it is important to be able to predict by computational methods how amino acid mutations in the receptor binding domain (RBD) of the coronavirus spike affect its stability, or binding affinity to the angiotensin converting enzyme receptor 2 (ACE2), and

[a)] Electronic mail: ovchinnv@georgetown.edu
[b)] Electronic mail: marci@tammy.harvard.edu

Simulation #	Solvation	#Mutants	Duration (ns)	c_P (p-value)	c_S (p-value)	Correct(%)[†]
RBD simulations to model $lMFI$						
1	cube	55	110	-0.36(7.8e-3)	-0.42(1.5e-3)	62.6
2	shell	557	110	-0.22(1.7e-7)	-0.29(5.5e-12)	59.5
3	shell	55	1010	-0.30(2.4e-2)	-0.31(2.2e-2)	60.5
4	shell[‡]	557	110	-0.34(2.1e-16)	-0.35(2.3e-17)	61.8
RBD/ACE simulations to model lK_D						
5[*]	shell	557	110	-0.17(4.4e-5)	-0.27(8.8e-11)	59.4
5[§]	–	–	–	0.01(7.7e-1)	-0.14(8.3e-4)	54.6

TABLE I. Summary of MD simulations performed in this study. For modeling $lMFI$, the correlations are between the protein residue RMSFs computed from simulation and the experimental $lMFI$; For modeling lK_D, the correlations are between the protein residue RMSFs computed from simulation and the experimental lK_D; [†] Indicates the proportion of all modeled pairs for which the modeled expression order within the pair matched the experimental order. [‡] These simulations were performed with conformational restraints to the initial structure. [*] Both the RBD and ACE2 were simulated in complex, but the average residue RMSF was computed using only the RBD atoms. [§] The data used to compute the statistics are from simulation 5, but the average residue RMSF was computed over both RBD and ACE protein atoms.

to antibodies. For example, stabilization of antigenic spikes in the prefusion conformation[4] has been used to improve vaccine efficacy. While the majority of information needed for rational protein mutagenesis derives from experiments, molecular dynamics (MD) simulations are increasingly able to provide dynamic information on temporal and/or spatial scales of protein motion that are smaller than the experimental resolution.

Here, we studied whether conventional MD simulations can be used to estimate relative expression fitness of RBD mutants. We chose a random subset of 557 RBD sequence mutants from the total set of ≃54,000 in the deep mutational scanning (DMS) study of Starr et al.[5] that were approximately uniformly distributed in the level of expression, and performed MD simulations in triplicate for at least 110ns for each mutant. The DMS dataset provides the logarithm of the mean fluorescence intensity ($lMFI$), which was used in the assay as a proxy for protein expression.[5]

RESULTS The results of the simulations are described below, and summarized in Table I and Fig. 1. The specific mutants chosen are listed in a supplementary data file.

First, we chose a small set of 55 mutants, immersed each mutant in a cubic box of explicit water solvent, and carried out 110-ns MD simulations in triplicate for each mutant. The initial 10ns were performed with harmonic restraints as part of simulation equilibration, and the remaining 100ns were used to compute statistics for correlation. As a proxy for mutant fitness, we computed (i) root-mean-square distance (RMSD) of the C_α protein atoms from the simulation structures to the corresponding initial structures, (ii) average root-mean-squared fluctuation (RMSF) of protein C_α atoms, and (iii) average RMSF of the centers-of-mass (COMs) of all residues; these metrics were averaged over the triplicates. The three metrics were chosen based on the assumption that mutants that express poorly do so because of a destabilization of the folded

