SUPPORTING INFORMATION

Hierarchy and extremes in selections from pools of randomized proteins

S. Boyer, D. Biswas, A. K. Soshee, N. Scaramozzino, C. Nizak, O. Rivoire

1 Supplementary experimental methods

1.1 Library construction

The library-specific parts of the frameworks, upstream of the variable CDR3 (Figure 1) are shown in Figure S17. 23 of these frameworks were designed based on the amino acid sequences of 23 natural V_H segments, with minor modifications to accommodate common restriction sites at the two ends of the CDR2 and CDR3. Out of these 23 frameworks, 20 were chosen to have minimal sequence similarities, and 3 are from a same human V_H segment: one is the germ-line (naïve) form, one results from limited maturation (85% sequence similarity to the germ line) and the other from extensive maturation (broadly neutralizing antibody against HIV [1] with only 65% sequence similarity to the germ line). A 24th framework was made exclusively of glycines to serve as a control. Downstream of the CDR3, the fixed part has amino acid sequence FDYWGQGTLVTVSSG in all libraries. The nucleotide sequences were optimized for *E. coli* codon usage and are provided as supplementary file (Dataset S1).

The 24 frameworks were obtained from Genewiz (South Plainfield, NJ) as synthetic genes with restriction sites flanking the CDRs to allow for the introduction of arbitrary sequences at the CDRs. In particular, the CDR3 region is flanked by BssHI and XhoI sites. These synthetic genes were cloned into a modified version of pIT2 phagemid (standard phage display vector) lacking V_L [2]. To randomize the CDR3 region, a degenerate oligonucleotide containing 12 random nucleotides (from Eurogentec, Angers, France) flanked by BssHI and XhoI sites was PCR-amplified, digested, and ligated into gel purified pIT2 phagemids harboring each of the 24 frameworks. Ligation products were purified and electroporated into TG1 *E. coli* (from Lucigen, Middleton, WI) at efficiencies exceeding 10^7 transformants (to ensure a >100-fold coverage of the 10^5 diversity), while keeping 100-fold lower efficiency in control electroporations of ligation product without insert (to minimize the occurence of empty vectors in libraries, below 1%).

1.2 Phage display screening

All chemicals were purchased from Sigma-Aldrich (St Louis, MO) unless otherwise specified. Deionized water of resistivity 16 MΩ.cm was produced with an ion exchange resin (Aquadem(R) system, Veolia, Lyon, France). 2xTY medium was prepared by dissolving 16 g tryptone, 10 g yeast extract, 5 g NaCl (tryptone and yeast extract from USBIO distributed by Euromedex, Strasbourg, France) in 1 L of deionized H_2O and autoclaving for 15 min at 120 °C.

The DNA target (PAGE purified, lyophilized) was resuspended in Tris-EDTA buffer at 400 μM. For each selection round, 50 μL of magnetic beads coated with streptavidin (Dynabeads(R) M-280 Streptavidin from Invitrogen Life Technologies SAS, Saint Aubin, France) were prepared according to the manufacturer's protocol. 10 μL of DNA stock solution were mixed with 50 μL of washed Dynabeads(R) and incubated for 10 minutes at room temperature using gentle rotation. The biotinylated hairpin DNA coated beads were separated with a strong magnet for 2-3 minutes and washed 2-3 times with a buffer containing 5 mM

Tris-HCl (pH 7.5), 0.5 mM EDTA and 1M NaCl.

Phage production is the same as described before [2] except that the infected TG1 culture was grown for 7 hours (instead of overnight) at 30°C in 2xTY + 100 μg/mL ampicillin + 50 μg/mL kanamycin.

During phage display experiments, the supernatant containing our library (around 10^{11} phages) in 2xTY, was adjusted to 10 mM NaPO4 pH = 7.4. Phages were first incubated against either naked magnetic beads or non-treated polystyrene 3 cm diameter Nunc Petri dish (Thermo Fisher Scientific, Waltham, MA) for negative selection. For DNA target selection, DNA LoBind tubes (Eppendorf AG, Hamburg, Germany) were used. Phages were incubated during 1 h without agitation and 30 min on a rocker at room temperature. The remaining phages were then incubated with either hairpin DNA or PVP targets. In the case of hairpin DNA, 50 μL of beads were incubated with an excess of DNA targets (around 10^{14}), washed according to the manufacturer's protocol, yielding on the order of 10^{13} immobilized DNA targets, at a 100-fold excess over available phages (10^{11}). Antibody selection was then performed against either DNA-coated beads or a PVP-functionalzed Petri dish for 90 mn on a rocker. 10 washing steps with 1xPBS + 0.1% Tween 20 were performed. Next, selected phages were eluted using 1 mL of fresh solution of 100 mM triethylamine for 20 min and neutralized with 500 μL of Tris/HCl buffer (1 M, pH 7.4). Eluted phages were rescued by infection of an excess of exponentially growing TG1 *E. coli* cells (14 mL of a 2xTY culture at O.D. 600 nm = 0.6) for titration and phage preparation for subsequent rounds of selection. Infected TG1 were then plated on 2xTY + ampicillin plates for overnight amplification at 37°C. Glycerol stocks were stored at −80°C.

1.3 Amplification biases

Each round of selection is followed by a round of amplification consisting in infecting the bacteria with the selected phages. Sequence-specific differences in amplification may arise from differences of growth rate of the bacteria carrying different phagemids, or differences in infectivity or display ability of the phages. We measured by sequencing both the differences between frameworks when considering a mixture of the 24 libraries (Figure S5) and between CDR3 when considering a library of given framework (Figure S15).

Between frameworks, only the S1 and CH1 libraries show significant enrichment upon amplification alone. Each of these two libraries dominates over the others in one experiment of selection with a mixture of libraries but, when the mixture of all 24 libraries is selected against the PVP target, they are dominated by another library, the HG library, which does not show any enrichment upon amplification. This observation, together with the strong correlation between frequencies before and after amplification (Figure S5B), are evidence that differences in library amplification are not responsible for the observed hierarchy between frameworks (Figure 2).

Within each library, a clear enrichment for the glutamine amino acid is observed, irrespectively of the framework (Figure S15). This bias has a simple interpretation: the *supE* strain of *E. Coli* that we use for phage display is a partial amber stop codon suppressor. In this strain, the amber codon codes about one third of the times for a glutamine and acts as a stop codon the two other thirds. The reduced production of antibodies due to the presence of an amber codon thus confers a growth advantage to the bacteria (antibody expression is costly for *E. coli*). Consistently with this interpretation, we verify that all the glutamines present in the data are associated with the amber codon. The results presented in the paper exclude sequences with an amber codon but, in most experiments with selection, glutamine does not appear in the selected consensus sequence and considering the amber codon as coding for an amino acid or for a stop codon has no incidence on the conclusions. Apart from glutamine amplification, no other significative pattern of amplification is visible or may plausibly explain the results of the experiments with selection.

2 Supplementary properties of extreme value distributions

2.1 Relations between parameters

The fit to an extreme value distribution with parameters (κ_0, τ_0) applies for selectivities above a threshold s_0^*. Fitting the data above a larger threshold $s_1^* > s_0^*$ must lead to the same shape parameter $\kappa_1 = \kappa_0$ (simply denoted κ) but to a different scaling parameter τ_1 given by [3]

$$\tau_1 = \tau_0 + \kappa(s_1^* - s_0^*). \tag{S1}$$

These are the relations verified in Figure 4A.

Another independent parameter, which depends on the bulk of the distribution, is the fraction ϕ_0 of the data above the threshold s_0^*, which obviously depends on the value of the threshold ($\phi_1 < \phi_0$ when $s_1^* > s_0^*$).

In total, four parameters are relevant: s^*, ϕ, κ and τ.

2.2 Spacings between extremes

We show here that if $s_1 > s_2 > \dots$ are drawn at random with a probability density $f_{\kappa,\tau,s^*}(s)$ given by Eq. (3) then their spacings defined by $\Delta_r = s_r - s_{r+1}$ scale as $\Delta_r/\Delta_1 \sim r^{-(\kappa+1)}$, where κ is the shape parameter. This follows from a more general result:

$$\Delta_r \sim \tau N^\kappa r^{-(\kappa+1)}, \tag{S2}$$

where τ is the scaling parameter and N the total number of samples.

The proof can be given in terms of the rescaled variable $x = (s - s^*)/\tau$ whose probability density $f_\kappa(x)$ is defined in Eq. (4), since $\Delta_r = \tau(x_r - x_{r+1})$. As indicated by Eq. (5), the rank r and the value x are related for large N by $r(x)/N \sim \int_x^\infty f_\kappa(u)\mathrm{d}u = (1+\kappa x)^{-1/\kappa}$. Inverting this relation gives $x_r = q(r/N)$ with $q(z) = (1 - z^{-\kappa})/\kappa$. In the limit of large N where the formalism applies, we have therefore $x_r - x_{r+1} \simeq -(1/N)q'(r/N)$ with the derivative of $q(z)$ given by $q'(z) = -z^{-(\kappa+1)}$. All together, this gives $x_r - x_{r+1} \simeq N^\kappa r^{-(\kappa+1)}$ and thus $\Delta_r \sim \tau N^\kappa r^{-(\kappa+1)}$.

