Supporting Information for

Combining molecular dynamics with Bayesian analysis to predict and evaluate ligand-binding mutations in influenza hemagglutinin

Peter M. Kasson, Daniel L. Ensign, and Vijay S. Pande

Departments of Chemistry and Structural Biology, Stanford University, Stanford CA 94305.

Complete references 8 and 10:

(8) Auewarakul, P.; Suptawiwat, O.; Kongchanagul, A.; Sangma, C.; Suzuki, Y.; Ungchusak, K.; Louisirirotchanakul, S.; Lerdsamran, H.; Pooruk, P.; Thitithanyanont, A.; Pittayawonganon, C.; Guo, C. T.; Hiramatsu, H.; Jampangern, W.; Chunsutthiwat, S.; Puthavathana, P. *J Virol* **2007**, *81*, 9950-5.

(10) Yamada, S.; Suzuki, Y.; Suzuki, T.; Le, M. Q.; Nidom, C. A.; Sakai-Tagawa, Y.; Muramoto, Y.; Ito, M.; Kiso, M.; Horimoto, T.; Shinya, K.; Sawada, T.; Usui, T.; Murata, T.; Lin, Y.; Hay, A.; Haire, L. F.; Stevens, D. J.; Russell, R. J.; Gamblin, S. J.; Skehel, J. J.; Kawaoka, Y. *Nature* **2006**.

Simulation and Analysis Methods:

Molecular dynamics simulations. All simulations were run using Gromacs 4.0,[1] the AMBER03 forcefield for proteins,[2] and the GLYCAM forcefield for carbohydrates.[3] Topologies were prepared using LEAP and converted to Gromacs format using the amb2gmx script by Eric Sorin. The crystal structure of hemagglutinin from H5N1 strain VN1194 (PDB code 2IBX)[4] was used as the starting point for simulations; all mutations are with respect to this strain. Because the available human H5 crystal structures did not contain ligand, starting coordinates for α2,3-sialyllactose were taken from an avian H5 structure bound to LSTA (PDB code 1JSN)[5] that was structurally aligned to the 2IBX structure using DaliLite[6]. Only the crystallographically resolvable glycans were included in these simulations. The wild-type VN1194 and each mutant tested were simulated for 100ns (approximately 900 processor-days per simulation) using 64 processors of the Stanford BioX2 cluster (8-core Cloverton nodes connected via Infiniband). An additional 200 simulations of each protein trimer tested were also performed using Folding@Home.

Simulations were run using TIP3P water and 150 mM NaCl. Box dimensions were 100A x 100A x 150 A, allowing for approximately 41,100 water molecules depending on the particular mutant simulated. Simulations were run with a 2 fs timestep, constraining hydrogen bonds using the LINCS algorithm.[7] Temperature coupling was performed using the Berendsen method with a target temperature of 300K and a coupling constant of 1 ps. Periodic boundary conditions were employed with reaction-field electrostatics and a 1.2 nm cutoff for both van der Waals and coulomb interactions. The dielectric constant of 60 used for reaction-field calculations corresponds to the dielectric of 150mM sodium chloride at 30C. Additional simulation parameters are listed in Table S3. A plot of root mean squared deviation (RMSD) from the 2IBX crystal structure versus time for the initial 100-ns simulation of H5N1 hemagglutinin is given in Figure S5.

Mutual information analysis of dynamics. Displacement magnitudes relative to the start of the simulation were calculated at 100 ps intervals and used as the input for mutual information calculations. Displacements $d_i(t) = |X_i(t) - X_i(0)|$ were calculated for all alpha carbons of the protein and all ligand atoms i based on simulation snapshots X(t) at each time t after rigid-body alignment to the starting structure using least-squares distance to the bound sialic acid residues as a target function. Mutual information was calculated using histogram estimates of the probability density function for $d_i(t)$ with 10 bins at even intervals from $\min(d_i(t), \forall t)$ to $\max(d_i(t), \forall t)$. Sensitivity analyses of scoring versus selection of alignment group and number of histogram bins are presented in Figure S6. Mutual information H(i,j) was calculated in Python based on Matlab code kindly

provided by Relly Brandman and Yigal Brandman. The normalized quantity, symmetric uncertainty

$$S(i,j) = \frac{H(i,j)}{H(i,i) + H(j,j)},$$

was used to measure relatedness in a fashion maximally independent from the magnitude of motion undergone by each residue. Each residue *i* was then scored by excess mutual information [8] to the ligand relative to the protein: $E_i = \overline{S(i,j)} - \overline{S(i,k)}$ for each protein residue *i*, all ligand atoms *j* and all protein residues $k \neq i$. Because hemagglutinin is trimeric, E_i was measured for each residue relative to the bound ligand in the corresponding monomer and averaged over monomers. The top 5% of residues were then selected for further analysis.

