Phase diagrams for biophysical fitness landscape design

Vaibhav Mohanty[1,2,3,*] and Eugene I. Shakhnovich[1,*]

[1]*Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138*
[2]*Harvard/MIT MD-PhD Program, Harvard Medical School, Boston, MA 02115 and Massachusetts Institute of Technology, Cambridge, MA 02139*
[3]*Program in Health Sciences and Technology, Harvard Medical School, Boston, MA 02115 and Massachusetts Institute of Technology, Cambridge, MA 02139*

Populations evolve toward the top of fitness landscapes in the way that physical particles experience forces toward the bottom of a potential energy landscape. We recently introduced fitness landscape design (FLD) as a means of customizing the shape of a protein's biophysical fitness landscape by using antibody sequences and concentrations as tunable control parameters. Until now, previous evidence for the feasibility of FLD has been numerical and simulated. Here, we derive an explicit analytical theory for the FLD phase diagram, providing tight bounds on the designability frontiers describing the extent to which fitnesses of different protein sequences can be tuned independently of each other. We then leverage a recently published experimental dataset of binding affinities between over 62,000 antibody variants and each of three glycoprotein antigens to construct experimental FLD phase diagrams, which show close agreement with theory. These results support the feasibility of engineering quantitatively programmable fitness landscapes for laboratory protein evolution experiments.

Introduction. Evolution of a population is often viewed as a semi-random "survival of the fittest" process in which natural selection contributes a driving force towards increased fitness over time, while genotype mutations facilitate exploration of the sequence space [1–4]. Mathematically, absolute fitness $F(\mathbf{s})$ is defined as the growth rate of a strain or species $\mathbf{s}$; i.e., $dn(\mathbf{s})/dt = F(\mathbf{s})n(\mathbf{s})$, where $n(\mathbf{s})$ is the number of individuals of type $\mathbf{s}$ [1]. First introduced by Sewall Wright more than 90 years ago [5], the mapping from genotype sequence $\mathbf{s}$ to fitness $F(\mathbf{s})$ is visually represented as a mountain-like structure termed a *fitness landscape*, which has served as a useful tool for understanding evolutionary trajectories. The structure of a fitness landscape can be affected by intrinsic factors, such as protein mutations which impact organism survivability, or interactions with the environment, such as protein-protein interactions between a cell surface protein and extracellular peptides [6–13].

We recently introduced *fitness landscape design* (FLD) [14], which is the task of molding a target protein's biophysical fitness landscape into a desired structure by adjusting the biophysical interactions between the target protein and a set of antibodies which bind to the target protein. Essentially, in FLD, antibodies are tunable control parameters which reshape a protein's biophysical fitness landscape according to a user-specified template landscape (Figure 1(a)). Some physically inspired studies have studied control of evolutionary dynamics purely from a dynamical systems perspective or by regulating drug dosage or chemical environments in real time [15–18], but FLD [14] is the first to suggest that the fitness landscape underlying evolution can itself be shaped into a quantitatively predictable desired structure. As an application, we showed that our FLD algorithms can discover cross-reactive antibodies which proactively suppress the fitness trajectories of viral escape mutants [14].

FLD relies on the idea that antibody-protein interactions have sufficient flexibility such that different target protein sequences can experience a wide range of possible fitnesses.

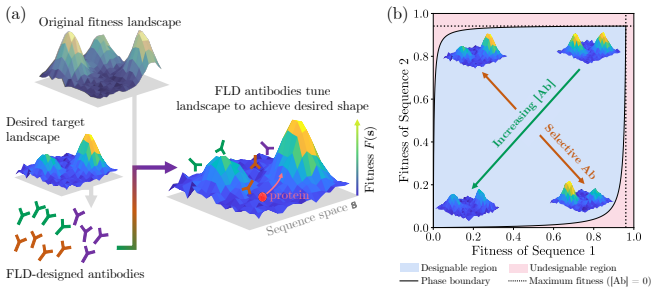

FIG. 1. (a) Schematic for fitness landscape design, which uses antibodies used as tunable control parameters to customize the biophysical fitness landscape of a target protein by suppressing different sequence fitnesses with user-defined penalties. (b) Schematic FLD phase diagram, showing how the fitnesses of two different sequences (tops of the two peaks) can be suppressed relative to one another.

For example, flooding a system with antibodies—even weakly binding ones—will generally lead to low fitnesses for all strains. Unsurprisingly, having zero antibodies in solution with the viral bulk culture will lead to the maximum possible fitneses for each strain. However, numerical evidence has shown that it is possible to discover selective antibodies which can suppress some fitnesses but not others, allowing for many degrees of freedom in tuning fitness landscape shapes by finding appropriate antibody sequences [14]. These possibilities are described by the schematic FLD phase diagram in Figure 1(b), which characterizes what fitness landscape assignments are "designable" or "undesignable."

In this work, we derive an explicit analytical theory for the shape of the FLD phase diagram, in agreement with previous numerical estimates [14]. We then leverage a recently published experimental dataset of over 62, 000 antibodies' binding affinities to each of three different influenza surface antigens to construct the first experimental FLD phase diagrams, demonstrating excellent agreement with theory and supporting the notion that FLD is experimentally realizable.

Analytical theory. We now develop an analytical theory for the boundary between the *designable* and *undesignable* regions of the FLD phase diagram. For viral surface protein evolution, biophysical properties—particularly, binding affinities—of viral-host cell binding and viral-antibody are quantitatively predictive of the viral strains' fitnesses. A statistical mechanics-based biophysical model of viral fitness which we showed can be analytically derived from chemical reaction kinetics between viral, host cell, and antibody proteins [14] aptly captures the empirical fitness data. The viral fitness can be analytically expressed as

$$F(\mathbf{s}) \approx \frac{k_{\text{rep}}(N_o - 1)N_{\text{ent}}[\text{H}]e^{-\beta\Delta G_{\text{H}}(\mathbf{s})}}{C_0 + [\text{H}]e^{-\beta\Delta G_{\text{H}}(\mathbf{s})} + \sum_n [\text{Ab}_n]e^{-\beta\Delta G_{\text{Ab}}^{(n)}(\mathbf{s})}}, \quad (1)$$

where k_{rep} is a virion's microscopic rate constant for cell entry and replication, N_o is the average number of offspring produced a single replication event, N_{ent} is the number of viral surface proteins which can potentially be used for cell entry, $[\text{H}]$ is the total concentration of host cell receptors, $[\text{Ab}_n]$ is the total concentration of the n-th type of antibody, $\Delta G_{\text{H}}(\mathbf{s})$ is the binding affinity between the viral entry protein and host cell receptor, $\Delta G_{\text{Ab}}^{(n)}(\mathbf{s})$ is the binding affinity between the entry protein of viral strain $\mathbf{s}$ and the n-th antibody, and β is the inverse temperature. This model has been extensively validated by simulations of viral serial dilution experiments [14, 19], by *in vitro* growth experiments in norovirus [20], and by epidemiological infectivity data for SARS-CoV-2 [21].

To make analytical calculations tractable, we work in the case where we have two target protein sequences ("sequence 1" and "sequence 2", denoted as $\mathbf{s}_1$ and $\mathbf{s}_2$) and only one antibody whose concentration $[\text{Ab}]$ and sequence serve as control parameters. We use rescaled fitnesses with constants $k_{\text{rep}}(N_o - 1)N_{\text{ent}} = 1$ for mathematical convenience. The biophysical fitness model simplifies to

$$F(\mathbf{s}) \approx \frac{[\text{H}]e^{-\beta\Delta G_{\text{H}}(\mathbf{s})}}{C_0 + [\text{H}]e^{-\beta\Delta G_{\text{H}}(\mathbf{s})} + [\text{Ab}]e^{-\beta\Delta G_{\text{Ab}}(\mathbf{s})}}, \quad (2)$$

for each of the two sequences.

To find the boundaries of the designable region, we ask what are the maximum and minimum fitnesses $F_2 \equiv F(\mathbf{s}_2)$ that sequence 2 could exhibit given a fixed fitness $F_1 \equiv F(\mathbf{s}_1)$ for sequence 1? To calculate the bounds on F_2 given F_1, we first rearrange eq. (2)

$$[\text{Ab}] = \frac{[\text{H}]e^{-\beta\Delta G_{\text{H}}(\mathbf{s})}}{e^{-\beta\Delta G_{\text{Ab}}(\mathbf{s})}}\left(\frac{1}{F(\mathbf{s})} - \frac{C_0}{[\text{H}]e^{-\beta\Delta G_{\text{H}}(\mathbf{s})}} - 1\right). \quad (3)$$

The above equation holds separately for sequences 1 and 2, which would both approximately experience the same antibody concentration in the limit where the antibody concentration great exceeds viral strain concentrations, which is a clinically relevant limit in viral infections such as SARS-CoV-2 in vaccinated individuals [14]. Equating the antibody concentrations gives us

$$\begin{aligned}&\frac{[\text{H}]e^{-\beta\Delta G_{\text{H}}(\mathbf{s}_1)}}{e^{-\beta\Delta G_{\text{Ab}}(\mathbf{s}_1)}}\left(\frac{1}{F_1} - \frac{C_0}{[\text{H}]e^{-\beta\Delta G_{\text{H}}(\mathbf{s}_1)}} - 1\right)\\ &= \frac{[\text{H}]e^{-\beta\Delta G_{\text{H}}(\mathbf{s}_2)}}{e^{-\beta\Delta G_{\text{Ab}}(\mathbf{s}_2)}}\left(\frac{1}{F_2} - \frac{C_0}{[\text{H}]e^{-\beta\Delta G_{\text{H}}(\mathbf{s}_2)}} - 1\right),\end{aligned} \quad (4)$$

and solving for F_2 in terms of F_1, we have

$$F_2(F_1) = \left[\frac{1}{F_2^{\text{free}}} + \rho e^{\beta\Delta}\left(\frac{1}{F_1} - \frac{1}{F_1^{\text{free}}}\right)\right]^{-1}, \quad (5)$$

where we have defined $F_i^{\text{free}} \equiv \frac{[\text{H}]e^{-\beta\Delta G_{\text{H}}(\mathbf{s}_i)}}{C_0+[\text{H}]e^{-\beta\Delta G_{\text{H}}(\mathbf{s}_i)}}$ as the fitness for the "free" protein when antibody concentration is zero, $\Delta \equiv \Delta G_{\text{Ab}}(\mathbf{s}_1) - \Delta G_{\text{Ab}}(\mathbf{s}_2)$ as the free energy difference between the two sequences' bound states to the antibody, and $\rho \equiv e^{-\beta(\Delta G_{\text{H}}(\mathbf{s}_1)-\Delta G_{\text{H}}(\mathbf{s}_2))}$, which is equivalent to the ratio of the two sequences' dissociation constants with the host protein.