2.3 Scaling of maxima with library sizes

We show here that if the maximum of N random variables drawn with a probability density given by Eq. (3) is s_1, then adding more elements to produce a library $m > 1$ times larger leads to a maximal value $s_1' \geq s_1$ satisfying

$$\mathbb{E}[s_1' - s_1] = \tau N^\kappa m^\kappa \mathrm{Li}_{\kappa+1}\left(1 - \frac{1}{m}\right), \tag{S3}$$

where $\mathrm{Li}_k(z) = \sum_{n=1}^\infty z^n/n^k$ defines the so-called polylogarithmic function. $\mathbb{E}[s_1' - s_1]$ is thus an increasing function of κ as illustrated in Figure S13, where Eq. (S3) is also compared to numerical simulations. The relevance of this formula rests on the assumption that sub-libraries of a library with shape parameter κ are characterized by the same κ, which finds support in the data (Figure S14). In practice, the extreme value distribution applies only to the fraction ϕN of the data above the threshold s^*. When expressed relative to the expected spacing between the top two values in the initial population of N variables, $\Delta_1 = \tau N^\kappa$, Eq. (S3) is however independent of N and τ: $\mathbb{E}[s_1' - s_1]/\Delta_1 = m^\kappa \mathrm{Li}_{\kappa+1}\left(1 - \frac{1}{m}\right)$.

To derive this formula, we consider an initial population of size mN whose maximum is s_1' and define a subpopulation of (approximate) size N by retaining with probability $1/m$ each of its elements. The maximum s_1 of this subpopulation has thus rank n in the initial population with probability $p_n = (1-1/m)^{n-1}1/m$ – the probability that none of $n-1$ top values are retained but that the n-th is. Following Eq. (S2), the

distance between $s_1 = s'_n$ and s'_1 is estimated as $\delta'_n = s'_1 - s'_n = \sum_{r=1}^{n-1} \Delta'_r = \tau(mN)^\kappa \sum_{r=1}^{n-1} r^{-(\kappa+1)}$. This leads to

$$\mathbb{E}[s'_1 - s_1] = \sum_{n=1}^{\infty} p_n \delta'_n = \frac{\tau(mN)^\kappa}{m} \sum_{n=1}^{\infty} \sum_{r=1}^{n-1} \left(1 - \frac{1}{m}\right)^{n-1} \frac{1}{r^{\kappa+1}} = \frac{\tau(mN)^\kappa}{m} \sum_{r=1}^{\infty} \sum_{n=r+1}^{\infty} \left(1 - \frac{1}{m}\right)^{n-1} \frac{1}{r^{\kappa+1}} \quad \text{(S4)}$$

and, after summing the geometric series $\sum_{n=r+1}^{\infty} (1 - 1/m)^{n-1} = (1 - 1/m)^r m$, to

$$\mathbb{E}[s'_1 - s_1] = \tau(mN)^\kappa \sum_{n=1}^{\infty} \left(1 - \frac{1}{m}\right)^{n-1} \frac{1}{r^{\kappa+1}}, \quad \text{(S5)}$$

which is equivalent to Eq. (S3).

References

[1] 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, and M C Nussenzweig. Somatic mutations of the immunoglobulin framework are generally required for broad and potent HIV-1 neutralization. *Cell*, 153(1):126–138, 2013.

[2] A K Soshee, S Zürcher, N D Spencer, A Halperin, and C Nizak. General in vitro method to analyze the interactions of synthetic polymers with human antibody repertoires. *Biomacromolecules*, 15(1):113–121, 2014.

[3] S Coles. *An introduction to statistical modeling of extreme values*. Springer, 2001.

3 Supplementary tables

	κ	τ	s^*	ϕ
S1/PVP	0.44 ± 0.22	$1.6 \times 10^{-4} \pm 10^{-5}$	0.37×10^{-3}	$\sim 6 \times 10^{-4}$
F3/PVP	0.07 ± 0.21	$3.1 \times 10^{-4} \pm 8 \times 10^{-5}$	1.2×10^{-3}	$\sim 6 \times 10^{-4}$
HG/DNA	0.26 ± 0.21	$5.7 \times 10^{-3} \pm 1.5 \times 10^{-3}$	0.7×10^{-3}	$\sim 6 \times 10^{-4}$
CH1/DNA	-0.62 ± 0.25	$2.5 \times 10^{-2} \pm 8 \times 10^{-3}$	2.5×10^{-3}	$\sim 3.6 \times 10^{-4}$

Table S1: Parameters κ, τ, s^* and ϕ describing the four experiments presented in Figure 3. Note that τ and ϕ depend on s^*, which may be chosen within a finite interval of values. However, the values of $\tau(s^*)$ and $\phi(s^*)$ at $s^* = s_0^*$ determine their values at $s^* = s_1^*$ as indicated in Eq. (S1) for τ.

	Fraction of sequences with >1 sequencing error / 12 bases
Mix24	0.043
Mix24 amplified	0.029
Mix24/PVP round 1	0.025
Mix24/PVP round 2	0.051
Mix24/PVP round 3	0.107
Mix24/DNA round 1	0.032
Mix24/DNA round 2	0.065
Mix24/DNA round 3	0.029
Duplicate Mix24/DNA round 3	0.024
Mix21/PVP round 2	4×10^{-3}
Mix21/PVP round 3	4×10^{-3}
Mix21/DNA round 1	0.027
Mix21/DNA round 2	0.046
Mix21/DNA round 3	0.106
F3/PVP round 1	0.034
F3/PVP round 2	0.029
F3/PVP round 3	0.04
F3/DNA round 1	0.027
F3/DNA round 2	0.048
F3/DNA round 3	0.085

Table S2: Estimation of sequencing errors – Fraction of the sequences with at least one error in the 12 bases immediately downstream of the 12 based of the CDR3 (errors estimated given the known sequence of the fixed part of the framework).

S1/PVP in Mix24 $n_0^2 = n_0^3 = 10$	$\kappa = 0.34 \pm 0.22$
S1/PVP in Mix24 $n_0^2 = 10$ and $n_0^3 = 25$	$\kappa = 0.42 \pm 0.25$
S1/PVP in Mix21 $n_0^2 = n_0^3 = 100$	$\kappa = 0.56 \pm 0.18$
S1/PVP in Mix21 sampled $n_0^2 = n_0^3 = 10$	$\kappa = 0.48 \pm 0.16$
S1/PVP in Mix21 sampled $n_0^2 = 10$ and $n_0^3 = 50$	$\kappa = 0.67 \pm 0.37$

Table S3: Robustness of the EVT analysis – The analysis presented in the main text retains only sequences present in sufficient number in the samples of the populations that are sequenced at the second and third rounds – namely $n_i^2 > n_0^2 = 10$ and $n_i^3 > n_0^3 = 10$. This table shows that varying the values of the thresholds n_0^2 and n_0^3 has little incidence on the value of the shape parameter κ inferred by EVT analysis. The sample of the S1 library against PVP in the mixture of 21 libraries (Mix21, see Figure 2) contained 10^6 sequences while the sample in the mixture of 24 libraries (Mix24) contained only 10^5; the two last rows of the table shows that further sampling at 1/10 the former to reach samples of comparable sizes have no incidence on the results.