Mutual information analysis of sequence data. Sequence mutual information was calculated as previously described[9] using a multiple sequence alignment of all available human and avian H5N1 hemagglutinin sequences as of July 2008. Mutual information was calculated in a discrete fashion between each position in the alignment, and symmetric uncertainty was used as the scoring metric. Pairwise mutual information was calculated each residue position in the H5N1 multiple sequence alignment as follows:

$$I(i,j) = H(i) + H(j) - H(i,j)$$

where

$$H(i) = \sum_{a \in A} -p(x_i = a) \log p(x_i = a)$$

and

$$H(i,j) = \sum_{a \in A} \sum_{b \in A} -p(x_i = a, x_j = b) \log p(x_i = a, x_j = b),$$

for sequence positions *i* and *j*, where the variable x_i represents the values of the multiple sequence alignment at position *i*. Use of a substitution matrix to give varying mismatch probabilities according to the chemical similarity of the amino acids involved may add further sensitivity.

Symmetric uncertainty[10] was used to normalize the pairwise mutual information matrix as follows: $U(i,j)=2*I(i,j) / (I(i,i) + I(j,j))$ for all positions *i* and *j*. Single-linkage hierarchical clustering was performed in MATLAB using U as the distance metric and guaranteeing $U(i,i) =1$, all *i*. The 99.9th percentile of all non-self symmetric uncertainty values was calculated, and the corresponding distance metric was used as a threshold for cluster identification. Other recent work has developed the quantity MIp as a measurement of mutual information corrected for phylogenetic conservation [11]. We computed MIp and its overlap with dynamics MI but found that for hemagglutinin MI yielded results that had a clearer physical interpretation than did MIp (Figure S7).

Generation of hemagglutinin point mutants. Point mutants were generated based on structures of the VN1194:α2,3-sialyllactose complex using Modeller[12]. Each mutant was energy-minimized, re-solvated, and then simulated in Gromacs as per the protocol

described above. Ligand dissociation was assessed by measuring the closest distance between the sialic-acid residue of the ligand and the sialic-acid binding pocket of the protein. Binding-pocket residues were defined as all residues within 5A of the sialic acid for >50% of the 100-ns simulation of the VN1194-α2,3-sialyllactose complex; a distance threshold of 10A was used for dissociation. As the protein and ligand did not have time to equilibrate properly in the unbound state, re-association was not assessed.

Bayesian analysis of dissociation rates. For each mutant *m* and each starting conformation X, we calculate the probability density function P(k_m | D, X, I) for the dissociation rate k_m given the simulation data D and the set of simulation conditions I. We approximate dissociation as a two-state reaction. The simulation data D can then be represented as follows: given N simulations, n of which lead to dissociation, we record the times of dissociation $\{t_1..t_n\}$ and the time intervals in which no dissociation was observed $\{T_1 .. T_{N-n}\}$. Then,

$$P(k_m \mid D,X,I) = \frac{\Theta^{n+1}}{n!} {k_m}^n \exp[-k_m\Theta], \text{ where } \Theta = \left[\sum_{i=1}^{n} t_i\right] + \left[\sum_{j=1}^{N-n} T_j\right]$$

and

$$P(k_2 > k_1 \mid D,X,I) = \frac{{\theta_2}^{n_2+1}}{n_2!}\frac{{\theta_1}^{n_1+1}}{n_1!}\int_{-\infty}^{0}\int_{0}^{\infty} {k_2}^{n_2} e^{-k_2\theta_2}\left(k_2-\delta\right)^{n_1} e^{-(k_2-\delta)\theta_1} dk_2 d\delta,$$

for any two protein mutants 1 and 2 (a more complete derivation follows). In addition to the 17 point mutants, we performed a set of "sham mutation" control experiments by running the mutation protocol given above using the wild-type VN1194 sequence; we use this dataset to evaluate the probability that each point mutation increases the dissociation rate relative to the wild-type complex. Bootstrap resampling was performed over simulation trajectories using 100 samples, and significance testing of the resulting distributions of P($k_m > k_{w.t.}$) was performed using the Kolmogorov-Smirnov test and a Bonferroni multiple-hypothesis correction for the number of mutants, averaging over starting conformations.