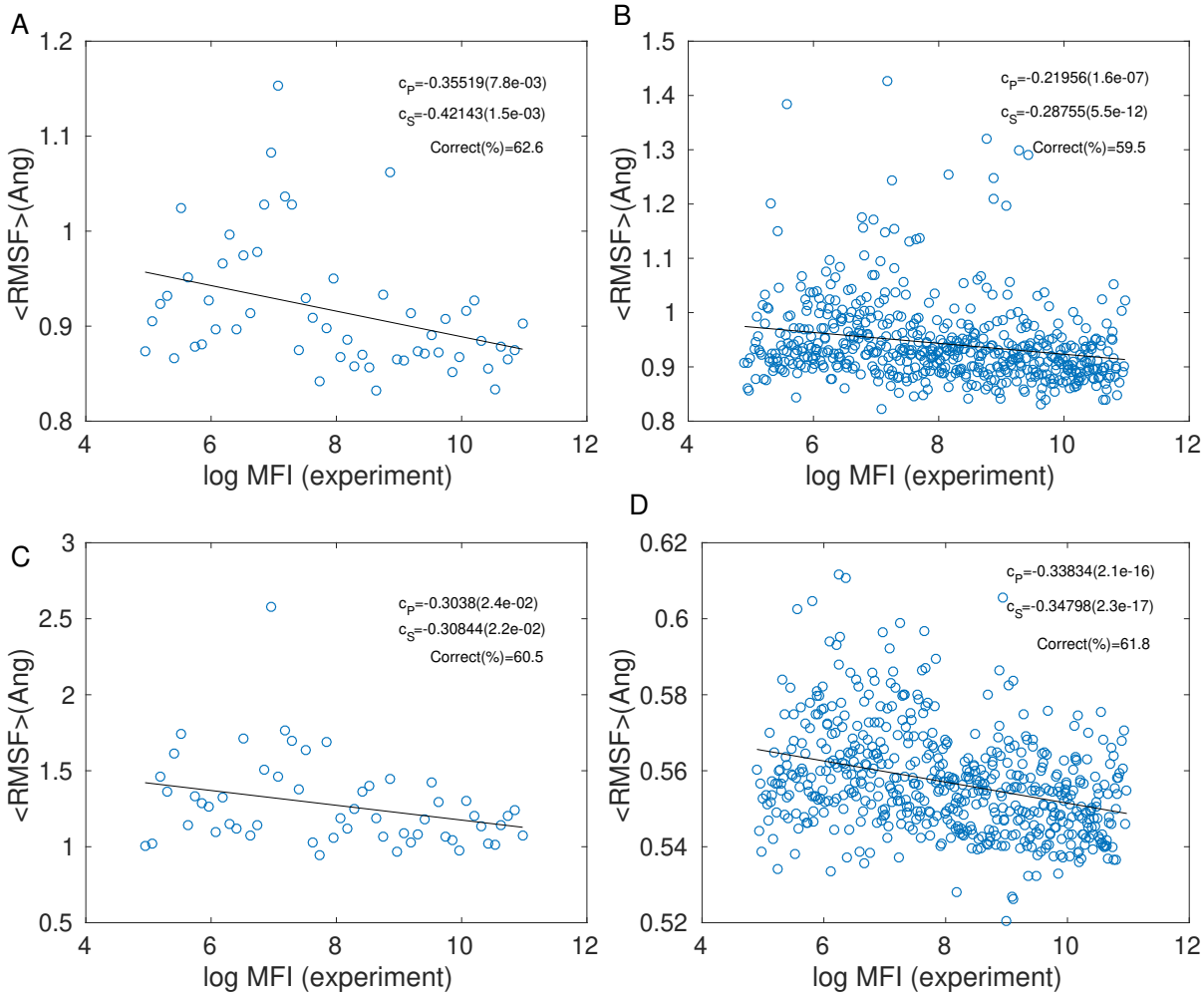

FIG. 1. Scatter plots of RMSFs from simulations *vs.* logarithm of experimental MFI for various simulation sets A. Simulations performed in a water cube; B. Simulations in a water shell; C. 1μs-long simulations in a water shell; D. Simulations in a water shell with conformational restraints.

conformation of the wild-type (WT/Wuhan) strain, which results in higher overall distance from, and/or larger fluctuations around, the WT structure. We note that this hypothesis predicts negative correlations between expression level and RMSDs or RMSFs, as is observed in most of our data. The three measures gave similar results (Tab. S1), so that in Table I and Fig. 1 we only show results with the RMSF-COM metric.

The Pearson and Spearman (rank) correlation coefficients between the RMSF-COM and the $lMFI$ are -0.36 and -0.42, respectively, indicating relatively low, but significant, correlation (respective p-values are 7.8e-3 and 1.5e-3). Because in protein design it is often of interest to predict mutants with improved binding or fitness, we also computed the fraction of predictions that correctly rank any two mutants, relative to the $lMFI$ data, at $\sim$63%. Since a completely uninformed model would predict this value to be 50%, the 'enrichment' obtained using MD simulations is $63/50 = 1.26$, or 26%.

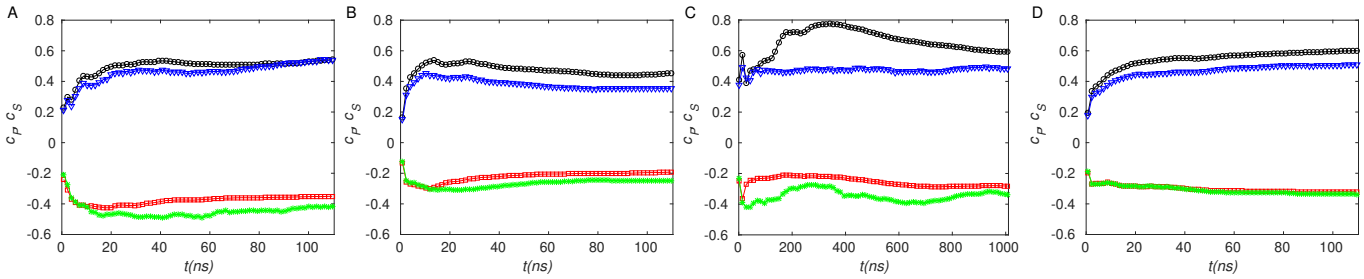

FIG. 2. Time-evolution of the Pearson and Spearman correlation coefficients between RMSDs and $lMFI$ (red □ and green *, respectively), and the average corelation between the RMSDs of simulation replicates (black ○ and blue ▽ , respectively), for various mutant sets. The legends for panels A–D are as in Fig. 1.

The above results were obtained from MD simulations in a cubic box of explicit solvent, which contain ∼70K atoms, and simulation speeds of ∼230 ns/day (on a workstation equipped with an NVIDIA RTX2080Ti graphics processor and a Ryzen 9 3900X CPU). Therefore, about 12 hours of simulation were needed for each mutant. Although simulations can be run in parallel on a supercomputer, it is desirable to explore more approximate MD methods that reduce simulation cost. To do this, we performed the next round of simulations using the shell solvation model developed by us.[6] With this model, each mutant protein is placed in a shell of solvent with thickness 8.5Å, and solvent evaporation is prevented using a half-harmonic potential. Further details can be found in Methods and in Ref. 6. The resulting simulation systems consisted of ∼10K atoms, and yielded ∼700 ns/day, *i.e.*, $> 3\times$ faster than the simulation in the cubic box. The increase in speed was due mainly to the reduced number of atoms, but also to the related fact that a shorter nonbonded cutoff (9Å) could be used. We note that the solvation shell model has known artifacts, such as artificial surface tension, and a small but controllable departure from equilibrium.[6]

With the increased speed, reduced storage requirements, and access to a supercomputer, we were able to generate more reliable statistics. We chose 557 mutants, which included the previous set of 55, and were uniformly distributed on the mutational landscape. For these 110-ns solvation shell simulations, the Pearson and Spearman correlation coefficients between the RMSF-COM and the $lMFI$ were -0.22 and -0.29, respectively, corresponding to a decrease in the predictive ability relative to the cubic solvation cell (see Fig. 1). The p-values of 1.6e-7 and 5.5e-12, respectively, were much lower than in the previous case, due to the larger simulation set size. The fraction of correct predictions was 60%, which corresponds to an enrichment factor of 20% over a random model. Thus, the use of a simplified solvation treatment resulted in deterioration of the fitness predictions.

