Parameters and determinants of responses to selection in antibody libraries

Steven Schulz[a], Sébastien Boyer[b], Matteo Smerlak[c], Simona Cocco[d], Rémi Monasson[d], Clément Nizak[e], and Olivier Rivoire[a]

[a] *Center for Interdisciplinary Research in Biology (CIRB), Collège de France, CNRS UMR 7241, INSERM U1050, PSL University, Paris, France*
[b] *Département de biochimie, Faculté de Médecine, Université de Montréal, Montréal, Canada*
[c] *Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany*
[d] *Laboratory of Physics of École Normale Supérieure, UMR 8023, CNRS & PSL University, Paris, France*
[e] *Chimie Biologie Innovation, ESPCI Paris, CNRS, PSL University, Paris, France*

Abstract

Antibody repertoires contain binders to nearly any target antigen. The sequences of these antibodies differ mostly at few sites located on the surface of a scaffold that itself consists of much less varied amino acids. What is the impact of this scaffold on the response to selection of a repertoire? To gauge this impact, we carried out quantitative phage display experiments with three antibody libraries based on distinct scaffolds harboring the same diversity at randomized sites, which we selected for binding to four arbitrary targets. We first show that the response to selection of an antibody library is captured by a simple and measurable parameter with direct physical and information-theoretic interpretations. Second, we identify a major determinant of this parameter which is encoded in the scaffold, its degree of evolutionary maturation. Antibodies undergo an accelerated evolutionary process, called affinity maturation, to improve their affinity to a given target antigen as part of the adaptive immune response. We find that libraries of antibodies built around such maturated scaffolds have a lower response to selection to other arbitrary targets than libraries built around naïve scaffolds of germline origin. Our results are a first step towards quantifying and controlling the evolutionary potential of biomolecules.

1. Introduction

The idea that evolution by natural selection is not only leading to adaptations but to a propensity to adapt, or "evolvability", has been repeatedly put forward [1, 2, 3]. As demonstrated by a number of mathematical models, evolvability can indeed emerge from evolutionary dynamics without any direct selection for it [4, 5, 6, 7]. Yet, theoretical insights have not translated into experimental assays for measuring and controlling evolvability in actual biological systems. Biomolecules as RNAs and proteins are ideal model systems for developing such assays as they are amenable to controlled experimental evolution [8]. For proteins, in particular, several biophysical and structural features have been proposed to correlate with their evolvability, most notably their thermal stability [9] and the modularity and polarity of their native fold [10]. A major limitation, however, is the absence of a measurable index of evolvability quantifying evolutionary responses to compare to biophysical or structural quantities.

Here, we present results of quantitative selection experiments with antibodies that address this issue. Antibodies are particularly well suited to devise and test new approaches to measure and control evolvability. They conveniently span a large phenotypic diversity, specific binding to virtually any molecular target, by

means of a limited genotypic diversity. Most of the diversity of natural antibody repertoires is indeed achieved by a few randomized loops that are displayed onto a structurally more conserved framework [11]. Further, well-established screening techniques are available for manipulating libraries of billions of diverse antibodies [12]. More fundamentally, antibodies are subject to two evolutionary processes on two distinct time scales: their frameworks evolve on the time scale of many generations of their host, as all other genes, and both frameworks and loops also evolve on a much shorter time scale as part of the immune response in the process of affinity maturation [13]. Importantly, affinity maturation-associated mutations are somatic and the sequences of maturated antibodies are not transmitted to subsequent generations. Germline antibody frameworks, whose transmitted sequences are the starting point of affinity maturation, are thus well positioned to be particularly evolvable, as evolving to increase their affinity to antigens is part of their physiological role.

As a first step towards quantifying and controlling the evolvability of antibodies, we previously characterized the response to selection of antibody libraries built around different structural scaffolds [14]. We took for these scaffolds the frameworks of heavy chains (V_H) of natural antibodies, and built libraries by introducing all combinations of amino acids at four consecutive sites in their complementary determining region (CDR3) loop, a part of their sequence known to determine their binding affinity and specificity [11]. Using phage display [15], we selected sequences from these libraries for their ability to bind different molecular targets and inferred the relative enrichment, or selectivity, of different antibody sequences by high-throughput sequencing [16]. Comparing experiments with libraries built on different scaffolds and selected against different targets led us to two conclusions. First, we quantified the variability of responses to selection of different sequences within a library and found this variability to differ widely across experiments involving different libraries and/or different targets. Second, we observed a hierarchy of selectivities between libraries, with multiple sequences from one particular library dominating selections involving a mixture of different libraries. These results raised two questions: (i) How to relate the hierarchies of selectivities between and within libraries? (ii) How to rationalize the differences between scaffolds that are all homologous?

Here, we answer these two questions through the presentation of new data and new analyses. First, we propose to characterize the hierarchies within and between libraries with two parameters for which we provide interpretations from the three standpoints of physics, information theory and sequence content. Second, we present new experimental results that identify the degree of maturation of an antibody scaffold as a control parameter for its selective potential. The results are, to our knowledge, the first demonstration based on quantifying the evolutionary responses to multiple selective pressures that long-term evolution has endowed germline antibody frameworks with a special ability to respond to selection.

2. Methods

2.1. Experimental design

In the absence of mutations, the outcome of an evolutionary process is determined by the properties of its initial population. In our experiments where antibodies are evolved in successive cycles of selection and amplification, the critical property of a sequence x present in the initial population is its selectivity $s(x)$, the factor by which it is enriched or depleted from one cycle to the next (see Box). Selection involves binding to a target, which is varied between experiments. Experiments are designed for the selectivity $s(x)$ to reflect the binding affinity of sequence x to this target (see Appendix 1.1). Inevitably, however, it can also depend

on affinity to non-target substrates and to sequence-dependent differences in amplification. Importantly, while the details of these "biases" are contingent to the experimental approach, their presence is a generic feature of any process of molecular evolution, including the natural process of antibody affinity maturation of which the experiments mimic the first step, prior to the introduction of any mutation.

Each of our libraries consists of sequences with a common part, which we call a scaffold, and 4 positions that are randomized to all $N = 20^4$ combinations, where 20 is the number of natural amino acids. The mapping $x \mapsto s_{L,T}(x)$ from 4-position sequences x to selectivities thus depends both on the scaffold that defines the library L and on the target T that defines the selective pressure. We are interested in properties of the scaffold that favor large values of selectivities, where "large" is considered either relative to other sequences within the same library (same scaffold) or relative to sequences from different libraries (different scaffolds).

Our previous experiments involved 24 different libraries, each built on a different scaffold consisting of a natural V_H fragment [14]. These fragments originate from the germline or the B cells of organisms of various species. Scaffolds from the germline encode naïve antibodies which have not been subject to any affinity maturation, while scaffolds from B cells encode maturated antibodies which have evolved from naïve antibodies to bind strongly to antigens encountered by the organisms. We previously performed experiments where the initial population consisted either of a single library or a mixture of different libraries [14]. In particular, in two experiments using very different targets (a neutral polymer and a DNA loop) we co-selected all 24 libraries together. Strikingly, while only 2 of the 24 libraries were built on germline scaffolds, the final population of one experiment was dominated by antibodies built on one of the two naïve scaffolds, and the second by the other one. This suggested us that germline scaffolds may have an intrinsically higher selective potential.

To investigate this hypothesis, we analyze here the selection against 4 different targets of 3 libraries with varying degrees of maturation. The scaffolds of the 3 libraries originate from Human V_H fragments and have evolved to different degrees as part of the immune response of patients infected by HIV (Fig. S1). The first scaffold (Germ) is taken from the germline and has not undergone any maturation. The second scaffold (Lim) has been subject to limited affinity maturation and differs from Germ, from which it originates, by 14 % of its amino acids. The third scaffold (Bnab) is a so-called broadly neutralizing antibody [17], which has evolved over many years to recognize a conserved part of the HIV virus [18]; it also originates from Germ, to which it differs by 34 % of its amino acids, and has evolved independently of Lim, to which it differs by 38 %. The 3 libraries, which are built around these scaffolds by introducing all combinations of amino acids at 4 positions of their CDR3 were part of the 24 libraries used in our previous experiments [14]. Here, to systematically compare the selective potential of these libraries, we present experiments where they are selected against four different targets, two DNA targets (DNA hairpins with a common stem but different loops, denoted DNA1 and DNA2, Fig. S2) and two homologous protein targets (the fluorescent proteins eGFP and mCherry, denoted prot1 and prot2), each unrelated to the HIV virus against which the Lim and Bnab scaffolds had been maturated.

2.2. Parametrization

To quantitatively compare the outcome of different experiments with different libraries and targets, we introduce here two parameters, σ and μ, which respectively quantify intra and inter-library differences in selectivities. These parameters derive from a statistical approach that considers only the distribution $P(s)$

of values that selectivities take across the different sequences of a library [19, 20, 21]. They correspond to the assumption that this distribution is log-normal,

$$P(s) = \frac{1}{\sqrt{2\pi}\sigma s} \exp\left(-\frac{(\ln s - \mu)^2}{2\sigma^2}\right). \tag{1}$$

The parameter σ captures intra-library differences in response to selection while the parameter μ provides the additional information required to describe inter-library differences.

The assumption that distributions of selectivities are log-normal has several justifications. First, it empirically provides a good fit of the data, not only in our experiments as we show below, but in a number of previous studies of antibody-antigen interactions [22] and protein-DNA interactions [23], including studies that had access to the complete distribution $P(s)$ [23]. Second, log-normal distributions are stable upon iteration of the evolutionary process: if two successive selections are performed so that $s = s_1 s_2$ with s_1 and s_2 independently described by log-normal distributions with parameters (σ_1, μ_1) and (σ_2, μ_2), s also follows a log-normal distribution with parameters $\sigma = (\sigma_1^2+\sigma_2^2)^{1/2}$ and $\mu = \ln[(\sigma_1^{-2}e^{\mu_1}+\sigma_2^{-2}e^{\mu_2})/\sigma^{-2}]$; more generally, log-normal distributions are attractors of evolutionary dynamics [24]. Third, log-normal distributions are physically justified from the simplest model of interaction, an additive model where the interaction energy between sequence $x = (x_1, \ldots, x_\ell)$ of length ℓ and its target takes the form $E(x) = \sum_{i=1}^{\ell} \epsilon_i(x_i)$ with contributions $\epsilon_i(x_i)$ from each position i and amino acid x_i, and thus its selectivity $s(x) \simeq \mathrm{e}^{-E(x)/k_B T}$, where T is the temperature and k_B the Boltzmann constant (Appendix 1.1). At thermal equilibrium and for sufficiently large ℓ, a log-normal distribution of the affinities is then expected with $\mu \sim -\ell\langle\epsilon\rangle/k_B T$ and $\sigma \sim \ell^{1/2}(\langle\epsilon^2\rangle - \langle\epsilon\rangle^2)^{1/2}/k_B T$, where $\langle\epsilon\rangle$ and $\langle\epsilon^2\rangle - \langle\epsilon\rangle^2$ are respectively the mean and variance of the values of binding energies per position $\epsilon_i(x_i)$.

2.3. Inference of parameters

Selectivities are measured as relative enrichments of sequences in two successive rounds of selection. We obtain the parameters σ and μ by fitting the values with truncated log-normal distributions (Fig. 1A and Appendix 3.3). This inference is complicated by two factors: only the upper tail of the distribution of selectivities is sampled in the experiments and enrichments provide selectivities only up to a multiplicative factor (see Box). While the parameter σ is independent of this multiplicative factor, comparing the parameters μ between libraries requires performing selections where different libraries are mixed in the initial population. To refine and validate our inference, we also performed selection experiments where we mixed a very small number of random and top selectivity sequences (Fig. 1B): as the random sequences typically reflect the mode of the distributions, (the most likely selectivity value), these experiments provide an independent estimation of μ that we can profitably use (see details in Appendix 3.3).

The values of σ and μ that we infer for the 3 libraries Germ, Lim and Bnab when selected against each of the 4 targets DNA1, DNA2, prot1 and prot2 are presented in Fig. 2A. We validated the quality of the fits by probability-probability and quantile-quantile plots (Figs. S16-S18). We also assessed the robustness of the inference by comparing replicate experiments (Figs. S16-S22), and comparing experiments where a library is selected either alone or in mixture with the other two (Fig. S19). Finally, we verified that the results are unchanged whether selectivities are measured from enrichments between the 2nd and 3rd cycles, or the 3rd and 4th ones (Figs. S20-S21).

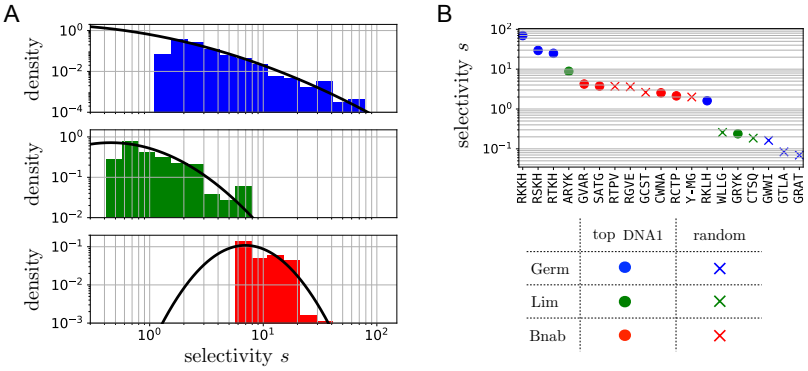


Figure 1: Fitting empirical distributions of selectivities with log-normal distributions. **A.** The selection of a library L against a target T provides the selectivities of the sequences in L that have top selectivity against T. Here, the histograms show the selectivities obtained from experiments where the Germ (in blue), Lim (in green) and Bnab (in red) libraries were selected against the DNA1 target. The black line is the best fit to a log-normal distribution. **B.** To locate precisely the mode of the distributions, we performed experiments where the initial population consists in a mixture of very few top (dots) and random (crosses) sequences. Top sequences are identified from A based on the largest selectivities against the target. Random sequences, on the other hand, are picked at random in the libraries and are expected to have typical selectivities located at the maxima of the black curves in A. Taken together, the results indicate that when selected against the DNA1 target, the Germ library has the highest σ and the Bnab library the highest μ. Similar results are obtained for selections against other targets (Fig. S8 and Table 1).

3. Results

3.1. Intra-library hierarchy

The hierarchy of selectivities within a library is quantified by the parameter σ: a small σ indicates that all sequences in the library are equally selected while a large σ indicates that the response to selection varies widely between sequences in the library. When comparing the $\sigma_{L,T}$ inferred from the selections of the 3 libraries L against each of the 4 targets T, a remarkable pattern emerges: the more a scaffold is maturated, the smaller is σ, $\sigma_{\mathrm{Germ},T} > \sigma_{\mathrm{Lim},T} \geq \sigma_{\mathrm{Bnab},T}$ for all targets T, and even $\min_T(\sigma_{\mathrm{Germ},T}) > \max_T(\sigma_{\mathrm{Lim},T}, \sigma_{\mathrm{Bnab},T})$ (Fig. 2A). Statistically, if considering the inequalities to be strict, the experiments to be independent and any result to be *a priori* equally likely, the probability of this finding is only $p = (3!)^{-4} \simeq 7.10^{-4}$.

Examining sequence logos shows that although selections of the Germ library are characterized by a similarly high value of σ for the 4 targets, the sequences that are selected against each target are different (Fig. 2B-C). The amino acids found to be enriched are consistent with the nature of the targets: selections against the DNA targets are dominated by positively charged amino acids (letters in blue) and selections against the two protein targets, which are close homologs, are dominated by similar amino acid motifs.

In contrast, sequences logos for the Bnab library show motifs that are less dependent on the target (Fig. 2B and Fig. S10). This observation is rationalized by an experiment where only the amplification step is performed, in the absence of any selection for binding. Sequence-specific amplification biases are then revealed, with sequence motifs that are similar to those observed when selection for binding is present (Fig. S10). With protein targets at least, the motifs are nevertheless sufficiently different to infer that selection for binding to the target contributes significantly to the selectivities (see also Fig. S6). Target-specific selection for binding, which is dominating the top selectivities in the Germ library (Fig. S11), is thus of the same order of magnitude as amplification biases for the top selectivities in the Bnab library.