Now, we note that host-viral binding affinities $\Delta G_{\text{H}}(\mathbf{s}_1)$ and $\Delta G_{\text{H}}(\mathbf{s}_2)$ are fixed (because we cannot manually mutate the host receptor protein), but we do have control over the choice of antibody, which is chosen from a library of M antibodies. We work within the approximation that the two antibody-antigen binding free energies $\Delta G_{\text{Ab}}(\mathbf{s}_1)$ and $\Delta G_{\text{Ab}}(\mathbf{s}_2)$ are chosen from a bi-variate Gaussian distribution:

$$\begin{pmatrix}\Delta G_{\text{Ab}}(\mathbf{s}_1)\\ \Delta G_{\text{Ab}}(\mathbf{s}_2)\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & r_{12}\sigma_1\sigma_2\\ r_{12}\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right), \quad (6)$$

where means μ_1 and μ_2, standard deviations σ_1 and σ_2, and the Pearson correlation $r_{12} \in [-1, 1]$ are taken to be be parameters fixed by the choice of the two target antigens $\mathbf{s}_1$ and $\mathbf{s}_2$. The antibody we choose is taken from a finite library of size M; for instance, if we consider all combinatorial amino acid variation at L paratope sites, $M = 20^L$. Like in Derrida's random energy model (REM) [22, 23], having a finite (even if a large) number of possible samples means that there will be a finite ground state energy. The frontiers of the designable phase will be determined by such extremal antibody samples from the library which maximize the magnitude $|\Delta| = |\Delta G_{\text{Ab}}(\mathbf{s}_1) - \Delta G_{\text{Ab}}(\mathbf{s}_2)|$, the difference between the two antibody-antigen binding free energies. Such antibodies are selective in that they tightly bind to one sequence but not the other, causing one sequence to have low fitness while the other is high.

From the Gaussian approximation, we know that Δ is also Gaussian distributed: $\Delta \sim \mathcal{N}(\mu_\Delta, \sigma_\Delta)$, where we define $\mu_\Delta \equiv \mu_1 - \mu_2$ and $\sigma_\Delta = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2 r_{12}}$. A standard result in classical extreme-value statistics [24] provides an approximation to the mean of the maximum (or minimum) value of Δ in the limit of large M samples:

$$\Delta_\pm \approx \mu_\Delta \pm \sigma_\Delta y_M, \quad (7)$$

where

$$y_M = \sqrt{2\log M} - \frac{\log(4\pi\log M) - 2\gamma}{2\sqrt{2\log M}} + O\left(\frac{1}{\log M}\right), \quad (8)$$

and γ is the Euler-Mascheroni constant. The leading order contribution to y_M is exactly the ground state energy of the REM [23], and the sub-leading order term which scales as $O(\log\log M/\sqrt{\log M})$ is related to finite-size corrections to the ground state energy of the REM [25]. Heuristically, we can understand that the Gaussian approximation employed here works because the antibody-antigen complex in the bound state can be treated as a heteropolymer in a folded state, with the variable monomers (or spins, in the REM analogy) consisting of L amino acids on the antibody surface. The equivalence of the spin-glass model of folded heteropolymers to the REM is well-known result in protein folding physics [26]. The antibody for which $\Delta = \Delta_-$ is the "ground state" antibody which selectively binds antigen sequence 1 more strongly than it binds to sequence 2, compared to any other antibody in the library. Conversely, the $\Delta = \Delta_+$ solution yields an antibody which selectively binds antigen sequence 2 more strongly than it binds to sequence 1, compared to any other antibody in the library.

We can now calculate sequence 2's maximum possible fitness $F_2^+(F_1)$ and minimum possible fitness $F_2^-(F_1)$ given the fitness of sequence 1, which will be obtained from the $\Delta = \Delta_-$ and $\Delta = \Delta_+$ antibodies, respectively

$$F_2^{\pm}(F_1) = \left[\frac{1}{F_2^{\text{free}}} + \rho e^{\beta\Delta_{\mp}}\left(\frac{1}{F_1} - \frac{1}{F_1^{\text{free}}}\right)\right]^{-1}. \tag{9}$$

These designability boundaries—a central result of this work—are valid on the sequence 1 fitness domain $F_1 \in (0, F_1^{\text{free}}]$, which corresponds to the sequence 2 fitness range $F_2 \in (0, F_2^{\text{free}}]$. The exact same boundaries could be obtained by repeating the derivation but instead maximizing or minimizing F_1 for fixed F_2, yielding relabeled but equivalent curves $F_1^+(F_2) = F_2^-(F_1)$ and $F_1^-(F_2) = F_2^+(F_1)$.

We plot theoretical phase diagrams using eq. (9) in Figure 2 across a range of antibody library sizes M and Pearson correlations r_{12}. The designable region (blue) is generally an asymmetric leaf-shaped region within the unit square whose sharp corners are placed at $(0,0)$ and $(F_1^{\text{free}}, F_2^{\text{free}})$. The two curves $F_2^+(F_1)$ and $F_2^-(F_1)$ (black solid lines) respectively outline the top and bottom borders of the designable region, and they converge at two points, representing limiting cases of the antibody concentration: in the limit of no antibodies ($[\text{Ab}] = 0$), both fitnesses adopt their "free" limits $F_1 = F_1^{\text{free}}$ and $F_2 = F_2^{\text{free}}$ (black dotted lines). Fitnesses above these "free" limits are unobtainable since antibodies do not enhance antigen-host binding. In the limit of high antibody concentration ($[\text{Ab}] \to \infty$), we see from the original fitness equation eq. (2) that both fitnesses $F_1, F_2 \to 0$. Outside of the designable region is the undesignable region (red), where fitness assignments are unreachable by any choice of antibody. Although we work in the single antibody limit, we can also obtain the estimated fitness curve (purple dashed line) for a large random ensemble of different antibodies with uniform concentrations by taking the annealed approximation $e^{\beta\Delta} \mapsto \mathbb{E}[e^{\beta\Delta}] = e^{\beta\mu_\Delta + \frac{1}{2}\beta\sigma_\Delta^2}$ within eq. (5).

We see in Figure 2 that the designable region expands with

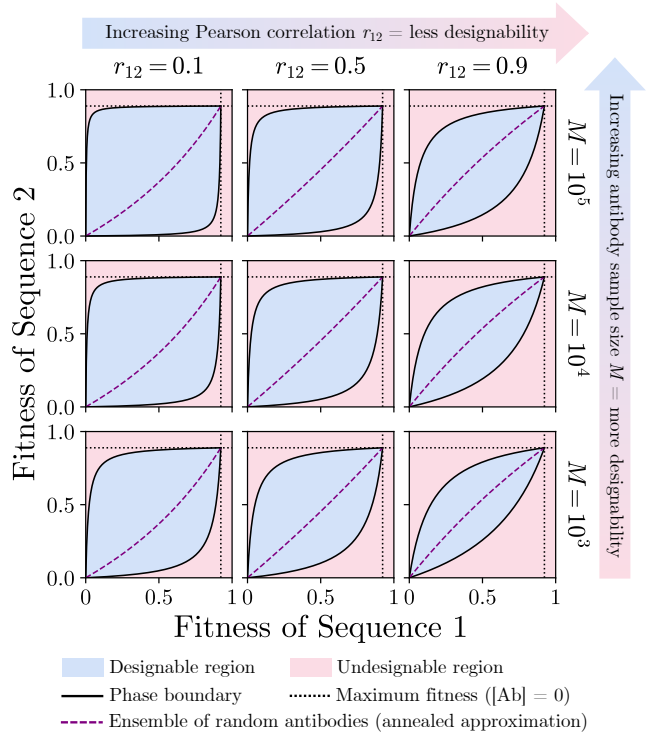

FIG. 2. Example theoretical FLD phase diagrams from eq. (9). The area of the designable region expands with decreasing Pearson correlation r_{12} and with increasing antibody library size M.

increasing antibody library size M and with decreasing Pearson correlation r_{12} between $\Delta G_{\text{Ab}}(\mathbf{s}_1)$ and $\Delta G_{\text{Ab}}(\mathbf{s}_2)$. Increasing M also increases the expected maximum/minimum of the free energy samples from the Gaussian distribution in eq. (6). The latter trend makes sense when we contextualize the biophysical meaning of r_{12}. Two sequences are likely to have high r_{12} when they are more chemically similar. For example, swapping a small nonpolar amino acid for another on the antigen is difficult for the antibody to chemically distinguish. The binding affinities are likely to be similar, so both sequences' fitnesses are likely to be similar, which decreases the area of the designable region. Conversely, swapping multiple nonpolar amino acids for charged amino acids can make r_{12} go to zero or even become negative. We now develop an analytical theory quantitatively explain this behavior of the designable region.