To determine the effect of MD simulation length on the prediction quality, we extended the solvation shell simulations of the 55 mutants described before to 1μs, and computed the RMSDs of the RBD structures for the three sets of simulations (see Fig. 2) *vs.* time. Although one might expect that longer simulation lengths

would result in better predictions, *e.g.*, because the additional time would facilitate structural relaxation toward low free energy conformations,[7] the results are generally at odds with this expectation. This may be related to the work of Raval et al.[8], who found that long MD simulations do not necessarily improve the refinement of protein structures built by homology modeling. For all three sets of simulations, the highest correlations (ignoring the negative sign) occur for $t \leq\sim 50ns$, *i.e.* relatively early in the simulation trajectories (Fig. 2A-C).

The best correlation to the $lMFI$ data corresponds to the fully solvated (cubic box) simulations at $t \sim 55$ns, at which time the Spearman coefficient is ~ -0.5. For the solvation shell simulations, the best correlation is found at $t \leq 25$ns, with the Spearman coefficient of ~ -0.33 and -0.44 for the sets of 110ns and 1010ns simulations, respectively. For the other simulations, the correlations tend to worsen gradually with time, though not necessarily monotonically.

While the physical origins for the gradual deterioration in the correlations over the simulation times are unclear, the observation that the motions most predictive of fitness occur early in the simulations are of practical use. First, it indicates that very long simulations are not necessarily beneficial. Second, the early simulation samples are generally in closer proximity to the starting simulation structures (here, the Xray crystal structure of the WT RBD[9] to which point mutations were applied). Sampling in the vicinity of the starting structure can be increased with the use of harmonic restraints.

To test whether such restraints would improve the fitness predictions, we repeated the 557 simulations using the shell model with a weak best-fit harmonic restraint on the RMSD between the simulation and the starting structure, applied to the heavy atoms for the entire 100ns. The results of this set of simulations show an improvement in both Pearson and Spearman correlations (-0.43 and -0.35, respectively), and in the enrichment (61.8%) over the unrestrained simulations. The temporal evolution of the correlation between the RMSD and $lMFI$ decreases slowly, but monotonically in the restrained simulations, indicating that the longer restrained equilibration increases the number of predictive simulation samples.

Overall, the present MD simulations of RBD mutants show a modest ability to predict expression fitness (absolute Spearman correlation of $0.3 - 0.4$ and an enrichment factor of 20%–25% over a random model. It is noteworthy that each prediction requires on the order of 100ns of simulation time, which can be generated in $\sim$4 hours on a single GPU workstation. Moreover, different simulations can be run in parallel.

The DMS dataset of Starr et al.[5] also contains measurements of the dissociation constant K_D for the RBD mutants and the ACE2 receptor. To check how the quality of our fitness predictions from MD simulation would compare to predictions of K_D, we performed unbiased simulations in triplicate of 559 mutant RBDs

in complex with ACE2 using the solvation shell model, and computed the correlations between the average RMSFs and the negative logarithm of K_D (lK_D). We note that these 559 mutants are different from the 557 chosen for computing expression correlations, because they were selected to sample uniformly the space of experimental lK_D rather than $lMFI$ (see Methods). The results are shown in Table I and Fig. S1. These simulations were substantially more computationally costly because they involved simulating ACE2 in addition to RBD, and because our focus in this study is on fitness of the RBD. Therefore, we did not perform any simulations in the cubic box.

The computed correlations between lK_D and the RMSFs are significantly lower than those for the $lMFI$ simulations (Table I and Fig. S1). When the fluctuations are averaged over both the RBD and ACE2 residues, the Pearson and Spearman correlation coefficients are 0.01 and -0.14, respectively. The p-value for the latter is 8.3e-4, indicating that the the relative ranking between mutants is significant, despite providing only a small advantage over a random model (an enrichment of 9.2%). If the RMSFs are averaged only over the RBD, the respective correlation coefficients improve to -0.17 and -0.27, and the fraction of correctly ordered predictions increases to 59.4% (an enrichment of 18.8%).

A possible explanation of the correlations between RMSFs and lK_D's is that higher RBD fluctuations reflect a larger entropy penalty that has to be overcome to achieve strong binding. The fact that the correlation improves when only the RBD residues are considered suggests that such an entropy loss is associated specifically with ordering the mutant RBD residues. To investigate this hypothesis, we computed an approximation to the conformational entropy of the RBD in the ACE2-bound and unbound states using quasiharmonic analysis[10,11] in coarse-grained residue coordinates (see Supporting Material for details). It is known that quasiharmonic entropy typically overestimates the true conformational entropy (see, *e.g.*, Tyka, Clarke, and Sessions[12]). However, because the the RBD mutants have nearly identical structures, the entropy values corresponding to different mutants are expected to have similar systematic errors, allowing a meaningful relative comparison. Scatter plots of the temperature-scaled RBD entropy in the bound and unbound states *vs.* lK_D are shown in Fig. S2. The correlations between the negative lK_D and the computed entropies were found to be in the range [-0.28, -0.24], with p-values of 7e-8 or less, indicating modest but statistically significant correlations between the ACE2-RBD binding free energy and entropy of the RBD, both in the bound and the unbound states (see Fig. S2). These calculations therefore support our interpretation of the anticorrelation between RMSFs and lK_D. We note that it is technically possible to compute free energies of binding between the RBD and ACE2 exactly within the framework of classical statistical mechanics and restrained MD simulations.[13] Since such simulations would very likely require long

simulation times (*e.g.*, several μs per complex), they are currently too computationally costly for routine high-throughput screening.