4 Supplementary figures

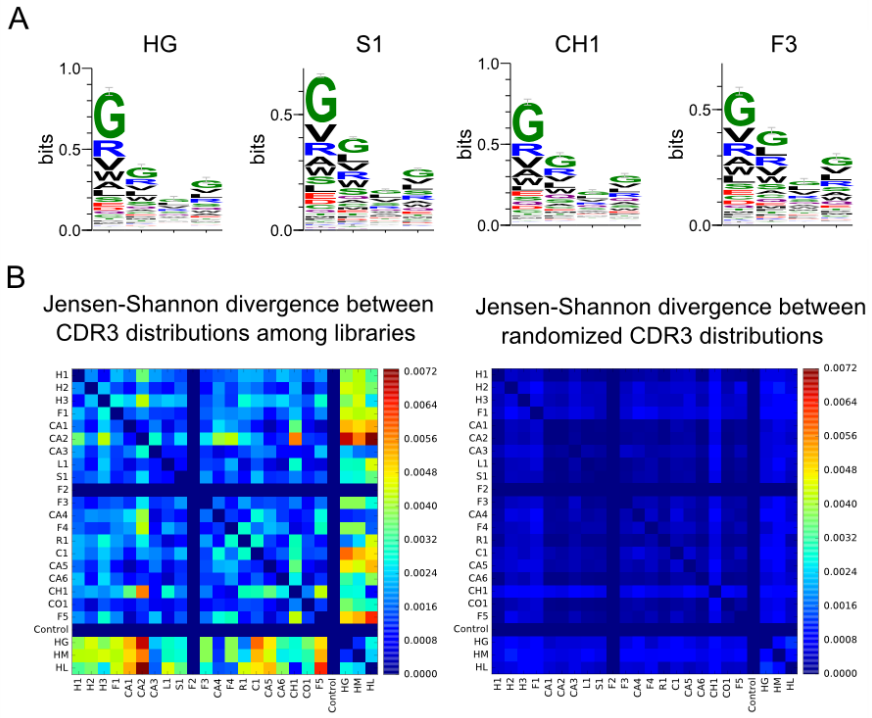


Figure S1: Diversity of the libraries – The different libraries are intended to harbor the same distribution of amino acids at the 4 varied positions. We measured these distributions by sequencing samples from the initial libraries. **A.** Sequence logos showing the entropies of the various amino acids at the four positions: the distribution is non-uniform but similar from one library to the next. **B.** More quantitatively, the distance between distributions is estimated using the Jensen-Shannon divergence: if q_a^ℓ is the frequency of amino acid a in the CDR3 of library ℓ, the Jensen-Shannon divergence between libraries k and ℓ is defined as $\sum_a q_a^k \ln(q_a^k/q_a^\ell) + \sum_a q_a^\ell \ln(q_a^\ell/q_a^k)$. This divergence is found to be 5 to 10 times larger than expected from sampling noise. This represents the experimental precision at which we were able to introduce the same diversity in each library. These differences of frequencies between initial libraries are, however, much smaller than the differences of frequencies before and after a round of selection within a same library.

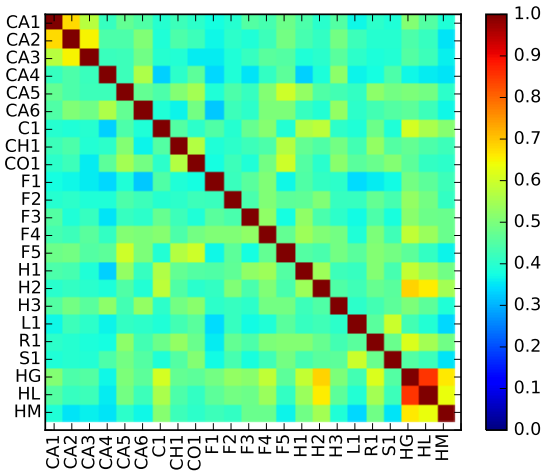


Figure S2: Sequence similarities between frameworks – Similarity between two frameworks is measured as the fraction of common amino acids in an alignment of their two sequences. Only the library-specific part of the frameworks (Figure 1) defined in Figure S17 is considered here. In most cases, the sequence similarity is in the range 30-60%.

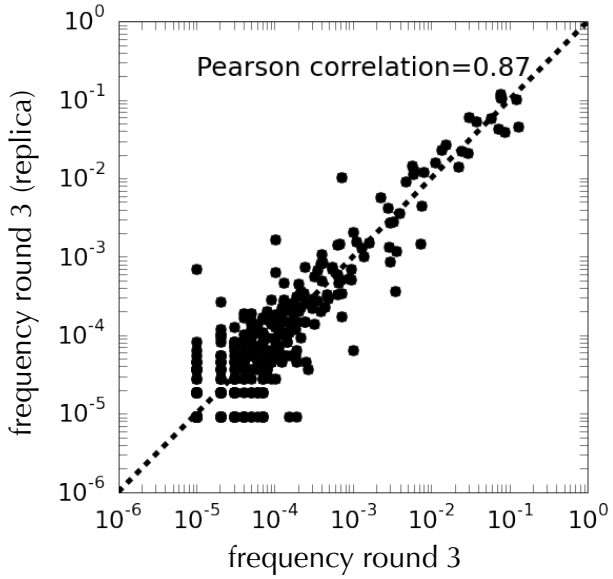


Figure S3: Reproducibility – To assay the reproducibility of the experiments, two independent selections of the mixture of 24 libraries were performed against the DNA target and the frequencies of the sequences were compared at the third round: the high correlation between the two results indicates high reproducibility.

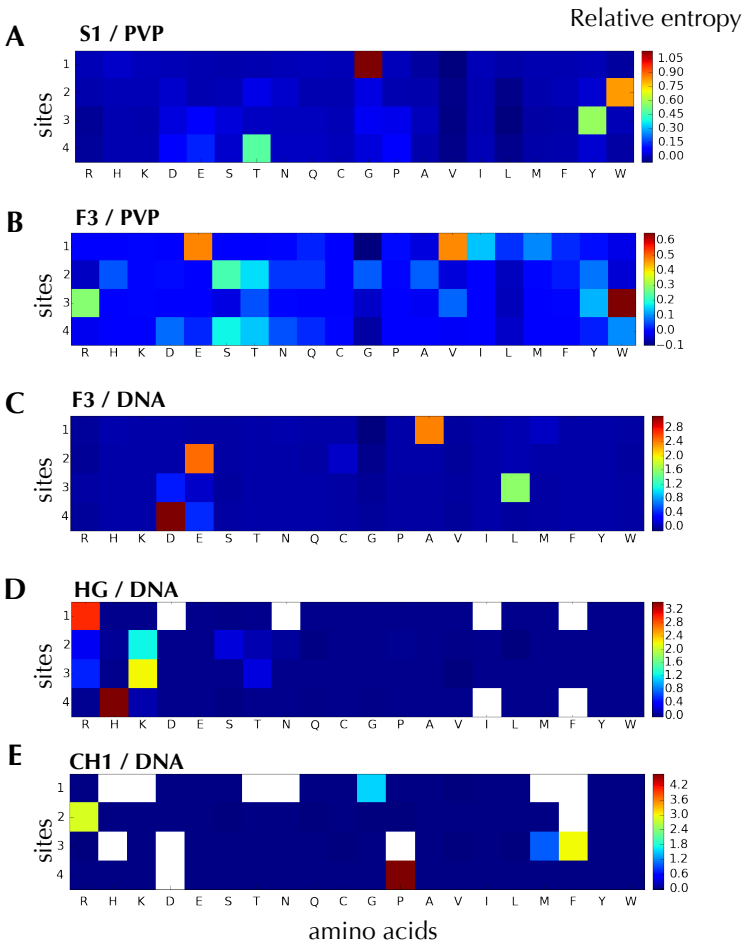


Figure S4: Library and target specificities – Relative entropies of the different amino acids at the third round of different experiments; the relative entropy is calculated per site as $f_i^a \ln(f_i^a/q_i^a)$ where f_i^a is the frequency of amino acid a at position i in the third round and q_i^a in the initial library (round 0). **A** and **B** show that the consensus sequence is framework dependent. **B** and **C** show that it is target specific. Finally, **C**, **D** and **E** provide further evidence of framework dependency. (White squares indicate amino acids not represented in the population).

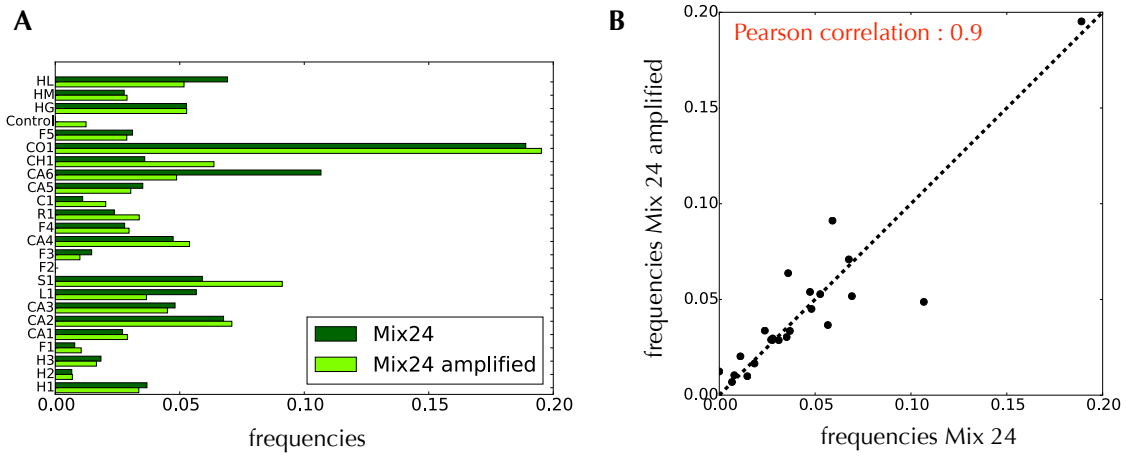


Figure S5: Biases in amplifications – Comparison of the composition of the mixture of 24 libraries before and after amplification in absence of selection. **A.** Differences of frequencies, showing that only the S1 and CH1 libraries are enriched. **B.** Correlations between the frequencies (same data).