The area of the designable region relative to the (trivial) area of the unit square is called the *codesignability score* $\text{CDS}(\mathbf{s}_1, \mathbf{s}_2)$ for the two sequences, as defined and computed numerically in ref. [14]. High codesignability score indicates increased flexibility for fitness assignment: an antibody can be found such that sequence 1 fitness can be high while sequence 2 fitness is low—or vice versa—and it is possible to modulate antibody concentration so that the fitnesses are both high or both low. We previously showed that codesignability scores for all pairs of sequences in a larger set are useful for understanding what groups of sequences may have their fitnesses designed together, though all of our computations of codes-

ignability scores were numerical [14]. Here, we can compute codesignability score analytically.

By definition, codesignability score is the area of the designable region, which is computed by finding the area between the $F_2^+(F_1)$ and $F_2^-(F_1)$ curves:

$$\mathrm{CDS}(\mathbf{s}_1, \mathbf{s}_2) = \int_0^{F_1^{\mathrm{free}}} \mathrm{d}F_1 \left[F_2^+(F_1) - F_2^-(F_1)\right], \tag{10}$$

which has an exact integral solution:

$$\mathrm{CDS}(\mathbf{s}_1, \mathbf{s}_2) = I\left(\rho e^{\beta\Delta_-}\right) - I\left(\rho e^{\beta\Delta_+}\right), \tag{11}$$

where

$$I(x) \equiv \frac{F_1^{\mathrm{free}} F_2^{\mathrm{free}}}{\left(1 - \frac{xF_2^{\mathrm{free}}}{F_1^{\mathrm{free}}}\right)^2} \left[1 - \frac{xF_2^{\mathrm{free}}}{F_1^{\mathrm{free}}}\left(1 - \log\frac{xF_2^{\mathrm{free}}}{F_1^{\mathrm{free}}}\right)\right]. \tag{12}$$

We prove analytically in the End Matter that $\partial\mathrm{CDS}/\partial r_{12} < 0$ for any valid r_{12}, so codesignability monotonically decreases with Pearson correlation r_{12}, suggesting that chemically similar protein antigens are harder to assign differing fitnesses.

Experimental designability phase diagrams for influenza antigens. The phase diagrams from theory qualitatively match the results from our original numerical FLD paper [14], where we used force field-based binding energies and support vector machines to construct simulated phase diagrams. We now show that the analytical theory also agrees with *in vitro* experimental results.

Ref. [27] reports experimental *in vitro* binding affinity measurements between $M = 2^{16} = 65,536$ mutational variants of the human CH65 antibody and each of three mutational variants of the influenza H1 glycoprotein antigen. They chose $L = 16$ amino acid sites which varied between the unmutated common ancestral antibody and the affinity-matured CH65 antibody and performed all combinatorial mutations between those two antibody types, meaning two possible amino acids at each of the 16 sites were tested. Their binding affinity measurements, obtained using the Tite-Seq teqchnique [28], are reported as log-dissociation constants $K_d(\mathbf{s}) \equiv e^{\beta\Delta G_{\mathrm{Ab}}(\mathbf{s})}$, where $\mathbf{s}$ is the antigen sequence label referring to one of three influenza strains: A/Massachusetts/1/1990 (MA90), A/Solomon Islands/3/2006 (SI06), and A/Massachusetts/1/1990 with the G189E mutation (G189E). Non-binding sequences were assigned the titration boundary value $-\log K_d = 6$. After filtering out poor quality measurements, their published dataset contained collectively over 193,000 binding affinities across the three antigens and all antibodies.

Here, we use ref. [27]'s rich dataset of reported binding affinities to construct the first-ever experimental FLD phase diagrams for each pair of influenza antigens, allowing direct comparison to our analytical calculations. Since host-antigen binding affinity measurements are unavailable, we work within the approximations that the three strains roughly equally bind the host receptor ($\rho \approx 1$) and that host concentrations are large ($[\mathrm{H}] \gg C_0$). This approximation yields $F_1^{\mathrm{free}} = F_2^{\mathrm{free}} \approx 1$,

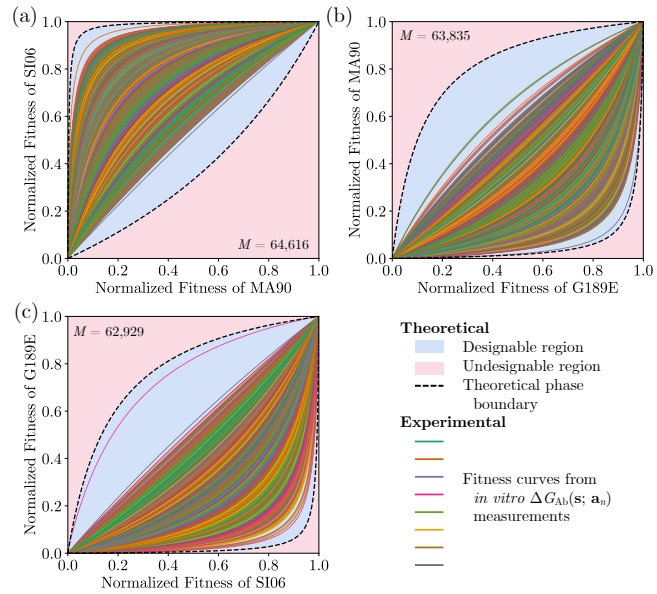

FIG. 3. Experimental FLD phase diagrams from binding affinity measurements of over 62,000 antibodies with three influenza glycoprotein antigens, and theoretical phase boundaries (black, dashed) from the Gaussian approximation eq. (9). Empirical measurements are shown as solid, colorful lines.

rescaling the rectangular region bounded by the dotted lines in Figure 2 to the unit square, which effectively renormalizes the fitnesses of the two antigens to the range $(0, 1]$. For each tested antibody, and for each pair of influenza antigens, we use the experimental binding affinity measurements from ref. [27] to calculate Δ and substitute it into eq. (5), which is plotted on the unit square. Theoretical phase diagram boundaries are calculated by computing the means, standard deviations, and Pearson correlations of the two-antigen $(\Delta G_{\mathrm{Ab}}(\mathbf{s}_1), \Delta G_{\mathrm{Ab}}(\mathbf{s}_2))$ joint distributions and substituting into eq. (9).

These experimental results and theoretical boundaries are plotted in Figure 3. The solid lines of various colors are the transformed experimental data, while the black dashed lines indicate the theoretical boundaries separating the predicted designable (blue) and undesignable (red) regions. Figure 3(a) shows that the fitness of SI06 is generally higher than the fitness of MA90, with the designable region spanning most of the upper triangle above the central diagonal of the unit square. Figure 3(b) and (c) have their designable regions primarily spanning the lower triangle of the unit square, with SI06 generally having equal or higher fitness than G189E, which tends to have equal or higher fitness than MA90. Notably, each of the three pairs of antigens showcase one or more prominent "outlier" antibodies: orange top line in Figure 3(a), green and brown top lines as well as gray bottom line in Figure 3(b), and the pink top line in Figure 3(b). These antibodies push the designability frontier by being more selective for the antigen which generally binds more weakly to the other antibody variants. A notable feature of all three figure panels is the tightness of the theoretical bounds from eq. (9) near these outlier antibodies.

The lower bound in Figure 3(a) and the upper bound in Figure 3(b) appear overly optimistic. This is likely due to the non-Gaussian distribution of the log dissociation constants in ref. [27]'s dataset. One contributing factor to non-Gaussianity is the assignment of $-\log K_d = 6$ to all antibody-antigen pairs which bind too weakly to be measured using Tite-Seq [27]. Another likely factor is the effect of evolutionary selection on the available antigens and antibodies. MA90 circulated in 1990 while SI06 circulated in 2006; its fitness may be expected to be higher than MA90 for most antibodies in the repertoire because it may indeed have evolved to escape related antibodies [27]. Similarly, G189E was specifically evolved in the laboratory to escape the universal common ancestral antibody, so its affinity to most of the antibodies in the repertoire are expected to be lower [27]. The extreme value statistics/REM results used in our theory rely on the $\Delta G_{\mathrm{Ab}}(\mathbf{s})$ samples being unbiased, i.i.d. draws from the Gaussian distribution. Since both evolutionary selection and experimental choice of amino acids played a role in the construction of the antibody library, ref. [27]'s dataset is not expected to be Gaussian.

In the End Matter, we address non-Gaussianity of the free energy difference $\Delta = \Delta G_{\mathrm{Ab}}(\mathbf{s}_1) - \Delta G_{\mathrm{Ab}}(\mathbf{s}_2)$ distribution by capturing tail behaviors with generalized Pareto distributions, since only tails behaviors impact extreme value results with finite samples. We subsequently show that this tail-fitting approach enables *scalability* of phase boundary prediction from a smaller set of experimental affinity measurements. End Matter Figure 4 shows that high-fidelity phase diagrams estimated for antibody libraries of size M can be obtained from only a training set of size $10\% \times M$.

