Parameters, inference, and maturation dependence of selection potentials in antibody scaffolds

Steven Schulz

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée au Collège de France

Parameters, inference, and maturation dependence of selection potentials in antibody scaffolds

Soutenue par
Steven Schulz
le 1er octobre 2020

École doctorale n°564
Physique en Île-de-France

Spécialité
physique statistique

Composition du jury :

Martine Ben Amar, Professeur École normale supérieure Paris, PSL	*Présidente*
Shenshen Wang, Professeur University of California, Los Angeles	*Rapportrice*
Andrea Pagnani, Professeur Politecnico di Torino	*Rapporteur*
Martin Weigt, Professeur Sorbonne Université	*Examinateur*
Olivier Rivoire, CR CNRS Collège de France, PSL	*Directeur de thèse*
Clément Nizak, CR CNRS ESPCI ParisTech, PSL	*Directeur de thèse*

COLLÈGE DE FRANCE
1530

current version compiled on: 2020-10-26 15:51:44+01:00
identical to: 2020-10-26 15:38:58+01:00, modulo external figures removed for copyright reasons

Steven Schulz: *Parameters, inference, and maturation dependence of selection potentials in antibody scaffolds*, 1er octobre 2020

"[. . . R]eality must take precedence over public relations,
for nature cannot be fooled."

Richard P. Feynman
in appendix F to the Rogers Commission Report
on the Space Shuttle Challenger Accident, 1986

"Nobody else took what I was doing seriously, so nobody would want to work with me. I was thought to be a bit eccentric and maybe cranky."

Peter W. Higgs
in an interview on The Life Scientific
on BBC Radio 4, 2014

「……懐かしい村に戻って来た太郎は、すぐ家の方に向かって走りました。けれども家が、自分の家がありません。太郎は辺りの人々に『浦島太郎の家を知りませんか』と尋ねましたが、誰も知りません。人々も見たこともない人ばかりです。村の様子もすっかり変わっています。太郎は会う人ごとに『浦島太郎の家を知りませんか』と尋ねました。だが太郎の名前さえ知っている人は誰もいません。疲れきって太郎は道端に座り込んでしまいました。この時一人のおじいさんが通りました。太郎はおじいさんに『この辺りに浦島太郎の家はありませんか』と尋ねました。おじいさんはしばらく考えてから『昔、そう言う名前の人いたと言うことを聞いたことがあります。でもその人は大昔の人です。だから家なんて残っているは図がありません。どの辺に住んでいたかももうはっきりとは分からないと思います』と答えました。太郎は耳を疑いました。訳が分からなくなりました。……」

excerpt from 浦島太郎
Japanese fairytale

Title

Parameters, inference, and maturation dependence of selection potentials in antibody scaffolds

Abstract

We characterize antibody "evolvability" by combining high-throughput techniques from molecular biology and tools from statistical physics and data science, an interdisciplinary approach already successfully applied in other biological contexts. Evolvability describes the ability of antibodies to evolve, *i.e.* the effect of mutation and selection on their phenotype. It is an essential property for the success of affinity maturation, an accelerated evolutionary process leading to antibodies with improved binding affinity to a given pathogen. Can we observe evolvability? Can we define a mathematical parameter that represents evolvability? Can we measure this parameter? What antibodies are promising starting points for affinity maturation? Here, we study the effect of evolution on binding affinity by mimicking the initial step of affinity maturation against various antigenic targets: We select for binding affinity from libraries of randomized antigen binding sites using phage display and high-throughput sequencing. Our libraries are built around human antibody scaffolds exhibiting different levels of previous maturation against a third-party target (HIV). We observe vast differences in their response to selection, 1) at the intra-library level with few, target-specific variants strongly dominating all others, 2) at the inter-library level with the naïve library systematically dominating mature libraries. Using statistical physics, we argue how these hierarchies are linked to selection potential, a component of evolvability that we define as the susceptibility to variation and selection. We establish that inter- and intra-library differences share a common origin captured by a single, library-dependent, generative parameter σ encoding for the variance of binding energies (Mathusian fitness) within libraries. Interestingly, highest selection potentials are systematically observed in the library based on a naïve antibody, suggesting a scenario of naïve antibodies being "evolved to evolve".

Keywords

evolution, evolvability, selection potential, *in vitro* evolution, high-throughput sequencing, antibody, affinity maturation

Titre

Paramètres, inférence et dépendance de la maturation des potentiels sélectifs dans les échafaudages d'anticorps

Résumé

Nous caractérisons l'«évoluabilité» des anticorps en combinant des techniques à haut débit en biologie moléculaire, des outils inspirés de physique statistique et les sciences des données, une approche interdisciplinaire déjà implantée dans d'autres contextes biologiques. L'évoluabilité décrit la capacité d'anticorps à évoluer, c'est-à-dire à sélectionner des phénotypes plus favorables sous l'effet de mutations aléatoires. Celle-ci est une propriété essentielle pour la maturation d'affinité qui est un processus évolutif permettant d'augmenter l'affinité des anticorps contre un pathogène donné. Peut-on observer l'évoluabilité ? Peut-on définir un paramètre mathématique qui représente l'évoluabilité ? Peut-on mesurer ce paramètre ? Quels anticorps sont des points de départ prometteurs pour la maturation d'affinité ? Ici, nous étudions l'effet de l'évolution sur l'affinité de liaison en imitant les premières étapes de la maturation d'affinité contre plusieurs cibles antigéniques : Nous sélectionnons l'affinité de liaison dans des banques d'anticorps randomisés sur leurs sites de liaison en utilisant le phage display et le séquençage à haut débit. Nos banques sont construites sur la base d'échafaudages d'anticorps humains possédant des niveaux différents de maturation antérieure contre une cible tierce (VIH). Nous observons des différences importantes dans leurs réponses face à la sélection, 1) au niveau intra-banque avec peu de variants spécifiques à la cible qui dominent tous les autres variants, 2) au niveau inter-banque la banque naïve dominant systématiquement les banques maturées. En utilisant la physique statistique, nous expliquons comment ces hiérarchies dérivent du potentiel sélectif, une composante de l'évoluabilité que nous définissons comme la susceptibilité à la variation et à la sélection. Nous élaborons que les hiérarchies inter- et intra-banques résultent d'une même origine décrite par un paramètre dépendant de la banque et génératif, σ qui encode pour la variance d'énergies de liaison (valeurs sélectives malthusiennes) dans les banques. Curieusement, le potentiel sélectif le plus élevé est observé systématiquement dans la banque basée sur un anticorps naïf ce qui suggère un scénario où les anticorps naïfs auraient été «evolués pour évoluer».

Mots-clés

évolution, évoluabilité, potentiel sélectif, évolution *in vitro*, séquençage à haut débit, anticorps, maturation d'affinité

Contents

List of Figures

List of Tables

Prolog

I am a theoretical physicist by education, but a considerable part of this PhD project and manuscript consists of molecular biology and experiments. It is in this respect, that the project was all in itself an experiment to me and, to take up the words of one of my supervisors, I "must have been completely out of mind" at the time I decided to go for it. Given my experience today, I can do nothing but agree, with all the positive and negative connotations that are associated to these words. Yet, the motivation to go this way is very clear: Today's research in physics, biology, and "data science" (if you want to call it this way) happens at the crossroads of these disciplines; physicists are working on biological data and biologists are using physical experimentation (yes, I found the word "antibody" and Planck's constant $\hbar$ within the same paper [2, 3, 4]). However, physicists oftentimes lack the understanding of where biological data comes from and how it is obtained, and *vice versa*. But critical information is oftentimes concealed in some hidden corner of scientific literature or simply unavailable. I had found myself facing this situation during my Master's internship and with no doubt, this PhD project gave me the opportunity to explore the opposite site, to take a look inside the black box, to learn molecular biology and experiments from scratch, and to gain a broader vision in addition to my prior theoretical knowledge. The downsides of such a career, however, have to be emphasized as well. People tend to praise interdisciplinarity, but the daily life experience is sometimes disillusioning: Communication and mutual understanding between people of different disciplines is oftentimes suboptimal. This is nowhere more problematic than when it comes to fighting with referees who represent the more traditional backgrounds and are trying to lobby for their stance. The weighing of "specialists" and "generalists" is the very topic of this manuscript but, as previous research concluded [5, 6], the path of a generalist is a narrow one, quenched in between the strong attractors of specialization and frustration. In summary, such an experience is scientifically incredibly rich, but strategically questionable. I do not want to miss this experience, but I find myself a little more conservative than before and I am not sure I would decide to go the same way again given that I have a preference to stick to the principle of least action.