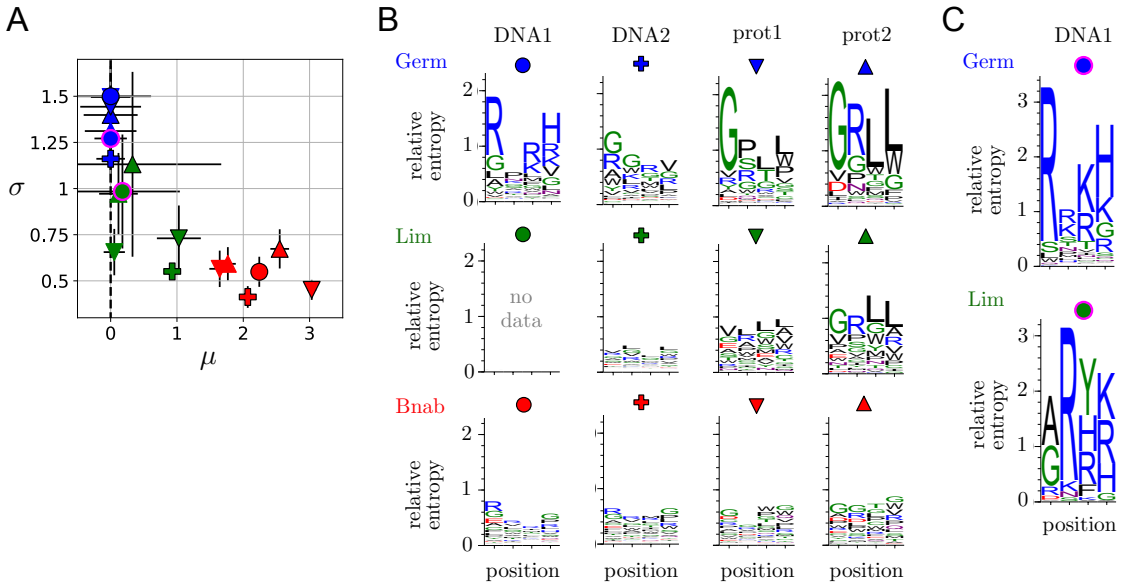


Figure 2: Comparing selections of libraries built on scaffolds with different degrees of maturation – **A.** Parameters (μ, σ) of the distributions of selectivities for our 3 libraries selected against 4 targets. The color of the symbols indicates the library (Germ, Lim or Bnab) and its shape the target (DNA1, DNA2, prot1 or prot2) with the conventions defined in B. Symbols with a black or no contour indicate results from replicate experiments where the 3 libraries are mixed in the initial population, and symbols with a magenta contour where a library is screened in isolation. $\mu_{\text{Germ},T}$ is conventionally set to $\mu_{\text{Germ},T} = 0$ for all targets T (Appendix 3.4). μ is generally more challenging to infer than σ and it shows here more variations across replicate experiments. **B.** Sequence logos for $\tilde{s}_i(a)$, which represent the contribution of the different amino acids to the selectivities (see Box), for the selections of the three libraries, Germ, Lim and Bnab against the two DNA targets (DNA1 and DNA2) and the two protein targets (prot1 and prot2). These results correspond to experiments where the 3 libraries are mixed in the initial population. The Lim library is outcompeted by the other two libraries when selected against the DNA1 target, which does not leave enough sequences to make a meaningful inference (see also Fig. S10 for more details on the sequence logos for the Bnab library). **C.** Sequence logos for $\tilde{s}_i(a)$ for the Germ and Lim libraries selected in isolation against the DNA1 target. For the Lim library, this palliates the absence of data in B. For the Germ library, it shows that essentially the same motifs are found whether the library is selected in a mixture as in B or on its own; the area under the logos is, however, different: it would be $\sigma^2/2$ with infinite sampling, but major deviations are caused by limited sampling (Fig. S9).

Remarkably, the Lim library behaves either like the Germ library or the BnAb library, depending on the target. In particular, a motif of positively charged amino acids emerges when selecting it against one of the two DNA targets (DNA1), but no clear motif emerges when selecting it against the other one (DNA2). Besides, when a clear motif emerges, it can be identical to the motif emerging from the Germ library as in case of a selection against the prot2 target, or different, as in the case of a selection against the DNA1 target (but with a similar selection of positively charged amino acids).

3.2. Inter-library hierarchy

The hierarchy of selectivities between libraries is quantified by the parameter μ. This parameter also shows a pattern that is independent of the target: $\mu_{\text{Germ},T} \simeq \mu_{\text{Lim},T} < \mu_{\text{Bnab},T}$ and even $\max_T(\mu_{\text{Germ},T}, \mu_{\text{Lim},T}) < \min_T(\mu_{\text{Bnab},T})$ (Fig. 2B). Inferring μ is more challenging than inferring σ and the differences observed between the Germ and Lim libraries are most likely not significant, as apparent from the observed variations between replicate experiments. The μ of the Bnab library is, on the other hand, systematically larger. The difference is explained by an experiment where selection is performed in the absence of DNA or protein targets but in the presence of streptavidin-coated magnetic beads to which these targets are usually attached.

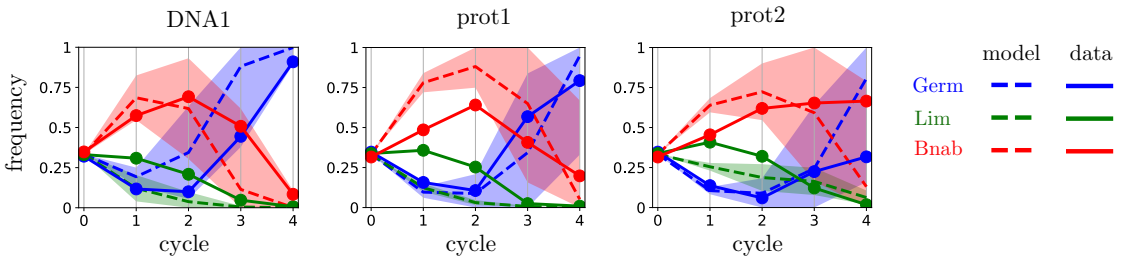


Figure 3: Dynamics of library frequencies – A mixture of the three libraries, Germ (blue), Lim (green) and Bnab (red) was subject to four successive cycles of selection and amplification against different targets. The full lines report the evolution of the relative frequencies of the three scaffolds. The dotted lines represent the estimated dynamics using the characterization of each library by a log-normal distribution with the parameters σ, μ estimated from the selection of the libraries against the same target (Appendix 1.4). The shaded area correspond to one standard deviation in the estimation of the parameters σ, μ. The model assumes that sequences are uniformly represented in each initial library, which is not the case in experiments and explains why the agreement with the data is only qualitative.

This experiment reproduces the differences in $\mu_{L,T}$, which indicates a small but significant affinity of the Bnab scaffold for the magnetic beads, independent of the sequence x (Fig. S12). While the differences in σ appear to be independent of the target, the differences in μ are thus related to a common feature of the targets. Given these different origins, the correlation between σ and μ that we observe may be fortuitous.

3.3. Implications for evolutionary dynamics

The different patterns of intra and inter-library hierarchies lead to a non-trivial evolutionary dynamics when selecting from an initial population that is composed of different libraries. In particular, a non-monotonic enrichment is expected when mixing two libraries characterized by (μ_1, σ_1) and (μ_2, σ_2) with $\mu_1 > \mu_2$ but $\sigma_1 < \sigma_2$: the library with largest μ dominates the first cycles while the one with largest σ dominates the later ones. This is indeed observed in experiments where different libraries are mixed in the initial population (Fig. 3). The dynamics of the relative frequencies of different libraries is globally predicted by a calculation of library frequencies in the mix based on the parameters (μ, σ) inferred for each library independently (Appendix 1.4), even though deviations are expected from the non-uniform sampling of sequences within each library, which is not encoded in σ or μ. Parametrizing the response to selection of a library by the two parameters (μ,σ) is thus not only useful to characterize its intrinsic response but also to rationalize the evolutionary dynamics of mixtures of libraries.

4. Discussion

How to interpret the result that intra-library diversity, as characterized by σ, decreases with the level of maturation of the scaffold? Here, we show that the parameter σ provides a characterization of intra-library diversity that is equivalent to three other approaches based on extreme value theory, information theory and sequence logos.

4.1. Extreme value statistics

In our previous work [14], we fitted the tail of the distribution of selectivities with generalized Pareto distributions, a family of distributions with two parameters, a shape parameter κ and a scaling parameter

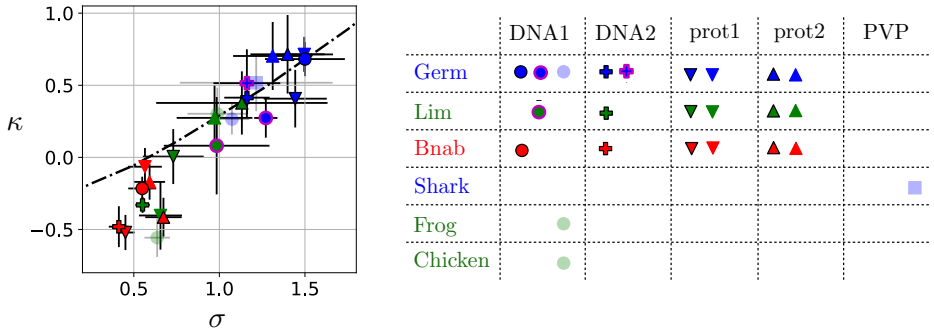


Figure 4: Shape parameter κ from fits of the selectivities to generalized Pareto distributions versus σ from fits to log-normal distributions – Results from different libraries selected against different targets are represented here with the same convention as in Figure 2: blue, green and red plain colors for the Germ, Lim and Bnab libraries, circle, cross, downward and upward triangles for the DNA1, DNA2, prot1 and prot2 targets. In addition, results from our previous work [14] are indicated in transparent blue if they involve a library built onto a germline scaffold and in transparent green if they involve a library built onto a maturated scaffold. The hierarchy indicated by κ is essentially the same as the hierarchy indicated by σ, consistent with the expected relationship between κ and σ (black dotted line, Fig. S14). By the two approaches, libraries built onto germline scaffolds are found to have a more diverse response to selection than libraries built onto maturated scaffolds irrespectively of the target (all values of σ and κ are given in Table 1).

τ. This was motivated by extreme value theory, which establishes that these parameters are sufficient to describe the tail of any distribution (Appendix 1.2). For different libraries L and different targets T, we found that generalized Pareto distributions provide a good fit of the upper tail of $P_{L,T}(s)$, with, depending on the scaffold L and target T either $\kappa > 0$ (heavy tail), $\kappa < 0$ (bounded tail) or $\kappa = 0$ (exponential tail). The origin of these different values of κ was, however, unclear.

Comparing probability-probability plots to assess the quality of the fits, our data appears equally well fitted by generalized Pareto distributions and log-normal distributions (Figs. S16-S22). This finding is at first sight puzzling as some of the fits with generalized Pareto distributions involve a non-zero shape parameter $\kappa \neq 0$ but extreme value theory states that the tail of log-normal distributions is asymptotically described by a shape parameter $\kappa = 0$ for all values of σ, μ [25]. Extreme value theory is, however, only valid in the double asymptotic limit $N \to \infty$ and $s^* \to \infty$, where N is the total number of samples and s^* the threshold above which these samples are considered. With finite data, determining whether this asymptotic regime is reached is notoriously difficult when the underlying distribution is log-normal [26]. More precisely, N points randomly sampled from a log-normal distribution with parameter σ are known to display an apparent $\kappa_N = \sigma/(2 \ln N)^{1/2}$ which tends to zero only very slowly with increasing values of N [26]. In fact, this relationship itself requires N (or σ) to be sufficiently large and finite size effects can even produce an apparent $\kappa_N < 0$ (Fig. S14).

While casting doubt on the practical applicability of extreme value theory, these statistical effects do not call into question the main conclusion of our previous work [14]: different combinations of scaffolds L and targets T exhibit different within-library hierarchies, which are quantified by the different values of their (apparent) shape parameter κ. Fits with a log-normal distribution provide another parameter σ that report essentially the same differences (Fig. 4). More importantly, we verify on our previous data, which partly involves different scaffolds and different targets, that libraries built on germline scaffolds have a higher σ than libraries built around maturated scaffolds (Fig. 4 and Table 1).

4.2. Informational interpretation

The parameter σ can also be given an information-theoretic interpretation. From a statistical standpoint, the specificity of selection of a population is naturally quantified by the relative entropy $D(f^1\|f^0) = \sum_x f^1(x) \ln\left[f^1(x)/f^0(x)\right]$ which compares the distribution $f^1(x)$ of sequences x after one cycle of selection-amplification to their initial distribution $f^0(x)$. When taking this initial distribution $f^0(x)$ to be a uniform distribution over the N possible sequences, $f^0(x) = N^{-1}$, $f^1(x)$ is nothing but the selectivity $\tilde{s}(x)$ obtained by choosing λ in Eq. (4) such that $\sum_x \tilde{s}(x) = 1$. The inverse of $D(\tilde{s}\|N^{-1})$ answers the following statistical question (Appendix 1.3): how large should the initial population be to infer from the outcome of an experiment that selection is at work? Assuming a large initial library with selectivity distribution $P(s)$, $D(\tilde{s}\|N^{-1}) = \sum_x s(x) \ln[s(x)N]$ can be rewritten as

$$D(\tilde{s}\|N^{-1}) = \left\langle \frac{s}{\langle s\rangle} \ln \frac{s}{\langle s\rangle} \right\rangle \tag{2}$$

where the average is taken with $P(s)$, i.e., $\langle\phi(s)\rangle = \int_0^\infty \mathrm{d}s P(s)\phi(s)$. When $P(s)$ is log-normal,

$$D(\tilde{s}\|N^{-1}) = \frac{\sigma^2}{2}, \tag{3}$$

showing that σ can be interpreted as quantifying the specificity of selection at the population level. This statistical viewpoint can be extended to characterize the specificity of arbitrary sets of binders and ligands (Appendix 1.3), generalizing a proposal to define specificity as an amount of information encoded in interactions [27].

4.3. Sequence motifs

Assuming that the different sites i along the sequence contribute independently to the selectivity, $\tilde{s}(x) = \prod_i \tilde{s}_i(x_i)$, the specificity $D(\tilde{s}\|N^{-1})$ is nothing but $\sum_i D(\tilde{s}_i\|A^{-1}) = \sum_i \sum_{a_i} \tilde{s}_i(a_i) \ln[\tilde{s}_i(a_i)A]$, the total area under the sequence logos of $\tilde{s}_i(a)$, where $A = 20$ is the total number of amino acids. By displaying both amino acid specificities and an overall measure of specificity of selection $D(\tilde{s}\|N^{-1})$, sequence logos thus provide a convenient summary of selection within a library.

This comes, however, with an important caveat when selectivities are available only for a small subset of $N' \ll N$ sequences, as it is the case in experiments. If ignoring unobserved sequences when computing $\tilde{s}_i(a_i)$, the empirically determined quantity $\sum_i D(\tilde{s}_i\|A^{-1})$ overestimates the true value of $D(\tilde{s}\|N^{-1})$, all the more as N' is smaller (Fig. S9). Because of this effect, the areas under the curve of the sequence logos based on $\tilde{s}_i(a)$ are not comparable to $\sigma^2/2$ as Eq. (3) would suggest. They are also not comparable across different experiments when the sampling sizes N' differ (Fig. 2B and C). Finally, even with $N' = N$, deviations between $\sum_i D(\tilde{s}_i\|A^{-1})$ and $D(\tilde{s}\|N^{-1})$ may arise if the contributions of the different positions are not additive.

5. Conclusion

In summary, we find that libraries built around germline antibody scaffolds have a response to selection that is quantitatively different from libraries built around maturated scaffolds: for arbitrary targets, they contain sequences with a wider range of affinities, including specific sequences with the strongest affinities.

This constitutes the first quantitative evidence that germline antibodies are endowed with a special evolutionary ability to generate selectable diversity. Our work was centered onto 3 libraries, one based on a germline scaffold and two based on scaffolds derived from this germline scaffold with different degrees of maturation, which we selected against 4 different targets. Assessing the generality of our conclusion will require further experiments with additional scaffolds and targets. The statistical framework that we introduced here provides the required tools to perform this analysis systematically and quantitatively. Beyond the 3 libraries studied here, our conclusions are also supported by our previous results [14], which involved a library built on another germline scaffold, 20 libraries built on other maturated scaffolds, and a completely different target (Fig. 4).

Which physical mechanisms may underly the differences in selective potential that we observe? A number of studies, ranging from structural biology to molecular dynamics simulations, have reported changes in antibody flexibility and target specificity over the course of affinity maturation [28, 29, 30, 31, 32, 33, 34, 35]. The emerging picture is that naïve antibodies are flexible and polyspecific and become more rigid and more specific as they undergo affinity maturation. An increase of structural rigidity in the course of evolution is also found in proteins unrelated to antibodies [36]. Germline scaffolds may thus be more flexible than maturated scaffolds. If this scenario is correct, how this structural flexibility translates into evolutionary diversity once different complementary determining regions (CDRs) are grafted onto the scaffolds remains to be explained.

Irrespectively of mechanisms, we described selectivity distributions of libraries with two statistical parameters, σ and μ, which we showed to be determined by different factors. Of these two parameters, σ, which has simple physical and information-theoretic interpretations, is candidate to serve as a general quantitative index of selective potential for biomolecules. Beyond selection, a next step is to extend this work to quantify evolvability, i.e., the response to successive cycles of selection and mutations. Yet, being able to quantify the selective potential of a scaffold by an index that is systematically reduced in the course of evolution already raises an interesting challenge: can we increase this index to design libraries with better response to selection?

Acknowledgments

This work was supported by FRM AJE20160635870 and by ANR-17-CE30-0021-02. It benefited from the expertise of the high-throughput sequencing platform at the Institut de Biologie Intégrative de la Cellule (I2BC) at Gif-sur-Yvette, France.