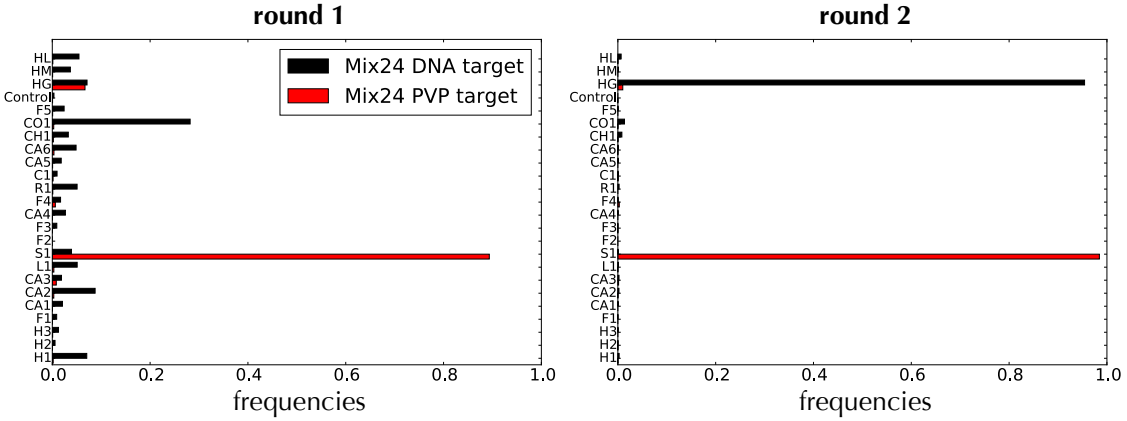


Figure S6: Target-dependent hierarchy – Figure 2 shows that a mixture of 24 libraries selected against the DNA target is dominated by the HG framework while a mixture of 21 libraries that excludes the HG library is dominated by the CH1 framework. As shown in this figure, when the same mixture of 24 libraries is selected against the PVP target, a different framework, the S1 framework, dominates (consistently, it also dominates when screening the mixture of 21 libraries, which includes S1).

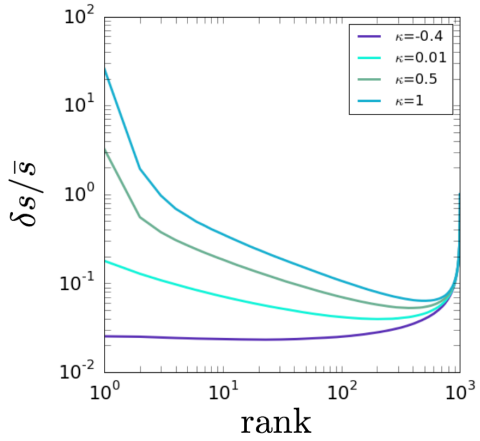


Figure S7: Variations in the values of extreme selectivities – When sampling N random variables from the extreme probability density $f_\kappa(x)$ given by Eq. (4), the value s_r of the variable of rank r is distributed with a mean $\bar{s}_r$ and standard deviation δs_r. The ratio $\delta s_r/\bar{s}_r$ is largest for the very top sequences, as shown here based on numerical simulations. This observation is consistent with deviations of the data from a power law observed for the very top selectivities even when $\kappa > 0$ and when the overall fit with an extreme value distribution is good (Figures 4A-B).

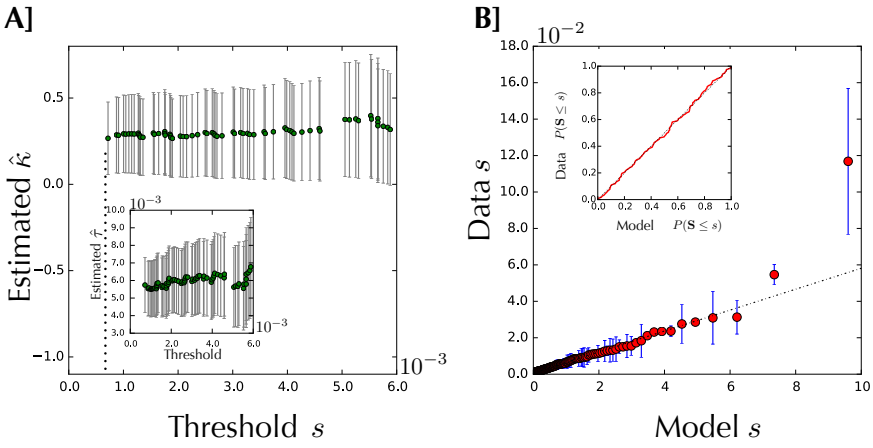


Figure S8: EVT analysis for the selection of the HG library against the DNA target (data shown in Figure 3B). A fit of the general model gives $\kappa = 0.26 \pm 0.21$, $\tau = 5.7 \times 10^{-3} \pm 1.5 \times 10^{-3}$ while a fit of the exponential model ($\kappa = 0$) gives $\tau_0 = 8 \times 10^{-3} \pm 1.4 \times 10^{-3}$; the exponential model is excluded with a p-value 1.4×10^{-3}, in favor of $\kappa > 0$. Note that the threshold $s^* = 10^{-3}$ above which the fit is stable and good is much below the value of the selectivity above which a power law is observed in Figure 3B, of the order of $s = 10^{-2}$.

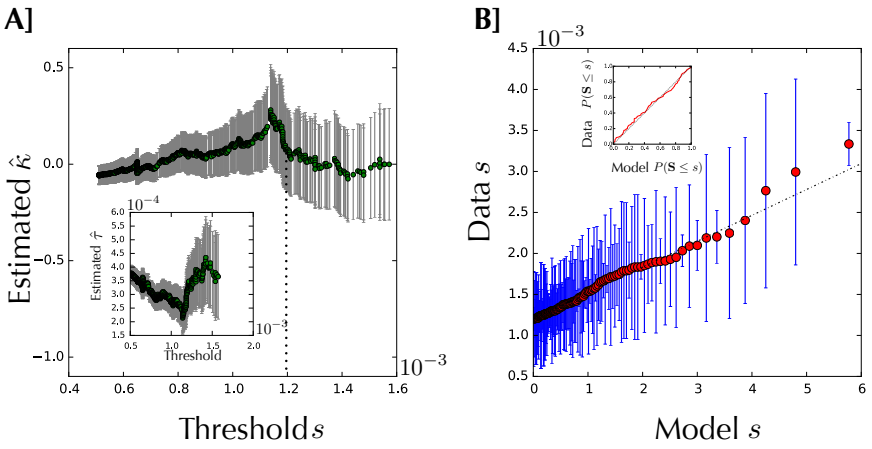


Figure S9: EVT analysis for the selection of the F3 library against the PVP target (data shown in Figure 3D). A fit of the general model gives $\kappa = 0.07 \pm 0.21$, $\tau = 3.1 \times 10^{-4} \pm 8 \times 10^{-5}$ while a fit of the exponential model ($\kappa = 0$) gives $\tau_0 = 3.4 \times 10^{-4} \pm 6 \times 10^{-5}$; the exponential model is excluded with a p-value 0.75, which is non significant. This data is therefore consistent with an exponential model $\kappa = 0$.

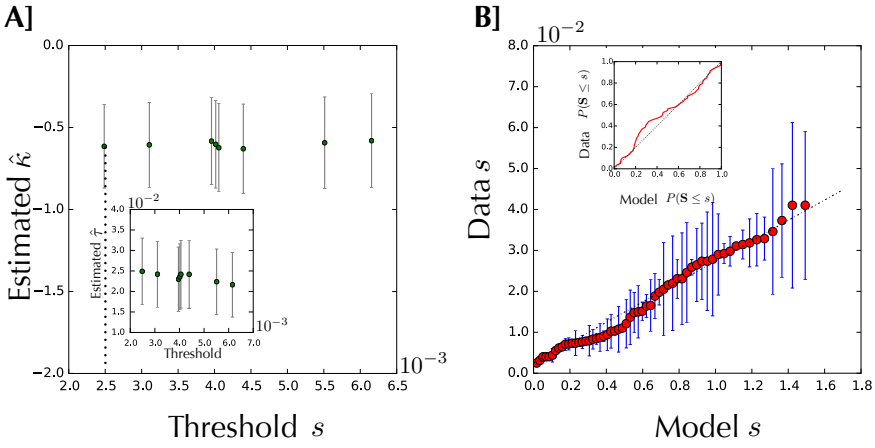


Figure S10: EVT analysis for the selection of the CH1 library against the DNA target (data shown in Figure 3C). A fit of the general model gives $\kappa = -0.62 \pm 0.25$, $\tau = 2.5 \times 10^{-2} \pm 8 \times 10^{-3}$ while a fit of the exponential model ($\kappa = 0$) gives $\tau_0 = 1.5 \times 10^{-2} \pm 4 \times 10^{-3}$; the exponential model is excluded with a p-value 10^{-2}, in favor of $\kappa < 0$.