The present manuscript attempts the definition, observation, and measurement of evolvability or selection potential on a system that occurs as much in nature (affinity maturation) as it does in clinical technology (vaccine and drug design, diagnosis), namely a diversity of antibodies that

faces selection for binding to a given target or pathogen. The manuscript contains four main chapters (excluding introduction and conclusion) that roughly divide it into the "philosophy of our problem", theory of the theory, theory of the experiments, and the analysis and interpretation of the experiments. In addition, I provide an extensive appendix with our experimental protocols, supplementary figures, tables, and computations, as well as python code used in our analyses, and a preprint. The premise of my writing was to make all the ingredients of the project accessible to everyone, that is, the molecular biology experiments to non-biologists and the theory with equations to non-physicists; and I hope I managed to achieve this goal more or less. We made high-throughput sequencing data from antibody library selections generated prior to and during this PhD available in unprocessed and preprocessed form through the NCBI Sequence Read Archive and a shared Dropbox folder, respectively.

A list of abbreviations can be found in table D.1 on page 188. A list of quantities and symbols that I use throughout the manuscript can be found in table D.2 on page 189. I owe an apology to statistical physicists who denote the configuration of a sequence of spins typically by σ. For some reason that I totally forgot, we here adopt the notation x for such configurations, and σ is assigned to the key quantity of selection potential. I also want to emphasize that my writing style naturally makes use of the first person plural ("we"), and so do I in this manuscript. This is a deeply rooted reflex that traces back to the very beginnings of my undergraduate studies: One of my professors (and later supervisor of my Bachelor thesis) said one day in an off-topic discussion during the Classical Mechanics class that the use of "we" would be a sign of good practice in scientific output, reflecting the fact that research is a collective labour. I have since then systematically stuck to his suggestion. It seems relevant to me to note this anecdotical fact here, because I have seen people use the first person singular in doctoral manuscripts; the idea behind being supposedly that the bare task of writing up a thesis is a rather solitary task (which is true and particularly true during coronavirus lockdown).

Get your popcorn ready and enjoy ...

Introduction

Evolution is the designer of the forms and functions of living matter that we can wonder at in nature today. Starting from scratch or a primitive ancestor, it has acted through hundreds of millions of years by variation and selection to shape, improve, and adapt new species to their environment(s). Beyond biology, evolutionary algorithms are also successfully applied *e.g.* in computational contexts. The simplest (zero-temperature) example is probably the optimization of an objective function through random search. Given that variation (mutation) and selection are the key mechanisms of evolution, as according to Charles Darwin, one may ask whether evolution can be effective inconditionally. Can whatever object or subject come up with a solution (aka adaptation or new function) to a given task (aka selective pressure), that is, be "evolvable"? On the one hand, every possible mutation on an unevolvable object would be neutral or destructive. On the other hand, some mutation on an evolvable object would be the path towards improved function. If evolvability turns out relevant, what does evolvability depend on and are certain objects in biology evolved to be evolvable? The literature contains a number of theoretical studies and mathematical models of evolvability in biological systems. In addition, evolvability in proteins has been proposed to be correlated with other biophysical and structural properties, such as thermodynamic stability and polarity (see chapter 1). A quantity directly encoding for the "strength" of an evolutionary response and its measurement, however, are missing. The goal of this PhD project was to find such a parameter.

Here, we define, measure, and reveal the factors of selection potentials which we introduce as a component of evolvability in a model system that, besides for the study of evolution, is ubiquitous in clinical contexts: Libraries of antibodies that are selected and evolved for binding to given target molecules. The rationale for the use of antibodies is multisided: In general, proteins allow for a practical definition of variation and fitness/phenotype in terms of (changes in) amino acid sequence and well-defined physical quantities such as binding affinity, respectively. In addition, antibodies are subject to a standalone, time-lapse evolutionary process in jawed vertebrates that allows the organism to specifically fight a plethora of potential foreign pathogens in case of encounter, while avoiding to target the organism itself. Taken together, antibodies represent a convenient model system for the study of (Darwinian) evolution in general, as molecular phenotypes and involved timescales allow for mathematical modeling and quantitative data. The design of our libraries

consists of random antibody binding site (notably CDR3) sequences in the context of a fixed antibody scaffold sequence. This is akin to primary repertoire formation upon the initialization of the adaptive immune response where high evolvability of the antibody is supposedly important: Choosing an antibody scaffold corresponds to recombination of V, D, and J gene fragments and CDR3 randomization corresponds to random insertions upon junction of the V, D, and J fragments to form the CDR3. On the theoretical side, the study and inference of certain classes of random biophysical models is required for the definition and measurement of selection potentials.

We generate selection trajectories for various combinations of antibody libraries and binding targets. Most notably, we find selection potentials that are independent of the binding target, but differ significantly between libraries. This represents to our knowledge a first direct observation of differences in evolvability in an *in vitro* biological system. This result suggests, that evolvability is a property of the library, irrespective of the selective task, and crucially depends on the antibody scaffold used for library construction. Interestingly, the evolvability appears to systematically decrease with increasing level of maturation of the antibody scaffold within the limited set of antibody libraries studied here. This is seemingly consistent with literature on antibody dynamics that oftentimes reports rigidification of initially flexible antibodies upon affinity maturation (see chapter 1). This also suggests that germline antibodies may have been evolved and selected to feature high evolvability in light of their task in the adaptive immune response.

Perspectives of our findings comprise (i) the study of the interdependence of our evolvability index with other protein phenotypes/properties, and (ii) reveal possible controls of this index which would be of interest *e.g.* in clinical applications. Regarding (i), our work suggests previous maturation as a key determinant of antibody evolvability. To systematically study correlations between evolvability and maturation level, protein dynamics, stability, and other properties, our experimental assessment of evolvability needs to be scaled up to allow for the testing of many different antibodies at many time points on a maturation trajectory. This scale-up resumes to speed-up and parallelization, and efforts in this direction are being made within the research group.