References

[1] G. P. Wagner, L. Altenberg, Perspective: complex adaptations and the evolution of evolvability, Evolution 50 (3) (1996) 967–976.

[2] M. Kirschner, J. Gerhart, Evolvability, Proceedings of the National Academy of Sciences 95 (15) (1998) 8420–8427.

[3] A. Wagner, Robustness and evolvability in living systems, Vol. 24, Princeton university press, 2013.

[4] L. Ancel Meyers, F. D. Ancel, M. Lachmann, Evolution of Genetic Potential, PLoS computational biology 1 (3) (2005) e32.

[5] M. Parter, N. Kashtan, U. Alon, Facilitated Variation: How Evolution Learns from Past Environments To Generalize to New Environments, PLoS computational biology 4 (11) (2008) e1000206.

[6] M. Hemery, O. Rivoire, Evolution of sparsity and modularity in a model of protein allostery., Physical review. E, Statistical, nonlinear, and soft matter physics 91 (4) (2015) 042704–10.

[7] A. Crombach, P. Hogeweg, Evolution of evolvability in gene regulatory networks, PLoS computational biology 4 (7) (2008) e1000112.

[8] P. A. Romero, F. H. Arnold, Exploring protein fitness landscapes by directed evolution, Nature reviews Molecular cell biology 10 (12) (2009) 866.

[9] J. D. Bloom, S. T. Labthavikul, C. R. Otey, F. H. Arnold, Protein stability promotes evolvability., Proceedings of the National Academy of Sciences 103 (15) (2006) 5869–5874.

[10] E. Dellus-Gur, Á. Tóth-Petróczy, M. Elias, D. S. Tawfik, What makes a protein fold amenable to functional innovation? Fold polarity and stability trade-offs., Journal of Molecular Biology 425 (14) (2013) 2609–2621.

[11] E. A. Padlan, Anatomy of the antibody molecule, Molecular immunology 31 (3) (1994) 169–217.

[12] H. R. Hoogenboom, Selecting and screening recombinant antibody libraries, Nature Biotechnology 23 (9) (2005) 1105–1116.

[13] H. N. Eisen, Affinity enhancement of antibodies: how low-affinity antibodies produced early in immune responses are followed by high-affinity antibodies later and in memory B-cell responses, Cancer immunology research 2 (5) (2014) 381–392.

[14] S. Boyer, D. Biswas, A. Kumar Soshee, N. Scaramozzino, C. Nizak, O. Rivoire, Hierarchy and extremes in selections from pools of randomized proteins., Proceedings of the National Academy of Sciences of the United States of America 113 (13) (2016) 3482–3487.

[15] G. P. Smith, V. A. Petrenko, Phage Display, Chemical Reviews 97 (2) (1997) 391–410.

[16] D. M. Fowler, C. L. Araya, S. J. Fleishman, E. H. Kellogg, J. J. Stephany, D. Baker, S. Fields, High-resolution mapping of protein sequence-function relationships, Nature Methods 7 (9) (2010) 741–746.

[17] D. R. Burton, P. Poignard, R. L. Stanfield, I. A. Wilson, Broadly neutralizing antibodies present new prospects to counter highly antigenically diverse viruses., Science 337 (6091) (2012) 183–186.

[18] F. Klein, R. Diskin, J. F. Scheid, C. Gaebler, H. Mouquet, I. S. Georgiev, M. Pancera, T. Zhou, R.-B. Incesu, B. Z. Fu, P. N. P. Gnanapragasam, T. Y. Oliveira, M. S. Seaman, P. D. Kwong, P. J. Bjorkman, M. C. Nussenzweig, Somatic Mutations of the Immunoglobulin Framework Are Generally Required for Broad and Potent HIV-1 Neutralization, Cell 153 (1) (2013) 126–138.

[19] L. Pauling, D. Pressman, A. L. Grossberg, The serological properties of simple substances. vii. a quantitative theory of the inhibition by haptens of the precipitation of heterogeneous antisera with antigens, and comparison with experimental results for polyhaptenic simple substances and for azoproteins, Journal of the American Chemical Society 66 (5) (1944) 784–792.

[20] A. Nisonoff, D. Pressman, Heterogeneity and average combining constants of antibodies from individual rabbits., Journal of immunology (Baltimore, Md. : 1950) 80 (6) (1958) 417–428.

[21] D. Lancet, E. Sadovsky, E. Seidemann, Probability model for molecular recognition in biological receptor repertoires: significance to the olfactory system., Proceedings of the National Academy of Sciences 90 (8) (1993) 3715–3719.

[22] S. Rosenwald, R. Kafri, D. Lancet, Test of a statistical model for molecular recognition in biological repertoires., Journal of theoretical biology 216 (3) (2002) 327–336.

[23] L. Wolf, O. K. Silander, E. van Nimwegen, Expression noise facilitates the evolution of gene regulation, Elife 4 (2015) e05856.

[24] M. Smerlak, A. Youssef, Limiting fitness distributions in evolutionary dynamics., Journal of theoretical biology 416 (2017) 68–80.

[25] E. J. Gümbel, Statistics of extremes, Columbia Univ. Press, 1958.

[26] R. Perline, Strong, weak and false inverse power laws, Statistical Science 20 (1) (2005) 66–88.

[27] M. H. Huntley, A. Murugan, M. P. Brenner, Information capacity of specific interactions., Proceedings of the National Academy of Sciences of the United States of America 113 (21) (2016) 5841–5846.

[28] G. J. Wedemayer, P. A. Patten, L. H. Wang, P. G. Schultz, R. C. Stevens, Structural insights into the evolution of an antibody combining site, Science 276 (5319) (1997) 1665–1669.

[29] J. Yin, A. E. Beuscher IV, S. E. Andryski, R. C. Stevens, P. G. Schultz, Structural Plasticity and the Evolution of Antibody Affinity and Specificity, Journal of Molecular Biology 330 (4) (2003) 651–656.

[30] J. R. Willis, B. S. Briney, S. L. DeLuca, J. E. Crowe, J. Meiler, Human germline antibody gene segments encode polyspecific antibodies., PLoS computational biology 9 (4) (2013) e1003045.

[31] A. M. Sevy, T. M. Jacobs, J. E. Crowe, J. Meiler, Design of Protein Multi-specificity Using an Independent Sequence

Search Reduces the Barrier to Low Energy Sequences., PLoS computational biology 11 (7) (2015) e1004300.

[32] V. Manivel, N. C. Sahoo, D. M. Salunke, K. V. Rao, Maturation of an antibody response is governed by modulations in flexibility of the antigen-combining site, Immunity 13 (5) (2000) 611–620.

[33] I. F. Thorpe, C. L. Brooks, Molecular evolution of affinity and flexibility in the immune system., Proceedings of the National Academy of Sciences 104 (21) (2007) 8821–8826.

[34] T. Li, M. B. Tracka, S. Uddin, J. Casas-Finet, D. J. Jacobs, D. R. Livesay, Rigidity Emerges during Antibody Evolution in Three Distinct Antibody Systems: Evidence from QSFR Analysis of Fab Fragments, PLoS computational biology 11 (7) (2015) e1004327–23.

[35] M. C. Thielges, J. Zimmermann, W. Yu, M. Oda, F. E. Romesberg, Exploring the Energy Landscape of Antibody-Antigen Complexes: Protein Dynamics, Flexibility, and Molecular Recognition, Biochemistry 47 (27) (2008) 7237–7247.

[36] E. C. Campbell, G. J. Correy, P. D. Mabbitt, A. M. Buckle, N. Tokuriki, C. J. Jackson, Laboratory evolution of protein conformational dynamics., Current Opinion in Structural Biology 50 (2018) 49–57.

[37] T. D. Schneider, R. M. Stephens, Sequence logos: a new way to display consensus sequences, Nucleic acids research 18 (20) (1990) 6097–6100.

[38] M. Djordjevic, A. M. Sengupta, Quantitative modeling and data analysis of selex experiments, Physical biology 3 (1) (2005) 13.

[39] C. Rastogi, H. T. Rube, J. F. Kribelbauer, J. Crocker, R. E. Loker, G. D. Martini, O. Laptenko, W. A. Freed-Pastor, C. Prives, D. L. Stern, R. S. Mann, H. J. Bussemaker, Accurate and sensitive quantification of protein-DNA binding affinity., Proceedings of the National Academy of Sciences of the United States of America 115 (16) (2018) E3692–E3701.

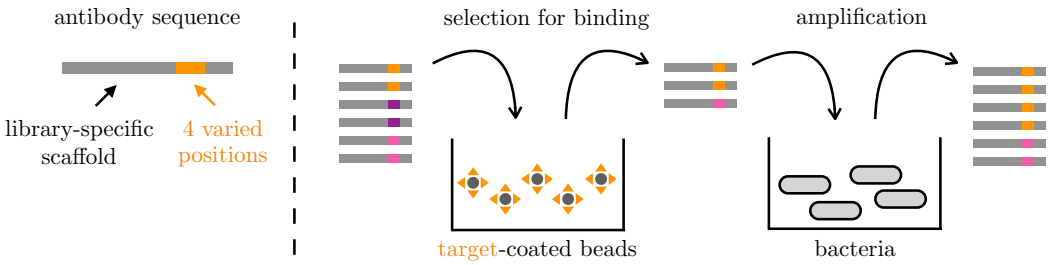
[40] S. Coles, J. Bawa, L. Trenner, P. Dorazio, An introduction to statistical modeling of extreme values, Vol. 208, Springer, 2001.

[41] T. M. Cover, J. A. Thomas, Elements of information theory, John Wiley & Sons, 2012.

BOX – Principles of antibody selection experiments

We perform phage display experiments with different libraries of antibodies as input and different molecular targets (DNA hairpins or proteins) as selective pressures [15]. Our antibodies are single domains from the variable part of the heavy chain (V_H) of natural antibodies. Antibodies in a library share a common scaffold of $\simeq 100$ amino acids and differ only at four consecutive sites of their third complementary determining region (CDR3), which is known to be important for binding affinity and specificity. A library comprises all combinations of amino acids at these four sites and therefore consists of a total of $N = 20^4 \simeq 10^5$ distinct sequences $x = (x_1, x_2, x_3, x_4)$. Initial populations include a total of 10^{11} sequences, corresponding to $\sim 10^6$ copies of each of the distinct $\sim 10^5$ sequences when a single library is considered. Physically, these populations are made of phages, each presenting at its surface one antibody and containing the corresponding sequence.

An experiment consists in a succession of cycles, each composed of two steps. In the first step, the phages are in solution with the targets, which are attached to magnetic beads and in excess relative to the phages to limit competitive binding (see Appendix 1.1). The beads are retrieved with a magnet and washed to retain the bound antibodies. In the second step, the selected phages are put in presence of bacteria which they infect to make new phages, thus amplifying retained sequences. A population of $\sim 10^{11}$ phages is thus reconstituted. Both the selection for binding to the target and the amplification can possibly depend on the sequence of the antibody.



We define the selectivity $s(x)$ of sequence x to be proportional to the probability for sequence x to pass one cycle. As the targets are in excess relative to the antibodies, selectivities are independent of the cycle c (see Appendix 1.1). In the limit of infinite population sizes, $s(x)$ is proportional to the relative enrichment $f^c(x)/f^{c-1}(x)$ of the frequencies $f^c(x)$ after any two successive cycles $c-1$ and c. To estimate these selectivities, about 10^6 sequences are sampled before and after a cycle and read by high-throughput sequencing. Given the counts $n^{c-1}(x)$ and $n^c(x)$ of sequence x before and after cycle c, we estimate the selectivity of x as

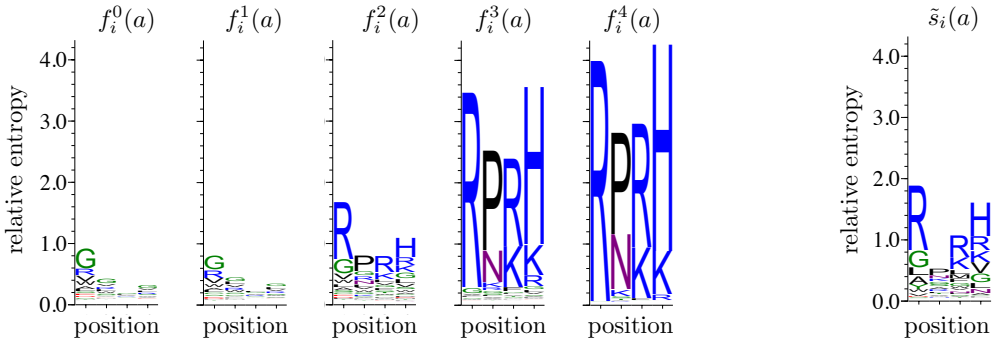
$$s(x) = \lambda \frac{n^c(x)}{n^{c-1}(x)} \tag{4}$$

where λ is an arbitrary multiplicative factor.

In practice, two types of noise must be taken into account when applying Eq. (4): an experimental noise, which implies that antibodies have a finite probability to pass a round of selection independently of their sequence, and a sampling noise, which arises from the limited number of sequence reads. This sampling noise is negligible if $n^{c-1}(x)$ and $n^c(x)$ are sufficiently large. This is generally not the case for any sequence

at the first cycle $c = 1$ where all $N = 20^4$ sequences are present in too small numbers but becomes the case at the third cycle $c = 3$ for the 100 to 1000 sequences with largest selectivities. We therefore compute $s(x)$ between the second and third cycles as $s(x) = \lambda n^3(x)/n^2(x)$ by restricting to sequences x that satisfy $n^2(x) \geq 10$ and $n^3(x) \geq 10$. Additionally, as the smallest selectivities are due to experimental noise, we retain only the sequences with $s(x) > s^*$ where s^* is determined self-consistently (Appendix 3.2 and Fig. S3). Selectivities $s(x)$ obtained by this procedure generally depend on the library (scaffold) L and the target T but are reproducible between independent experiments using the same library and the same target (Fig. S4).

To visualize the sequence dependence of selectivities, we use sequence logos [37]. In this representation, for each position i along the sequence, a bar of total height $\sum_a f_i^c(a) \ln [20 f_i^c(a)]$ is divided into letters, where each letter represents one of the 20 amino acids a with a size proportional to $f_i^c(a)$, the frequency of a at position i in the population after cycle c; for instance, $f_2^c(a) = \sum_{x_1=1}^{20} \sum_{x_3=1}^{20} \sum_{x_4=1}^{20} f^c(x_1, a, x_3, x_4)$; finally, the letters are colored by chemical properties: polar in green, neutral in purple, basic in blue, acidic in red and hydrophobic in black. It illustrates how some motifs are progressively enriched over successions of selective cycles. This representation is, however, dependent on the frequencies $f^0(x)$ of sequences in the initial population. To eliminate this dependency, we define an effective frequency $\tilde{s}_i(a)$ per position i and amino acid a as $\tilde{s}_i(a) = \sum_x s(x)\delta(x_i, a)/\sum_x s(x)$, which would correspond to the frequency of a at position i after one round of selection if all sequences x were uniformly distributed in the initial population. It can also be represented by a sequence logo but depends only on $s(x)$, as illustrated here by the Germ library selected against the DNA1 target (see Figs. S5-S7 for other cases):



SUPPLEMENTARY INFORMATION

1. Theoretical methods

1.1. Physics of selection

1.1.1. Selectivities and binding energies

When assuming that selection is controlled by equilibrium binding to the target, the distribution of selectivities is constrained by physical principles. Starting with a population of identical antibodies A and a single target T in excess relative to antibodies, $[T]_{\rm tot} \gg [A]_{\rm tot}$, the probability for an antibody to be bound to a target is

$$s_{AT} = \frac{[AT]_{\rm eq}}{[AT]_{\rm eq} + [A]_{\rm eq}} = \frac{1}{1 + K_{AT}[T]_{\rm eq}^{-1}} \simeq \frac{1}{1 + K_{AT}[T]_{\rm tot}^{-1}} \tag{5}$$

where $[AT]_{\rm eq}$ and $[A]_{\rm eq}$ are, respectively, the equilibrium concentration of bound and free antibodies and where $K_{AT} = [A]_{\rm eq}[T]_{\rm eq}/[AT]_{\rm eq}$ is the dissociation constant that characterizes the equilibrium. We used here the fact that most of the targets are unbound so that $[T]_{\rm eq} = [T]_{\rm tot} - [AT]_{\rm eq} \simeq [T]_{\rm tot}$, which is justified for our experiments where the total number of targets far exceeds the total number of antibodies, $[AT]_{\rm eq} < [A]_{\rm tot} \ll [T]_{\rm tot}$. The dissociation constant can also be written as $K_{AT} = k_-/k_+$, where k_+ and k_- denote respectively the association and dissociation rates of an antibody-target pair.