As done in the above case for modeling fitness, we also computed the correlation of lK_D and the RMSD between the simulation and starting structures of the RBD/ACE2 complexes as a function of time (Fig. S3). The best correlation, corresponding to a Spearman coefficient of ~ -0.14, was at $\sim$42ns of simulation time.

Overall, unbiased MD simulations of RBD/ACE2 complexes using the solvation shell model to accelerate the calculations, resulted in smaller correlations, relative to the $lMFI$ data, between residue fluctuations of the RBD and the experimental lK_D data. Although simulating the complexes in a cubic box could increase the correlations, as in the $lMFI$ cases above, we have not done so because of the high computational cost involved ($\sim$31K atoms for the shell model *vs.* $\sim$255K atoms for the cubic box).

Fitness prediction methods grounded in physics-based simulations, such as the MD approach evaluated here, have the advantage that they generally do not require tuning to fit a specific protein, since they are derived from general principles, *e.g.*, an energy function that is applicable to any protein without highly specialized chemical modifications.[14–17] Their main disadvantages are that (1) some structural information is required, *e.g.*, the Xray crystal structure used here, (2) the predictions may not be very accurate, *e.g.*, if they are based on statistical samples that require long simulations to converge (though this is not observed here, as described above),[18] or (3) they fail to incorporate some important factor that affects fitness, *e.g.*, the experimental protein expression system has a protease that tends to cleave particular amino-acid sequences,[19] independently of the thermodynamic stability of the folded protein.

By contrast, prediction methods based on machine learning (ML) do not require protein structures (although some algorithms benefit from them[20,21]), but rely on a preexisting data set for model training.[22,23] If the training data set is sufficiently representative of the set for which predictions are desired, ML models can automatically correct for systematic biases.

To compare the prediction quality of our MD-based fitness metrics with that from an ML model, we parametrized a simple sequence-based linear regression model to compute either RBD mutant fitness ($lMFI$) or binding affinity to ACE2 (lK_D), as described in *Methods*.

The main results of the ML model are shown in Figs. 3 and S4 for $lMFI$ and lK_D, respectively, and summarized in Table II. With 50% of the data split between the training and testing sets, the Pearson and Spearman correlation coefficients for modeling $lMFI$ are ~ 0.75 and ~ 0.78, respectively. The error bars for the correlation coefficients are generally less than $\sim$0.005, and represent twice the standard deviation, or 95% confidence limits. They are shown in Fig. 3C, in which the various model performance metrics

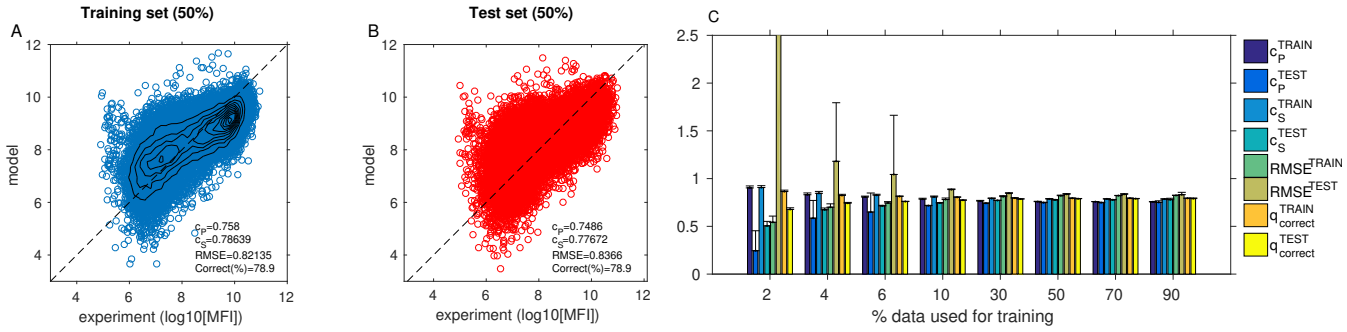

FIG. 3. Results of the linear regression model to predict $lMFI$ expression. A & B: scatter plots of model predictions *vs.* experiment; A: training set; B: testing set; 50% of the data was assigned to each set. In A, ten equispaced contours of the 2D joint probability density function (PDF) of modeled and experimental $lMFI$'s are drawn, spanning the range of the PDF; the majority of the $lMFI$ values fall near ∼10 which cannot be seen from the scatter plot alone because of the large amount of the data. 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 text). The training data were chosen randomly, and the error bars in C correspond to twice the standard deviation computed from ten training realizations.

	Training set				Test set			
Model	c_P	c_S	RMSE	Correct(%)	c_P	c_S	RMSE	Correct(%)
$lMFI$	0.76	0.79	0.82	79.5	0.75	0.78	0.83	78.9
lK_D	0.73	0.77	1.4	78.8	0.72	0.76	1.4	78.4

TABLE II. Summary of linear regression model results. The training set contained 27192 sequences, corresponding to 50% of the DMS dataset, chosen randomly with uniform probability; the remaining 27192 sequences comprised the test set. p-value for the correlation coefficients are not explicitly shown, because they were zero to within machine precision, which was due to the large number of samples. The statistical uncertainties in the correlation coefficients are indicated in Fig. 3. RMSE denotes the root-mean-square error in the units of experiment.

are displayed as a function of the percentage of data used for training *vs.* testing. The similarity of the correlations for the training and testing sets (as long as the training sample size is above 10% of the total sample size) indicates that the model shows no siginficant overfitting to the training data, due in part to the simplicity of the model.