The structure of the manuscript is as follows: In chapter 1, we provide a brief recap of the basics of our question and approach: Darwin's theory of evolution, the role and structure of the antibody and its affinity maturation, as well as the definition and current knowledge on evolvability in light of the literature. The following chapters 2 and 3 approach our question from the theoretical and experimental viewpoints, respectively. Starting from the kinetic equations and a class of random models for biophysical interactions widely used in the modeling of protein evolution and elsewhere, we establish in chapter 2 the distribution of fitness values, or enrichments, when the selective pressure is defined by equilibrium binding. This leads to the lognormal distribution with parameters μ and σ. The definition of selection potential in the form of a single scalar quantity, which is precisely given by σ, is motivated from physical and information-theoretic viewpoints and its implications in light of the random biophysical models are discussed. In chapter 3, we present in detail the choice, construction, and cloning of antibody libraries, as well as their expression and

in vitro selection for binding to well-defined target molecules by phage display and biopanning. Moreover, the high-throughput sequencing strategy, as well as the pipeline for measurement of sequence frequencies and enrichments from the sequencing of selected and unselected antibody libraries are explained. The chapters 4 and 5 confront experimental data from chapter 3 with the simple models from chapter 2, *i.e.* the lognormal distribution with our evolvability index as a model parameter (chapter 4) and simple biophysical models of binding (chapter 5). Chapter 4 discusses general features of the selection data, before focusing on the inference procedure, assessment of fit quality and predictive power, and the comparison of model parameters including selection potential across many combinations of antibody libraries and binding targets. We also propose in chapter 4 a re-analysis of the same selection data in light of the more complicated, yet still very simple independent-site model using an inference method based on multi-type branching processes. The results are still preliminary and we content ourselves with a general discussion. In particular, a reanalysis should allow for a test of the independent-site assumption within the antibody binding site that goes into the lognormal distribution of enrichments. Finally, we briefly summarize our overall results, discuss their implications, and sketch ideas for future research on their basis in chapter 5. Supplementary figures, tables, computations, and experimental methods, as well as python code used for the inference of lognormal and additive models, and the simulation of selection experiments are provided.

Chapter 1

Towards quantifying evolvability

Owing to a large number of theoretical contributions from various viewpoints, the notion of "evolvability" now resides on a robust conceptual basis. However, none of these have materialized in the actual observation and measurement of evolvability in real biological systems as yet, which defines the goal of this project. The focus of this introductory chapter will be the definition of "evolvability", a review of current knowledge, and the introduction of a model system that we think is the ideal candidate for an experimental approach to evolvability: the antibody, a key agent of the adaptive immune system. First, we will derive the notion of evolvability from Darwin's first principles of natural evolution based on a separation of timescales. (See section 1.1.) Our viewpoint towards the problem is the one of quantitative biology: Since Darwin and his shooting of birds on the Galápagos islands, modern observation techniques, such as high-throughput sequencing, extended our look at biological systems to both larger and smaller scales, such as individual proteins and libraries (populations) of proteins. Here, the traditional notions of "species", "fitness", and "variation" become meaningful and mathematical models, in combination with large-scale biological data, become useful. (See section 1.2.) We will argue why evolvability is presumably critical and observable in antibodies which evolve on two different timescales: between generations and within a generation during the adaptive immune response. (See section 1.3.) Previous insights into evolvability based on mainly theoretical and computational studies of protein evolution will be briefly discussed. (See section 1.4.) Finally, and for the sake of reference, we provide biological details about the short-term evolution of antibodies, the so-called affinity maturation. These may be helpful to understand our notion of "maturation degree" of an antibody. (See chapter B.) A list of notions that we introduce in this chapter is provided in table 1.1.

term	definition
evolvability	ability of an object to yield improvement (or functional innovation when the objects are biomolecules) upon an evolutionary optimization process; see subsection 1.2.2 for more, somehow equivalent definitions
selection potential	susceptibility to selection/efficiency of response to selection of a library or repertoire of random objects
antibody scaffold region	part of the antibody sequence that is germline-encoded (before somatic hyper-mutation) and thus subject to the antibody's long-term evolution; comprises notably FWR1, 2, 3, 4, CDR1, 2
antibody non-scaffold region	part of the antibody sequence that is *not* germline-encoded (before somatic hyper-mutation); comprises most of CDR3
affinity maturation	short-term evolution of the antibody as part of the adaptive immune response by a process in which it cyclically acquires random somatic mutations and is selected for improved binding to a pathogen
maturation degree	"amount" of somatic-mutational history of an antibody, as captured by the time since start of the affinity maturation, or the number of fixed somatic mutations

Tab. 1.1: Summary of a few notions introduced in this chapter and used throughout the manuscript, as well as their respective definitions.

1.1 Evolution and evolvability

Here, we briefly review Darwin's theory of natural evolution, which unites selection, variation, and inheritance as the three ingredients that take evolution forward (subsection 1.1.1). We then explain the notion of "evolvability" which is not part of this theory, but may be expected as a direct consequence (subsection 1.1.2). We introduce "selection potential" as a factor of evolvability that we seek to mathematically define and measure in later chapters (subsection 1.1.3).

1.1.1 Evolution à la Darwin: mutation, selection, and inheritance

Darwin formulated his ideas on why life is the way it is in the second half of the 19^{th} century in his book *On the Origin of Species* [7], and his theory of natural evolution is the one most accepted and estimated most relevant nowadays. It prevails over a number of other more or less similar, competing non-Darwinian theories, such as a variation of Lamarck's theory and orthogenesis. According to Darwin's theory, evolution is governed by (i) variation, (ii) natural selection, and (iii) inheritance. (i) Variation terms the fact that no offspring is an identical copy of the parent and the mechanisms behind variation are subject to another theory in itself (genetic variation,

Mendelian inheritance). (ii) In a population, variation gives rise to a diversity of individuals with varied features and properties that may provide improvement, deterioration, or unaltered performance with respect to the needs in a given environment. The performance of a variant i in a fixed environment is measured by its "fitness" s_i that may represent its probability of survival or of the number of offspring. Selection designates the process of enriching certain variants over others in a population, *i.e.* increasing their number of copies n_i relative to those of others, according to their performance or adaptation to the environment which is precisely encoded in the (distribution of) fitness values s_i. Selection is required by the finiteness of ressources in the environment, which introduces interaction (competition) between individuals and variants: Those variants able to use these ressources better than others, are the ones favored and enriched upon selection. While ressources can mean space, food, and other physical factors, the most basic finite ressource is frequency $f_i = \frac{n_i}{\sum_j n_j}$, which is purely mathematical and must always sum to one. An important additional evolutionary factor not accounted for in Darwin's theory, but acute in nature, is genetic drift, which captures random changes (fluctuations) in frequencies f_i in finite populations with $\sum_i n_i < \infty$ that are not a consequence of fitness differences and selection of one over the other. (iii) Inheritance describes the fact that the phenotype of the parent is handed down to the offspring.

1.1.2 Evolvability: the propensity to evolve

Although selection, variation, and inheritance are today generally accepted as the three pillars of natural evolution, Darwin's theory is conceptually neither complete, nor precise, and comes with a number of open questions/problems. In particular, if we start from the expectation that "everything" in nature is under permanent pressure for improvement, then the mechanisms of evolution must themselves be subject to evolution. In this way, these mechanisms become recursively defined rather than preset once and for all: Evolution *of* these mechanisms at longer timescales thus manipulates evolution *by* these mechanisms at shorter timescales. Overall, evolution not only leads to adaptation, but also to the propensity to adapt. This motivates the notion of "evolvability", which is defined as the capacity of an object/subject to evolve and yield functional novelty [8]. It encodes for the efficacy of short-term evolution as arranged for (or not) by the long-term evolution. Theoretical studies, mostly based on mathematical models of protein evolution, disclosed the existence and relevance of evolvability (see last section 1.4 for details), most importantly showing that it can be selected for (although not directly, but via a combination of other selective pressures) [9, 10, 11, 12, 13] and that it thus has the status of a property/phenotype (like binding affinity, catalytic activity, *etc.*), a fact that has been under debate for some time [14, 8, 15, 16]. While evolvability is theoretically well-established, the challenge of defining, observing, quantifying, and ultimately controlling evolvability in experimental systems has not yet been achieved. Ultimately, the mere observation of evolvability does not represent a goal in itself, but is rather a first step towards a potential control of evolvability. The ability to do so should have implications in numerous biological and non-biological contexts that rely on experimental

directed evolution of biomolecules and where evolvability of the starting points of evolutionary trajectories is usually implicitly assumed. For instance, in vaccine and drug design, one wishes to push the defender (*e.g.* the immune response and its antibodies) towards high evolvability, whereas the aggressor (*e.g.* microbes with potential for multi-drug resistance) should be guided towards low evolvability [13]. This manuscript attempts to point the way towards this goal.