We can equivalently write

$$s_{AT} = \frac{1}{1 + {\rm e}^{\beta(\Delta G_{AT} - \mu)}} \tag{6}$$

by introducing a binding free energy $\Delta G_{AT} = \beta^{-1} \ln K_{AT}$ and a chemical potential $\mu = \beta^{-1} \ln [T]_{\rm tot}$, where β sets the energy scale [38]. This Fermi-Dirac statistics is approximated by a Boltzmann statistics

$$s_{AT} \simeq {\rm e}^{-\beta(\Delta G_{AT} - \mu)}. \tag{7}$$

when $\Delta G_{AT} \gg \mu$. This approximation is justified when $[T]_{\rm tot} \ll K_{AT}$ or, equivalently, $[AT]_{\rm eq} \ll [A]_{\rm eq}$, i.e., when the concentration of the targets or the binding affinity are sufficiently low for most of the antibodies to be unbound. Working in this regime is important for the selectivities to reflect binding free energies. Otherwise, the targets are saturating, which cause antibodies to be bound with high probability irrespectively of their dissociation constant.

These conclusions are unchanged when considering a population consisting of different antibodies A with different dissociation contants K_{AT} and binding free energies $\Delta G_{AT} = \beta^{-1} \ln K_{AT}$. In summary, when considering different antibodies A, each with its own dissociation constant K_{AT}, the choice of the target concentration $[T]_{\rm tot}$ is subject to the two constraints

$$\sum_A [A]_{\rm tot} \ll [T]_{\rm tot} \ll \min_A K_{AT}. \tag{8}$$

The first constraint $\sum_A [A]_{\rm tot} \ll [T]_{\rm tot}$ guarantees an absence of competition between antibodies so that the selectivities s_{AT} are intrinsic properties of the sequences of A, independent of the composition of the population and therefore independent of the round c when successive cycles of selection are performed; formally, $[T]_{\rm eq}$, which depends on all A present, can then be replaced by $[T]_{\rm tot}$ in Eq. (5). The second

constraint $[T]_{\rm tot} \ll \min_A K_{AT}$ guarantees that even the best binders are not in a saturation regime with $s_A \simeq 1$ independently of differences in their dissociation constants K_{AT}. In our phage display experiments, $\sum_A [A]_{\rm tot} \simeq 10^{11}\,{\rm mL}^{-1}$ and $[T]_{\rm tot} \simeq 10^{14}\,{\rm mL}^{-1}$, which satisfies the first constraint. The concentration $\sum_A [AT]_{\rm eq}$ of selected antibodies before amplification is estimated between $10^5\,{\rm mL}^{-1}$ at the first round of selection and $10^7 - 10^8\,{\rm mL}^{-1}$ at the fourth. Considering this last number to reflect properties of the best binders, we estimate that $\min_A K_{AT}/[T]_{\rm tot} \simeq \sum_A [AT]_{\rm eq}/\sum_A [A]_{\rm tot} \simeq 10^3$, which satisfies the second constraint.

1.1.2. Justification and limitations of log-normal distributions

Assuming an additive model for the interaction where the binding energy between sequence $x = (x_1, \ldots, x_\ell)$ and its target takes is of the form $\Delta G(x) = \sum_{i=1}^{\ell} \epsilon_i(x_i)$ with the $\epsilon_i(x_i)$ taking random values, the central limit theorem indicates that for sufficiently large ℓ the energies $\Delta G(x)$ are distributed normally with a mean $\mu \simeq -\ell\langle\epsilon\rangle$ and a variance $\sigma^2 \simeq \ell(\langle\epsilon^2\rangle - \langle\epsilon\rangle^2)$, where $\langle\epsilon\rangle$ and $\langle\epsilon^2\rangle - \langle\epsilon\rangle^2$ are respectively the mean and variance of the values of binding energies per position $\epsilon_i(x_i)$. Given Eq. (7), this leads to a log-normal distribution for the selectivities $s(x) \propto e^{-\beta\Delta G(x)}$.

The assumptions involved in this derivation may not be justified, starting from the assumption that selectivity can be equated to binding affinity. However, essentially all deviations from this model, sequence-dependent amplification differences, saturation of the targets, multiple binding sites or non-additive interactions, can be incorporated in a more refined model, at the expense of introducing additional parameters [39]. Deviations from a log-normal distribution of selectivities can therefore, at least in principle, be systematically analyzed and understood.

1.2. Statistics of selection

1.2.1. Extreme value statistics

Extreme value theory states that for any probability distribution $\mathbb{P}(S)$, the probability to have $S = s \geq s^\star$ conditioned to $S \geq s^\star$ converges to a generalized Pareto distribution $f_{\kappa,s^\star,\tau}(s) = \tau^{-1} f_\kappa\left((s - s^\star)/\tau\right)$ as $s^\star \to \infty$ [40], where

$$f_\kappa(x) = \begin{cases} (1+\kappa x)^{-\left(1+\frac{1}{\kappa}\right)} & \text{if} \quad \kappa \neq 0, \\ {\rm e}^{-x} & \text{if} \quad \kappa = 0. \end{cases} \tag{9}$$

The shape parameter κ is determined by the tail of the distribution of S. In particular, $\kappa < 0$ for bounded distributions and $\kappa = 0$ for distributions with exponentially decreasing tails, including log-normal distributions. On the other hand, $\kappa > 0$ for distributions whose tail decays as a power-law. For such distributions, when considering a large number N of random values $s_1 > s_2 > \cdots > s_N$, $s_r \sim s_1 r^{-\kappa}$ for $r \ll N$, which is represented in a log-log plot of s_r versus the rank r by the linear relationship $\ln(s_r/s_1) \sim -\kappa \ln r$ for the smallest values of r.

1.2.2. Effective shape parameter of log-normal distributions

In the asymptotic limit where $N \to \infty$ followed by $s^\star \to \infty$, log-normal distributions are described by a shape parameter $\kappa = 0$, but their tail decays only slowly. As a result, a large but finite number N of random values drawn from a log-normal distribution may appear to be drawn from a distribution with a non-zero shape parameter $\kappa_N \neq 0$.

More precisely, it can be shown that N values $s_1 > s_2 > \cdots > s_N$ drawn from a log-normal distribution with parameters σ, μ satisfy for $r \ll N$ the relation

$$\mathbb{E}[\ln s_r] \simeq \mu + \sigma(2\ln N)^{1/2} - \sigma(2\ln N)^{-1/2}\ln r, \tag{10}$$

which corresponds to an apparent shape parameter $\kappa_N = \sigma(2\ln N)^{-1/2}$ [26]. As κ_N vanishes only very slowly with N, it is difficult to determine whether N data points arise from a log-normal distribution or from a distribution with a shape parameter $\kappa > 0$. For instance, increasing the sample size from $N = 10^5$ to $N = 10^6$ changes κ_N by only 8 %.

Eq. (10) itself assumes that N is large enough. Numerically, we observe that for a given value of N, it breaks down when σ is below some value $\sigma^\star$. In such cases, the data may appear to arise from a bounded distribution with $\kappa_N < 0$. Fig. S14 shows the relationship between κ_N and σ obtained from numerical simulations when fixing $N = 10^4$ and $\mu = 0$, in which case $\sigma^\star \simeq 0.5$. The same relationship appears as a black dotted line in Fig. 4.

1.3. Information theory of selection

1.3.1. Relative entropies

A general statistical approach to quantify how random variables drawn from a probability P^1 are consistent with a reference probability distribution P^0 is to use their relative entropy $D(P^1\|P^0)$, also known as the Kullback-Leibler divergence [41], which is defined by

$$D(P^1\|P^0) = \sum_x P^1(x)\ln\frac{P^1(x)}{P^0(x)}. \tag{11}$$

The inverse of this quantity corresponds roughly to the number of samples required to discriminate P^1 from P^0. More precisely, the probability under P^0 of N samples drawn from P^1 scales as $e^{-ND(P^1\|P^0)}$ [41].

1.3.2. Information theory of specific interactions

The problem of quantifying specificity arises when two classes of objects or properties A and T may be associated. If this association is described by the probability $P^1(A,T)$ that A is associated with T, a natural measure of specificity is $D(P^1\|P^0)$ where $P^0(A,T)$ represents the expectation from random associations. If $P^0(A,T) = P^1(A)P^1(T)$ where $P^1(A) = \sum_T P^1(A,T)$ and $P^1(T) = \sum_A P^1(A,T)$ are the marginal distributions of A and T, $D(P^1\|P^0)$ corresponds to the mutual information $I(A;T)$ between the random variables A and T [41]. This choice of P^0, however, generally does not reflect the expectation from random associations and the relevant measure of specificity is therefore generically not captured by a mutual information but by the more general relative entropy $D(P^1\|P^0)$.

In the case of association between a set of ligands A and a set of targets T controlled by equilibrium binding, the probability $P^1(A,T)$ to find A bound to T is

$$P^1(A,T) = \frac{[AT]_{\rm eq}}{[A]_{\rm eq} + \sum_{T'}[AT']_{\rm eq}} \simeq \frac{[AT]_{\rm eq}}{[A]_{\rm eq}} = K_{AT}^{-1}[T]_{\rm eq} \simeq K_{AT}^{-1}[T]_{\rm tot} \tag{12}$$

where K_{AT} is the dissociation constant between A and T and where the approximations are justified in

Appendix 1.1. A random association is defined here by considering equal dissociation constants,

$$P^0(A,T) = \frac{[A]_{\rm tot}[T]_{\rm tot}}{\sum_{A',T'}[A']_{\rm tot}[T']_{\rm tot}}. \tag{13}$$

This distribution generally differs from $P^1(A)P^1(T)$.

A selectivity s_{AT} can be defined for each pair A, T as $s_{AT} = P^1(A,T)/P^0(A,T)$ so that

$$D(P^1\|P^0) = \left\langle \ln\left(\frac{P^1}{P^0}\right)\right\rangle_1 = \sum_{A,T} P^1(A,T)\ln\frac{P^1(A,T)}{P^0(A,T)} = \sum_{A,T} P^0(A,T)s_{AT}\ln s_{AT} = \langle s\ln s\rangle_0 \tag{14}$$

where $\langle\cdot\rangle_0$ and $\langle\cdot\rangle_1$ denote averages taken with $P^0(A,T)$ and $P^1(A,T)$ respectively.

More generally, $s_{AT} = \lambda P^1(A,T)/P^0(A,T)$ with an arbitrary multiplicative constant λ that can always be written $\lambda = \langle s\rangle_0$. This corresponds to replacing s by $s/\langle s\rangle_0$ in the previous formula,

$$D(P^1\|P^0) = \left\langle \frac{s}{\langle s\rangle_0}\ln\frac{s}{\langle s\rangle_0}\right\rangle_0 \tag{15}$$

This is equivalent to Eq. (2) where a single target T is considered and where $P^0(A,T) = 1/N$ and $P^1(A,T) = s(x)$ with x representing the sequence of A. This relationship is valid for any initial distribution $f^0(x)$ as long as $f^1(x) \propto s(x)f^0(x)$.

The identity

$$D(s\|N^{-1}) = \left\langle \frac{s}{\langle s\rangle}\ln\frac{s}{\langle s\rangle}\right\rangle \tag{16}$$

where averages $\langle\cdot\rangle$ are now taken with a distribution $P(s)$ of the selectivities over the different sequences x is, however, valid only when considering as initial distribution a uniform distribution over the sequences. The notation $D(s\|N^{-1})$ assumes, besides, that $\sum_x s(x) = 1$ so that $s(x)$ can be interpreted as a probability distribution.

1.3.3. Equivalence with the parameter σ

If further assuming that $P(s)$ is a log-normal distribution with parameters σ and μ, $\langle s\rangle = e^{\mu+\sigma^2/2}$ and $\langle s\ln s\rangle = \langle s\rangle(\mu+\sigma^2)$ so that

$$D(s\|N^{-1}) = \frac{\sigma^2}{2} \tag{17}$$

irrespectively of the value of μ. This reflects the fact that specificity quantifies only relative differences in binding free energies between different ligands.

A previous study proposed the mutual information as a measure of specificity [27]. It is justified, however, only within the special model considered in [27] where, because of the overall symmetry of the interactions between the M locks A and M keys T, $P^1(A) \simeq P^1(T) \simeq 1/M$, and therefore $P^0(A,T) = 1/M^2 \simeq P^1(A)P^1(T)$.

1.4. Dynamics of selection

1.4.1. Recursion for the sequence frequencies

If $N^c(x)$ denotes the number of copies of sequence x at cycle c, the dynamics of selection satisfies the recursion

$$N^c(x) = \alpha_c s(x) N^{c-1}(x) \tag{18}$$

where α_c represents an amplification factor to reach at every round the same total population size N, i.e., $\sum_x N^c(x) = N$ independent of c. In terms of frequencies $f^c(x) = N^c(x)/N$, this gives $\alpha_c = (\sum_x s(x) f^c(x))^{-1}$ and

$$f^c(x) = \frac{s(x) f^{c-1}(x)}{\sum_{x'} s(x') f^{c-1}(x')} = \frac{(s(x))^c f^0(x)}{\sum_{x'} (s(x'))^c f^0(x')}. \tag{19}$$

These recursions assume a large N, so that the frequencies $f^c(x) = N^c(x)/N$ are meaningful; in particular, they assume that no sequence disappears.

Note the similarity with a Boltzmann distribution with the cycle c playing the role of an inverse temperature.

1.4.2. Recursion for the library frequencies

When considering a population consisting of an equal mix of different libraries L, the frequency $f^c(L) = \sum_{x \in L} f^c(x)$ of library L satisfies the recursion

$$f^c(L) = \frac{\langle s^c \rangle_L}{\sum_{L'} \langle s^c \rangle_{L'}} \tag{20}$$

with

$$\langle s^c \rangle_L = \sum_{x \in L} (s(x))^c f^0(x) = \int_0^\infty \mathrm{d}s \, P_L(s) s^c = \exp\left(c\mu_L + \frac{c^2 \sigma_L^2}{2} \right). \tag{21}$$

Here, the first equality defines the average $\langle \cdot \rangle_L$ within each library L. The second equality, on the other hand, makes two assumptions: first, that selectivities s within library L are described by a distribution of selectivities $P_S(s)$ and, second, that sequences within a library are uniformly represented in the initial population. The third equality makes the additional assumption that $P_L(s)$ is a log-normal distribution with parameters σ_L and μ_L.

Under these different assumptions, the frequency of library L at cycle c is given by

$$f^c(L) = \left(\sum_{L'} \mathrm{e}^{c(\mu_{L'} - \mu_L) + c^2 \left(\sigma_{L'}^2 - \sigma_L^2 \right)/2} \right)^{-1}. \tag{22}$$

This shows that for small c, the dynamics is controlled by the μ_L, with in limit $c \to 0$, $(f^c(L) - f^0(L))/f^0(L) \simeq c(\mu_L - \langle \mu \rangle)$, i.e., at the first cycle, the frequency of library L increases if its μ_S exceeds the average $\langle \mu \rangle$ across libraries and it decreases otherwise. For large c, on the other hand, the dynamics is controlled by the σ_Ls with $f^c(L) \to 1$ for the library L that has largest σ_L, regardless of the values of μ_L.

These calculations rely on several assumption, in particular the assumption that sequences within a library have initially uniform frequencies, which is not satisfied in the experiments. This explains the differences between the model and the data in Fig. 3.

2. Experimental methods

Experimental methods are as in our previous work [14], except for target immobilization and sequencing data analysis as summarized below.

2.1. Phage production

Production of antibody-displaying phage was performed through infection of library cells (TG1 strain) with M13KO7 helper phage and growth at 30°C for 7 h in selective 2xYT medium containing 100 μg/mL ampicillin (Sigma-Aldrich, Saint-Louis, MO, USA). Cells were then centrifuged and the supernatant containing displaying phages was kept and stored at 4°C overnight. All selections were performed on the day immediately following the phage production step.

2.2. Target immobilization

Target molecules were immobilized on streptavidin-coated magnetic beads (Dynabeads(R) M-280 Streptavidin) purchased from Invitrogen Life Technologies (Carlsbad, CA, USA). The DNA hairpin targets (DNA1 and DNA2) in fusion with a biotin at their 5′ end were purchased from IDT (Leuven, Belgium) diluted in MilliQ water and stored at −20°C. The genes of protein targets (eGFP and mCherry, corresponding respectively to PDB IDs 2Y0G and 2H5Q) in fusion with a SBP tag were kindly provided by Sandrine Moutel (Institut Curie, Paris, France). They were produced in liquid T7 Express *E. Coli* cultures induced at $OD_{600} = 0.5$ with 300 μM Isopropyl β-D-1-thiogalactopyranoside (IPTG, Sigma-Aldrich, Saint-Louis, MO, USA) final and incubated overnight at 30°C. The proteins were harvested by threefold flash freezing in liquid nitrogen and quick thawing in a water bath at 42°C, followed by incubation with 50 μg/mL lyzozyme final and 2.5 U/mL DNase I final at 30°C for 15 minutes and centrifugation at 15,000 g and 4°C for 30 minutes. The supernatant was aliquoted in protein low-bind tubes (Protein LoBind, Eppendorf, Hamburg, Germany), flash frozen in liquid nitrogen and stored at −80°C until use.