Derivation of probability that k_{off} for any given mutant 1 is faster than k_{off} for mutant 2.
We approximate dissociation as a two-state reaction, so it can be represented as a single-exponential process with probabilities:

$$P_{dissociate}(t = \tau \mid k,X,I) = ke^{-k\tau}$$

$$P_{stay}(t = 0..\tau \mid k,X,I) = e^{-k\tau}$$

We encapsulate a set of simulation data D = $(N,n,\{t_{i,d}\},\{T_i\})$ as a group of N simulations, n of which lead to dissociation with times of dissociation $\{t_1..t_n\}$, and N-n of which do not dissociate over the observation time intervals $\{T_1 .. T_{N-n}\}$. Then,

$$P(D \mid k,X,I) = \prod_{i=1}^{n} ke^{-kt_{i,d}} \prod_{j=1}^{N-n} e^{-kT_j}$$

$$= k^n e^{-k\Theta}, \text{ where } \Theta = \sum_{i=1}^{n} t_{i,d} + \sum_{j=1}^{N-1} T_j$$

Using a uniform prior for *k*, we obtain:

$$P(k \mid D,X,I) = \frac{\Theta^{n+1}}{n!} k^n \exp[-k\Theta]$$

Now consider two dissociation reactions, 1 and 2, with datasets D_1 and D_2. We wish to predict the probability that $k_2 > k_1$. We define a variable $\delta = k_1 - k_2 < 0$.

$$\begin{aligned} P(\delta \mid D,X,I) &= \int_0^\infty P(k_2 \mid D_2,X,I) P(k_1 \mid D_1,X,I) dk_2 \\ &= \int_0^\infty P(k_2 \mid D_2,X,I) P(k_2 - \delta \mid D_1,X,I) dk_2 \\ &= \int_0^\infty \frac{\theta_2^{\,n_2+1}}{n_2!} k_2^{\,n_2} e^{-k_2\theta_2} \frac{\theta_1^{\,n_1+1}}{n_1!} (k_2 - \delta)^{n_1} e^{-(k_2-\delta)\theta_1} dk_2 \end{aligned}$$

and $P(k_2 > k_1) = P(\delta < 0)$:

$$P(k_2 > k_1 \mid D,X,I) = \int_{-\infty}^{0} P(\delta \mid D,X,I) = \frac{\theta_2^{\,n_2+1}}{n_2!} \frac{\theta_1^{\,n_1+1}}{n_1!} \int_{-\infty}^{0} \int_0^\infty k_2^{\,n_2} e^{-k_2\theta_2} (k_2 - \delta)^{n_1} e^{-(k_2-\delta)\theta_1} dk_2 d\delta$$

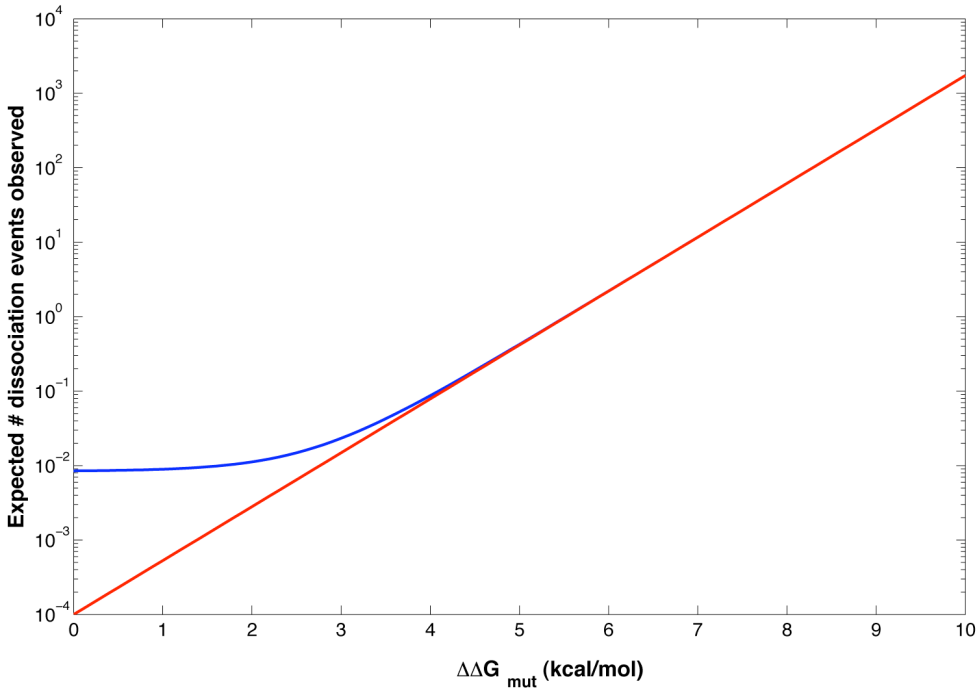

Figure S1. Expected number of dissociation events detected. The expected number of dissociation events detected in our simulations are plotted as a function of the ΔΔG‡ of a given mutant. The plot in blue assumes a wild-type k_{off} of 84 s^{-1}, and the plot in red assumes a k_{off} of 10^{-4} s^{-1}. Values are derived as follows:

$P(d) = 1-\exp(-1*(k_{off,wt} + \exp(\Delta\Delta G_{mut} / kT) * t)$
for small time,
$P(d) \sim= (k_{off,wt} + \exp(\Delta\Delta Gmut / kT) * t)$
MLE # of events for n simulations each of length $\tau = n*\tau*$ $(k_{off,wt} + \exp((\Delta\Delta G_{mut} / 0.6)$)
$n * \tau$ is taken as $1x10^{-4}$ s (e.g. 1000 samples of 100 ns). The average value for $n * \tau$ in our mutant dataset is slightly lower, at $6.2x10^{-5}$ s.

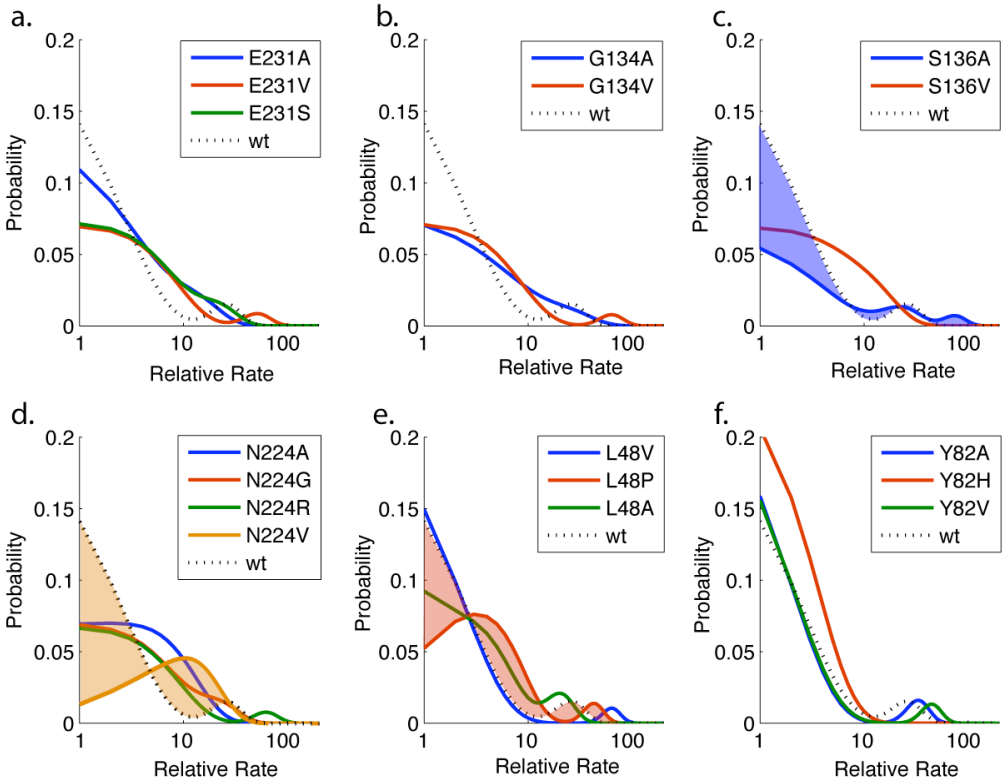

Figure S2. Probability distribution functions for mutant dissociation rates. Plotted in (**a-f**) are probability distribution functions for the dissociation rate of each mutant tested, compared against the wild-type VN1194 hemagglutinin. Dissociation rates are plotted relative to the maximum-likelihood estimate for the wild-type dissociation rate. The difference in probability distribution functions between the wild type and each of the three most significant mutants is highlighted.