Discussion. In summary, we have introduced analytical foundations for biophysical FLD to customize the evolutionary fitnesses of different protein sequences, grounding our numerical work [14] in a theoretical basis. By working in the two-sequence, one-antibody case, we have found analytically tractable closed-form expressions for the boundaries of the designable region of the designability phase diagrams. We quantitatively justified the intuition that the area of the designable region—the codesignability score—should increase when the antibody-antigen binding free energy distributions for different antigens are decorrelated. Lastly, we validated the theoretical bounds on the designability phase diagrams by employing experimental *in vitro* binding affinities for over 62, 000 antibodies across three influenza antigens. Our theoretical bounds demonstrated successful ability to capture outlier antibodies which are responsible for expanding the designability frontier. We also showed that the theoretical bounds could be improved upon by fitting the tails of the free energy difference distributions directly, even when the antibody dataset was downsampled by 90%.

FLD is now building a solid experimental foundation, though there are current limitations and many directions for future efforts. In recent years, biophysical fitness formulas of the form in eq. (1) used as the basis of FLD have been validated *in vitro* [20] and with epidemiological data [21]. Now, the present work uses experimental binding affinity measurements [27] to support the notion that the space of all possible fitness landscapes separates into designable and undesignable regions, enabling the design of custom fitness landscapes for protein evolution—an idea we first proposed in ref. [14]. Although experimental binding affinities are substituted into an experimentally validated fitness formula, a limitation of this study is that true end-to-end experimental validation of FLD should involve performing direct fitness measurements of different strains in bulk culture in the presence of FLD-designed antibodies. This is a focus of our ongoing work.

Biophysical FLD opens the door not only to improved pandemic preparedness and biosecurity via proactive vaccine design [14] by discovering proactive antibodies that simultaneously inhibit wildtype and escape antigens, but also to possible cancer therapeutics such as engineered CAR-T cells or small peptides which may prevent immune escape of cancer cells. The widespread use of phages for directed evolution [29–31] is also an active target for applications of FLD; we suggest that FLD principles may be used for building experimental toolkits to engineer custom fitness landscape environments for experimental evolution and fundamental population genetics research.

Acknowledgments This work was supported by a Hertz Foundation Fellowship (to V.M.) and by award T32GM14427 from the National Institute of General Medical Sciences (to Harvard/MIT MD-PhD Program). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health. The authors declare no conflict of interest.

* Correspondence: mohanty@hms.harvard.edu (V.M.) and shakhnovich@chemistry.harvard.edu (E.I.S.)

[1] J. F. Crow and M. Kimura, *An Introduction to Population Genetics Theory* (Harper and Row, New York, 1970).

[2] D. L. Hartl, *Principles of population genetics* (Sinauer Associates, Sunderland, Mass, 1980).

[3] J. H. Gillespie, en*Population Genetics: A Concise Guide* (JHU Press, 2004).

[4] W. J. Ewens, *Mathematical population genetics: I: Theoretical introduction*, second edition ed., Interdisciplinary Applied Mathematics No. Volume 27 (Springer, New York, 2004).

[5] S. Wright, The Roles of Mutation, Inbreeding, crossbreeding and Selection in Evolution, Proceedings of the XI International Congress of Genetics **8**, 209 (1932).

[6] E. I. Shakhnovich and A. M. Gutin, Implications of thermodynamics of protein folding for evolution of primary sequences, Nature **346**, 773 (1990).

[7] K. B. Zeldovich, P. Chen, and E. I. Shakhnovich, Protein stability imposes limits on organism complexity and speed of molecular evolution, Proceedings of the National Academy of Sciences of the United States of America **104**, 16152 (2007).

[8] C. S. Wylie and E. I. Shakhnovich, A biophysical protein folding model accounts for most mutational fitness effects in viruses, Proceedings of the National Academy of Sciences **108**, 9916 (2011).

[9] A. E. Lobkovsky, Y. I. Wolf, and E. V. Koonin, Predictability of Evolutionary Trajectories in Fitness Landscapes, PLoS Computational Biology **7**, e1002302 (2011).

[10] A. Serohijos, Z. Rimas, and E. Shakhnovich, enProtein Biophysics Explains Why Highly Abundant Proteins Evolve Slowly, Cell Reports **2**, 249 (2012).

[11] T. A. Hopf, C. P. I. Schärfe, J. P. G. L. M. Rodrigues, A. G. Green, O. Kohlbacher, C. Sander, A. M. J. J. Bonvin, and D. S. Marks, Sequence co-evolution gives 3D contacts and structures of protein complexes, eLife **3**, e03430 (2014).

[12] T. Sikosek and H. S. Chan, Biophysics of protein evolution and evolutionary protein biophysics, Journal of The Royal Society Interface **11**, 20140419 (2014).

[13] C. Norn and I. André, Atomistic simulation of protein evolution reveals sequence covariation and time-dependent fluctuations of site-specific substitution rates, PLOS Computational Biology **19**, e1010262 (2023).

[14] V. Mohanty and E. I. Shakhnovich, Biophysical fitness landscape design traps viral evolution (2025), bioRxiv: 2025.03.30.646233.

[15] A. Fischer, I. Vázquez-García, and V. Mustonen, The value of monitoring to control evolving populations, Proceedings of the National Academy of Sciences **112**, 1007 (2015).

[16] M. Lässig and V. Mustonen, Eco-evolutionary control of pathogens, Proceedings of the National Academy of Sciences **117**, 19694 (2020).

[17] S. Iram, E. Dolson, J. Chiel, J. Pelesko, N. Krishnan, Güngör, B. Kuznets-Speck, S. Deffner, E. Ilker, J. G. Scott, and M. Hinczewski, Controlling the speed and trajectory of evolution with counterdiabatic driving, Nature Physics **17**, 135 (2021).

[18] L. Milocco and T. Uller, Utilizing developmental dynamics for evolutionary prediction and control, Proceedings of the National Academy of Sciences **121**, e2320413121 (2024).

[19] N. Chéron, A. W. Serohijos, J.-M. Choi, and E. I. Shakhnovich, Evolutionary dynamics of viral escape under antibodies stress: A biophysical model, Protein Science **25**, 1332 (2016).

[20] A. Rotem, A. W. R. Serohijos, C. B. Chang, J. T. Wolfe, A. E. Fischer, T. S. Mehoke, H. Zhang, Y. Tao, W. Lloyd Ung, J.-M. Choi, J. V. Rodrigues, A. O. Kolawole, S. A. Koehler, S. Wu, P. M. Thielen, N. Cui, P. A. Demirev, N. S. Giacobbi, T. R. Julian, K. Schwab, J. S. Lin, T. J. Smith, J. M. Pipas, C. E. Wobus, A. B. Feldman, D. A. Weitz, and E. I. Shakhnovich, enEvolution on the Biophysical Fitness Landscape of an RNA Virus, Molecular Biology and Evolution **35**, 2390 (2018).

[21] D. Wang, M. Huot, V. Mohanty, and E. I. Shakhnovich, enBiophysical principles predict fitness of SARS-CoV-2 variants, Proceedings of the National Academy of Sciences **121**, e2314518121 (2024).

[22] B. Derrida, Random-Energy Model: Limit of a Family of Disordered Models, Physical Review Letters **45**, 79 (1980).

[23] B. Derrida, Random-energy model: An exactly solvable model of disordered systems, Physical Review B **24**, 2613 (1981).

[24] H. Cramér, *Mathematical Methods of Statistics* (Princeton University Press, 1946).

[25] B. Derrida and P. Mottishaw, Finite size corrections in the random energy model and the replica approach, Journal of Statistical Mechanics: Theory and Experiment **2015**, P01021 (2015).

[26] E. I. Shakhnovich and A. M. Gutin, Formation of unique structure in polypeptide chains: Theoretical investigation with the aid of a replica approach, Biophysical Chemistry **34**, 187 (1989).

[27] A. M. Phillips, D. P. Maurer, C. Brooks, T. Dupic, A. G. Schmidt, and M. M. Desai, Hierarchical sequence-affinity landscapes shape the evolution of breadth in an anti-influenza receptor binding site antibody, eLife **12**, e83628 (2023).

[28] R. M. Adams, T. Mora, A. M. Walczak, and J. B. Kinney, Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves, eLife **5**, e23156 (2016).

[29] H. Pedersen, S. Hölder, D. P. Sutherlin, U. Schwitter, D. S. King, and P. G. Schultz, A method for directed evolution and functional cloning of enzymes, Proceedings of the National Academy of Sciences **95**, 10523 (1998).

[30] A. Fernandez-Gacio, M. Uguen, and J. Fastrez, Phage display as a tool for the directed evolution of enzymes, Trends in Biotechnology **21**, 408 (2003).

[31] K. M. Esvelt, J. C. Carlson, and D. R. Liu, A System for the Continuous Directed Evolution of Biomolecules, Nature **472**, 499 (2011).

[32] S. Coles, en*An Introduction to Statistical Modeling of Extreme Values*, Springer Series in Statistics (Springer London, London, 2001).

[33] L. de Haan and A. Ferreira, en*Extreme value theory: an introduction*, Springer series in operations research (Springer, New York ; London, 2006).

[34] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, and P. van Mulbregt, enSciPy 1.0: fundamental algorithms for scientific computing in Python, Nature Methods **17**, 261 (2020), publisher: Nature Publishing Group.