Binding of target molecules to streptavidin-coated magnetic beads was performed in DNA low-bind tubes (DNA LoBind tubes, Eppendorf, Hamburg, Germany) for the DNA targets or protein low-bind tubes (Protein LoBind tubes, Eppendorf, Hamburg, Germany) for the protein targets. Beads and targets were incubated in 0.5x PBS at ambient temperature on a rocker for 15 min, followed by removal of all liquid and 3 washing steps: addition of 500 μL washing solution, vortexing, separation of beads using a magnet and removal of all liquid. Finally, the beads were stored in washing buffer at 4°C for use on the following day. Bw1X buffer (1 M NaCl, 5 mM Trizma at pH = 7.4, 0.5 mM EDTA) was used as washing buffer for DNA targets (to screen electrostatic interactions), 1x PBS with 0.1 % Tween20 for protein targets (to screen hydrophobic interactions). The same procedure was followed for negative/null selection tubes, with MilliQ water instead of target solutions.

Successful immobilization of protein targets was confirmed by fluorescence measurements of treated beads against untreated and MilliQ water-treated beads as negative controls.

2.3. Phage display selection

The selection protocol is as previously published in [14]. The washing buffer was removed from the target-covered beads. Then, 1 mL of culture supernatant from the phage production step containing $\approx 10^{11}$ phages was added to the negative selection tube (containing no targets) and incubated for 90 minutes at

ambient temperature, shaking. The beads were separated by a magnet and the liquid was transferred to the positive selection tube (containing the targets) and incubated for 90 minutes at room temperature, shaking. Finally, all liquid containing unbound phage was removed and the beads were subjected to a 10-fold washing using 10 mL of 1x PBS with 0.1 % Tween20. Bound phage were eluted from beads with 1.4 % triethylamine (Sigma-Aldrich, Saint-Louis, MO, USA) in MilliQ water and used for infection of fresh exponential TG1 cells to obtain the selected library.

2.4. Illumina sequencing

Glycerol stocks of library cells at relevant selection cycles were defrosted and plasmids were extracted using purification kits from Macherey-Nagel (Düren, Germany). No liquid culture was performed prior to plasmid extraction to avoid potential additional biases from growing an overnight culture beforehand. Resulting plasmids were used as input for Illumina sequencing preparation PCR: a first reaction using primer sequences common to all three libraries downstream CDR_3 (`GCTCGAGACGGTAACCAGG`, forward) and halfway inside V_H (`ACAACCCGTCTCTTAAGTCTCGT`, reverse) added random barcodes of length 5 nt to discriminate between neighboring clusters. A second reaction added P5 and P7 indices to identify library, target and selection round corresponding to each cluster, as well as the adapter for the sequencing procedure. Illumina sequencing and demultiplexing were performed at I2BC, Gif-sur-Yvette, France.

3. Data analysis

3.1. Preprocessing

The Illumina sequencing yields for each sample (i.e., each library, target and selection round) between 10^5 and 5.10^6 sequencing clusters. The data files contain the entirely overlapping forward and reverse reads for all clusters of a given sample. Each cluster was accepted or discarded based on the following procedure: Both the forward and reverse reads were screened for the presence of the primer sequences (up to 4 nt mismatch accepted for each) and cut to keep only the part between the primers (including the primers). Either one was discarded if the primer search was unsuccessful. We then checked if the remaining forward and/or reverse sequence fragments have the expected length of 170 nt, corresponding to the region of interest. If only one direction had the expected length, only this direction was kept and the other one was discarded. If both directions did not have expected length, the complete cluster was discarded. Finally, if both reads had expected length, a consensus sequence was generated by taking on each position with disagreement between both reads the nucleotide measured with highest quality read. A final check was performed for (i) a sufficient average quality read over the whole region of interest ($\langle Q \rangle \geq 59$) and (ii) the restriction sites immediately up- and downstream CDR_3 (`TGTGCGCGC` and `TTCGACTAC`) are located at their expected positions (108-116 and 129-137 in reverse direction; up to 4 nt mismatch accepted for each). The cluster was discarded if either of these two criteria was not fulfilled.

After completion of this procedure, (i) the framework (Germ, Lim or Bnab) and (ii) the CDR3 sequence for all remaining sequencing reads in the full-library experiments were identified. Step (i) was performed by measuring the Hamming distance of the visible library-specific framework part upstream the CDR3 of the read (of length 116 nt) to all three framework reference sequences. The read was assigned to the nearest framework if the Hamming distance to the nearest framework was ≤ 7 nt *and* the difference in Hamming distance to the nearest and next-nearest frameworks was ≥ 3 nt. For step (ii), the CDR3 sequence was

simply extracted from the read for the full-library experiments. For the selections with reduced diversity a similar procedure as for the framework part was applied: the measured CDR3 sequence was assigned to the nearest among ~ 20 reference sequences if the Hamming distance was $\leq 3\,\mathrm{nt}$ and the difference in Hamming distance between nearest and next-nearest was $\geq 1\,\mathrm{nt}$. After assessment of the sequence identity of all clusters in a dataset, the CDR3 sequences were translated into amino acids and the number of occurrences of each clone (determined by its framework and its CDR3 sequence) was counted.

The nucleotide sequences of the visible framework parts upstream the CDR3 of all three libraries as well as the Hamming distances d_H between the pairs is as follows:

Germ:

`ACAACCCGTCTCTTAAGTCTCGTGTTACCATCTCTGTTGACACCTCTAAAAACCAGTT...`

`CTCTCTGAAACTGTCTTCTGTTACTGCGGCGGACACTGCGGTTTACTACTGTGCGCGC`

Lim:

`ACAACCCGTCTCTTAAGTCTCGTGTTACCATCTCTATCGACACCTCTAAAAACCACTT...`

`CTCTCTGCGTCTGATCTCTGTTACTGCGGCGGACACTGCGGTTTACCACTGTGCGCGC`

Bnab:

`ACAACCCGTCTCTTAAGTCTCGTCTGACCCTGGCGCTGGACACCCCGAAAAACCTGGT...`

`TTTCCTGAAACTGAACTCTGTTACTGCGGCGGACACCGCGACCTACTACTGTGCGCGC`

$d_H(\mathrm{Germ}, \mathrm{Lim}) = 10\,\mathrm{nt}$, $d_H(\mathrm{Lim}, \mathrm{Bnab}) = 25\,\mathrm{nt}$ and $d_H(\mathrm{Germ}, \mathrm{Bnab}) = 22\,\mathrm{nt}$.

For the mixed full-library selections, final data files contain three columns: 1) framework identity ('`germ`' for Germline, '`lmtd`' for Limited, '`bnAb`' for Bnab, '`????`' if framework inference failed), 2) CDR3 identity given by the sequence of 4 amino acids or the sequence of 12 nucleotides or by '`????`' if the CDR3 readout failed, 3) number of occurrences in the dataset. The preprocessed data from the experiments reported in this paper is made available in this format.

We checked that the results are unaffected by the choice of the parameters in the preprocessing procedure described here.

3.2. Noise cleaning

Selectivities are computed from sequencing counts as indicated in Eq. 4. To account for sampling noise, only sequences whose count is ≥ 10 both at round c and $c{+}1$ are considered. Moreover, we ignore selectivities $s(x)$ below a threshold $s^\star$, which arise from unspecific binding. Unspecific binding modifies the expression for the selectivity of sequence x to include a sequence-independent unspecific binding energy ΔG_{us},

$$s(x) = \frac{\mathrm{e}^{-\beta\Delta G(x)} + \mathrm{e}^{-\beta\Delta G_{\mathrm{us}}}}{1 + \mathrm{e}^{-\beta\Delta G(x)} + \mathrm{e}^{-\beta\Delta G_{\mathrm{us}}}}. \tag{23}$$

It sets a lower bound for the selectivity given by

$$s_{\mathrm{us}} = \frac{\mathrm{e}^{-\beta\Delta G_{\mathrm{us}}}}{1 + \mathrm{e}^{-\beta\Delta G_{\mathrm{us}}}} = \frac{1}{1 + \mathrm{e}^{\beta\Delta G_{\mathrm{us}}}}. \tag{24}$$

The argument for log-normality of selectivity distributions applies only when the specific binding contribution $\Delta G(x)$ dominates the selectivity. We therefore eliminate the selectivities dominated by unspecific binding.

This is done by introducing a cut-off s^*. The choice is made such that (i) the values of the inferred parameters $\hat{\sigma}$ and $\hat{\mu}$ are approximately constant for all $s \geq s^*$ and (ii) s^* is large enough to eliminate

selectivities due to unspecific binding. This last condition is implemented by plotting the counts $n^2(x)$ and $n^3(x)$ at the two successive cycles, as illustrated in Figure S3: sequences with $s = s_{\mathrm{us}}$ appear in the diagonal with a variance that decreases with increasing counts, as expected from sampling noise, and s^* is chosen so as to exclude these sequences. In cases where specific binding to the target is very strong, sequences selected for unspecific binding are not present (Fig. S15A), while in cases where specific binding is too weak, only sequences selected for unspecific binding are present (Fig. S15F).

The same criteria apply when fitting to generalized Pareto distributions to infer the parameter κ but criterion (i) may lead to a higher value of s^* if the measured selectivities extend beyond the tail of the distribution. In our previous work [14], we only considered criterion (i). In one case (Frog3 against DNA1), the s^* that we define here by accounting for (ii) differs from the s^* that had previously defined (Fig. S15), which leads to a significantly different estimation of κ: $\hat{\kappa} = -0.53 \pm 0.19$ instead of $\hat{\kappa} = 0.97 \pm 0.38$. In the other cases, we recover essentially the same results. The new analysis provides, however, additional insights; in the case of Frog3 against PVP, it thus appear that the vanishing value of κ can be attributed to the selectivities being dominated by unspecific binding (Fig. S15).

3.3. Fit to log-normal distributions

To infer from experimental data the parameters σ and μ of a log-normal distribution, as given by Eq. (1), we focus on the best available selectivities $s_i > s^*$, the log-normal distribution is under-sampled. In practice, it is more convenient to work with the log of the selectivities, $y_i = \ln s_i$, and to fit them with a normal distribution. If restricting to values y_i larger than a given threshold y^*, the probability $\mathbb{P}[Y = y|Y \geq y^*]$ of observing y_i given that $y_i \geq y^*$ is

$$\mathbb{P}[Y = y|Y \geq y^*] = \frac{\mathbb{P}[Y = y]}{\mathbb{P}[Y \geq y^*]} = \sqrt{\frac{2}{\pi}} \frac{\mathrm{e}^{-\frac{(y-\mu)^2}{2\sigma^2}}}{\sigma \left[1 - \mathrm{erf}\left(\frac{y^*-\mu}{\sqrt{2}\sigma}\right)\right]}, \tag{25}$$

where $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{e}^{-\xi^2} \mathrm{d}\xi$ is the Gauss error function. The log-likelihood $\mathcal{L}(\mu, \sigma, y^*)$ then verifies

$$-\frac{1}{N}\mathcal{L}(\mu, \sigma, y^*) = -\frac{1}{N}\sum_{i=1}^{N} \ln \mathbb{P}[Y = y_i|Y \geq y^*] = \ln(\sigma) + \ln\left[1 - \mathrm{erf}\left(\frac{y^* - \mu}{\sqrt{2}\sigma}\right)\right] + \frac{1}{2\sigma^2 N}\sum_{i=1}^{N}(y_i - \mu)^2, \tag{26}$$

up to irrelevant additive constants independent of the parameters μ and σ. For a given y^*, we minimize this quantity with respect to the parameters σ and μ to obtain $\hat{\sigma}(y^*)$ and $\hat{\mu}(y^*)$ and then chose y^* such that for any $y \geq y^*$ both $\hat{\sigma}(y)$ and $\hat{\mu}(y)$ are nearly constant (criterion (i) in Appendix 3.2). Finally, we obtain a lower bound on the uncertainty of the parameter values using the Fisher information matrix and the Cramer-Rao bound. To assess the quality of fit, we produce P-P plots comparing the cumulative distribution of data to

$$z = F(y|y^*) = \mathbb{P}[Y \geq y|Y \geq y^*] = \frac{\mathrm{erf}\left(\frac{y-\mu}{\sqrt{2}\sigma}\right) - \mathrm{erf}\left(\frac{y^*-\mu}{\sqrt{2}\sigma}\right)}{1 - \mathrm{erf}\left(\frac{y^*-\mu}{\sqrt{2}\sigma}\right)} \tag{27}$$

where z is the fraction of the data above $y \geq y^*$ according to the model, and Q-Q plots comparing the data to the inverse distribution function $y = F^{-1}(z|y^*)$.

3.4. Normalization of μ across libraries

The selection of a library L against a target T yields only the values of the highest selectivities $s(x)$ up to an unknown multiplicative constant λ (see Box). The parameter $\sigma = \sigma_{L,T}$ is independent of λ but not the parameter $\mu = \mu_{L,T}$. The relative values of $\mu_{L,T}$ for different libraries L selected against the same target T are determined by performing selections where the different libraries are mixed in the initial population: this leaves undetermined one overall multiplicative constant per target. Finally, we fix them by setting $\mu_{\mathrm{Germ},T} = 0$ for each target T. In practice, inferring μ from the tail of $P(s)$ is challenging, even more so when different libraries are mixed, as one library often dominates the population after a few cycles. To overcome this limitation, we can separately measure the selectivities of random sequences, which typically belong to the mode of the distribution $P(s)$, located at $m = \mu - \sigma^2$.

For a given target, our approach is thus to first perform 3 cycles of selection with each library, Germ, Lim and Bnab. Using the results from cycles 2 and 3, we estimate as many selectivities $s_{L,T}(x)$ as possible (see Box and Fig. 1A). We then identify 2 to 4 sequences with largest selectivity from each library, which we mix with 2 to 4 random sequences from each library, and perform one round of selection of the mixture of these ~ 20 sequences. From the results of this experiment, we estimate with high precision the relative selectivities of top and typical sequences from the different libraries (Fig. 1B). We typically find that the random sequences from a same library have a similar selectivity which we use to define the relative modes $m_{L,T}$ of the 3 libraries. Given these modes $m_{L,T}$, we then infer from the available values of $s_{L,T}(x)$ the parameter $\sigma_{L,T}$ by maximum likelihood, using the relationship $\mu_{L,T} = m_{L,T} + \sigma^2_{L,T}$. Finally, we fix the remaining overall multiplicative constant by setting $\mu_{\mathrm{Germ},T} = 0$.

In practice, to reduce the total number of experiments, we performed the selection of the full libraries in mixtures; as we verified with one target, the results are equivalent to those obtained from separate selections (Fig. S8). We also found unecessary to estimate the selectivities of typical sequences against all targets once we understood that these values are not controlled by the target.

		Mix3 (rounds 2, 3)				Mix3 (rounds 3, 4)		separate		Mix21 or Mix24	
		σ	κ	μ	τ	σ	κ	σ	κ	σ	κ
Germ	DNA1	1.50 ± 0.23	0.68 ± 0.12	0.00 ± 0.61	2.12 ± 0.27	1.38 ± 0.13	0.49 ± 0.11	1.27 ± 0.07	0.27 ± 0.14	1.07 ± 0.10	0.27 ± 0.11
	DNA2	1.16 ± 0.13	0.41 ± 0.11	0.00 ± 0.22	1.22 ± 0.17			1.16 ± 0.20	0.51 ± 0.23		
	prot1	1.44 ± 0.18	0.41 ± 0.20	0.00 ± 0.46	8.75 ± 2.07						
		1.50 ± 0.17	0.71 ± 0.12	0.00 ± 0.30	1.42 ± 0.19	1.13 ± 0.09	0.40 ± 0.13				
	prot2	1.40 ± 0.22	0.71 ± 0.27	0.00 ± 0.41	2.97 ± 0.88	1.07 ± 0.14	0.29 ± 0.13				
		1.31 ± 0.23	0.70 ± 0.24	0.00 ± 0.39	1.45 ± 0.38						
Lim	DNA1	0.56 ± 0.05	-0.68 ± 0.10	1.27 ± 0.06	4.40 ± 0.56	N/A	N/A	0.98 ± 0.31	0.08 ± 0.34	N/A	N/A
	DNA2	0.55 ± 0.04	-0.33 ± 0.06	0.93 ± 0.06	2.36 ± 0.22			N/A	N/A		
	prot1	0.73 ± 0.18	0.01 ± 0.19	1.03 ± 0.33	2.74 ± 0.67						
		0.66 ± 0.13	-0.40 ± 0.24	0.05 ± 0.16	1.01 ± 0.33	N/A	N/A				
	prot2	1.13 ± 0.50	0.38 ± 0.22	0.33 ± 1.34	2.15 ± 0.60	N/A	N/A				
		0.97 ± 0.22	0.27 ± 0.23	0.12 ± 0.29	1.22 ± 0.37						
Bnab	DNA1	0.55 ± 0.08	-0.22 ± 0.08	2.24 ± 0.12	8.09 ± 1.29	0.50 ± 0.03	-0.09 ± 0.08	N/A	N/A	N/A	N/A
	DNA2	0.41 ± 0.06	-0.48 ± 0.14	2.07 ± 0.08	3.55 ± 0.74			N/A	N/A		
	prot1	0.45 ± 0.05	-0.52 ± 0.12	3.03 ± 0.07	21.98 ± 3.67						
		0.45 ± 0.05	-0.52 ± 0.12	1.51 ± 0.07	4.82 ± 0.80	0.45 ± 0.05	0.31 ± 0.23				
	prot2	0.67 ± 0.11	-0.41 ± 0.13	2.55 ± 0.14	16.00 ± 3.46	0.57 ± 0.04	-0.05 ± 0.08				
		0.59 ± 0.09	-0.17 ± 0.12	1.77 ± 0.12	5.63 ± 1.04						
Chicken1	DNA1									0.64 ± 0.07	-0.56 ± 0.13
NurseShark1	PVP									1.22 ± 0.45	0.52 ± 0.20
Frog3	DNA1							0.61 ± 0.13	-0.53 ± 0.19	N/A	N/A
Frog3	PVP								0.04 ± 0.05	N/A	N/A

Table 1: Parameters obtained from fits of the distribution of selectivities to generalized Pareto distributions (κ, τ) and log-normal distributions (σ, μ) for experiments presented here and in our previous work [14]. N/A indicates that the data was insufficient to make a meaningful fit. For selectivities against the protein targets between rounds $c = 2$ and $c + 1 = 3$, values are given for two independent replica of the experiment. The given uncertainties correspond to a single standard deviation around the maximum likelihood estimate as given by the Cramer-Rao bound. In the case of Frog3 against DNA1, and only in this case, the value of κ differs from the one reported in our previous work [14] for reasons explained in Appendix 3.2 and Figure S15.