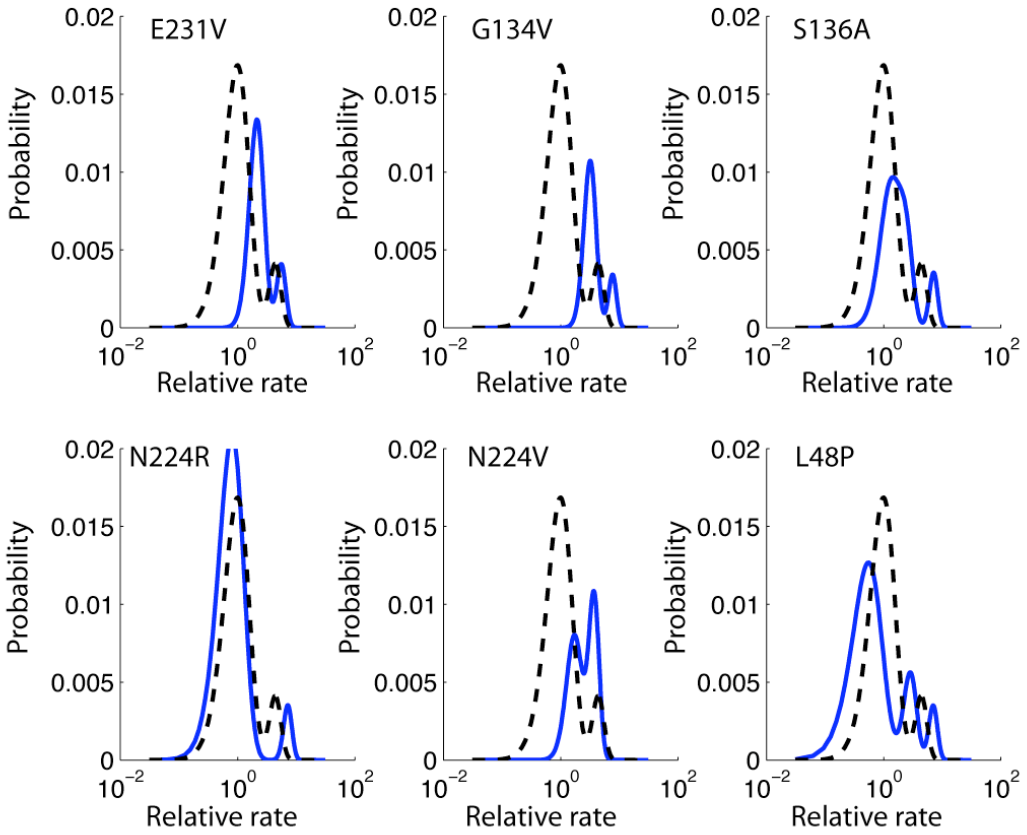

Figure S3. Probability distribution functions for additional sampling of mutant dissociation. Plotted are probability distribution functions for the dissociation rate of each mutant (blue lines) based on 33 different starting conformations each (3 from the crystal structure and 30 additional conformations selected at random from a 100-ns simulation. Dissociation rates are plotted relative to the maximum-likelihood estimate of the wild-type dissociation rate. Each of these plots represents sampling of 11-fold more starting conformations than the initial dataset, but analysis of $p(k_{off,mutant} > k_{off,\ wild\text{-}type})$ yields similar results.

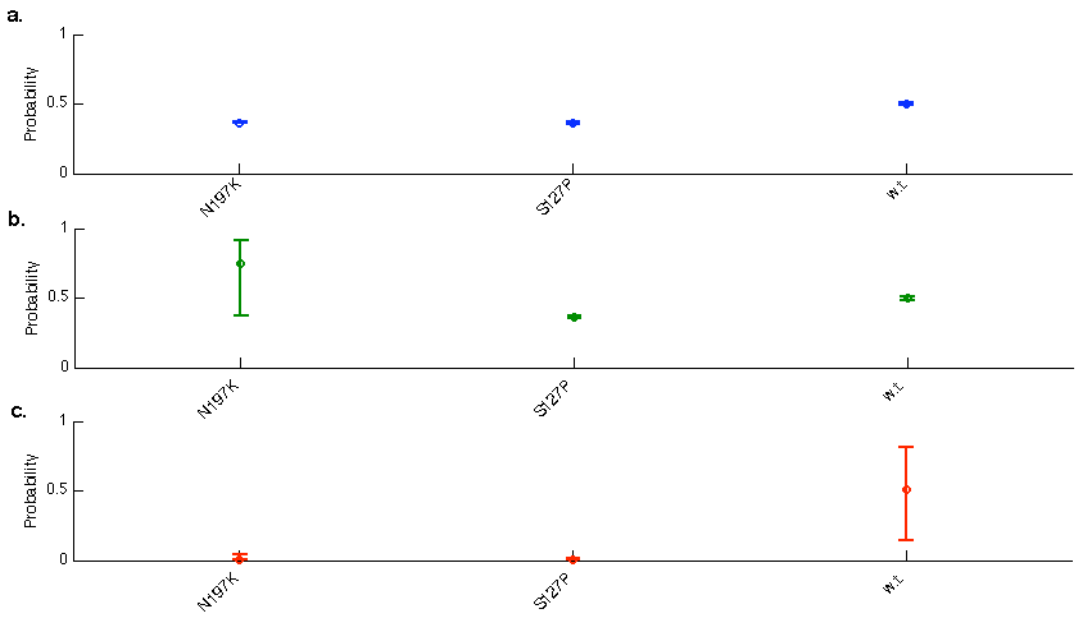

Figure S4. Negative controls: estimates of dissociation rate acceleration. For each starting conformation (**a-c**) and each mutant, the probability that the mutant k_{off} is faster than the wild-type VN1194 is plotted. Bars represent 90% confidence intervals from bootstrap resampling. As expected from experimental measurements of whole-virus binding to sialylglycopolymers, neither mutant N197K nor S127P shows substantial acceleration of k_{off}.), compared against the wild-type VN1194 hemagglutinin (black dotted lines).