END MATTER

Proof that codesignability score monotonically decreases with Pearson correlation r_{12}

Starting from eq. (10), we note

$$\frac{\partial \mathrm{CDS}}{\partial \sigma_\Delta} = \int_0^{F_1^{\mathrm{free}}} \mathrm{d}F_1 \left[\frac{\partial F_2^+(F_1)}{\partial \sigma_\Delta} - \frac{\partial F_2^-(F_1)}{\partial \sigma_\Delta} \right].$$
$$= \int_0^{F_1^{\mathrm{free}}} \mathrm{d}F_1 \left(\frac{\beta y_M}{F_1} - \frac{\beta y_M}{F_1^{\mathrm{free}}} \right) \left[\frac{e^{\beta\Delta_-}}{F_2^+(F_1)^2} + \frac{e^{\beta\Delta_+}}{F_2^-(F_1)^2} \right]. \tag{13}$$

Since $0 < F_1 \leq F_1^{\mathrm{free}}$, the term within the parentheses is positive. Since exponentials in the numerators and quadratic terms in the denominators are also positive, the entire integrand is positive, which means $\partial \mathrm{CDS}/\partial \sigma_\Delta > 0$. Noting that $\partial \sigma_\Delta / \partial r_{12} = -\sigma_1 \sigma_2 / \sigma_\Delta < 0$, it follows from chain rule that

$$\frac{\partial \mathrm{CDS}}{\partial r_{12}} = \frac{\partial \mathrm{CDS}}{\partial \sigma_\Delta} \frac{\partial \sigma_\Delta}{\partial r_{12}} < 0, \tag{14}$$

which completes the proof.

Experimental scalability of phase diagram prediction with tail-fitting and extreme value theory

To address non-Gaussianity, we also construct semi-empirical phase diagrams by fitting the tails of the distribution for the free energy difference, $\Delta = \Delta G_{\mathrm{Ab}}(\mathbf{s}_1) - \Delta G_{\mathrm{Ab}}(\mathbf{s}_2)$, for each pair of antibodies. In extreme value theory, only the tails of the distribution and the sample size (i.e. antibody library size M) affect the expected extrema. The theoretical phase diagrams from the Gaussian distribution are constructed by fitting only two parameters: the mean μ_Δ and standard deviation σ_Δ. In Supplementary Appendix A, we show that by fitting each upper and lower tail using the generalized Pareto distribution (GPD) with three parameters per tail—still small compared to the number of samples—we can capture the empirically observed phase boundaries nearly perfectly (Supplementary Appendix A; Figure S4). This is expected, since the GPD fits are influenced by extremal values in the data distribution.

However, we find remarkably that even after massively downsampling the antibody dataset by $\sim 90\%$, fitting the tails to the subsampled distribution are still predictive of the phase boundaries for the full antibody dataset containing over 62,000 antibodies, as shown in Figure 4 (see Supplementary Appendix A for step-by-step construction). We downsampled the complete antibody dataset to 10% of its original size in 10,000 independent trials and computed the phase boundaries. Figure 4 shows 25th, 50th (median), and 75th percentiles for these phase boundaries aggregated over the 10,000 trials. All predicted median phase boundaries (Figure 4: black, dashed lines) are tighter than the ones from the Gaussian approximation in Figure 3, despite having been constructed from roughly 10% of the original data. However, any extreme value prediction which relies on a smaller set of finite samples is less likely to capture outliers which may strongly impact the GPD fits, as seen in Figure 4(c), where the pink antibody outlier is not captured by the phase boundaries from the downsampled fits. The quartile phase boundaries (Figure 4: gray, dotted lines) show that some boundary predictions are better than others, depending on the presence/absence of outliers drawn from the antibody library (see Figure S7 for complete distributions of expected extrema from the 10,000 trials). Using larger antibody libraries for tail fitting fixes these discrepancies, as mentioned previously (Supplementary Appendix A; Figure S4). Overall, the downsampling approach shows promise for laboratory scalability of FLD phase diagram construction by reducing the number of antibodies needed to calculate boundaries, provided that tails are captured aptly.

We also emphasize that the phase diagrams computed here and in the main text are limited in that completely random antibody sequence sampling was not conducted, and only around $2^{16} \sim 6.6 \times 10^4$ out of $20^{16} \sim 6.6 \times 10^{20}$ possible antibody mutations at the $L = 16$ sites were explored, so broader amino acid variation as well as expansion of the paratope sites may yield expanded designable (blue) regions. Also, we add that eq. (1) suggests that apparent "gaps" between the outlier antibodies and the bulk of the distribution (e.g. between the pink outlier and bulk in Figure 4(c)) may be explored by using combinations of antibodies at different concentrations—an idea supported by our random antibody ensemble sampling results in the original FLD paper [14].

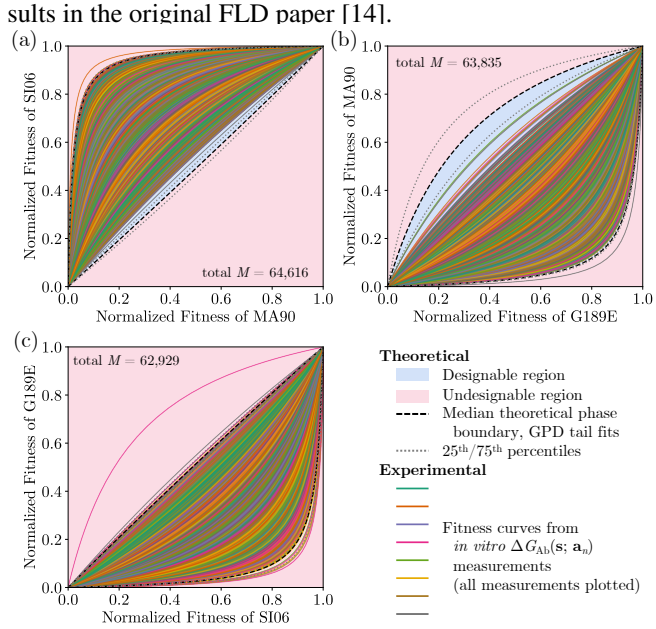

FIG. 4. Experimental FLD phase diagrams from binding affinity measurements of over 62,000 antibodies with three influenza glycoprotein antigens, and semi-empiricial phase boundaries from fitting the tails of downsampled antibody datasets which are approximately 10% of the size of the original dataset. Empirical measurements are shown as solid, colorful lines. Estimated median phase boundaries from 10,000 independent downsampling trials are shown as black, dashed lines, and 25th/75th percentile phase boundaries are shown as gray, dotted lines.

Supplementary Information for "Phase diagrams for biophysical fitness landscape design"

Vaibhav Mohanty[1,2,3,*] and Eugene I. Shakhnovich[1,*]
[1]*Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138*
[2]*Harvard/MIT MD-PhD Program, Harvard Medical School, Boston, MA 02115 and Massachusetts Institute of Technology, Cambridge, MA 02139*
[3]*Program in Health Sciences and Technology, Harvard Medical School, Boston, MA 02115 and Massachusetts Institute of Technology, Cambridge, MA 02139*
[*]Correspondence: mohanty@hms.harvard.edu (V.M.); shakhnovich@chemistry.harvard.edu (E.I.S.).

APPENDIX A: CONSTRUCTING SEMI-EMPIRICAL PHASE DIAGRAMS FROM TAIL FITTING

Here, we construct semi-empirical FLD phase diagrams for the three pairs of influenza antigens without using the Gaussian approximation. Instead, we fit the tails of the distribution of free energy differences in order to obtain more accurate boundary values from extreme value statistics. Finally, we show that downsampling the antibody dataset by 90% still yields tail fits which are predictive of the phase boundaries for the full antibody dataset.

Binding affinity distributions

First, in Figure S1 we reproduce the $\log K_d$ distributions (equivalent to the ΔG distributions) for the three influenza antigens, where each datapoint represents one of up to up to 65, 535 antibodies whose binding affinities to each antigen were measured by [27]. The distributions are directly reproduced from the dataset; not all antibodies have binding affinites reported (and are stored as NaN values), which is why each antigen has a different number of antibody datapoints. Notably, the Tite-seq technique [28] used by [27] was limited to $\log K_d$ values stronger (i.e. more negative) than -6, so antibodies which bound more weakly (i.e. $\log K_d \geq -6$) were automatically assigned a value of -6, resulting in large spikes in two of the antigens' binding affinity distributions (Figure S1(a-b)).

Binding affinity difference distributions

Next, we compute the differences between each antibody's $\log K_d$ for pairs of antigens:

$$\log K_{d,1} - \log K_{d,2} = \beta[\Delta G(\mathbf{s}_1) - \Delta G(\mathbf{s}_2)]. \tag{15}$$

These difference distributions are plotted in Figure S2. As can be readily seen in the histograms, these distributions are non-Gaussian, so the approximation applied in the main text is not expected to hold necessarily. To obtain better estimates of phase diagram boundaries (relative to those obtained from the Gaussian approximation), we only need to capture the tail behaviors for each distribution.