Although far more sophisticated and accurate models can be parametrized, such as natural language,[22] gaussian process,[20] or deep learning[23] models, our present objective is a qualitative comparison of the accuracy of MD-based *vs.* ML-based models, for which the modeling used here is sufficient. If enough training data are available, which in our case corresponds to about 5K mutations, or about 5000/200=25 mutations for each residue in the spike RBD (ca. residues 331–531), the present ML models provide superior predictions of both fitness and binding affinity.

We note that the model used here is linear in *residue space*, which implies that the contribution to the $lMFI$ or lK_D of each residue is additive. However, residues, and therefore mutations, obviously interact with each other *via* the protein structure, which is manifested through epistatic shifts,[1,5] and can be modeled by parametrizing additional residue *interaction* weights[24] (couplings in the Potts model formalism),[25] rather

than only the residue weights, as done here. This obvious drawback of the present model is illustrated in Figs. S5 and S6, which shows that the ML model performance is best for single-residue mutants, and gradually deteriorates for higher-order mutants. However, the overall correlation between experiment and model of > 0.7 for both fitness and binding affinity indicates that even simple linear regression models outperform molecular dynamics-based criteria, if sufficient model training data are available.

CONCLUSION We have investigated the accuracy of two related structural metrics obtained from classical MD simulations, RMSD and RMSF, in predicting changes in (1) fitness of residue mutants of the SARS-CoV2 RBD and (2) binding of the RBD to the ACE2 receptor. We quantified the predictive ability of MD using Pearson and Spearman correlations between RMSFs and RMSDs computed from >0.7ms of total simulation time with the deep mutagenesis data of Starr et al.[1] (Table I).

The computed correlations with the logarithm of the experimental mean fluorescence intensity are modest, in the range 0.3 – 0.4, with the highest correlations observed for the RBDs solvated in a cubic box. The less computationally intensive simulations with RBDs solvated in a solvent shell with restraints to the starting coordinates of the protein heavy atoms produced a Spearman correlation of ∼0.35. Interpreting the simulation data in terms of the percentage of correct ranking of all mutant pairs, we observe at most 62.6% correctness, which indicates a modest advantage over a random prediction of 50% (a 25% enrichment over the random model). We emphasize that the correlations to MD data computed here, although modest, are statistically significant, as indicated by the p-values that accompany all correlation values in Table I. In particular, the simulations of the larger mutant set (557 mutants) yielded p-values that were all below 2e-7, with some p-values being zero within the limit of machine precision (∼2e-16). However, although the statistical significance of the correlations is robust, demonstrating the robustness required testing the large number of mutants. A practical consequence is that obtaining useful predictions with the RMSD and RMSF metrics could require a large number of samples. Thus, these metrics are unlikely to be helpful in situations where, *e.g.*, a small set of mutations are to be tested.

The correlations between the fluctuations in the RBD/ACE2 complex and the experimental lK_D, are lower, being optimal at -0.27 for the Spearman correlation, when only the RBD is considered (*i.e.* ACE2 fluctuations are ignored). Because lK_D is linearly proportional to the RBD/ACE2 binding free energy (bFE), we speculated that the contribution to the bFE reflected in the RMSFs corresponds to the entropic cost of ordering the RBD that is necessary for binding. This hypothesis appears to be supported by quasi-harmonic analysis of the MD trajectories in coarse-grained residue coordinates, which showed a negative correlation between classical qhasiharmonic entropy of the RBD in isolation, as well as in complex with

ACE2, and binding strength.

MD simulations have been used previously to evaluate the stability of protein structures.[26–28] Bhardwaj et al.[26] incorporated short MD simulations into design screening, but it was not clear how essential this step was for the efficiency of the design process. Radom, Plückthun, and Paci[27] evaluated alternative structures of two protein complexes modeled with RosettaDock[29] and observed that "decoy" complexes were much more likely to dissociate than the corresponding correct structure. Buchko et al.[28] were able to predict the most stable designs in a trial set, but not to "rescue" the failed designs (those that that did not fold) *via* rational mutation engineering. Related applications of MD simulations to the refinement of approximate structures to approach atomic accuracy[7,8] met with mixed success, emphasizing the importance of an accurate energy function[7,8] or long simulation times to overcome slow conformational transitions.[7] With respect to the last point, a noteworthy and somewhat surprising finding from our unrestrained simulations is that optimal correlations with both the $lMFI$ and lK_D are obtained at relatively short simulation times (25–50ns); longer simulations generally did not lead to improvement. However, the differences in the protein systems studied, as well as in the study goals (refining a structure, rather than predicting expression) could explain the apparently conflicting findings.

The present results were obtained using only a single protein system, *i.e.* SARS-CoV-2 RBD bound or unbound to ACE2, because the required computational expense (736μs of simulation time) did not allow us to investigate other realistic proteins. However, we hope that it will stand out in its statistics-based conclusions, *vs.* the previous, more anecdotal, case studies. Further work will be needed to determine how well the predictive ability of the simple RMS metrics studied here applies to other proteins.

Finally, we note that it is possible to optimize the selection of various fitness metrics from structure ensembles generated by MD simulation or other methods, such as homology modeling, to achieve higher correlations with desired experimental observables (*e.g.*, level of protein expression, or binding affinity) than obtained here using RMSFs and RMSD. Efforts towards this goal were recently described by us elsewhere,[18,30] and were not undertaken here, as our focus was specifically on the predictive ability of the two widely used fluctuation metrics.