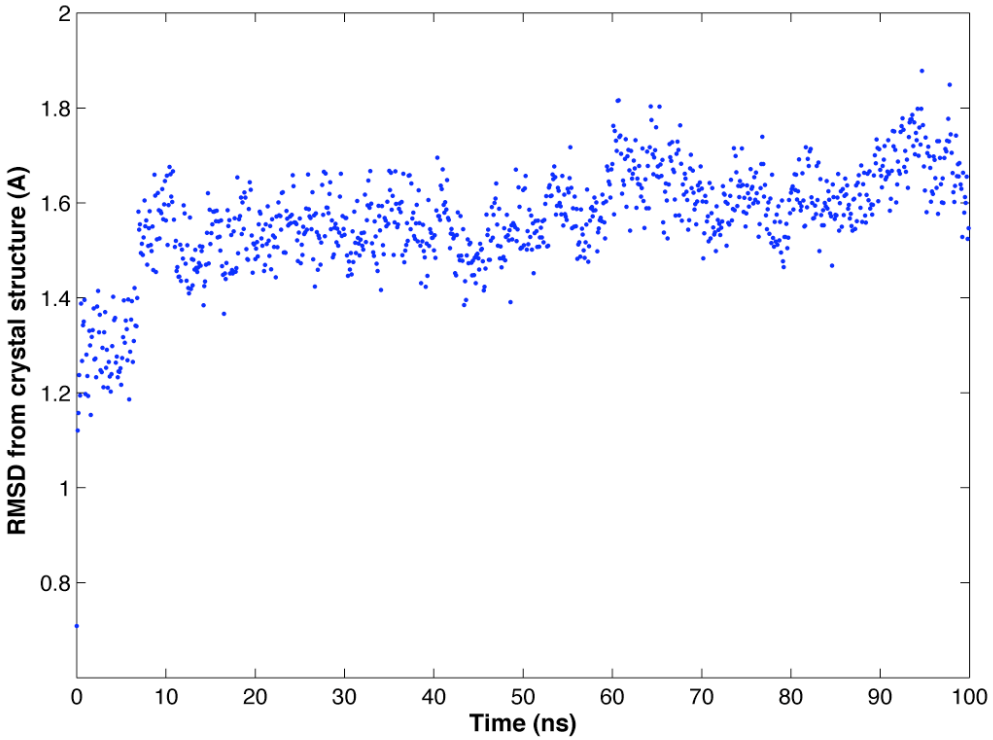

Figure S5. Root mean squared deviation from the crystal structure in a simulation of H5N1 hemagglutinin. The root mean squared deviation is plotted as a function of time for a 100-ns simulation of the H5N1 VN1194 hemagglutinin trimer. This simulation trajectory was used for calculation of mutual information values. RMSD was calculated for all alpha-carbons in the trimer.