Extreme value theory for expected bound on distribution tails with finite samples

In extreme value statistics, the generalized Pareto distribution (GPD) is commonly used for this tail fitting task [32, 33], and it enables an analytically tractable estimate for the expectation of the tail's maximum for a certain number of samples. The GPD's probability density function is given by

$$p(x;\xi,\mu,\sigma) = \frac{1}{\sigma}\left(1 + \frac{\xi(x-\mu)}{\sigma}\right)^{-(1+1/\xi)}, \tag{16}$$

where μ is called the location parameter (*not* necessarily equal to the distribution's mean), σ is the scale parameter, and ξ is the shape parameter. Following standard extreme value theory [32, 33], we first construct a cumulative distribution function for a random variable $X = \log K_{d,1} - \log K_{d,2}$ which represents a datapoint drawn from the tail of the binding affinity difference distribution modeled by the GPD. The location parameter is set to $\mu = 0$ because the distribution will be shifted through the

introduction of a threshold. As is commonly done in extreme value theory, we shift the data distribution so that the tail begins at a threshold value u, which is moved to the origin. So, we define $Y = X - u$, and the CDF can be written

$$\mathbb{P}(Y \leq y \,|\, Y > 0) = \int_0^y \mathrm{d}y' \, \frac{1}{\sigma} \left(1 + \frac{\xi y'}{\sigma}\right)^{-(1+1/\xi)} = 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi}. \tag{17}$$

Undoing the affine shift, we can write

$$\mathbb{P}(X - u \leq x \,|\, X > u) = 1 - \left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-1/\xi}. \tag{18}$$

The probability of a datapoint being greater than a certain value x *within the tail* is given by the product of probabilities of drawing a datapoint from the tail the conditional probability above:

$$\mathbb{P}(X > x) = \mathbb{P}(X - u > x \,|\, X > u)\mathbb{P}(X > u) = q\left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-1/\xi}, \tag{19}$$

where we let $q \equiv \mathbb{P}(X > u)$ be a pre-defined threshold quantile to which fixes the start of the tail.

Now, we define

$$X_{\max} \equiv \max\{X_1, \ldots, X_M\} \tag{20}$$

to be a random variable representing the maximum of M i.i.d. datapoints. The maximum will be less than some value x only if all drawn variables are also less than x:

$$\begin{aligned} \mathbb{P}(X_{\max} \leq x) &= \prod_{i=1}^{M} \mathbb{P}(X_i \leq x) = [\mathbb{P}(X \leq x)]^M \\ &= [1 - \mathbb{P}(X_{\max} > x)]^M \\ &= \left[1 - q\left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-1/\xi}\right]^M. \end{aligned} \tag{21}$$

In the large M (many samples) limit, we take the typical Poisson limit commonly invoked for independent events:

$$\mathbb{P}(X_{\max} \leq x) = \left[1 - q\left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-1/\xi}\right]^M \approx \exp\left[-Mq\left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-1/\xi}\right]. \tag{22}$$

The CDF above defines a generalized extreme value (GEV) distribution from which we can quickly compute the expectation of the extremal value by integrating within the tail:

$$\mathbb{E}[X_{\max}] = \int_u^\infty \mathrm{d}x \, x \mathbb{P}(X_{\max} = x). \tag{23}$$

The PDF is computed by taking a derivative of the CDF with respect to x:

$$\mathbb{P}(X_{\max} = x) = \frac{1}{\sigma} Mq\left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-1-\frac{1}{\xi}} \exp\left[-Mq\left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-1/\xi}\right]. \tag{24}$$

Performing the substitution

$$t = Mq\left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-1/\xi} \tag{25}$$

also gives us

$$x = u + \frac{\sigma}{\xi}\left((Mq)^{\xi} t^{-\xi} - 1\right), \tag{26}$$

and

$$\mathrm{d}x = -\sigma(Mq)^{\xi} t^{-\xi-1}\ \mathrm{d}t\,. \tag{27}$$

Now, the integral becomes

$$\begin{aligned}\mathbb{E}[X_{\max}] &= \int_{t=Mq}^{0} \mathrm{d}t\ \frac{-\sigma(Mq)^{\xi} t^{-\xi-1}}{\sigma(Mq)^{\xi}} t^{\xi+1} \left[u + \frac{\sigma}{\xi}\left((Mq)^{\xi} t^{-\xi} - 1\right)\right] e^{-t} \\ &= \int_{0}^{Mq} \mathrm{d}t \left[u + \frac{\sigma}{\xi}\left((Mq)^{\xi} t^{-\xi} - 1\right)\right] e^{-t}.\end{aligned} \tag{28}$$

In the limit of large M (which we used in the Poisson limit earlier), we take $Mq \to \infty$, so

$$\begin{aligned}\mathbb{E}[X_{\max}] &\approx \int_{0}^{\infty} \mathrm{d}t\, u e^{-t} + \frac{\sigma}{\xi}\int_{0}^{\infty} \mathrm{d}t \left((Mq)^{\xi} t^{-\xi} - 1\right) e^{-t} \\ &= u + \frac{\sigma}{\xi}\left[-1 + (Mq)^{\xi}\int_{0}^{\infty} \mathrm{d}t\, t^{-\xi} e^{-t}\right] \\ &= u + \frac{\sigma}{\xi}\left[-1 + (Mq)^{\xi}\Gamma(1-\xi)\right],\end{aligned} \tag{29}$$

where we have used the Gamma function Γ and have assumed $\xi < 1$, which we will find later holds for all of the datasets studied here. By fitting the GPD PDF (eq. (16)) with $\mu = 0$ to the tails of our $\log K_d$ difference distributions, we can obtain estimates for σ and ξ and then substitute them into eq. (29) in order to predict the expected bound on the data distribution for finite samples M. For an antibody library of size M, this is perhaps tautological, but below we will also show that we can downsample the antibody library to size $0.10 \times M$ and still make predictions on the boundaries for the full library of size M.

Fitting tails of binding affinity difference distributions

We now return to our empirical study of the binding affinity distributions. Although various methods exist for deciding where tail thresholds should be optimally placed [32, 33], we choose threshold quantiles arbitrarily to be the extremal 5% of the data distribution on either tail, *except* for the lower tail of the $\log K_{d,\mathrm{SI06}} - \log K_{d,G189E}$ distribution, for which we choose the bottom 20% in order to capture the large spike which is caused by the Tite-seq detection boundary.

After truncation based on the quantile thresholds described above, we reflect and/or affine shift the tails so that they are all in the non-negative real domain. Then, we fit the GPD to the tails using maximum likelihood estimation, as pre-implemented using SciPy's `scipy.stats.genpareto.fit` [34]. From the estimated $\hat{\xi}$ and $\hat{\sigma}$ values, we calculated the expected bounds on the tails using eq. (29). In Figure S3, we show the tails with the GPD fits, the estimated bounds from extreme value theory, and the actual experimentally observed bounds; the tails have been re-reflected and re-shifted to their original domains and orientations.

Semi-empirical phase diagrams without the Gaussian approximation: tail fitting

Substituting the estimated bounds from eq. (29) into main text eq. (9), we are able to construct semi-empirical phase diagrams (Figure S4) as an alternative to the phase diagrams computed from the Gaussian approximation in main text Figure 3. Compared to the Gaussian approximation, capturing the tail behaviors accurately yield excellent theoretical estimates. This is to be expected since we are essentially using the extremal data values from the empirical distribution to estimate the boundary, albeit somewhat indirectly through fitting ξ and σ.

A more powerful consequence of this tail fitting-based approach, however, is that *experimental scalability* is enabled, which we subsequently investigate by downsampling the antibody library.

Scalability of phase boundary inference: downsampling the antibody library

We now downsample the antibody library to roughly 10% of its original size. We randomly choose 10% of the 65,535 indices; however, some antibodies have NaN reported as their binding affinity to particular antigens. So, the exact number of antibodies comprising the downsampled binding affinity difference distributions (Figure S5) is not exactly 10% of the full distributions from

Figure S2. In general, the sample size of 10% captures the general shape of the histogram of the full distribution, but of course outliers are less likely to be sampled.

We then isolate tails using the same quantile cutoffs used for the full distribution, reflect and/or affine shift, and fit the GPD. Then, we calculate the expected phase diagram boundary using eq. (29), but with M set to the *full* antibody dataset size (i.e. between $62,000$ and $65,000$, depending on how many NaN values were present). This means we are using the fitted parameters from the downsampled dataset but are *predicting* the phase boundaries for the full datset. Example tail fits from a single trial of the sampling procedure are shown in Figure S6. The downsampling procedure is conducted for $10,000$ trials, and the distributions for the expected extrema $\mathbb{E}[\min(\log K_{d,1} - \log K_{d,2})]$ (for lower tails) or $\mathbb{E}[\max(\log K_{d,1} - \log K_{d,2})]$ (for upper tails), constructed from those 10,000 trials, are shown in Figure S7, along with quartile boundaries, means, and the true extrema for comparison. The phase diagrams arising from the downsampling procedure are shown and discussed in main text Figure 4, with quartile boundaries plotted as well. The ability to use a smaller antibody sample to predict the phase boundaries for the full antibody library demonstrates the scalability of the experimental approach for elucidating FLD phase diagrams.

However, it should be noted that the GPD fits and predicted maxima are heavily dependent on outlier values. As we see in the main text, sometimes the phase boundary is underestimated if outliers are not captured in a particular sample. This is naturally always a limitation of extreme value prediction using smaller samples.

Supplementary Figures

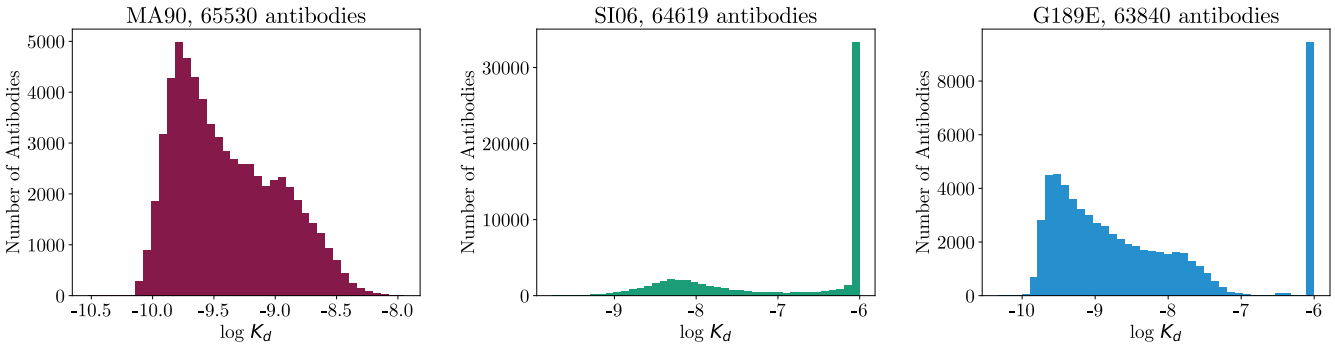

FIG. S1. Distribution of log dissociation constants between over 62,000 antibodies and each of three influenza antigens, directly reproduced from the public dataset from [27]. Spikes in the distribution are due to weak binders whose $\log K_d \geq -6$ being binned to a value of -6.