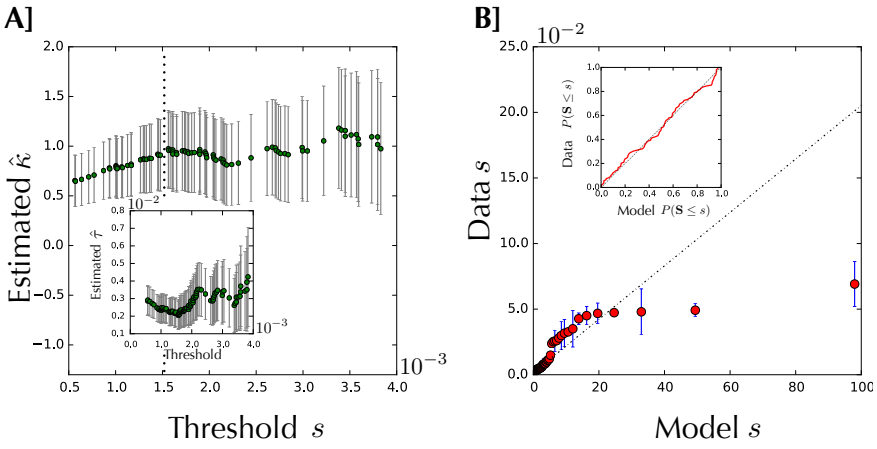


Figure S11: EVT analysis for the selection of the F3 library against the DNA target (data not shown in the main text). A fit of the general model gives $\kappa = 0.97 \pm 0.38$, $\tau = 2 \times 10^{-3} \pm 8 \times 10^{-4}$ while a fit of the exponential model ($\kappa = 0$) gives $\tau_0 = 7.5 \times 10^{-3} \pm 10^{-3}$; the exponential model is excluded with a p-value $< 10^{-4}$, in favor of $\kappa > 0$.

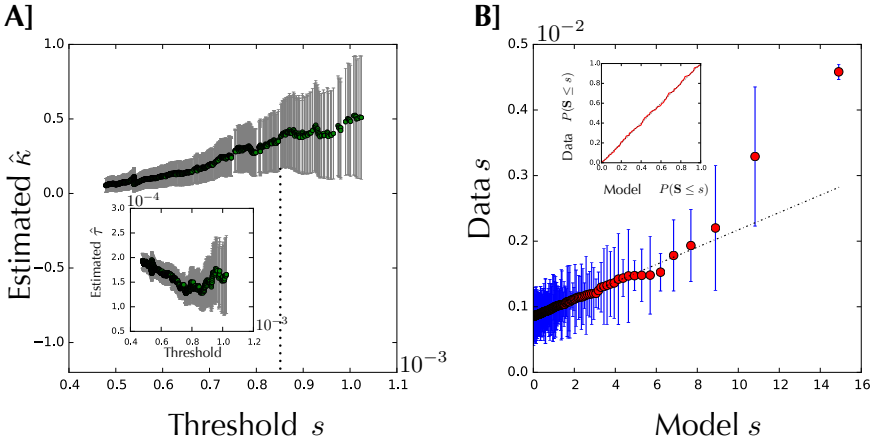


Figure S12: EVT analysis for the selection of the N1 library in the mixture of 24 libraries against the PVP target (data not shown in the main text). A fit of the general model gives $\kappa = 0.38 \pm 0.21$, $\tau = 1.3 \times 10^{-4} \pm 3 \times 10^{-5}$ while a fit of the exponential model ($\kappa = 0$) gives $\tau_0 = 2.2 \times 10^{-4} \pm 3 \times 10^{-5}$; the exponential model is excluded with a p-value $< 10^{-4}$, in favor of $\kappa > 0$.

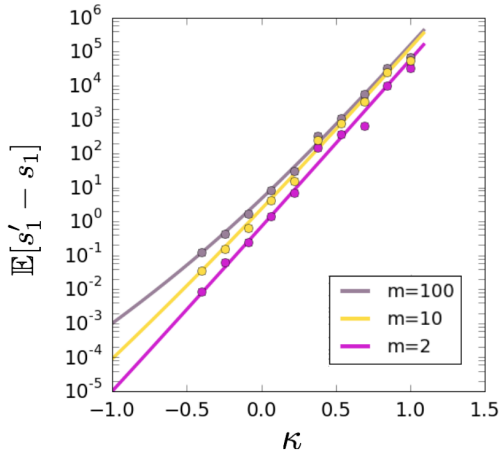


Figure S13: Scaling of the best binder with the library size – To estimate the gain that sampling m times more the same library may provide as a function of the shape parameter κ, we show here the expected difference $\mathbb{E}[s_1' - s_1]$ between the maximum s_1' of mN samples drawn with probability density $f_\kappa(x)$ from Eq. (4) and the maximum s_1 of N sub-samples. The plain lines are based on Eq. (S3) with $\tau = 1$ and $N = 10^5$ and the dots are the results of numerical simulations (averaged over many draws), showing a good agreement between the two. Note how $\mathbb{E}[s_1' - s_1]$ depends more strongly on κ than on m.

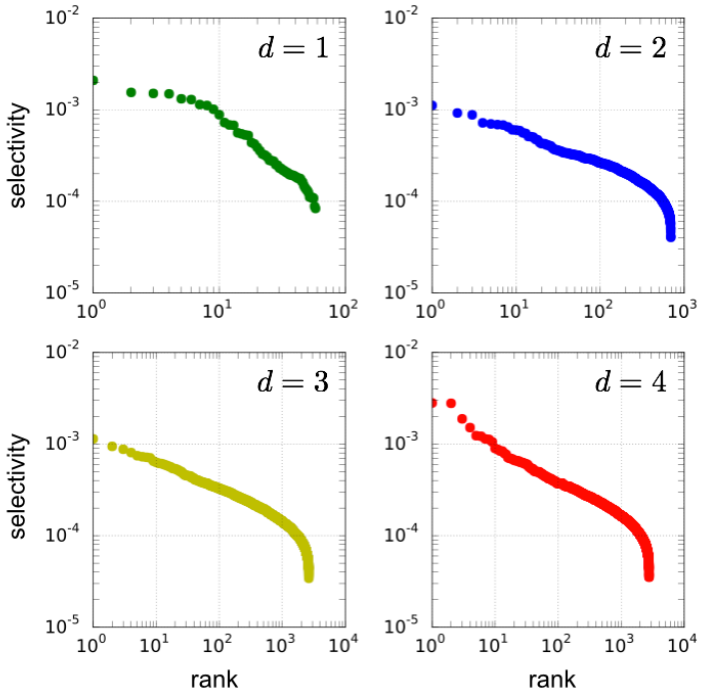


Figure S14: Stability of the shape parameter κ for non-random sub-samples of a same library – To test whether non-random sub-libraries may be expected to be described by the same shape parameter as the library from which they originate, we consider here the results of the selection of library S1 against PVP (Figure 3A), for which the consensus CDR3 has amino acids sequence GWYT and we make four non-overlapping sub-libraries consisting of sequences with CDR3 at distance $d = 1$ to 4 from this consensus, where the distance just counts the number of amino acid differences (number of mutations). This figure shows the selectivity versus the rank of the sequences in these sub-libraries. An EVT analysis indicates that $\kappa(d = 1) = 0.33 \pm 0.39$, $\kappa(d = 2) = 0.40 \pm 0.26$, $\kappa(d = 3) = 0.30 \pm 0.23$, $\kappa(d = 4) = 0.53 \pm 0.22$: all these values are comparable to the value $\kappa = 0.44 \pm 0.22$ of the shape parameter for the full library.

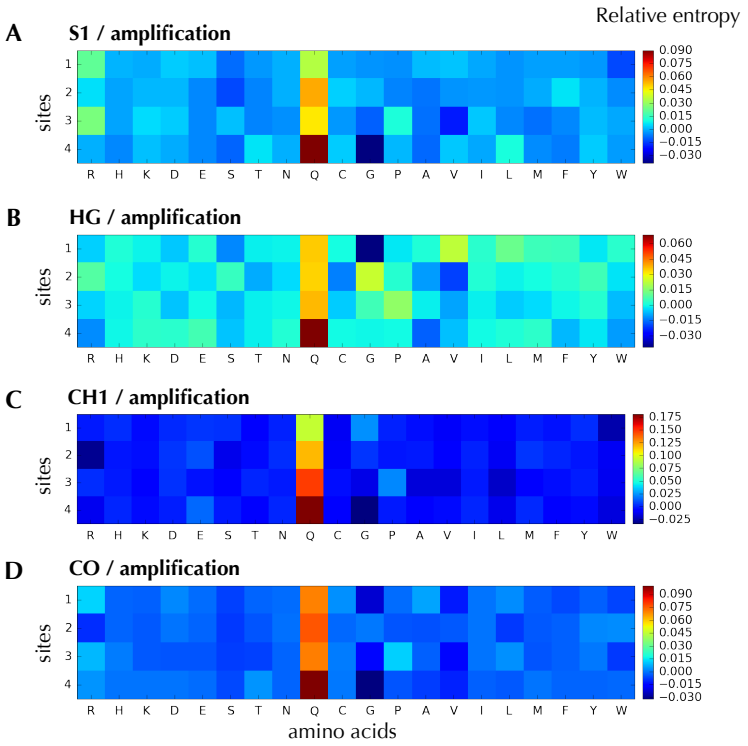


Figure S15: Amplification bias – Relative entropy between frequencies of CDR3 sequences before and after amplification without selection, showing an enrichment in glutamine (represented by the letter Q). The results presented in the paper exclude sequences with an amber codon, which is responsible for this effect (see supplementary experimental methods), but, in most experiments with selection, glutamine does not appear in the selected consensus sequence and considering the amber code as coding for an amino acid or for a stop codon has no incidence on the conclusions.