1.1.3 Selection potential

In nature, variation (or mutation) and selection presumably occur simultaneously and continuously. Experimentally, such as in directed evolution, the evolutionary process is generally such that these subprocesses are separated into disjoint time intervals: An elementary step, which consists of a period Δt of mutagenesis, followed by a period Δt of selection, is cycled many times; in the limit $\Delta t \to 0$, this should reproduce the continuous process. Consider the onset of selection after a mutation step that gave rise to a diverse population of many variants i with differences in their fitness values s_i: In the absence of mutations, the outcome of this selection step is determined by the properties of the initial population encoded in the list of fitness values $\{s_i\}_{i=1,\dots,N}$ and initial frequencies $\{f_i(t=0)\}_{i=1,\dots,N}$. Depending on whether the mutation step gives rise to minor or important differences in and absolute values of fitness, *i.e.* how strongly genotypic variation gives rise to phenotypic variation, the population will respond to selection more or less efficiently, *i.e.* with minor or drastic rises and falls in frequencies of variants. Selection will act efficiently, and quickly enrich high-fitness variants in populations featuring large phenotypic differences. This is the basis for our definition of "selection potential" as a factor of evolvability and is formalized by Fisher's fundamental theorem of natural selection, see section 1.2.2.

1.2 Quantitative approaches to evolution and evolvability

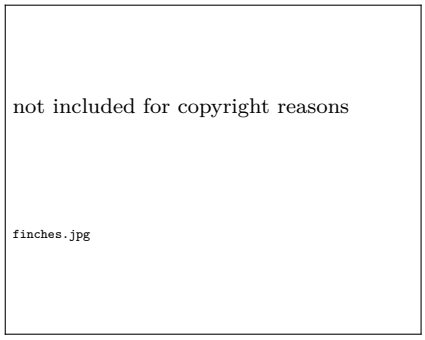
Darwin, as well as the previous section 1.1, relied on qualitative reasoning. However, the study of evolution has turned increasingly quantitative in the recent decades, due to the ever increasing availability of biological data, notably structures and sequences of biomolecules. We will seize the opportunity to quantitatively understand aspects of evolution in our experimental approach towards evolvability. There are several tracks for the quantitative study of evolution in general, and evolvability in particular: On the one hand, mathematical models and simulations of protein evolution are used to theoretically study aspects of evolvability in bottom-up approaches. On the other hand, quantitative experiments in combination with statistical modeling of the underlying systems are the ultimate test to any theory and closer to applications. Here, we travel from the macroscopic scale, where Darwin's finches live, to the microscopic world of proteins, where evolution is equally relevant, but where "quantitative biology" becomes possible (subsection 1.2.1). We present a simple mathematical model of selection and mutation in terms of "species", "fitness", "frequency"

that leads to Fisher's fundamental theorem of natural selection and the "quasi-species", and that will be relevant later and recur in chapter 2 in a time-discrete version (subsection 1.2.2).

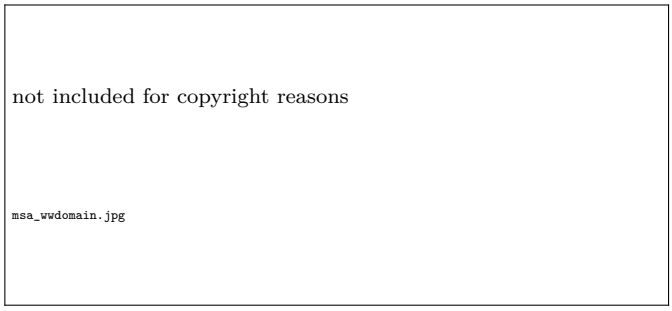
1.2.1 From Darwin's finches to the molecular level

Charles Darwin got close to the truth despite being limited in the scope of his observations. His thoughts were guided by mere animal observation, such as the discovery of "Darwin's finches", see figure 1.1(a): Upon sampling endemic birds on the Galápagos islands during his expedition on the HMS Beagle from 1831 to 1836, he (and John Gould, ornithologist) noticed a remarkable diversity of beak shapes and functionality within a group of otherwise similar bird species (*e.g.* similar song), that were altogether similar to a bird species on the American mainland. These could be correlated with different places and islands in the archipelago, as well as different food sources available at these places. Different beak shapes appeared to be adapted for the use of different food sources, *e.g.* large and strong beaks to crack nuts and small beaks to take up small seed. He ultimately concluded that different species must have branched and diverged from a common ancestor and shaped their beaks towards the observed adapted shapes, under the selective pressure of accessing different food sources and upon the "trial and fail" of (random) variation to the beak shape. As the progress in research since Darwin allows for observation and understanding of living matter at larger and smaller scales, it became clear that evolution is relevant across the scales, from populations of organisms down to proteins, which are functional heteropolymers that represent the molecular building blocks of life. Sequencing of the genomes of various species and construction of multiple-sequence alignments of protein homologs (the same protein in the context of different species) reveals differences in sequence, which are the result of branching events in the evolutionary past, see figure 1.1(b) for one example. In the end, Darwin's postulates of variation and natural selection provide a qualitative picture of the mechanisms of evolution. But to understand factors, mechanisms, and implications of Darwin's theory of evolution, quantitative descriptions and measurements are needed.

While the concepts of variation and selection can be easily translated into the language of mathematics in terms of frequencies f_i and fitness values s_i (see next subsection 1.1.2 for a simple mathematical model), these quantities, as well as the notion of "species" (or "variant") i itself are generally ill-defined and/or not measurable. At the lowest level, the level of single proteins, however, evolution takes a rather practical form and all these quantities and notions can be readily defined: (i) Proteins are uniquely defined by their sequence x of amino acids, $x = (x_1, x_2, \ldots, x_L)$ where x_i is the amino acid on site i, $i = 1, 2, \ldots, L$, of a sequence of length L, which can take on an alphabet of q values ($q = 20$ for amino acids), and each sequence x can be defined as a protein species i. (ii) Variation can be introduced by replacing amino acids, or even adding or removing them, upon erroneous replication (mutation). This defines variation as a random walk in sequence space, which can be pictured as a hyper cubic grid of length L and edge length q where each of the q^L nodes represents a unique sequence and which is



(a)



(b)

Fig. 1.1: Outcome of natural evolution at different levels. **(a)** At the developmental level: Darwin's finches featuring a variety of beak shapes and sizes, which have been selected for adaptation to different food sources. Taken from [7]. **(b)** At the molecular level: (Extract of an) Example of a multiple-sequence alignment (MSA) showing homologous protein sequences (*i.e.* evolutionarily related and with similar function) collected from many individuals across various species. Differences in amino acid sequences are the consequence of branching processes in the past and evolution into different directions. The protein shown here is the WW domain (`PF00397`); taken from [17].

endowed with the Hamming distance $d(x,y) = \sum_{i=1}^{L} \delta(x_i, y_i)$ as a non-Euclidean metric ($\delta(a,b)$ is the Kronecker Delta with $\delta(a,b) = 1$ if $a = b$ and $\delta(a,b) = 0$ otherwise). (iii) Proteins have biochemical properties (or phenotypes or functions), such as *e.g.* binding affinity K_D, catalytic activity z, and stability ΔG_{fold}. These property values differ for different sequences *a priori* and, as the sequence carries in principle all information about a phenotype, this defines a mapping $x \mapsto (K_D, z, \Delta G_{\text{fold}}, \dots)$ called the genotype-phenotype map. Yet the functional form of this mapping, or "landscape", is generally unknown. If in addition the mapping from phenotype to fitness, $(K_D, z, \Delta G_{\text{fold}}, \dots) \mapsto s$, is known, then the so-called fitness landscape $x \mapsto s(x)$ is known. In the case we realize experimentally here, selection involves equilibrium binding in a regime where the (Malthusian) fitness is given by $s(x) = \exp(-\beta \Delta G(x))$ and where ΔG is the free energy of binding. Of course, mutation and selection at the molecular scale are the path towards variation and selection at the organisms' scale and beyond, but it is unclear how the notions of variant, frequency, and fitness need to be renormalized in order to obtain their correspondances at these higher levels.