SUPPLEMENTARY FIGURES

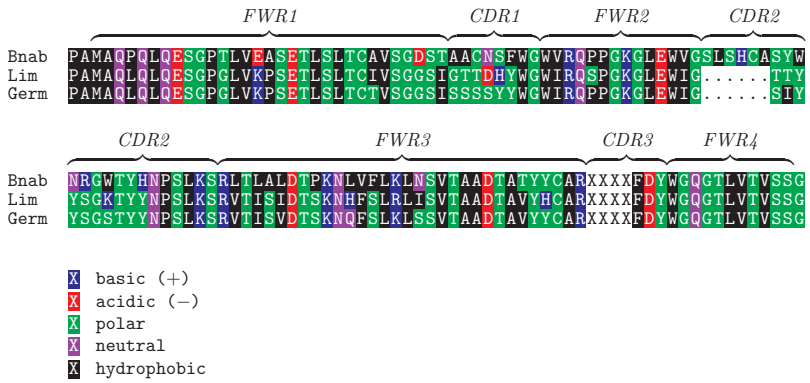


Figure S1: Alignment of the sequences of the three scaffolds, Bnab, Lim and Germ. The 4 randomized positions correspond to the part of the CDR3 indicated by `XXXX`.

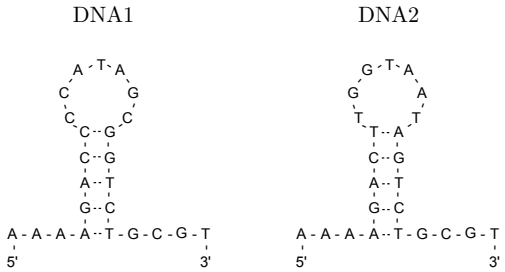


Figure S2: DNA1 and DNA2 binding targets. The targets display a hairpin structure at room temperature. They share a common stem sequence but the sequence of their loop differ. A biotin is placed at the 5' ends to allow for immobilization on streptavidin-coated magnetic beads.

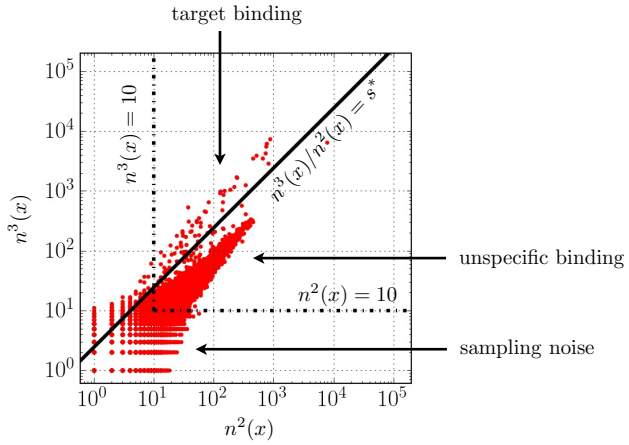


Figure S3: Illustration of the choice of the cutoff s^* below which measured selectivities are attributed to unspecific binding. The number $n^3(x)$ of counts in the sequencing data at round $c = 3$ is plotted against the number $n^2(x)$ of counts at round $c - 1 = 2$ for a selection of the Bnab library mixed with the two other libraries against the DNA1 target. An accumulation of sequences with similar selectivities is observed along the diagonal, with larger variance for smaller values as expected from an increased sampling noise. This is interpreted as arising from unspecific binding, associated with a selectivity s_{us} independent of the sequence. We define a cut-off s^* such that sequences x with $s = n^3(x)/n^2(x) \geq s^*$ cannot be attributed to unspecific binding. In addition, we restrict to sequences x with $n^2(x) \geq 10$ and $n^3(x) \geq 10$, as represented by the vertical and horizontal lines, to ensure that the inferred selectivities are not dominated by sampling noise.

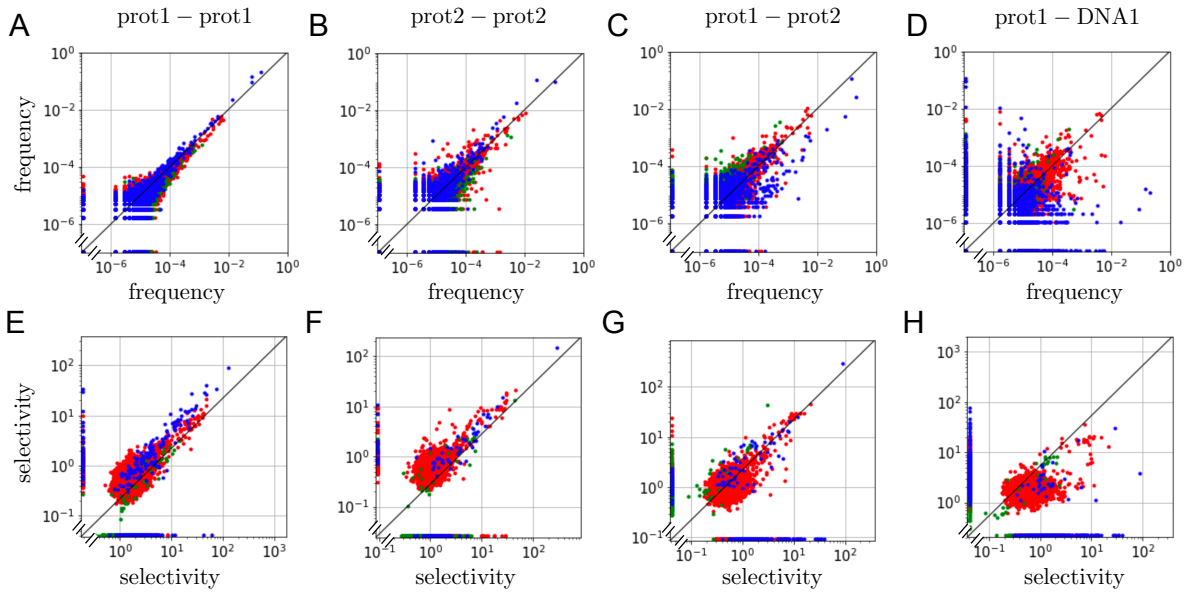


Figure S4: Comparisons between results of replicate and non-replicate experiments. **A.** Comparison of the frequencies $f^3(x) = n^3(x)/\sum_{x'} n^3(x')$ computed after the third cycle ($c = 3$) between two independent replicate experiments where a mixture of the Germ (in blue), Lim (in green) and Bnab (in red) libraries is selected against the protein target prot1. Due to stochastic sampling, some sequences x are well represented in one experiment ($n^3(x) \geq 10$) but not in the other; they are represented by the points along the two axes. As expected, the frequencies of the most prevalent sequences are the most reproducible. **B.** As in A but for protein target prot2. **C.** Comparing an experiment with prot1 as target with another with prot2 as target: common sequences are enriched in the two cases, although with not exactly the same frequencies. **D.** Comparing an experiment with prot1 as target with another with DNA1 as target, showing that different sequences are enriched in each case. In particular, the most frequent sequences when selecting against one target are absent in the third round when selecting against the other (points along the axes). **E,F,G,H.** Comparison of selectivities $s(x)$ calculated from the frequencies between the second and third rounds as $s(x) = \lambda n^3(x)/n^2(x)$. Points along the axes correspond to sequences for which the selectivity could be estimated only for one of the two experiments. We verify that in cases E,F,G where the targets are similar the same top selectivities are recovered (up to a multiplicative constant corresponding to a shift in log-log plots). Beyond stochastic effects, reproducibility is mainly limited by the differences in the production of the targets, as shown in Fig. S12.

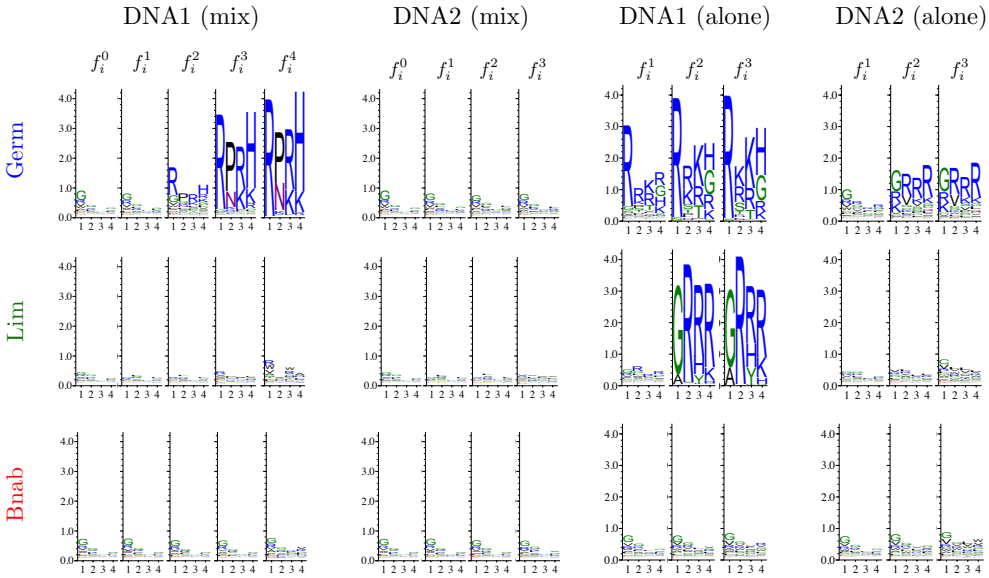


Figure S5: Extension of the figure in the Box to the 3 libraries Germ, Lim, Bnab selected either in a mixture (mix) or on their own (alone) against the DNA1 and DNA2 targets. The sequences logos represent the frequencies $f_i^c(a)$ of amino acids at each successive cycle $c = 0, 1, 2, 3, 4$.

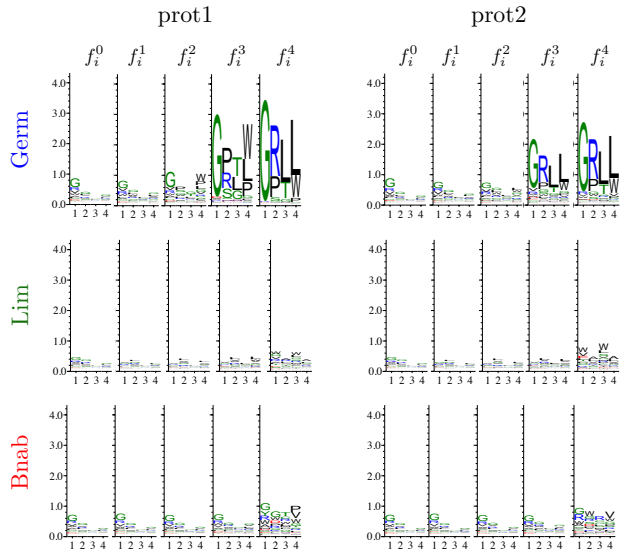


Figure S6: Extension of the figure in the Box to the 3 libraries Germ, Lim, Bnab selected in mixture against the prot1 and prot2 targets. The sequences logos represent the frequencies $f_i^c(a)$ of amino acids at each successive cycle $c = 0, 1, 2, 3, 4$.

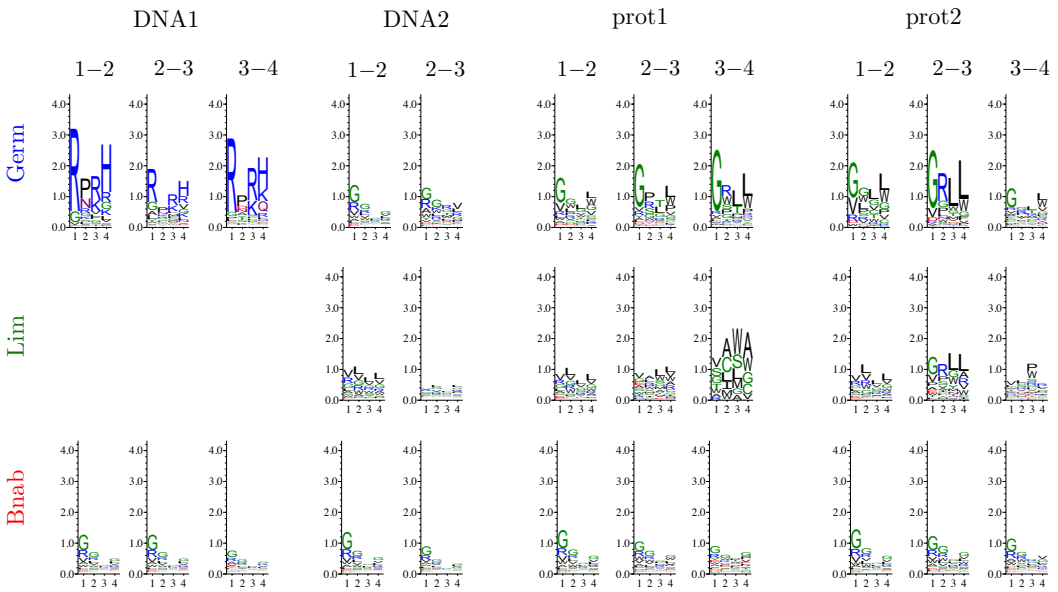


Figure S7: Sequence logos for the selectivities $\tilde{s}(x)$ computed between two successive rounds (1-2, 2-3 or 3-4). The differences between rounds reflect sampling fluctuations.

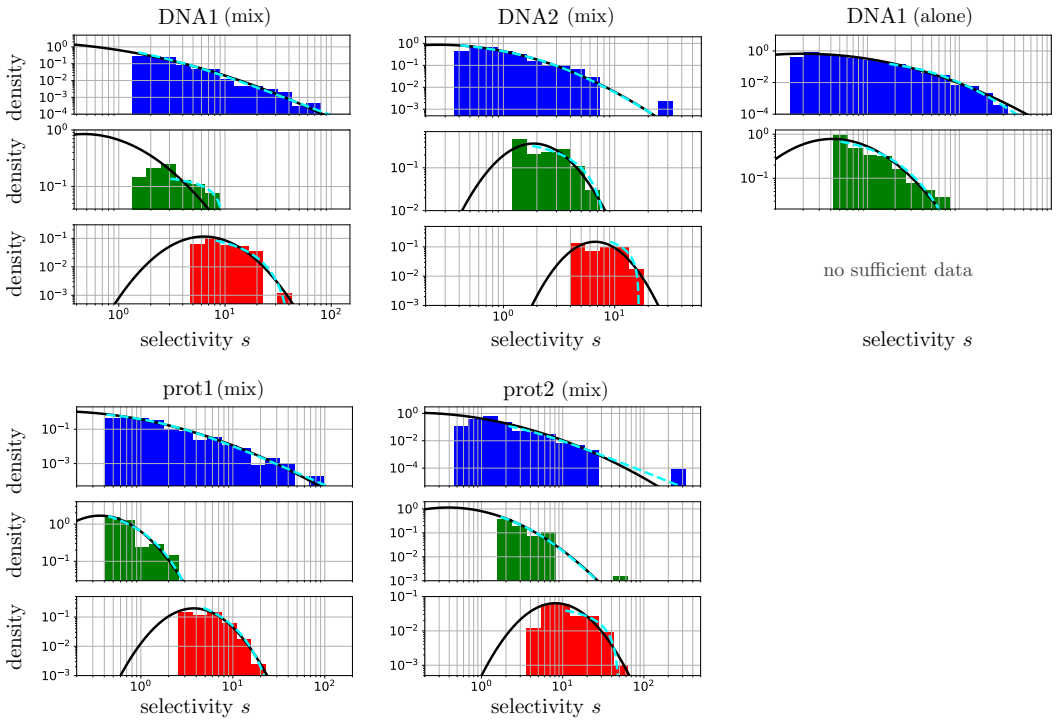


Figure S8: Distributions of selectivities of the three libraries (Germ in blue, Lim in red, Bnab in red) when selected either in a mixture (mix) or on their own (separate) against the different targets. This figure extends Fig. 1A that reports the selection against the DNA1 target of the Germ and Bnab libraries in mixture and of the Lim library on its own. In addition to the best fits to a log-normal distribution (black curves), the best fits to generalized Pareto distributions are also shown (cyan dotted curves). The selection of the Bnab library alone against the DNA1 target yielded insufficient data for a meaningful analysis.