Methods

Molecular dynamics simulations Initial coordinates for all molecular dynamics (MD) simulations were prepared starting from the Xray crystal structure of the SARS2-CoV receptor binding domain (RBD) in complex with the angiotensin-converting enzyme receptor II (ACE2), taken from Protein Data Bank file 6M0J.[9] In the RBD, the N-terminal residues 331–332 and C-terminal residues 527–531 were unresolved, and

were modeled using default internal coordinate tables in CHARMM[31]. Coordinates for all mutant residues were built by deleting all of the side chain atoms of the original residue, except for C_β, and building the corresponding mutant side chain from default internal coordinate tables. To avoid the possibility of steric clashes, the bond lengths in the side chains were set to 50% of their original size; the smaller side chains are restored to their original size during subsequent coordinate minimization.[32] The simple coordinate mutation protocol outlined above was adequate because most mutants involved only 1-2 mutations of noninteracting residues, raher than long residue loops. For simulations with ACE2, the unresolved C-terminal residues 616–621 were not modeled because they are ∼53Å away from the RBD binding site, and were deemed unlikely to differentially impact the binding affinity of ACE2 to RBD mutants.

To decrease simulation cost and reduce storage demand, we performed the remaining simulations using a quasi-equilibrium solvation shell model,[6] in which the RBD mutants (and ACE2, if included) were surrounded by a thin layer of solvent, whose evaporation was prevented using restraint potentials implemented as a plugin for the MD library OpenMM.[33] The solvation model has been validated in unbiased MD as well as free energy simulations, and found to introduce relatively minor artifacts, in the form of a slight but controllable departure from equilibrium dynamics, and increased solvent pressure on the protein.[6] Comparisons of potentials of mean force (PMFs) profiles of antigen-antibody separation produced with the shell model *vs.* periodic solvent box showed that the differences between them were small.[34] Sodium and chloride ions were added to the solvent shell to achieve a total ion concentration of 150mM.

The use of a solvent shell of 10Å thickness resulted in total atom counts in the range 9.65K–10.1K for the RBD mutants in isolation, and in the range 31K–31.1K for RBDs complexed with ACE2; the exact count depended on the RBD mutant simulated, and allowing MD simulation speeds of ∼700ns/day and ∼300ns/day, respectively, using an NVIDIA RTX2080Ti GPU and a Ryzen 3900X CPU.

All solvated systems were simulated using the OpenMM[33] library with the CHARMM36 energy function[14,35] in TIP3 water.[36] The MD simulation setup included particle-mesh Ewald electrostatics, a 9Å nonbonded cutoff for Lennard-Jones interactions, and hydrogen-mass repartitioning, which transfers 3 a.m.u. to every hydrogen from the parent heavy atom, allowing the use of a 4fs simulation step. Covalent bonds to hydrogens were held rigid, as were the water molecules, and the equations of motion were integrated using a Langevin dynamics integrator at 300K with the atomic friction set to 1ps^{-1}. In addition, for the shell simulations, we used (1) a density controller to maintain solvent density in the outer solvation shell at ≃1g/mL by dynamically adjusting the solvent layer thickness, and (2) a rigid-body restraint with force constant k=10kcal/mol/Å^2 to prevent rigid-body motion of the proteins.[6]

The first 10ns of each simulation were performed with a harmonic restraint on the root-mean-square deviations (RMSDs) of all protein heavy atoms from their respective initial positions, as part of equilibration. In one set of simulations (see main text and Table I), the restraints were retained for 1010ns. The RMSD restraint force constant was 100kcal/mol/Å^2.

RMSDs were computed using Visual Molecular Dynamics.[37] Residue RMS fluctuations, and classical quasi-harmonic entropies were computed in Matlab[38] (see Supporting Material for details).

SELECTION OF MUTANT STRAINS Because simulating all ≃54,000 RBD mutants[1] would be computationally prohibitive, two subsets were selected, distributed uniformly in the experimental log-mean fluorescence intensity ($lMFI$). The first, smaller, subset was chosen by computing a histogram of $lMFI$'s with bin width $\Delta lMFI$= 0.105, which assigns each mutant to a unique bin, and choosing at random one mutant sequence from each bin. If a bin was empty, as can happen near the tails of the histogram, it was omitted. This subset contained 55 mutants. The second subset was chosen similarly, but with the bin width set to $\Delta lMFI$= 0.01, and with the requirement that this subset contain the first. This set contained in 557 mutants. The members of each subset are listed in a supplementary data file. A similar subselection was performed for choosing RBD mutants bound to ACE2 for correlating MD simulation data with experimental lK_D's, which was done on the basis of uniform lK_D rather than $lMFI$. The bin width was ΔlK_D= 0.01, which resulted in a set of 559 mutants. We note that the numbers of mutants chosen based on $lMFI$ *vs.* lK_D were similar (557 and 559, respectively) because the experimental signal range in logarithm space was about 5.5 units for each dataset (see Fig. S1 in Ref. 5), even though the sets contained different mutants.

SEQUENCE COORDINATE NOTATION Let $\widehat{\boldsymbol{S}} \in M_{N_s \times N_r}$ denote a multiple alignment matrix (MSA) of N_s antigen sequences, with each row $\boldsymbol{S}_i$ of $\widehat{\boldsymbol{S}}$ corresponding to a unique sequence i of length N_r (Fig. S7A). $\widehat{\boldsymbol{S}}$ is represented by the standard 20-amino acid alphabet (A). We did not consider insertions or deletions and therefore the MSA did not have gaps. However, to accommodate more general MSAs, the alphabet could be extended to include the gap symbol ($G \equiv$ '–'), with a corresponding numerical representation.