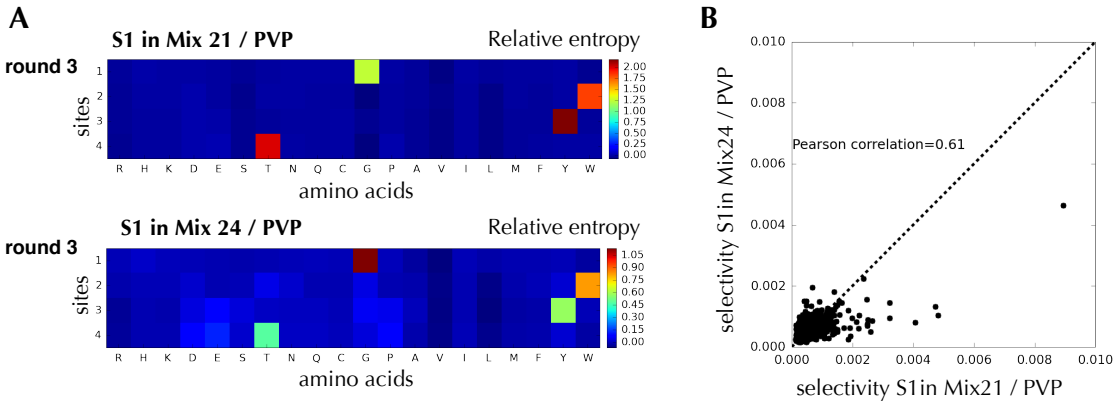


Figure S16: Reproducibility of selections against the PVP target – The results of two experiments of selection against the PVP target, one starting from a mixture of 24 libraries and the other from a subset of 21 libraries, which each are dominated by the S1 library, not only lead to an identical consensus sequence (panel A) but to reproducible results by EVT analysis (Table S3). In this case, not only are the initial populations different, but also potentially the targets since the experiments were performed 1.5 year apart and PVP is subject to aging: this may explain the imperfect correlations between frequencies (panel B; by contrast, the selection of the F3 library against PVP was performed at the same time than the selection of the mixture of 24 libraries and differences in consensus sequences cannot be due to differences of the targets in this case).

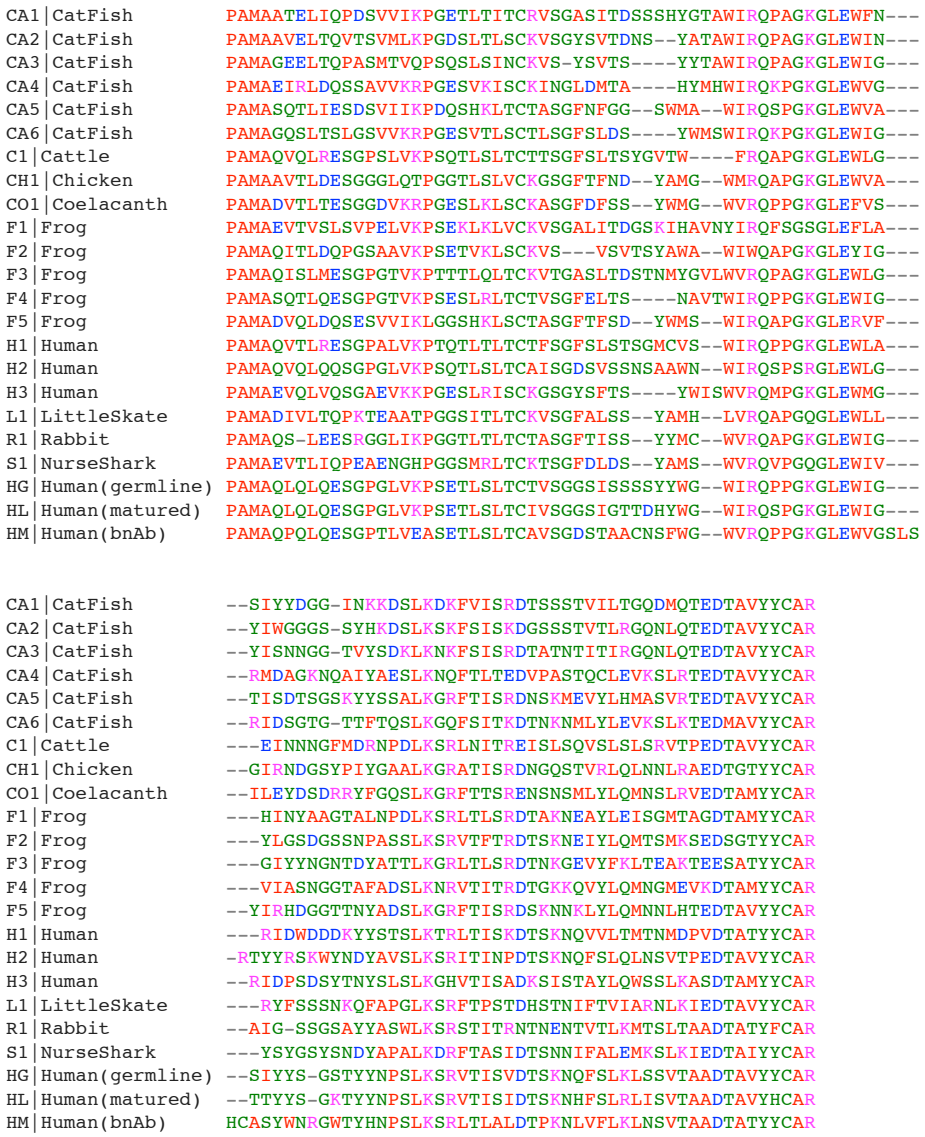


Figure S17: Amino acid sequences of the frameworks – Multiple sequence alignment of the library-specific part of the frameworks (Figure 1). The organism from which the sequence originate is indicated. See Dataset S1 for the nucleotide sequences.

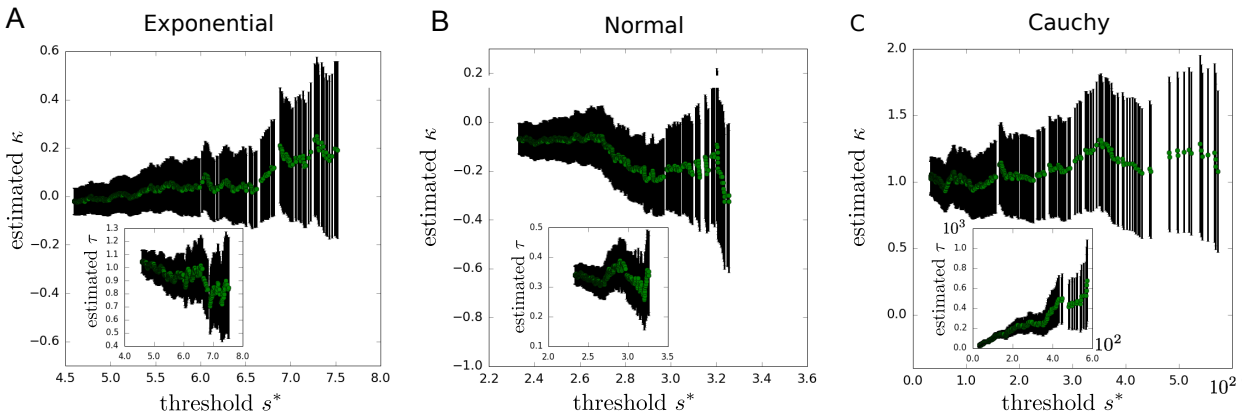


Figure S18: Illustration of the EVT analysis with simulations from standard distributions – We numerically drew 10^5 samples from the exponential, normal and Cauchy distributions and analyzed in each case the top 10^3 values by the same procedure that we used to analyze our data in Figure 4. **A.** For the exponential distribution, a fit of the general model gives $\kappa = -0.02 \pm 0.05$, $\tau = 1.04 \pm 0.08$ while a fit of the exponential model ($\kappa = 0$) gives $\tau_0 = 1.02 \pm 0.06$; the exponential model is excluded with a p-value 0.81, which is non significant. These results are consistent with the analytical result $\kappa = 0$. **B.** For the normal distribution, a fit of the general model gives $\kappa = -0.07 \pm 0.06$, $\tau = 0.34 \pm 0.03$ while a fit of the exponential model ($\kappa = 0$) gives $\tau_0 = 0.31 \pm 0.2$; the exponential model is excluded with a p-value 0.14, which is non significant. These results are consistent with the analytical result $\kappa = 0$. The exponential and normal distributions are examples of two different distributions that belong to the same class of extreme value statistics. **C.** For the Cauchy distribution, a fit of the general model gives $\kappa = 1.03 \pm 0.12$, $\tau = 30.3 \pm 3.8$ and the exponential model is excluded with a p-value $< 10^{-4}$, which is significant. These results are consistent with the analytical result $\kappa = 1$.