1.2.2 A simple mathematical model of Darwinian selection (and mutation)

To translate Darwin's concept of evolution into the more intuitive language of mathematics, let us consider a simple model of Darwinian selection (and later including mutation) with N different possible species, *e.g.* the $N = q^L$ unique sequences of length L. This model of selection will be realized experimentally in later chapters and, including the mutations, possibly in future

experiments within the group.

In general, the notion of "species" is rather ill-defined, as it is not clear how much difference between any two organisms is required to consider them as belonging to distinct species, or, how much difference is allowed to still consider them the same species. At the protein level, however, the relevance of a "sequence space" allows for a more rigorous definition of "species" (or "variant" or "mutant"): Proteins are (folded) chains of covalently bound amino acids. The sequence of amino acids fully characterizes and contains in principle all the information of a protein; each possible sequence gives rise to *a priori* different folds and phenotypes and differences between any two proteins must be the result of different amino acid sequences. A meaningful definition in this case thus considers each amino acid sequence as a different protein variant. In the case of sequences of length L, where each among the L positions can take on an alphabet of size q ($q = 20$ for amino acids, $q = 4$ for nucleotides), we thus have $N = q^L$ different variants. Here, we experimentally realize the full sequence space of $L = 4$ positions in the antibody binding site defining 20^4 antibody variants, each of which displays different binding pockets with different binding affinities to a certain target.

A population is fully determined by the vector of numbers of copies n_i of each variant i, $i \in \{1, 2, \ldots, N\}$. Alternatively, if population size is irrelevant, the population is also fully determined by frequencies $f_i = \frac{n_i}{\sum_{j=1}^{N} n_j} \in [0, 1]$, *i.e.* the fraction of the population that consists of species i. These frequencies satisfy by definition the normalization $\sum_{i=1}^{N} f_i = 1$. A complete stochastic description of finite-size populations in terms of the number of copies n_i per variant i is generally achieved through the language of multi-type branching processes (see chapter 4). Upon assuming the continuum limit of infinite population size, $N \to \infty$, and continuous time, the frequencies $f_i \in [0, 1]$ become continuous variables and the selection (and mutation) dynamics simplifies to deterministic differential equations for the f_i, that we will restrict to for our discussion here and modeling later.

Each variant i is associated with a fitness s_i, which describes its performance with regard to a certain selective pressure, either in absolute terms or relative to other variants. It may represent (equivalently) the probability of variant i to survive selection or, equivalently, the probability or rate at which variant i produces offspring. The quantity s_i (encoding fitness) will be of central interest throughout the rest of the manuscript, where we choose s_i to be the survival probability under selection for binding affinity (sequences with higher binding affinity will have higher fitness, see sections 2.1 and 2.2) and seek to study the distribution of s_is within the sequence space of an antibody binding site (see chapters 3, 4).

With only selection, this dynamics is given by

$$\dot{f}_i = s_i f_i - \phi f_i, \tag{1.1}$$

where the first term accounts for the offspring of variant i due to its fitness s_i, but the second

term accounts for the competition of variant i with all other variants in terms of frequencies. Here, ϕ must be chosen to satisfy the constraint $\sum_{i=1}^{N} f_i = 1$, which holds by definition. Summing equation (1.1) over i, we obtain

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\left(\sum_{i=1}^{N} f_i\right) = \sum_{i=1}^{N} \dot{f}_i = \sum_{i=1}^{N} s_i f_i - \phi \underbrace{\sum_{i=1}^{N} f_i}_{=1}, \tag{1.2}$$

and thus $\phi = \langle s \rangle_{\mathrm{pop}} = \sum_{i=1}^{N} f_i s_i$, which represents the mean fitness of the population ($\langle \cdot \rangle_{\mathrm{pop}}$ denotes population mean). The dynamics thus becomes

$$\dot{f}_i = f_i \left(s_i - \langle s \rangle_{\mathrm{pop}}\right). \tag{1.3}$$

This dynamics is non-linear in the frequencies f_i through the population mean $\langle s \rangle_{\mathrm{pop}}$. The stationary solution with $\dot{f}_i \equiv 0$ for all variants i requires for each variant i either $f_i = 0$ or $s_i = \langle s \rangle_{\mathrm{pop}}$. This implies that a single variant m will eventually be present with frequency one, *i.e.* $f_m = 1$ for a certain $i = m$ and $f_i = 0$ for all $i \neq m$. It depends on the initial conditions $f_i(t = 0)$ which variant will eventually invade the population: Among the variants i with non-zero initial frequencies, $f_i(t = 0) > 0$, it is the one which maximizes fitness, *i.e.* $m = \arg\max_{i:\, f_i(t=0)>0} s_i$. A discrete-time equivalent of this selection dynamics will be studied and realized experimentally later (see section 2.4 and chapter 3).

The mean fitness of the population ϕ satisfies an interesting relation under this dynamics: Taking its derivative with respect to time t and reinjecting equation (1.1) yields

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \sum_{i=1}^{N} s_i \dot{f}_i = \sum_{i=1}^{N} s_i^2 f_i - s_i \langle s \rangle_{\mathrm{pop}} f_i = \langle s^2 \rangle_{\mathrm{pop}} - \langle s \rangle_{\mathrm{pop}}^2 = \mathrm{var}(s)_{\mathrm{pop}}. \tag{1.4}$$

Thus, the increase in mean fitness of a population due to selection at a given timepoint directly relates to the population variance of fitness at this timepoint. Equation (1.4) is known as Fisher's fundamental theorem of natural selection and conveys our definition of selection potential: Large differences in fitness (*i.e.* large $\mathrm{var}(s)_{\mathrm{pop}}$) imply rapid selection for high-fitness variants and increase in population fitness. At the other extreme, it is intuitive that selection has no effect in a homogeneous population which consists of variants with identical fitness (*i.e.* $\mathrm{var}(s)_{\mathrm{pop}} = 0$); there is no variance in fitness and, according to equation (1.4), thus no increase in population fitness upon selection.