To be able to perform standard mathematical operations on $\widehat{\boldsymbol{S}}$, we embed $\widehat{\boldsymbol{S}}$ in an n-dimensional (nD) vector space, as follows. To each residue s_{ij} at position j in $\boldsymbol{S}_i$, we assign a vector $\boldsymbol{x}_{ij} \in \mathbb{R}_n \equiv \boldsymbol{x}(s_{ij})$ (see Fig. S7B). A number of such embeddings (also called encodings[39]) can be found in the literature,[39–42] which have been used in machine learning models of sequence-activity relationships.[42,43] Previous studies have found that the choice of embedding can impact model quality.[39] For the present calculations, we used the simple 3D embedding model of Grantham [40] based only on the amino-acid physico-chemical properties,

i.e. such that each of the 20 amino acid types is represented by the same vector, regardless of its position in the sequence. Specifically, we represented the MSA with the 3D coordinates x_{ij}^k with k=1,2,3, from Ref. 40. This choice proved adequate, as it yielded model predictions of expression and binding affinity with relatively high Pearson correlation coefficients (in the range $\simeq$ 0.7-0.8) (see *Results*).

LINEAR REGRESSION MODEL We assume that the logarithm of the mean fluorescence intensity ($lMFI$) as well as the logarithm of the dissociation constant (lK_D) can be modeled as a weighted sum over the embedded residue coordinates. Specifically, for a sequence indexed by i, the model value y^{mod}, with $y \in \{lMFI, lK_D\}$ is

$$y_i^{mod} = \sum_{j=1}^{N_r} \sum_{k=1}^{3} x_{ij}^k w_j^{mod}, \tag{1}$$

where w_j^{mod} is an optimized weight for the residue at position j, specific to either the $y \equiv lMFI$ model or the $y \equiv lK_D$ model.

The optimized weights are obtained by minimizing the total squared error over a pre-defined *training* data set $\{\boldsymbol{X}_i; y_i^{exp}\}, i \in I_{tr}$,

$$\begin{aligned} \boldsymbol{W}^{mod} &= \arg\min_{w_j \in \mathbb{R}} \|\boldsymbol{y}^{mod} - \boldsymbol{y}^{exp}\|^2 \\ &= \left(\widehat{\boldsymbol{X}}^T \widehat{\boldsymbol{X}}\right)^{-1} \widehat{\boldsymbol{X}}^T \boldsymbol{y}^{exp} \\ &\equiv \widetilde{\boldsymbol{X}}^{-1} \boldsymbol{y}^{exp}, \end{aligned} \tag{2}$$

where we defined $\widetilde{\boldsymbol{X}}^{-1}$ as the pseudo-inverse of $\widehat{\boldsymbol{X}}$.[44] Different training sets are discussed in the Results and the Supporting Material.

To estimate statistical errors in the model predictions, we sampled N_{tr} training sets randomly from the total DMS data set with uniform probability, and computed averages and standard deviations over the N_{tr} sets of model samples. For the data in Figs. 3 and S4, N_{tr} was set to 10. The model was implemented in Matlab.[38]

ACKNOWLEDGMENTS VO thanks Drs. Simone Conti and Aravinda Munasinghe for stimulating discussions. Financial support was provided by the CHARMM Development Project at Harvard University, and by the Bill & Melinda Gates Foundation and Flu Lab under joint grant opportunity OPP1214161. Computer resources were provided by National Energy Resource Scientific Computing Center (NERSC), which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and by Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725. The findings and conclusions contained within

are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation.

SUPPORTING MATERIAL Supporting Text, Table S1, Figures S1–S6, and a data file with the mutants selected for MD simulation accompany this article, and can be freely accessed on-line. The computer analysis code for this paper can be found online at `https://github.com/ovchinnv/cov-dms-simulations2023.git`.

[1] T. Starr, A. Greaney, W. Hannon, A. Loes, K. Hauser, J. Dillen, E. Ferri, A. Farrell, B. Dadonaite, M. McCallum, K. Matreyek, D. Corti, D. Veesler, G. Snell, and J. Bloom, Science (New York, N.Y.) **377**, 420 (2022).

[2] N. Andrews, J. Stowe, F. Kirsebom, S. Toffa, T. Rickeard, E. Gallagher, C. Gower, M. Kall, N. Groves, A.-M. O'Connell, D. Simons, P. B. Blomquist, A. Zaidi, S. Nash, N. Iwani Binti Abdul Aziz, S. Thelwall, G. Dabrera, R. Myers, G. Amirthalingam, S. Gharbia, J. C. Barrett, R. Elson, S. N. Ladhani, N. Ferguson, M. Zambon, C. N. Campbell, K. Brown, S. Hopkins, M. Chand, M. Ramsay, and J. Lopez Bernal, New England Journal of Medicine **in press** (2022), 10.1056/NEJMoa2119451.

[3] J. Shaman and M. Galanti, Science (New York, N.Y.) **370**, 527 (2020).

[4] P. Byrne and J. McLellan, Current opinion in immunology **77**, 102209 (2022).

[5] T. N. Starr, A. J. Greaney, S. K. Hilton, D. Ellis, K. H. Crawford, A. S. Dingens, M. J. Navarro, J. E. Bowen, M. A. Tortorici, A. C. Walls, N. P. King, D. Veesler, and J. D. Bloom, Cell **182**, 1295 (2020).

[6] V. Ovchinnikov, S. Conti, E. Lau, F. Lightstone, and M. Karplus, J. Chem. Theor. Comput. **16**, 1866 (2020).

[7] L. Heo and M. Feig, Proc. Natl. Acad. Sci. USA **115**, 13276 (2018).