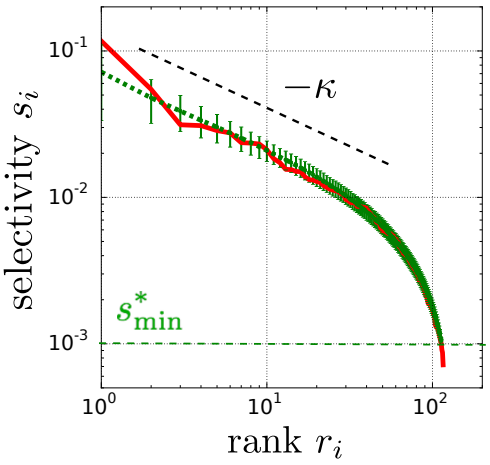


Figure S19: Fit of the generalized Pareto distribution versus a power law distribution – The data shown in red is the same as in Figure 3B (HG library against DNA target). The green dotted line indicates the best fit to a generalized Pareto distribution, using the values of κ and τ inferred in Figure S8. This graph shows that the fit extends far beyond the range of selectivities that may be fitted by a power law (black dotted line).

5 Supplementary code

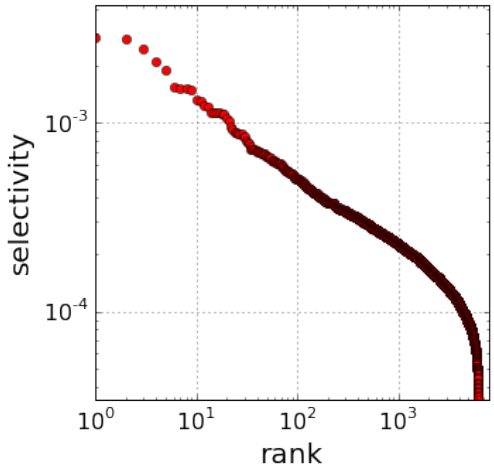
The following code is in the format of an IPython notebook. It assumes that the data has been processed and formatted into a file (here named `Data/S1_PVP.dat` and containing the results of the selection of the S1 library against the PVP target as in Figures 3A and 4) with 3 columns separated by tabs containing respectively, a sequence or a label of the sequence, an estimation of the selectivity of this sequence, and an estimation of the error in the value of the selectivity. One parameter, `s_star`, must be set by hand in cell `[6]` based on the plots obtained in cell `[5]`.

```
In [1]: %matplotlib inline
        import matplotlib.pyplot as plt
        import numpy as np
        import scipy
        import scipy.stats as ss
```

```
In [2]: # Path to the input file:
        data_file = 'Data/S1_PVP.dat'

        # Reading the data into dictionaries for selectivities and errors:
        sel_dict, err_dict = dict(), dict()
        for line in open(data_file , 'r'):
            seq, sel, err = line.split('\t')
            sel_dict[seq] = float(sel)
            err_dict[seq] = float(err)
```

5.1 Selectivity versus rank

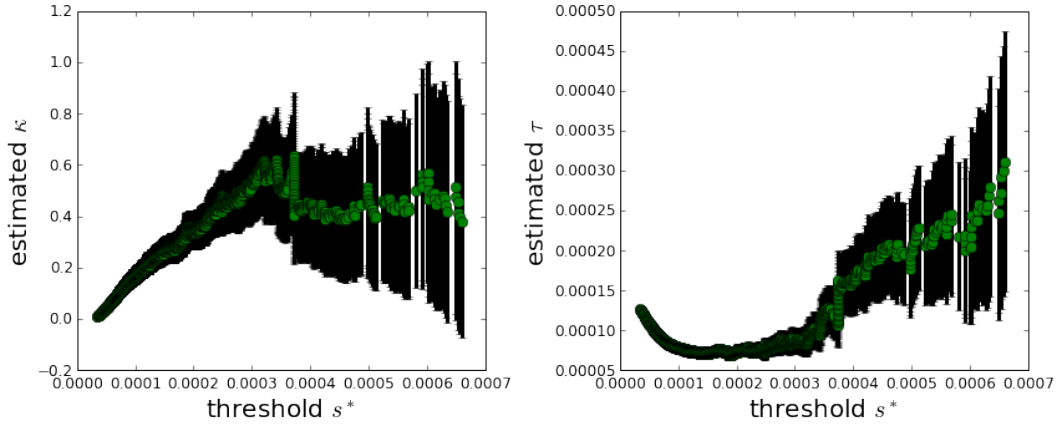
```
In [3]: # Sorting the data by decreasing values of selectivies:
        seq_sorted = sorted(sel_dict, key=lambda s: -sel_dict[s])
        sel_sorted = [sel_dict[s] for s in seq_sorted]
        err_sorted = [err_dict[s] for s in seq_sorted]

        # Figure 3A:
        plt.rcParams['figure.figsize'] = 5, 5; plt.rc('font', size=16)
        plt.loglog(range(len(sel_sorted)), sel_sorted,'or', lw = 2);
        plt.xlabel('rank', fontsize=20); plt.ylabel('selectivity', fontsize=20)
        plt.axis([0, 1.3*len(sel_sorted), sel_sorted[-1], 1.3*sel_sorted[0]]); plt.grid('on')
```



In [4]:
```
def log_likelihood_fct_exp(para, data_sorted):
    ''' Log-likelihood function for the exponential model '''
    tau, mu = para[0], min(data_sorted)
    return -sum([np.log(np.exp(-(x-mu)/tau)/tau) for x in data_sorted[:-1]])

def log_likelihood_fct_evt(para, data_sorted):
    ''' Log-likelihood function for the general model '''
    kappa, tau, mu = float(para[0]), para[1], min(data_sorted)
    return -sum([np.log((1+(x-mu)*(kappa/tau))**(-((kappa+1))/kappa)/tau)\
                 for x in data_sorted[:-1]])

def info_mat_exp(tau, data_sorted):
    ''' Observed information matrix for the exponential model '''
    mu = min(data_sorted)
    data = [x-mu for x in data_sorted[:-1]]
    return -1/sum([(tau-2*x)/tau**3 for x in data])

def info_mat_evt(para, data_sorted):
    ''' Observed information matrix for the general model '''
    matrix = np.zeros((2,2))
    kappa, tau, mu = para[0], para[1], min(data_sorted)
    data = [x-mu for x in data_sorted[:-1]]
    matrix[0][0] = -sum([(-kappa*x**2+tau**2-2*tau*x)/(tau*(kappa*x+tau))**2 for x in data])
    matrix[0][1] = -sum([x*(tau-x)/(tau*(kappa*x+tau)**2) for x in data])
    matrix[1][0] = -sum([x*(tau-x)/(tau*(kappa*x+tau)**2) for x in data])
    matrix[1][1] = -sum([(kappa*x*(kappa*(kappa+3)*x +2*tau)\
        -2*(kappa*x+tau)**2*np.log(1+kappa*x/tau))/(kappa**3*(kappa*x+tau)**2) for x in data])
    return np.linalg.inv(matrix)

def threshold_scan(sel_sorted, min_points, para=[1,0.001]):
    ''' Maximum likelihood estimation of kappa, tau for different values of the threshold.
    min_points sets the minimum number of points to be kept '''
    para_list, err_list = list(), list()
    for i in range(len(sel_sorted), min_points, -1):
        para = scipy.optimize.fmin(log_likelihood_fct_evt, para,\
                                   args=(sel_sorted[:i],), disp=False, maxiter=1000)
        para_list.append(para)
        err = info_mat_evt(para, sel_sorted[:i])
        err_list.append([1.96*np.sqrt(err[1][1]), 1.96*np.sqrt(err[0][0])])
    return para_list, err_list
```

In [5]:
```
min_points = 50
para_list, err_list = threshold_scan(sel_sorted, min_points)

# Figure 4A:
plt.rcParams['figure.figsize'] = 12, 5; plt.rc('font', size=12)
plt.subplot(121)
plt.errorbar(sel_sorted[min_points:][::-1],[p[0] for p in para_list],\
             [e[0] for e in err_list],fmt='k.',linewidth=3)
plt.plot(sel_sorted[min_points:][::-1],[p[0] for p in para_list],'go',markersize=8)
plt.xlabel(r'threshold $s^*$',fontsize=20); plt.ylabel(r'estimated $\kappa$',fontsize=20)
plt.subplot(122)
```

```
        plt.errorbar(sel_sorted[min_points:][::-1],[p[1] for p in para_list],\
                     [e[1] for e in err_list],fmt='k.',linewidth=3)
        plt.plot(sel_sorted[min_points:][::-1],[p[1] for p in para_list],'go',markersize=8)
        plt.xlabel(r'threshold $s^*$',fontsize=20); plt.ylabel(r'estimated $\tau$',fontsize=20)
        plt.tight_layout();
```




```
In [6]: # Choice of the treshold based on previous plots:
        s_star = .0004
```