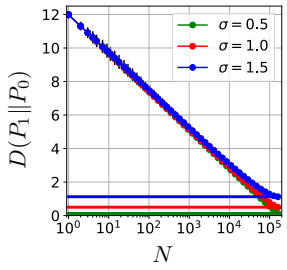


Figure S9: How the estimation of the entropy is biased by finite sampling. 10^5 values were drawn from a log-normal distribution with parameters $\mu = 0$ and $\sigma = 0.5$ (green), 1 (red) and 1.5 (blue). The relative entropy $D(P_1\|P_0)$ was then estimated using a random subsample of size N. For any $N < 10^5$, this leads to an overestimation of $D(P_1\|P_0)$ whose actual value $\sigma^2/2$ (see Eq. (3)) is represented by the horizontal lines at the bottom.

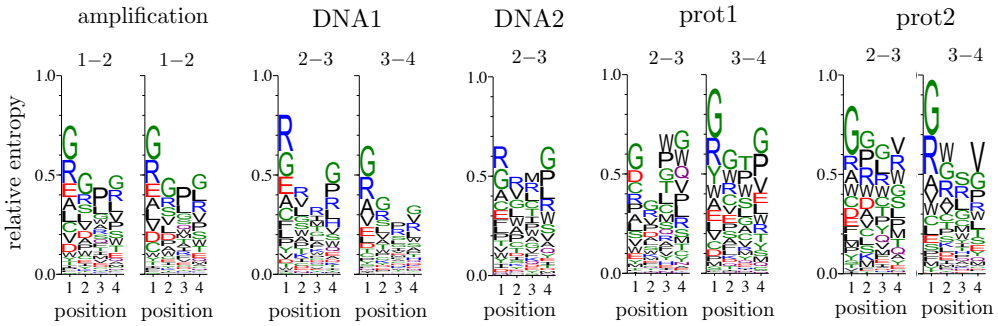


Figure S10: Sequence logos for the selectivities $\tilde{s}_i(a)$ of the Bnab library subject to either amplification only or to amplification and selection for binding against the DNA1, DNA2, prot1 or prot2 targets. The selectivities are computed between the first and second cycles (1-2) or between the third and fourth cycles (3-4); for amplification only, the results of two replicate experiments are shown. The sequence logos of selectivities calculated between rounds 2 and 3 are the same as those shown in Fig. 2 (Bnab library), except for the scale along the y-axis. All sequences logos share common patterns reflecting a common contribution from amplification biases. Sequence logos against the protein targets show, however, an enrichment for tryptophane (symbol `W`) that is not observed when selection involves amplification only. Selections of the Bnab library thus have a target-dependent contribution from binding affinity of similar order of magnitude as a common target-independent contribution from amplification biases.

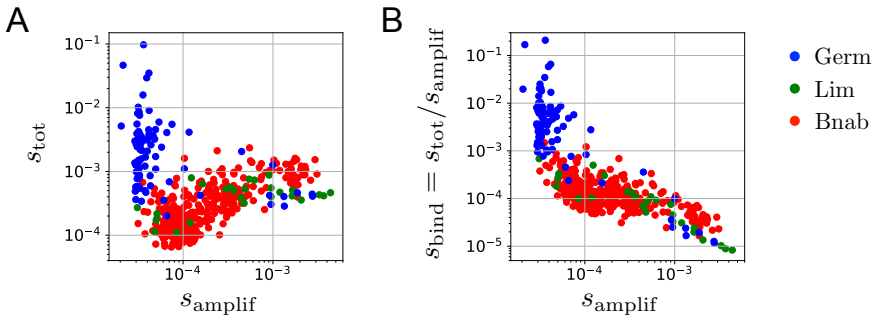


Figure S11: Contribution of amplification biases to the selectivities in selection against the DNA1 target. A separate experiment without any selection for binding was performed to estimate the difference of selectivities arising from the amplification step alone. **A.** The resulting $s_{\rm amplif}$ is here compared to the selectivities $s_{\rm tot}$ from an experiment including a selection for binding. The sequences with top $s_{\rm tot}$, which all belong to the Germ library (in blue), are among the sequences with lowest $s_{\rm amplif}$, which indicate that they are selected for binding with no contribution from the amplification bias. On the other hand, the sequences with top $s_{\rm tot}$ from the Lim and Bnab libraries (respectively in green and red), have also top $s_{\rm amplif}$, which indicate a significant contribution from amplification biases. **B.** The ratio $s_{\rm tot}/s_{\rm amplif}$ represents the contribution to selectivity of binding alone. The two selective pressures, binding and amplification, appear here to be orthogonal.

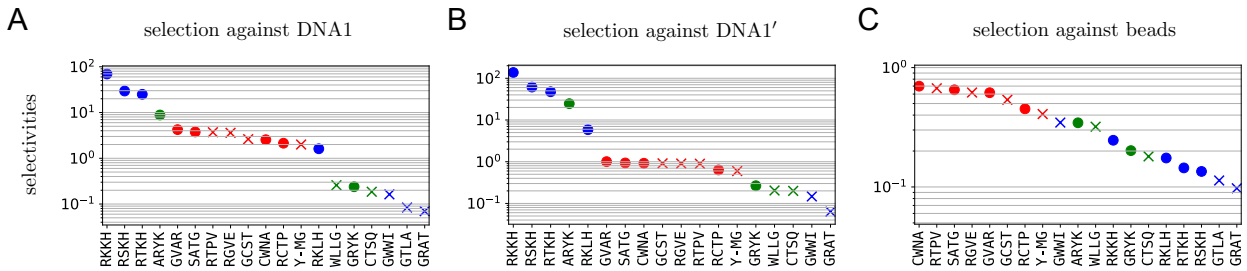


Figure S12: Supplementary experiments with minimal libraries. **A.** Selectivities of top and random sequences from the three libraries, Germ (in blue), Lim (in green) and Bnab (in red), against DNA1. This graph is identical to Fig. 1B. **B.** Results from a replicate experiment using a different stock of beads, showing that the selectivities are reproduced except for the Bnab sequences (in red), which have a systematically higher selectivity. **C.** Similar to A, but when selecting for binding to the beads in absence of the DNA1 target. The top selectivities are from the Bnab sequences (in red), indicating that they bind to the beads, a finding consistent with the discrepancy between A and B. Here, the differences in selectivities are also coming from differences of selectivity during amplification (Fig. S11). Consistent with Fig. S11, the top Germ sequences (blue dots) have in absence of the DNA1 target the worst selectivities.

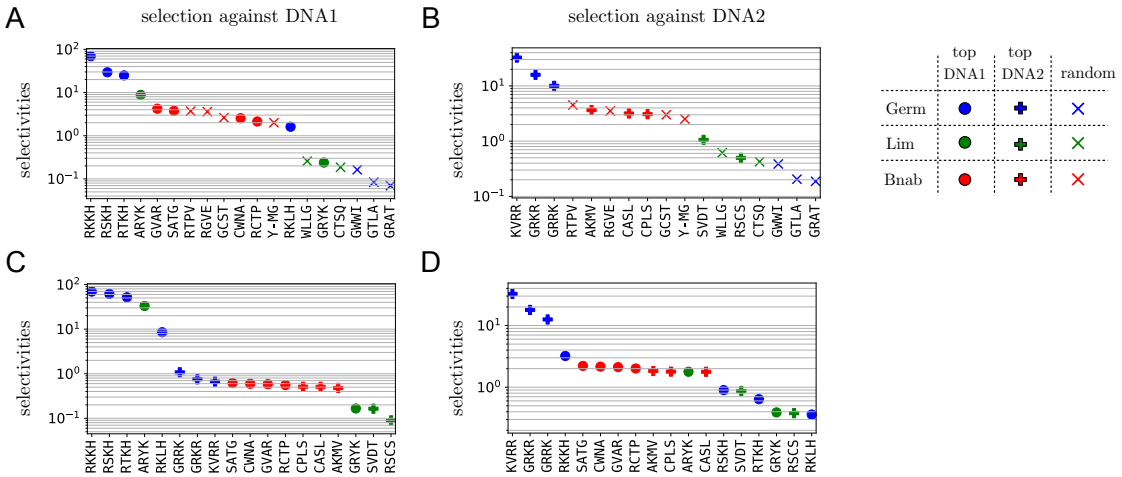


Figure S13: Cross selections with minimal libraries consisting of mixtures of top sequences against the DNA1 target (full circles) and top sequences against the DNA2 target (full crosses). **A,C.** Selection against the DNA1 target (same as Fig. 1B). **B,D.** Selection against the DNA2 target. The results confirm that some sequences from the Germ and Lim libraries bind specifically to the DNA1 target (blue dots and one of the green dots) and some sequences from the Germ library to the DNA2 target (blue crosses).

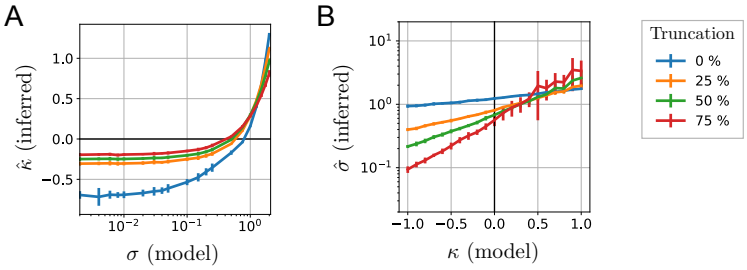


Figure S14: Relation between the parameter σ from log-normal fits and the parameter κ_N from generalized Pareto fits from numerical simulations. **A.** $N = 10^4$ values were drawn from a log-normal distribution with parameters $\mu = 0$ and varying σ (x-axis). The largest 25, 50, 75, 100 % of these values (i.e., 75, 50, 25, 0 % truncation) were fitted to a Pareto model with parameters κ and τ. The plot shows the estimation $\hat{\kappa}$ as a function of σ. Averages and standard deviations are taken over 25 independent realizations of the numerical experiment. It shows that limited sampling may cause a $\hat{\kappa} < 0$ to be inferred from values drawn from a log-normal distribution when σ is small, here $\sigma < 0.5$. **B.** Inverse simulation: A truncated log-normal model is fitted to the largest 25, 50, 75, 100 % among 500 values (i.e., 75, 50, 25, 0 % truncation) drawn from a Pareto model with parameters $\tau = 0.115$, $s^* = 0.001$ and varying κ (x-axis).

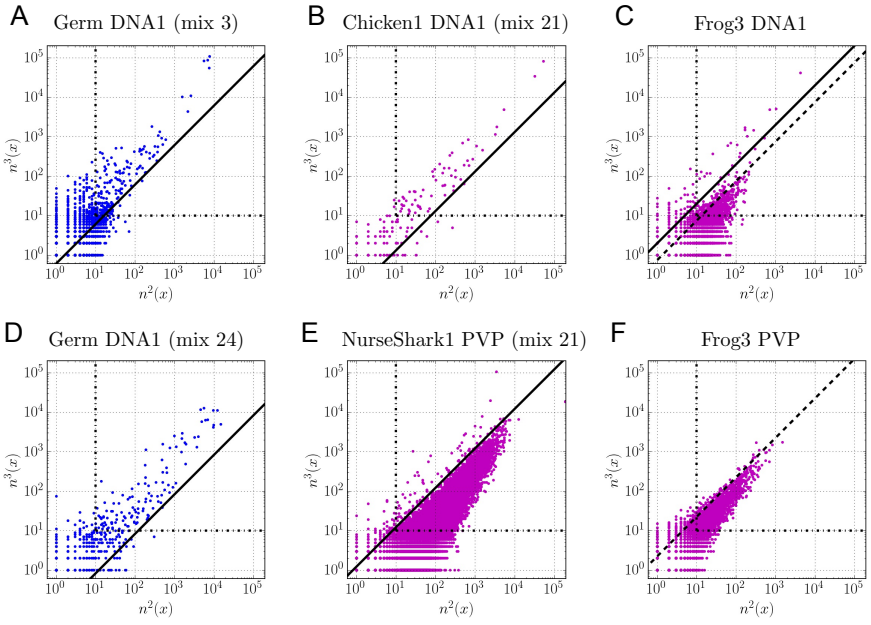


Figure S15: Definition of the threshold s^* above which selectivities s are considered for the experimental results reported here (A) and in Ref. [14] (B-F). As in Figure S3, the definition is based on a comparison between counts at the 2nd and 3rd cycles. The horizontal and vertical lines correspond to the criteria $n^2(x) \geq 10$ and $n^3(x) \geq 10$. The plain oblique line corresponds to the definition of s^* in this work. In the case of the selection of the Frog3 library against the DNA1 target, it differs from the value of s^* used in our previous work [14] (dotted oblique line) which failed to discard many selectivities coming from unspecific binding. In the case of the selection of the Frog3 library against the PVP target, all measured selectivities may be attributed to unspecific binding and we are therefore not including the inferred values of σ and κ in Fig. 4.

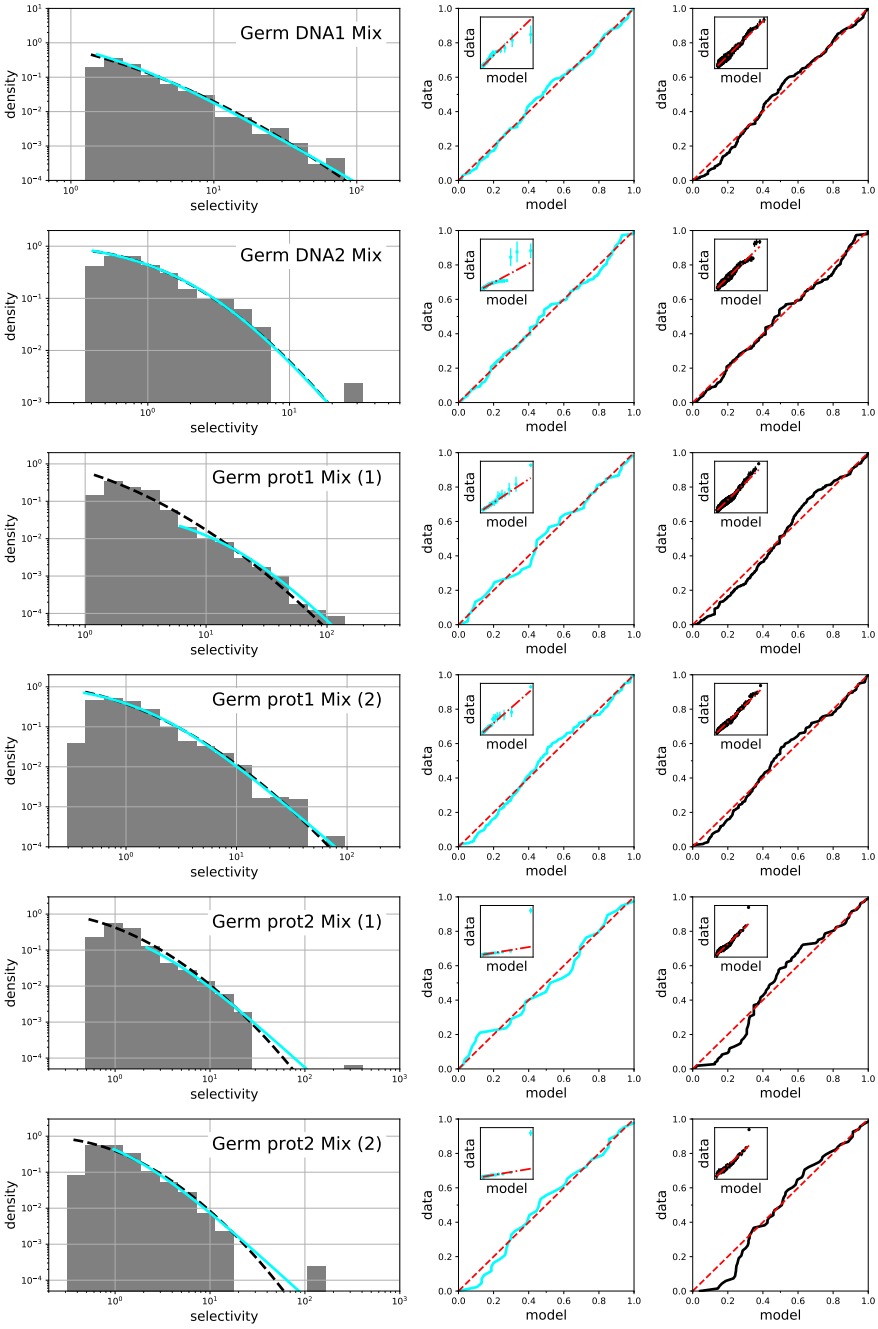


Figure S16: Assessments of the qualities of the fits of the selectivities to generalized Pareto distributions (cyan) and to log-normal distributions (black) for selections of the Germ library. The different graphs correspond to selections against different targets. For the protein targets prot1 and prot2, results from two replicate experiments are presented. All selectivities are computed by comparing the frequencies at the 2nd and 3rd cycle. The graphs on the right show the P-P and Q-Q (inset) plots for each fit. Perfect fits would correspond to the red dotted lines.