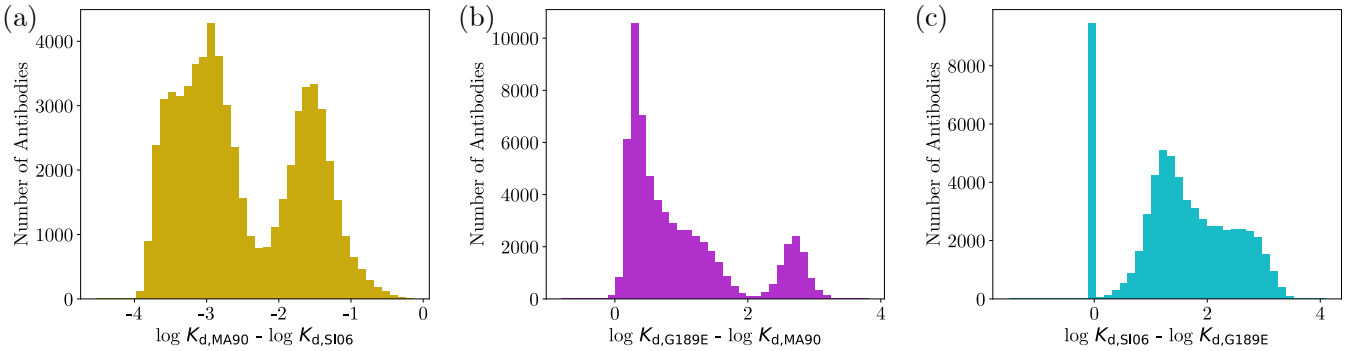

FIG. S2. Distributions of the difference in binding affinity each antibody has to two different influenza antigens, as described by eq. (15). Antibody count M differs for each distribution based on the number of NaN values in the dataset. (a) $M = 64,616$, (b) $M = 63,835$, and (c) $M = 62,929$.

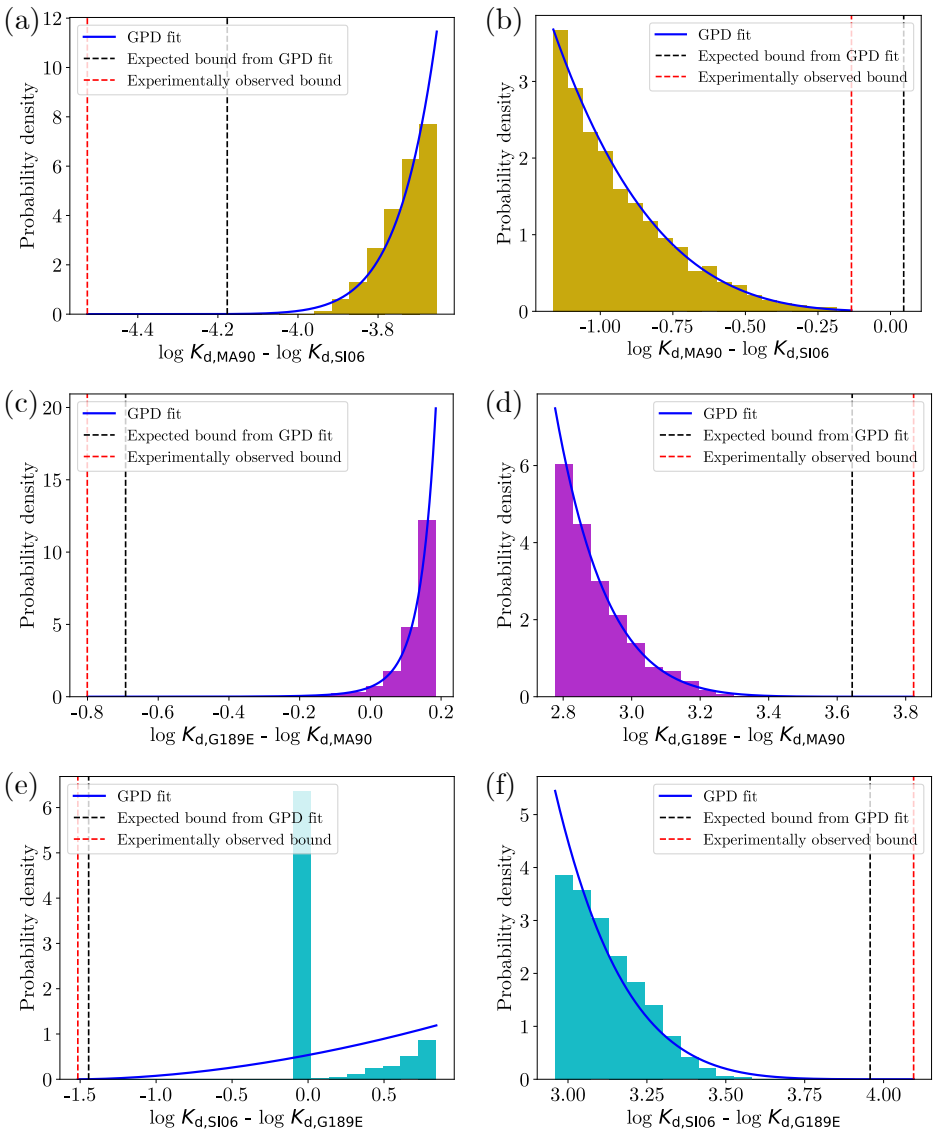

FIG. S3. Upper and lower tails of the distributions in Figure S2 fit using the generalized Pareto distrbution in eq. (16), with location parameter set to $\mu = 0$. (a) Lower tail of Figure S1(a), (b) upper tail of Figure S1(a), (c) lower tail of Figure S1(b), (d) upper tail of Figure S1(b), (e) lower tail of Figure S1(c), (f) upper tail of Figure S1(c). 5% quantiles were used to threshold the tails for (a-d,f), and the lower 20% quantile was used to threshold the tail for (e), in order to capture the spike.

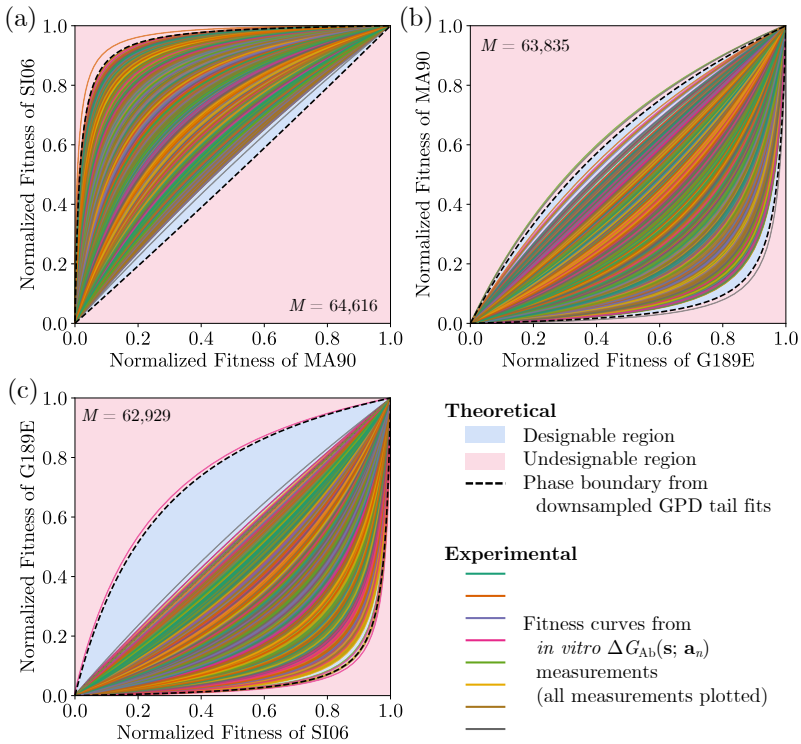

FIG. S4. Semi-empirical phase diagrams constructed by substituting the estimated bounds on the binding affinity difference distributions (i.e. black dashed lines in fig. S3) into main text eq. (9). The theoretical phase boundaries show excellent agreement with the experimental extremal antibodies, as expected since the full data range is used to estimate the bounds.