This model can be easily generalized to take into account mutation. Upon mutation, variant i can be mapped into another variant j by erroneous replication in the offspring production step. In this case, equation (1.1) generalizes to

$$\dot{f}_i = \sum_{j=1}^{N} Q_{ij} s_j f_j - \phi f_i, \tag{1.5}$$

where Q_{ij} is a stochastic matrix, *i.e.* $\sum_{i=1}^{N} Q_{ij} = 1$, $\forall j = 1, 2, \ldots, N$, whose entries represent the probabilities to produce variant j upon replication of variant i. Equation (1.1) with pure selection is recovered upon setting $Q_{ij} = \delta_{ij}$ with δ_{ij} being the Kronecker Delta. As before, ϕ encodes for the mean fitness of the population, as we must satisfy

$$0 = \sum_{j=1}^{N} \underbrace{\sum_{i=1}^{N} Q_{ij}}_{=1} s_j f_j - \phi \underbrace{\sum_{i=1}^{N} f_i}_{=1}. \tag{1.6}$$

Alternatively, upon reparametrizing $Q_{ij} = \delta_{ij} + \frac{\mu_{ij}}{s_j}$, where μ_{ij} is a constant mutation rate from variant j to variant i, equation (1.5) reads

$$\dot{f}_i = (s_i - \langle s \rangle_{\text{pop}})\, f_i + \sum_{j=1}^{N} \mu_{ij} f_j. \tag{1.7}$$

This form decomposes the effect of selection and mutation into two additive terms. Equations (1.5) can be turned into linear equations by a non-linear transform from f_i to g_i including an integrating factor that eliminates ϕ,

$$g_i(t) = f_i(t) \exp\left(\int_0^t \phi(\tau) \mathrm{d}\tau \right). \tag{1.8}$$

Taking the time derivative of g_i and injecting equation (1.5) leads to

$$\dot{g}_i = \sum_{j=1}^{N} Q_{ij} s_j g_j \tag{1.9}$$

which is a system of coupled linear differential equations, $\dot{g}_i = \sum_{j=1}^{N} W_{ij} g_j$ with $W_{ij} = Q_{ij} s_j$, and is therefore exactly solvable in terms of the eigenvalues of the matrix $\mathbf{W}$. But more interestingly, the selection-mutation dynamics in equation (1.5) can be reformulated again as a pure selection problem (equation (1.1)): In terms of the matrix $\mathbf{W}$, equation (1.5) reads

$$\dot{f}_i = \sum_{j=1}^{N} W_{ij} f_j - \phi f_i. \tag{1.10}$$

Assume that λ_k and $\mathbf{w}_k$, $k = 1, 2, \ldots, N$ are the eigenvalues and eigenvectors of $\mathbf{W}$, *i.e.* $\mathbf{W}\mathbf{w}_k = \lambda_k \mathbf{w}_k$ and $\sum_{j=1}^{N} W_{ij} w_{jk} = \lambda_k w_{ik}$. Upon expanding f_i in the basis of eigenvectors $\mathbf{w}_k$, $f_i = \sum_{j=1}^{N} h_j w_{ij}$, equation (1.10) becomes

$$\dot{f}_i = \sum_{j=1}^{N} h_j w_{ij} = \underbrace{\sum_{j,k=1}^{N} W_{ik} h_j w_{kj}}_{=\sum_{j=1}^{N} \lambda_j h_j w_{ij}} - \phi \sum_{j=1}^{N} h_j w_{ij}. \tag{1.11}$$

Upon comparing coefficients (which we are allowed to do as the $\mathbf{w}_k$ constitute a basis), we obtain

$$\dot{h_i} = \lambda_i h_i - \phi h_i \tag{1.12}$$

with

$$\phi = \sum_{i,j=1}^{N} W_{ij} f_j = \sum_{i,j,k=1}^{N} W_{ik} w_{kj} h_j = \sum_{i,j=1}^{N} \lambda_j w_{ij} h_j = \sum_{j=1}^{N} \lambda_j h_j. \tag{1.13}$$

This implies in particular that $\sum_{i=1}^{N} h_i \equiv 1$. Equation (1.12) is formally identical to equation (1.1), where f_i is replaced by h_i and s_i by λ_i. In the stationary solution of the mutation-selection dynamics, the eigenvector $\mathbf{w}_m$, which maximizes fitness $m = \arg\max_i \lambda_i$, *i.e.* with maximal eigenvalue λ_m, will eventually be present with frequency one, $h_m = 1$. This stationary solution is called the "quasi-species", in which all true species i with different fitness values s_i coexist with (potentially) non-zero frequencies $1 \geq f_i \geq 0$, $\forall i = 1, 2, \ldots, N$.

1.3 Evolvability: the antibody as a model system

In the search for a convenient model system to study evolvability, we decide to turn towards the antibody, a key protein of the adaptive immune response. The antibody can incite an immune response and evolve binding affinity to virtually any foreign target within the process of affinity maturation; what makes it so "evolvable"? We argue that the answer resides in the existence of two orthogonal evolutions at different timescales where one evolution optimizes the other (subsection 1.3.1). Thus, besides its potential for mathematical modeling (see previous section 1.2) and loads of biological (sequence and structural) data, the antibody is one, and to our knowledge the only, molecular realization of the separation of evolutionary timescales mentioned in section 1.1. We will transfer the notion of selection potential to the antibody "scaffold", a part of the antibody inherited across generations and encoding for the prospensity of the antibody to affinity-maturate (subsection 1.3.2). For reference purposes only, we also discuss the structure of the antibody and introduce notions that will appear later in the manuscript, such as framework (FWR) and complementarity-determining (CDR) regions (subsection 1.3.3). Equally for reference purposes, biological details of antibody affinity maturation are provided as appendix in chapter B, but are not required for the understanding of the manuscript.

1.3.1 Antibodies: long-term *versus* time-lapse Darwinian evolution

Structurally, the variable (V) antibody region which is responsible for binding is subdivided into (a heavy chain (V_H) and a light chain (V_L), and each of these into) 4 framework (FWR) and 3 complementarity-determining (CDR) regions which alternate along the sequence. The FWRs

define the core of the binding region and the CDRs define easily accessible surface loops on the antibody displayed by the FWRs and most likely in contact with the binding target (see last subsection 1.3.3 for more details). A major part comprising all FWRs, CDR1, 2, as well as the middle part of CDR3 of the organism's "default" (or "naïve") antibody repertoire is encoded for in the genome/germline and is partitioned into gene fragments. The remaining part of the CDR3 sequence is not germline-encoded and randomly determined by adding random junction sequences upon recombination of the gene fragments. On the one hand, the antibody evolves over millions of years by handing the genome/germline-encoded part of the naïve antibodies down to following generations. On the other hand, the antibody also evolves in the so-called affinity maturation process [18] within a generation which is initiated as part of the adaptive immune response upon pathogen encounter; affinity maturation modifies the naïve antibody by randomly introducing somatic mutations and selecting for strong binding to epitopes of the pathogen (see last section for details). The process typically concludes with "mature" antibodies that specifically recognize and bind to a pathogen within the course of only weeks, months [19], maybe years [20]. Most of the (fixed) somatic mutations occur in surface loops and in particular CDR3, where diversity is generally found to be most useful [21] and most essential for antibody specificities [22]. Importantly, mature antibodies are not inherited; only the germline-encoded part of the naïve antibodies is. Instead of subdividing the antibody into FWRs and CDRs based on structural considerations, we propose to subdivide it into a "scaffold" part (which is inherited, everything but CDR3) and a non-scaffold part (most of CDR3) based on evolutionary arguments. The antibody is thus subject to two orthogonal evolutions, one acting on the scaffold on long timescales between generations, as well as one acting on the entire antibody, but mostly on the non-scaffold part, on short timescales within an individual as part of its adaptive immune response. Evolvability of the scaffold part is presumably critical for the success of affinity maturation and a property imparted by the long-term evolution (as individuals with more evolvable naïve antibodies allow for more efficient immune responses, therefore have higher fitness, and are positively selected).

1.3.2 Selection potential of antibody repertoires or libraries

We propose that the antibody is a convenient model system for the study of evolvability due to the separation of timescales in its overall evolution. Moreover, biomolecules in general are convenient model objects for the study of aspects of evolution as they are amenable to experimental, controlled evolution and measurement and traceability *e.g.* via high-throughput sequencing (see section 1.2).