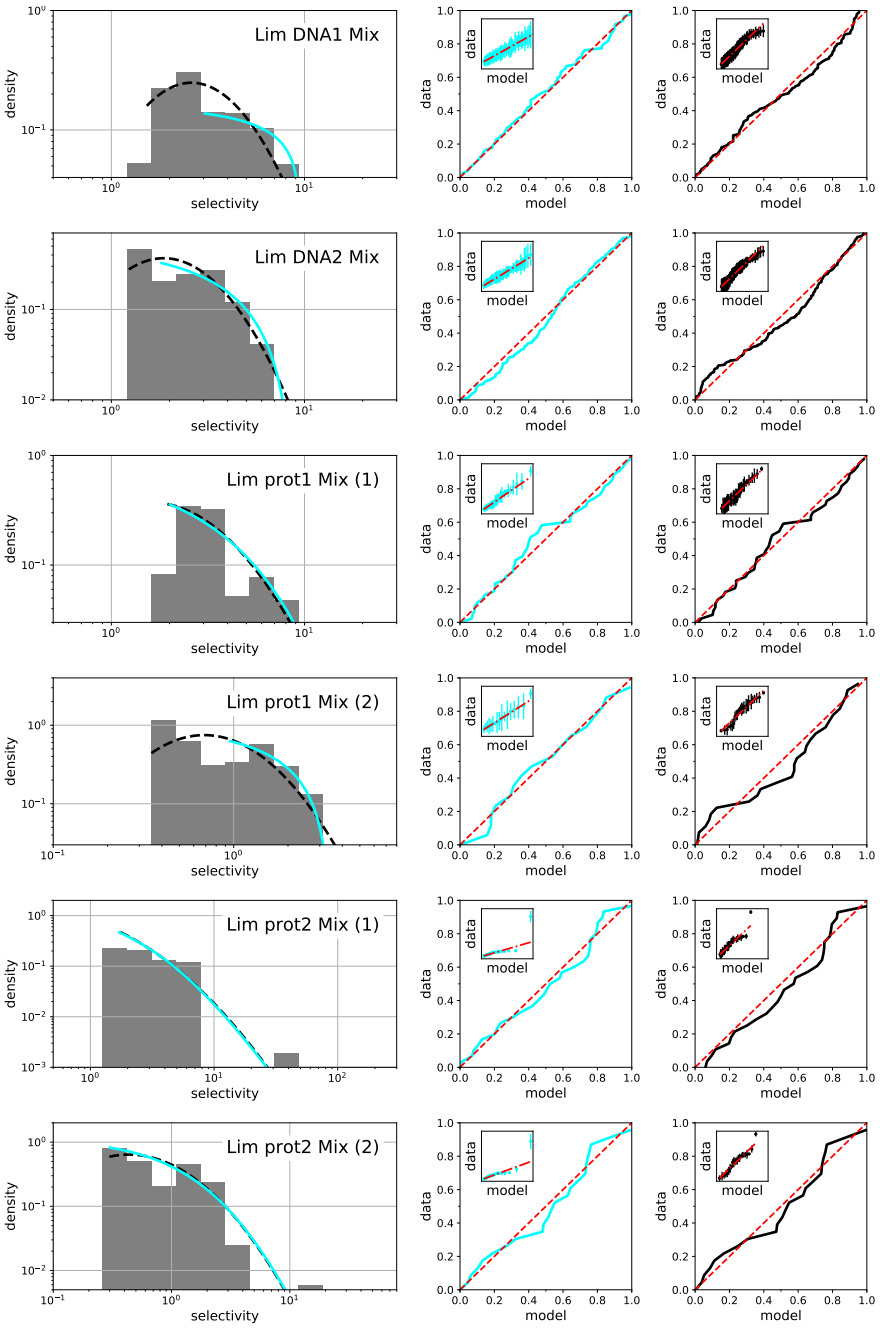


Figure S17: Same as Fig. S16 but for the Lim library instead of the Germ library.

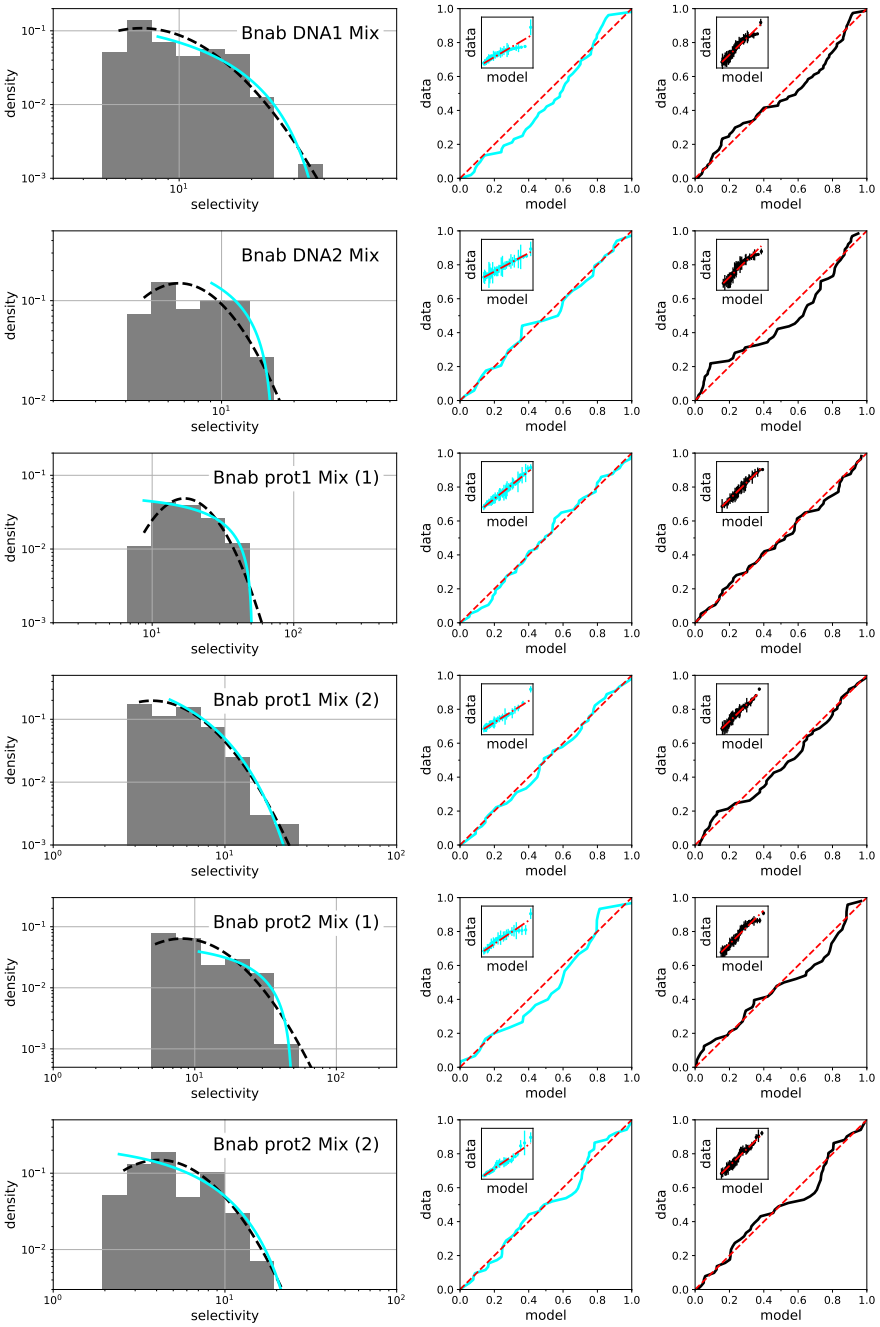


Figure S18: Same as Fig. S16 but for the Bnab library instead of the Germ library.

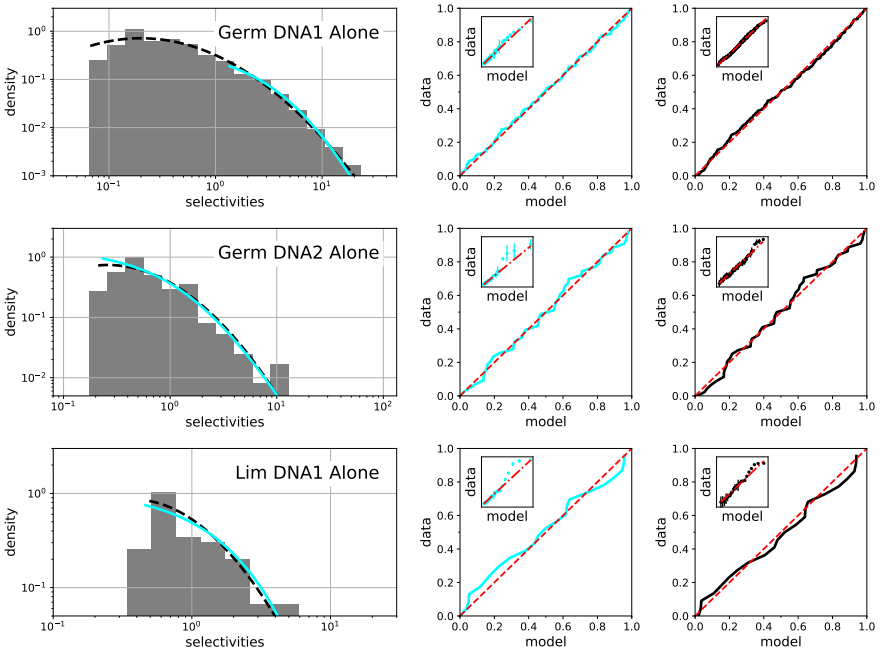


Figure S19: Same as Fig. S16 for the Germ library selected in isolation rather in a mixture with the two other libraries.

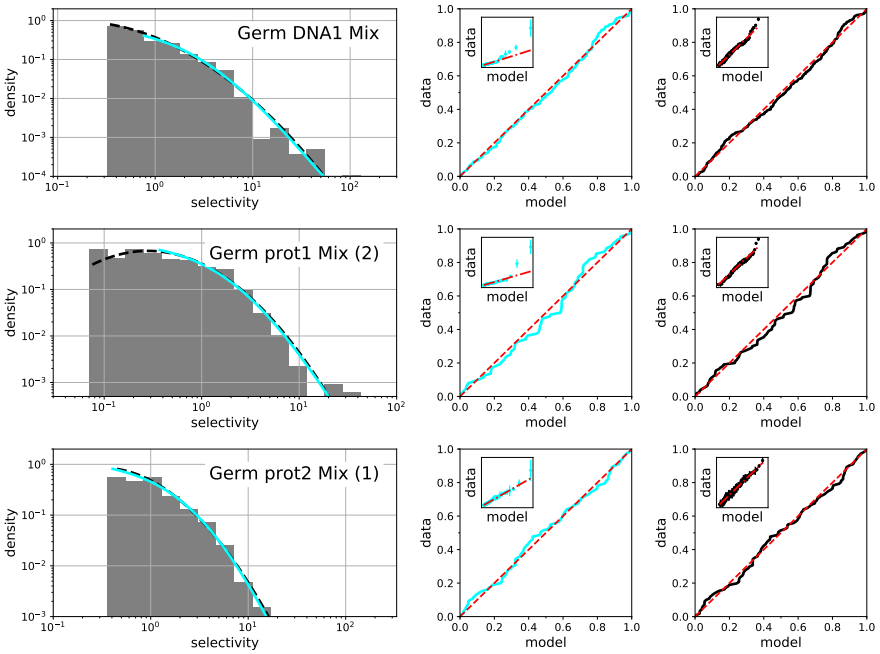


Figure S20: Same as Fig. S16 but for selectivities computed from a comparison between the 3rd and 4th cycle instead of the 2nd and 3rd cycle.

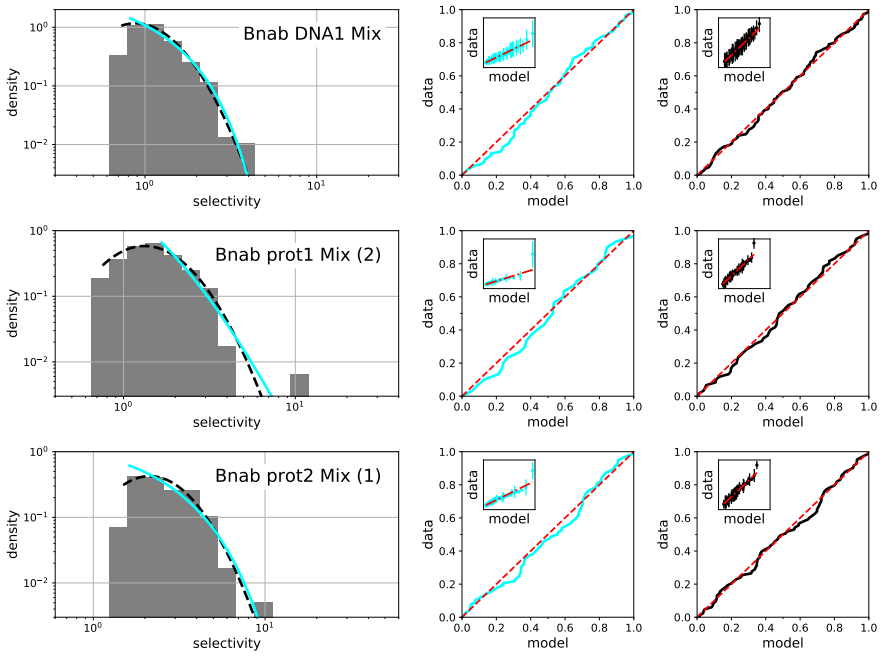


Figure S21: Same as Fig. S20 (selectivities computed from a comparison between the 3rd and 4th cycle) but for the Bnab library instead of the Germ library.

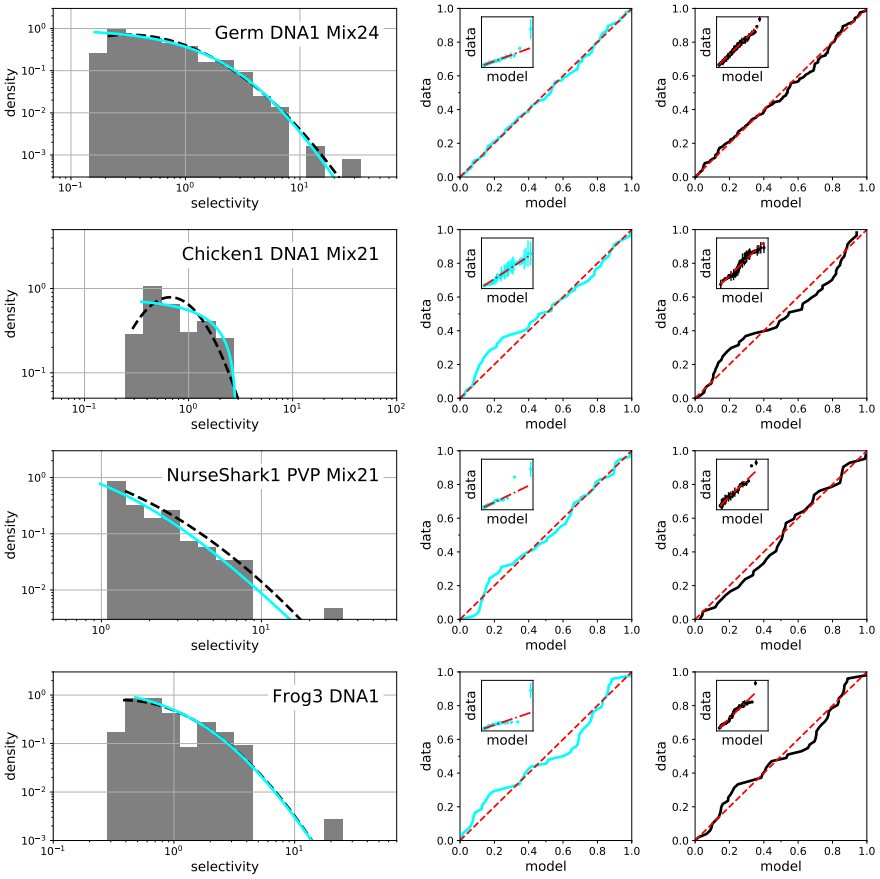


Figure S22: Same as Fig. S20 but for the experimental results reported in Ref. [14].