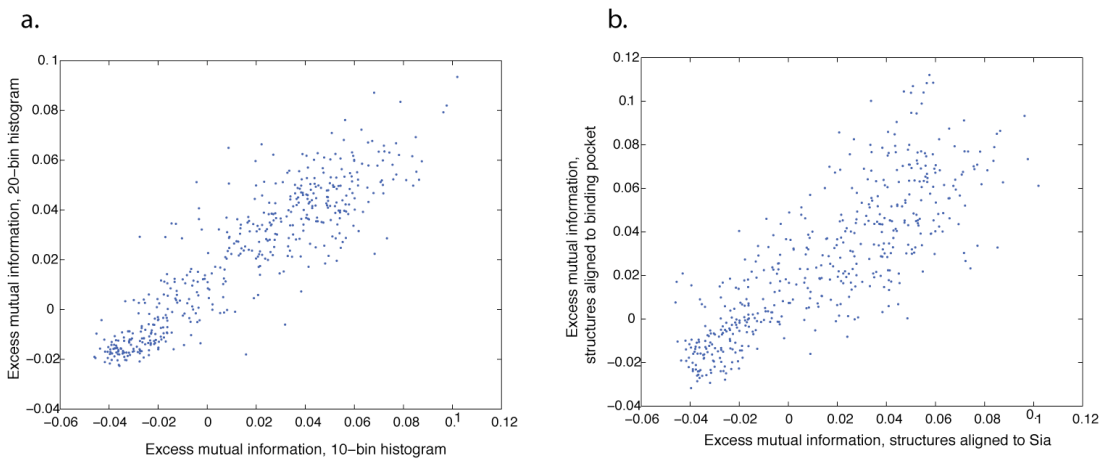

Figure S6. Sensitivity analysis of dynamics-mutual-information scoring to choice of alignment and number of histogram bins. Plotted in (**a**) are excess mutual information values calculated using 10-bin histogram estimates versus 20-bin estimates. The correlation coefficient is 0.90. Plotted in (**b**) are excess mutual information values calculated using rigid-body alignment to the bound sialic acids versus to the protein residues comprising the binding pocket. The correlation coefficient is 0.82.

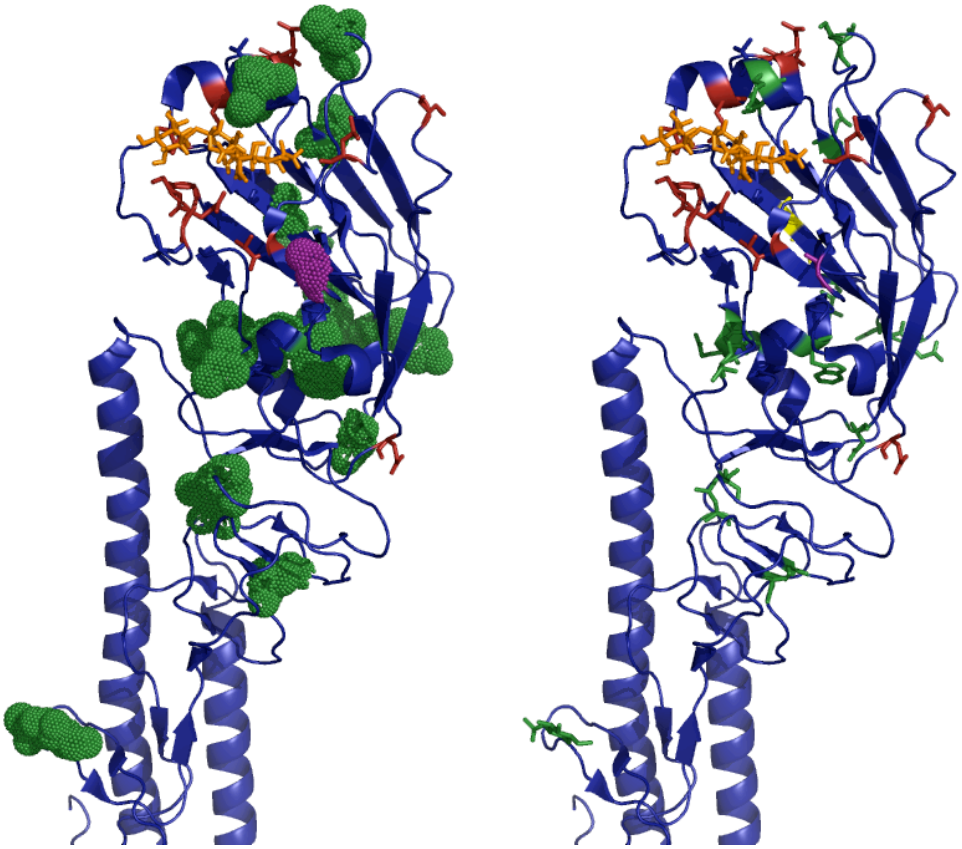

Figure S7. Identification of coordinated mutations using the MIp metric. In panel (**a**), residues scoring among the top 0.1% of MIp pairs are rendered in green. In panel (**b**), the two residues scoring both among the top 0.1% of MIp pairs and the top 5% via analysis of dynamics are rendered in yellow. Experimentally identified ligand specificity mutation sites are rendered in red, the remainder of the hemagglutinin protein is rendered in blue, and a bound α2,3-sialyllactose is modeled in orange. The one experimentally identified mutant also identified by MIp is rendered in magenta. The crystal structure of VN1194, PDB code 2IBX, was used for the protein coordinates.