To probe the evolvability of antibodies, we propose to study the response to selection for binding affinity *in vitro* of antibody libraries built around fixed scaffolds and harboring identical non-scaffold sequence diversities (with "scaffold" and "non-scaffold" as defined in the previous subsection 1.3.2). Such an experimental proceeding is mimicking the initialization of the *in vivo* immune response, *i.e.* initial repertoire formation by sampling from the pool of available scaffold segments and introducing random junction sequences between these scaffold segments, followed

by selection for binding affinity to a pathogen. Such selection trajectories, in combination with high-throughput sequencing of the libraries, should provide information about the component of evolvability that is linked to the existence and selectivity of (relatively) high-affinity sequences in the library/initial repertoire and which we refer to as the "selection potential" of a library/repertoire and the underlying scaffold: Libraries/repertoires thus have high selection potential if they feature few highly-performant sequences over otherwise poorly-performant sequences within their sequence diversity. This assures that few promising starting points for the ongoing solution can be efficiently selected for. On the contrary, libraries/repertoires with equally poorly or moderately-performant sequences have low selection potential as selection would not be able to identify relevant starting sequences. Selection potential thus encodes for how genotypic diversity translates (or not) into phenotypic diversity. In general, while evolvability is oftentimes defined with regard to a successful (or not) end-product of an evolutionary trajectory, selection potential encodes for the susceptibility to the onset of a new selective pressure in the initial stages of an evolutionary trajectory.

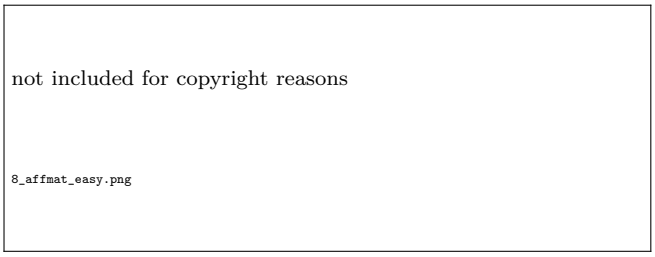
Furthermore, we propose to construct and select several such antibody libraries, either separately or in mixture, built around various scaffolds and holding identical sequence diversity in their non-scaffold regions. The choice of scaffolds can be possibly based on differences in their maturation degrees, *i.e.* differences in evolutionary history from previous affinity maturation to an unrelated target. Such an experimental scenario is again mimicking the *in vivo* original as both naïve B cells and memory B cells can in principle serve as input for a new maturation trajectory (see chapter B). Here, the question about selection potentials and evolvability may be as follows: The use of which scaffolds gives rise to more promising and selectable libraries/repertoires than others? Introducing identical genotypic diversity in the non-scaffold part (CDR3), which is also a crucial part of the antibody's binding pocket, allows to compare selection potentials between several libraries and to gauge the impact of the scaffold part; how much phenotypic diversity is introduced for a given binding pocket sequence diversity displayed in the context of any one among several given scaffold sequences?

The experimental basis for our observation of selection potentials are quantitative selection experiments based on phage display of antibody libraries and high-throughput sequencing.

1.3.3 The structure and role of antibodies in the adaptive immune system

The adaptive immune system in jawed vertebrates (*gnathostomata*) is responsible for the effective combating and elimination of foreign pathogens, in coordination with the organism's innate immune system. The main conceptual difference of innate and adaptive immunity is that the former applies a default (*non-specific*) clearance protocol upon infection with an arbitrary pathogen, whereas the latter is capable of engineering a *specific* response to almost any non-self molecular

(a)



(b)

Fig. 1.2: The antibody and its rapid evolution through affinity maturation. **(a)** Antibody structure which takes a "Y"-like shape, here rotated by π. Taken and adapted from [24]. **(b)** Simplified schematic of affinity maturation. The naïve antibody, or naïve B cell receptor (BCR), is affinity-matured through cycles of somatic hyper-mutation and selection for antigen binding, before being secreted into the blood. Taken and adapted from [25].

objects. This task is non-trivial, as the number of potential such non-self, pathogenic objects exceeds by far the number of specific remedies that an organism can simultaneously hold at any one time. The innate immune system itself is evolutionarily older and present in an even larger group of species; the additional adaptive system may be explained by its optimality for low-probability pathogens (*versus* optimality of the innate system for ubiquitous pathogens) [23]. In the chronology of an infection, the adaptive immune response is initiated after onset of and mediation by the innate response, as explained in chapter B.

One of the key actors of the adaptive immune system is the antibody [26], which will be the object of interest throughout the whole manuscript. The antibody is a "Y"-shaped protein, see figure 1.2(a), that is responsible for specific binding to epitopes on the encountered pathogen (under avoidance of binding to any self epitopes in the organism) in order to neutralize it, *i.e.* making it non-infective by blocking or sterically excluding interactions with the host cell surface, and/or to trigger downstream processes for pathogen clearance again involving other actors from the innate system, such as macrophages. Within the protein, these two functions are separately organized into the Fab regions which constitute the 2 upper arms of the "Y" (2 copies of the Fab region, one per arm; binding) and the Fc region, which constitutes the stem (clearance). Each half of the "Y" (with respect to the vertical symmetry axis) is composed of a heavy (H) chain that makes up both the upper and lower part of the half-"Y", and a light (L) chain that is fused via disulphide bonds (between cysteins) and non-covalent interactions to the upper part of the H chain. All H and L chains are again subdivided into variable (V_H, V_L) and constant (C_H, C_L) regions, with their names indicating that these regions are respectively prone or not to changes in amino acid sequence upon optimization of the antibody for binding capacity, see chapter B. The V_H and V_L regions are each one approximately 100 aa in length and further subdivided alternately into

4 framework (FWR) and 3 complementarity-determining (CDR) regions, numbered from FWR1 through FWR4 and CDR1 through CDR3, respectively. Structurally, the CDRs form loops in the antibody fold that are exposed at the two upper tips of the "Y". These regions are most likely to interact directly with epitopes, as well as most tolerant towards mutation and modification to accommodate for complementarity with epitopes (hence their name). The six CDRs together can be regarded as forming a "binding site" of the antibody, although the CDRs are not necessarily all equally important or even relevant. Among the CDRs, the CDR3 is particular in a way described below, as is the V_H compared to V_L. The FWRs of the V_H and V_L chains can be regarded as scaffolding these binding sites, but they are sometimes themselves directly involved in the interaction with an epitope.

In vivo, the antibody as described above is displayed as a receptor on the surface of B cells (BCR) or are secreted into the blood in various isotypes, notably IgM and IgG representing pentamers and monomers of the "Y"-shaped protein, respectively. Each B cell encodes *a priori* for a different sequence at the level of the antibody variable (V) chains. The presence of 2 or more binding sites on a single molecule gives rise to avidity effects, *i.e.* a higher apparent binding affinity as compared to the single-binding site affinity. Several antibody binding sites on the same molecule can bind to different epitopes on a single copy of the pathogen, or to identical epitopes on different copies of the pathogen (if not forbidden by steric exclusion).

Interestingly and importantly, a V_H-V_L fusion in absence of all other components is sufficient to maintain the binding properties of the full antibody, and thus its mere function of binding. This fact is profitably used in all contexts (from clinical and diagnostic to purely academic, such as here), where the search for functional antibody binding sites is of interest, as it allows to accommodate for construction, expression, screen, and analysis of large pools of variable regions, see chapter 3. Candidate variable regions can then be simply grafted back on the remaining antibody components. As the disulphide bonds are between the constant regions C_H and C_L and thus absent in this reduced construct, an artificial covalent, flexible bonding between V_H and V_L is typically achieved by glycine-rich linker sequences. Even more strikingly, a V_H in absence of a paired V_L may still retain binding specificities of the antibody. In fact, it has been shown that the CDR3 of V_H is sufficient for a large number of binding specificities [21]; the CDR3 of V_H is the most variable of all 6 CDR in the antibody, see chapter B.