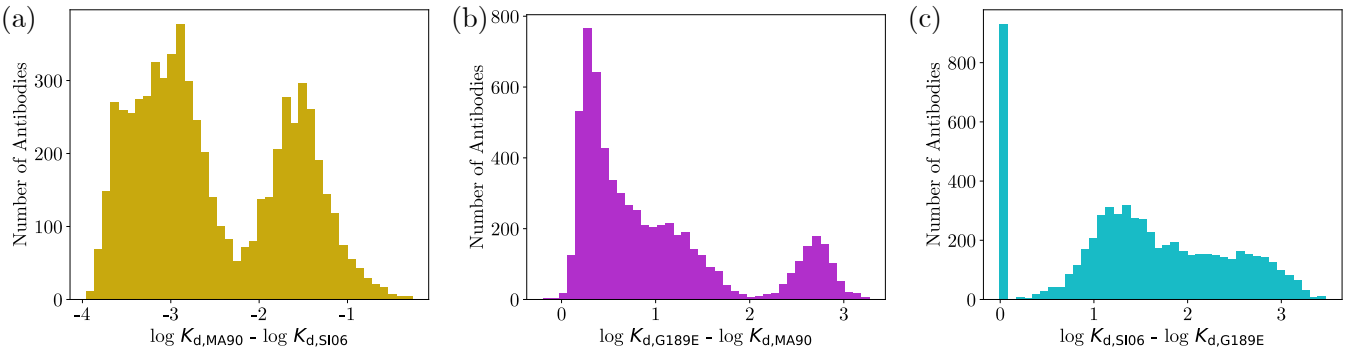

FIG. S5. A representative example of binding affinity difference distributions for a single downsampling trial, using approximately 10% of the original antibody library. Antibody count M differs for each distribution (and may not be exactly 10% of the original datasets) based on the number of NaN values in the dataset. (a) $M = 6,469$, (b) $M = 6,396$, and (c) $M = 6,314$.

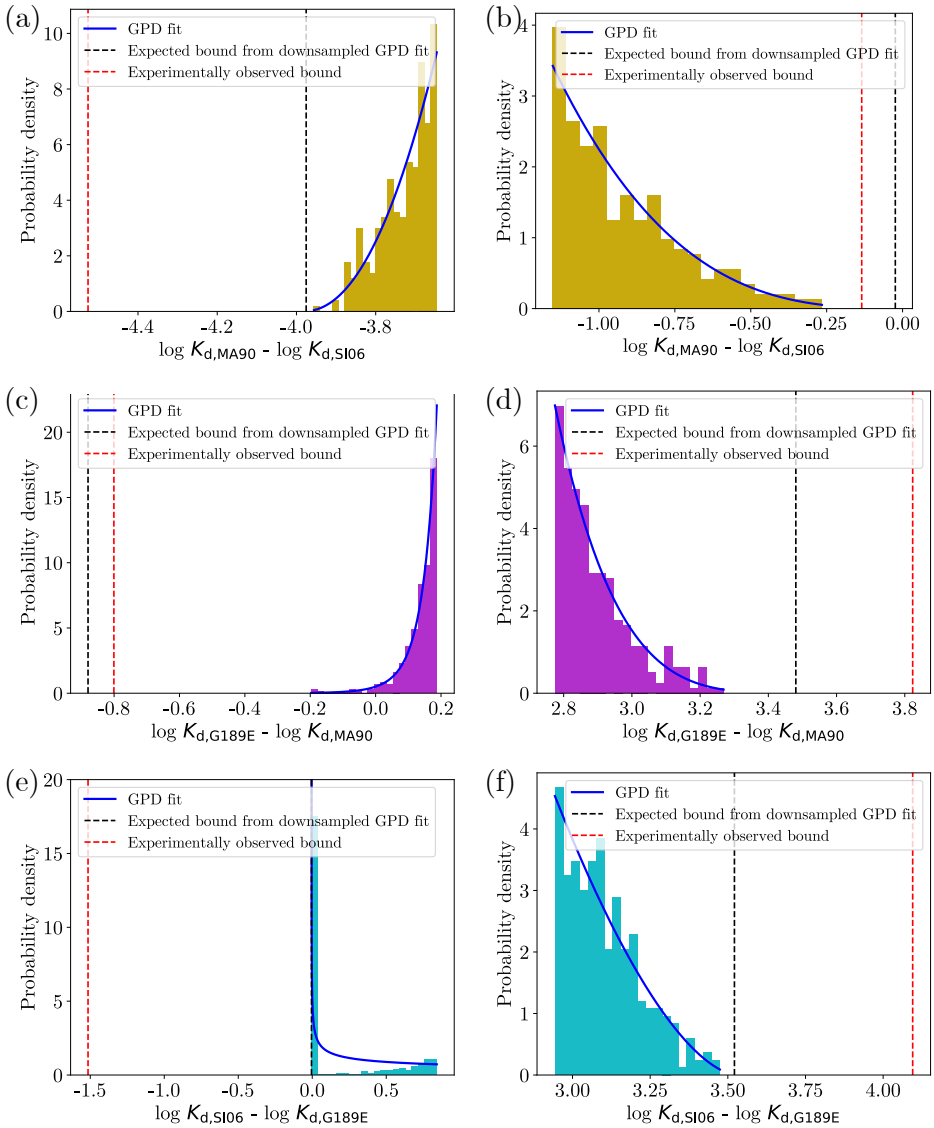

FIG. S6. Upper and lower tails of the distributions in Figure S5 fit using the generalized Pareto distrbution in eq. (16), with location parameter set to $\mu = 0$. (a) Lower tail of Figure S5(a), (b) upper tail of Figure S5(a), (c) lower tail of Figure S5(b), (d) upper tail of Figure S5(b), (e) lower tail of Figure S5(c), (f) upper tail of Figure S5(c). 5% quantiles were used to threshold the tails for (a-d,f), and the lower 20% quantile was used to threshold the tail for (e), in order to capture the spike.

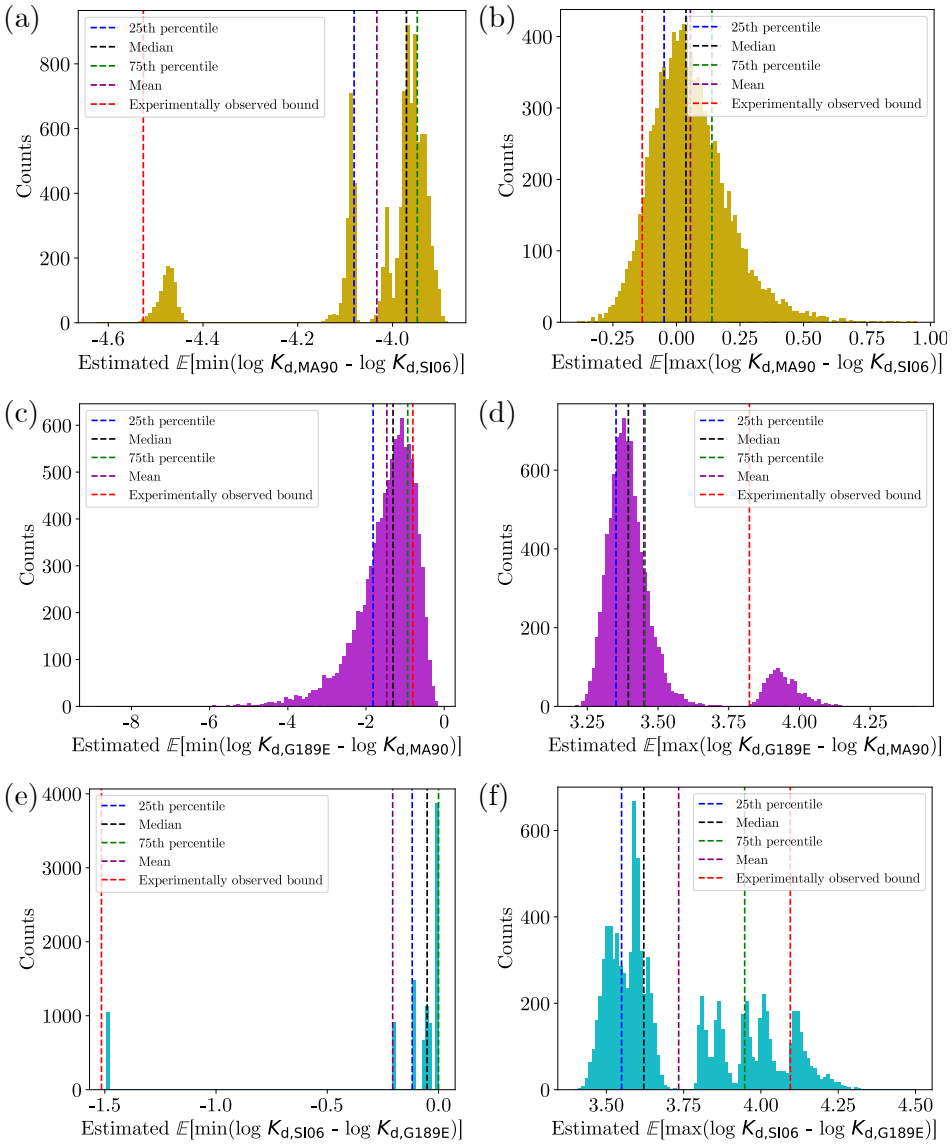

FIG. S7. Distributions for the estimated extremal values for each tail, where each distribution has been built from 10,000 trials of the downsampling procedure. For lower tails, $\mathbb{E}[\min(\log K_{d,1} - \log K_{d,2})]$ is calculated for antigens 1 and 2, and for upper tails $\mathbb{E}[\max(\log K_{d,1} - \log K_{d,2})]$ is calculated for antigens 1 and 2. (a) Lower tail for MA90 vs. SI06, (b) upper tail for MA90 vs. SI06, (c) lower tail for G189E vs. MA90, (d) lower tail for G189E vs. MA90, (e) lower tail for SI06 vs. G189E (f) upper tail for SI06 vs. G189E. Vertical lines indicate 25th, 50th (i.e. median), and 75th percentiles as well as the empirical mean and the experimentally observed extrema.