[8] A. Raval, S. Piana, M. Eastwood, R. Dror, and D. Shaw, Proteins **80**, 2071 (2012).

[9] J. Lan, J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan, Q. Zhang, X. Shi, Q. Wang, L. Zhang, and X. Wang, Nature **581**, 215 (2020).

[10] B. Brooks, D. Janežič, and M. Karplus, J. Comput. Chem. **16**, 1522 (1995).

[11] I. Andricioaei and M. Karplus, J. Chem. Phys. **115**, 6289 (2001).

[12] M. Tyka, A. Clarke, and R. Sessions, J. Phys. Chem. B **111**, 9571 (2007).

[13] H. Fu, W. Cai, J. H'enin, B. Roux, and C. Chipot, J. Chem. Theor. Comput. **13**, 5173 (2017).

[14] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. de Groot, H. Grubmüller, and A. MacKerell, Nature methods **14**, 71 (2017).

[15] D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman, Computer Physics Communications **91**, 1 (1995).

[16] D. Shivakumar, E. Harder, W. Damm, R. A. Friesner, and W. Sherman, Journal of Chemical Theory and Computation **8**, 2553 (2012).

[17] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, Journal of Chemical Theory and Computation **4**, 435 (2008), http://pubs.acs.org/doi/pdf/10.1021/ct700301q.

[18] S. Conti, V. Ovchinnikov, and M. Karplus, J. Comput. Chem. **43**, 1747 (2022).

[19] B. Ryan and G. Henehan, Current protocols in protein science **Chapter 5**, Unit5.25 (2013).

[20] P. Romero, E. Stone, C. Lamb, L. Chantranupong, A. Krause, A. Miklos, R. Hughes, B. Fechtel, A. Ellington, F. Arnold, and G. Georgiou, ACS synthetic biology **1**, 221 (2012).

[21] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. Kohl, A. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, D. Silver, O. Vinyals, A. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, Proteins **89**, 1711 (2021).

[22] B. Hie, E. Zhong, B. Berger, and B. Bryson, Science (New York, N.Y.) **371**, 284 (2021).

[23] S. Biswas, G. Khimulya, E. Alley, K. Esvelt, and G. Church, Nature methods **18**, 389 (2021).

[24] R. Louie, K. Kaczorowski, J. Barton, A. Chakraborty, and M. McKay, Proc. Natl. Acad. Sci. USA **115**, E564 (2018).

[25] R. Levy, A. Haldane, and W. Flynn, Curr. Opin. Struct. Bio. **43**, 55 (2017).

[26] G. Bhardwaj, V. Mulligan, C. Bahl, J. Gilmore, P. Harvey, O. Cheneval, G. Buchko, S. Pulavarti, Q. Kaas, A. Eletsky, P. Huang, W. Johnsen, P. Greisen, G. Rocklin, Y. Song, T. Linsky, A. Watkins, S. Rettie, X. Xu, L. Carter, R. Bonneau, J. Olson, E. Coutsias, C. Correnti, T. Szyperski, D. Craik, and D. Baker, Nature **538**, 329 (2016).

[27] F. Radom, A. Plückthun, and E. Paci, PLoS computational biology **14**, e1006182 (2018).

[28] G. W. Buchko, S. Pulavarti, V. Ovchinnikov, E. A. Shaw, S. A. Rettie, P. J. Myler, M. Karplus, T. S. D. Baker, and C. D. Bahl, Prot. Sci. **27**, 1611 (2018).

[29] J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. Rohl, and D. Baker, Journal of molecular biology **331**, 281 (2003).

[30] S. Conti, E. Lau, and V. Ovchinnikov, Antibodies **11**, 1 (2022).

[31] B. Brooks, C. Brooks III, A. Mackerell Jr., L. Nilsson, R. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, and et. al, J. Comput. Chem. **30**, 1545 (2009), pMC2810661.

[32] F. Noé, F. Ille, J. Smith, and S. Fischer, Proteins **59**, 534 (2005).

[33] M. Friedrichs, P. Eastman, V. Vaidyanathan, M. Houston, S. Legrand, A. Beberg, D. Ensign, C. Bruns, and V. Pande, J. Comput. Chem. **30**, 864 (2009).

[34] V. Ovchinnikov and M. Karplus, J. Phys. Chem. B , in press (2023).

[35] O. Guvench, S. Mallajosyula, E. Raman, E. Hatcher, K. Vanommeslaeghe, T. Foster, F. Jamison, and A. Mackerell, J. Chem. Theor. Comput. **7**, 3162 (2011).

[36] W. Jorgensen and J. Tirado-Rives, Proc. Natl. Acad. Sci. USA **102**, 6665 (2005).

[37] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).

[38] MATLAB, Version 7.10.0 (R2010a) (The MathWorks Inc., Natick, Massachusetts, 2010).

[39] L. Nanni and A. Lumini, Expert Systems with Applications **38**, 3185 (2011).

[40] R. Grantham, Science (New York, N.Y.) **185**, 862 (1974).

[41] W. Atchley, J. Zhao, A. Fernandes, and T. Drüke, Proc. Natl. Acad. Sci. USA **102**, 6395 (2005).
[42] H. Mei, Z. Liao, Y. Zhou, and S. Li, Biopolymers **80**, 775 (2005).
[43] G. Liu, H. Zeng, J. Mueller, B. Carter, Z. Wang, J. Schilz, G. Horny, M. Birnbaum, S. Ewert, and D. Gifford, Bioinformatics (Oxford, England) **36**, 2126 (2020).
[44] A. J. Chorin and O. H. Hald, Stochastic Tools in Mathematics and Science, 3rd ed. (Springer, New York, NY, 2013).