Residue	Score
48	0.0719
49	0.0716
82	0.0755
100	0.0875
115	0.0713
133	0.0758*
134	0.0836
136	0.0850
138	0.0976*
139	0.1020
140	0.0788
141	0.0743
181	0.0716
182	0.0853
184	0.0720
201	0.0811
208	0.0840
213	0.0782
214	0.0770
224	0.0864
225	0.0963*
231	0.0723
250	0.0771
325	0.0733

Table S1. Top 5% of residues scored by excess mutual information to the ligand. Residues also identified experimentally are marked by an asterix. Residues 184, 208, and 231 also lie within the trimer interface region, which we define as residues having having a neighbor from another monomer within 5 A.

Mutation	Log-likelihood of significance	$P(k_{off} > k_{wild\text{-}type})$
E231A	2.0359	0.5317
E231V	71.5566	0.7984
E231S	0.0415	0.6105
G134A	0	0.6728
G134V	89.8379	0.8101
S136A	103.8709	0.8687
S136V	0	0.5494
N224A	0	0.5190
N224G	0.1662	0.6249
N224R	95.7274	0.8190
N224V	38.6503	0.6836
L48V	58.4274	0.6418
L48P	95.7274	0.8169
L48A	0.6648	0.5047
Y82A	1.4957	0.5180
Y82H	0	0.2857
Y82V	23.9318	0.5952
wild type	0	0.5000

Table S2. Log-likelihood values for significance of acceleration of k_{off} by hemagglutinin mutants. Because the dissociation rate of α2,3-sialyllactose from wild-type VN1194 hemagglutinin is slow, very few dissociation events are sampled in the simulation dataset but even a small number of events can represent an important acceleration of k_{off}. To evaluate this tradeoff quantitatively, bootstrap resampling was performed on the simulation trajectory dataset for each mutant, and the posterior probability $P(k_{mutant} > k_{wild\text{-}type})$ was evaluated for each of the samples. Kolmogorov-Smirnov significance testing was performed on the average of these probabilities across starting conformations, and the significance values were transformed into log-likelihood scores.

integrator	= md
tinit	= 0.0
dt	= 0.002
nstlist	= 10
ns_type	= grid
pbc	= xyz
rlist	= 1.2
coulombtype	= Reaction-Field
rcoulomb_switch	= 0.9
rcoulomb	= 1.2
epsilon_r	= 60
vdw_type	= Shift
rvdw_switch	= 0.9
rvdw	= 1.2
DispCorr	= EnerPres
tcoupl	= Berendsen
tc-grps	= Glycoprotein Ligand Solvent
tau_t	= 1.0 1.0 1.0
ref_t	= 300 300 300
Pcoupl	= No
constraints	= h-bonds
constraint_algorithm	= Lincs
unconstrained_start	= no
lincs_order	= 4

Table S3. Simulation parameters. Listed are simulation parameters in the Gromacs MDP file format. The dielectric constant of 60 used for reaction-field calculations corresponds to the dielectric of 150mM sodium chloride at 30C.

References:

(1) Hess, B.; Kutzner, C.; vanderSpoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435-447.

(2) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J Comput Chem* **2003**, *24*, 1999-2012.

(3) Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; Gonzalez- …, J. In *J Comput Chem* 2007.

(4) Yamada, S.; Suzuki, Y.; Suzuki, T.; Le, M. Q.; Nidom, C. A.; Sakai-Tagawa, Y.; Muramoto, Y.; Ito, M.; Kiso, M.; Horimoto, T.; Shinya, K.; Sawada, T.; Kiso, M.; Usui, T.; Murata, T.; Lin, Y.; Hay, A.; Haire, L. F.; Stevens, D. J.; Russell, R. J.; Gamblin, S. J.; Skehel, J. J.; Kawaoka, Y. *Nature* **2006**, *444*, 378-82.

(5) Ha, Y.; Stevens, D. J.; Skehel, J. J.; Wiley, D. C. *Proc Natl Acad Sci U S A* **2001**, *98*, 11181-6.

(6) Holm, L.; Park, J. *Bioinformatics* **2000**, *16*, 566-7.

(7) Hess, B. *J. Chem. Theory Comput* **2008**, *4*, 116–122.

(8) Cline, M. S.; Karplus, K.; Lathrop, R. H.; Smith, T. F.; Rogers, R. G., Jr.; Haussler, D. *Proteins* **2002**, *49*, 7-14.

(9) Kasson, P. M.; Pande, V. S. *Pac Symp Biocomput* **2009**, 492-503.

(10) Witten, I. H.; Frank, E. *Data mining : practical machine learning tools and techniques*; 2nd ed.; Morgan Kaufman: Amsterdam ; Boston, MA, 2005.

(11) Dunn, S. D.; Wahl, L. M.; Gloor, G. B. *Bioinformatics* **2008**, *24*, 333-40.

(12) Fiser, A.; Sali, A. *Methods Enzymol* **2003**, *374*, 461-91.