5.2 Parameters of the best fit

```
In [7]: # Truncation of the data given s_star:
        sel_trunc = [s for s in sel_sorted if s > s_star]
        mu = min(sel_trunc)
        N_samples = len(sel_trunc)

        # Maximum likelihood estimation of tau under the exponential model:
        print 'Exponential fit:'
        para = scipy.optimize.fmin(log_likelihood_fct_exp,[.01],args=(sel_trunc,), maxiter=10000)
        tauexp = para[0]
        print 'tau_exp = %.5f\n' % tauexp

        # Maximum likelihood estimation of kappa, tau under the general model:
        print 'General fit:'
        para = scipy.optimize.fmin(log_likelihood_fct_evt, [.5,.01], args=(sel_trunc,), maxiter=10000)
        kappa, tau = para[0], para[1]
        print 'kappa = %.3f, tau = %.5f' % (kappa, tau)
```

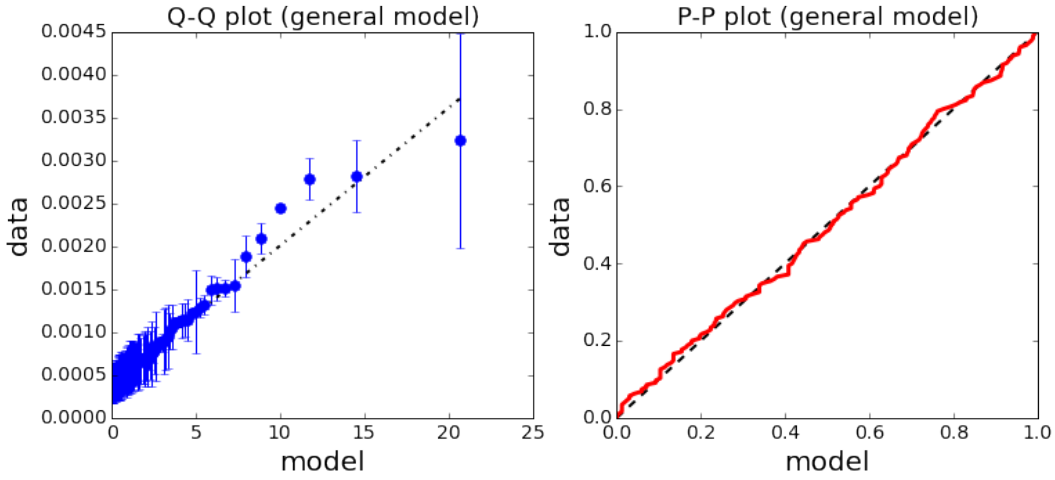
```
Exponential fit:
Optimization terminated successfully.
         Current function value: -1249.910350
         Iterations: 21
         Function evaluations: 42
tau_exp = 0.00028

General fit:
Optimization terminated successfully.
```

```
         Current function value: -1267.475534
         Iterations: 49
         Function evaluations: 95
kappa = 0.453, tau = 0.00016
```

5.3 Diagnosis plots for the fit

In [8]:

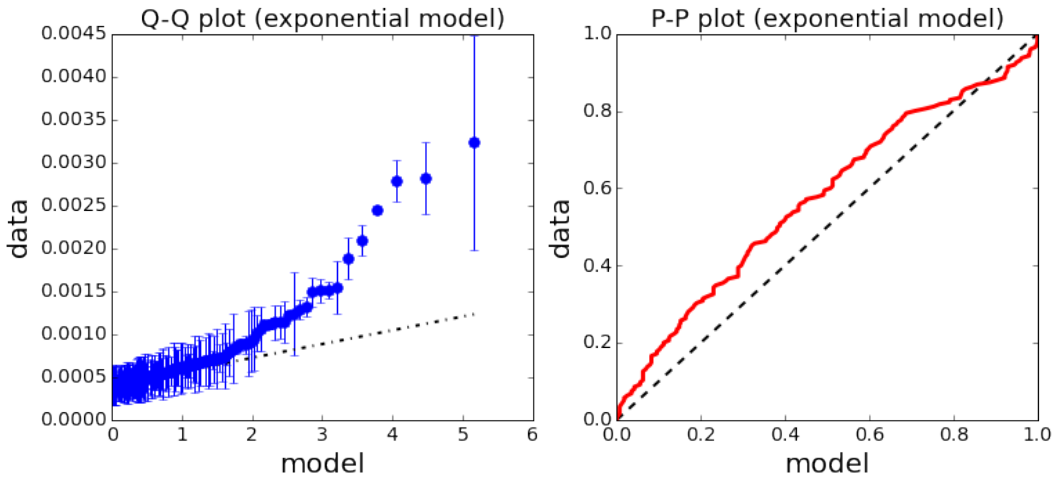
```
# Q-Q plot and P-P plots with the general model:
x_range = np.linspace(0, 1-1./N_samples, num=N_samples)
        sel_model = [((1-x)**(-kappa)-1)/kappa for x in x_range]
        sel_pp = [1-(1+kappa*(s-mu)/tau)**(-1/kappa) for s in sel_trunc[::-1]]

        # Figure 4B:
        plt.rc('font', size=14)
        plt.subplot(121); plt.title('Q-Q plot (general model)', fontsize=18); # Q-Q plot
        plt.plot([0,max(sel_model)],[mu,mu+tau*max(sel_model)],'k-.',linewidth=2);
        plt.errorbar(sel_model, sel_trunc[::-1], yerr=err_sorted[:N_samples][::-1],\
                     fmt='b.', markersize=15)
        plt.xlabel('model', fontsize=20); plt.ylabel('data', fontsize=20)
        plt.subplot(122); plt.title('P-P plot (general model)', fontsize=18); # P-P plot
        plt.plot([0,1],[0,1],'k--', lw=2)
        plt.plot(sel_pp, x_range ,'r',lw=3)
        plt.xlabel('model', fontsize=20); plt.ylabel('data', fontsize=20);
```



In [9]:

```
# Q-Q plot and P-P plots with the exponential model:
        plt.subplot(121); plt.title('Q-Q plot (exponential model)', fontsize=18); # Q-Q plot
        sel_model = [-np.log(1-x) for x in x_range]
        plt.plot([0,max(sel_model)],[mu,mu+tau*max(sel_model)],'k-.',linewidth=2);
        plt.errorbar(sel_model, sel_trunc[::-1], yerr=err_sorted[:N_samples][::-1],\
                     fmt='b.', markersize=15);
        plt.xlabel('model', fontsize=20); plt.ylabel('data', fontsize=20);
        plt.subplot(122); plt.title('P-P plot (exponential model)', fontsize=18); # P-P plot
        sel_pp = [1-np.exp(-(s-mu)/tauexp) for s in sel_trunc[::-1]]
        plt.plot([0,1],[0,1],'k--', lw=2)
```

```
plt.plot(sel_pp, x_range ,'r',lw=3);
plt.xlabel('model', fontsize=20); plt.ylabel('data', fontsize=20);
```



5.4 Significance of $\kappa \neq 0$

```
In [10]: def distr_LikeRatioTest(tauexp, N_samples, N_draws):
             ''' Making a null distribution for a likelihood ratio test of kappa not 0'''
             distr = list()
             for n in range(N_draws):
                 # Drawing samples from the exponential model:
                 sel_rand = np.random.exponential(tauexp, N_samples)
                 # Fitting them by maximum likelihood estimation with the general model:
                 para = scipy.optimize.fmin(log_likelihood_fct_evt, [.5,.01],\
                                            args=(sel_rand,), disp=False, maxiter=1000)
                 distr.append(2*(log_likelihood_fct_exp([tauexp], sel_rand)\
                                 - log_likelihood_fct_evt(para, sel_rand)))
             return distr

In [11]: # null distribution:
         N_draws = 10000
         distr = distr_LikeRatioTest(tauexp, N_samples)
         # experimental value:
         dL_sel = 2*(log_likelihood_fct_exp([tauexp], sel_trunc)\
                     - log_likelihood_fct_evt([kappa,tau], sel_trunc))
         # p-value:
         p_val = sum([1 for dL in distr if dL > dL_sel])/(1.*len(distr))
         if p_val > 0:
             print 'p-value against the exponential model: %.e' % p_val
         else:
             print 'p-value against the exponential model: < %.e' % 1./N_draws

p-value against the exponential model < 1e-04

In [12]:
# Representation of the estimated kappa from experimental measurements
# versus the distribution of estimated kappa for samples from an exponential model:
```

```
plt.hist(distr, 100, label=r'$\kappa = 0$')
         plt.axvline(dL_sel, color='r', lw=3, label='data');
         plt.legend(bbox_to_anchor=(1.2,1.03));
         plt.xlabel('test statistics', fontsize=20); plt.ylabel('counts', fontsize=20);
```