1.4 Evolvability in protein systems: state of the art

We will review a few theoretical ideas and results about evolvability that illustrate its potential relevance in biology and beyond, most notably that it can be targeted by selection (subsection 1.4.1). In proteins, a number of biophysical and structural properties, notably stability, molecular dynamics, and polarity, have been proposed to correlate with, or encode for evolvability. In the antibody, these properties are oftentimes proposed to accommodate for its ability to affinity-maturate and

provide a "solution" to almost any "task" (subsection 1.4.2).

1.4.1 Results from theoretical models of protein evolution

In order to evolve and achieve improvement with respect to a certain goal upon an evolutionary algorithm (variation and selection), the object under selective pressure must have the ability to do so, which is referred to as "evolvability". In the biological context, evolvability is consistently defined as the "propensity to evolve" (or to adapt) [15, 13], "the ability of random mutations to sometimes produce improvement or to not always be deleterious" [13, 16], the "capacity to generate heritable phenotypic variation"[8], or the "rate of selectable phenotypic variation", which embodies a general agreement on the notion of evolvability. The general consensus about evolvability in the context of evolutionary biology, however, ceased already here until recently [15]. Albeit being repeatedly put forward [14, 8, 16], the concept and relevance of evolvability is generally considered elusive, the question being whether evolvability can be considered a property or phenotype in line with *e.g.* binding affinity and catalytic activity of biomolecules. This status stands and falls depending on whether it can be (directly or indirectly) targeted by selection and evolution, *i.e.* is selection for and evolution of evolvability possible and what are the evolutionary/selective forces and physics behind?

As a matter of fact, evolvability is indeed different from other standard phenotypes in the following sense: Rather than being an actual property of any given *status quo*, it can be regarded as a variational and anticipatory property; it encodes for future, rather than present benefit of a genotype [8, 16]. This observation is best pictured by an analogy with virtual displacements in classical (analytical) mechanics [27]. Based hereon, the idea that evolvability could be selected for is generally challenged by an argument of causality [13], stating that selection would be required to act not on the fitness of the current genotype, but rather on the structure of the fitness landscape in its neighborhood, which is however not physically realized (yet). In a related note, evolvability is more appropriately rephrased as variability rather than variation, as variation refers to the physically realized end-product of an evolutionary process and variability to the potential to yield such variation. In addition, one may naïvely observe that all currently existing living matter is the product of evolution and is still subject to evolution, suggesting that the capacity to evolve may be trivially and inherently present in nature, and casting doubt on the relevance of the notion of evolvability. In line with this idea is the fact that directed evolution [28] is generally successful on evolving many biomolecular templates or biological systems towards many target properties. Finally, slow protein evolution may be (partly) explained by other factors that do not require the notion of evolvability [29].

In spite of these seemingly conceptual issues, a number of mathematical models demonstrated that evolvability is a selectable trait in time-dependent environments, albeit never being directly selected for and that it can emerge as a by-product of evolutionary dynamics [30, 9, 10, 31, 32, 11,

33, 12, 13]. In [13], for instance, numerical simulations of protein evolution in randomly changing rugged landscapes with exponentially many local optima, where selection does not explicitly act on evolvability, are studied and reveal that rapid environmental change puts selective pressure on evolvability. In [10], the evolution of spin-glass models of proteins in fluctuating environments is studied and the connection between evolutionary history in fluctuating environments on the one hand and modularity (disjoint organization of protein functions in a protein or sectors) and evolvability (propensity to new evolutionary tasks) on the other hand is established. Here, selection occurs for allostery, yet emergence of evolvability is also observed: The effect of evolution in changing environments restricts to sparse interactions between residues/spins, which, as a mere consequence of locality in high-dimensional spaces, are close to solutions to other selective pressures. Similar findings for RNA structures and logic circuits rather than proteins are reported in [32]. The problem of inducing evolvability can be restricted to the problem of inducing generalists [30, 34, 6], *i.e.* genotypes that are fit across a finite universe of environments as opposed to specialists that are very fit in one or a few, but unviable in the majority of environments. This probably differs from evolvability by the set of environments considered: a finite number of environments in the problem of generalists *versus* an infinite and undersampled set of environments in the problem of evolvability.

Although these studies demonstrate selection for evolvability or generalists upon evolution in fluctuating environments, such an outcome seems notoriously difficult to target: It was highlighted that selection for evolvability/generalists is restricted to intermediate phases featuring intermediate alternation frequencies between environments, intermediate ruggedness, and intermediate correlation between environments [34] and must be guided towards hidden interfaces between neutral networks [33]. This intermediate phase is sandwiched between phases of specialization [5], canalization [33] at low frequencies and frustrated selective pressures [5] and flexibility [33] at high frequencies ("high" and "low" frequencies as compared to mutation rates). One practical example of this difficulty to generate generalists are so-called "broadly-neutralizing antibodies" [5, 6] (see section B.3), which are capable of neutralizing various strains of highly diverse pathogens, such as HIV or influenza, but rare. In another study, the beneficial effect of tuned, non-constant frequencies over time on the induction of such generalists is highlighted [30], which may also be a starting point for the control of evolvability: Starting from low frequencies and increasing frequency with time can focus the evolution and increase the likelihood of take-over by generalists.

It was suggested that phenotype evolvability should be encoded within the genotype-phenotype (genotype-fitness) map [16]. This mapping associates each genotype (*e.g.* DNA sequence) to the phenotype(s) (*e.g.* of the expressed protein) that it is encoding, geno $\mapsto$ pheno. Evolvability relates to how genotypic variation (mutation) translates into phenotypic variation in a constant environment, which can be expressed in a "cartoonish" way as

$$\left.\frac{\partial \text{pheno}}{\partial \text{geno}}\right|_{\text{environment}} . \tag{1.14}$$

This is alternatively termed "variability" [35]). If the distribution of changes in phenotype Δpheno

within a set of possible changes in genotype Δgeno is large, selection can efficiently enrich for high phenotype values over low phenotypes and, thus, makes evolution effective. Increasing evolvability amounts to increasing the volume in (potentially multidimensional) phenotype space accessible, *i.e.* maximizing $\Delta\text{pheno} = \frac{\partial\text{pheno}}{\partial\text{geno}}\Big|_{\text{environment}} \cdot \Delta\text{geno}$ in a Taylor expansion-like notation, which can be achieved either by a "steeper" landscape, *i.e.* larger $\frac{\partial\text{pheno}}{\partial\text{geno}}\Big|_{\text{environment}}$, or by increasing the volume of accessible genotype space Δgeno itself.

Thus, it appears that the above definition allows for two different mechanisms to achieve evolvability; (i) through mutation rate with defines the volume of genotype space that can be probed within fixed time, and increasing the genotypic space accessible upon mutation potentially increases accessible phenotypic space volume; (ii) through the mutational effect of a fixed set of mutations which defines the volume of phenotypic space covered by a fixed genotypic space volume. Changing the context in which in given mutation occurs in a way that increases the phenotypic effect of this mutation also increases evolvability. It is important to stress the difference between these two factors behind evolvability; (i) defines genotypic mutation rate, whereas (i) and (ii) together define phenotypic mutation rate [33]. There is a myriad of mutational processes realized in nature that contribute to mechanism (i): These range from local moves in genotype space by point mutations that may occur at rates determined by error-prone DNA polymerase or codon usage optimized for non-synonymous mutations [13], to large-scale moves in sequence by DNA exchange (recombination) [13]. To integrate these mutational processes and their rates into a single quantity representing evolvability as to mechanism (i), Earl *et al.* proposed a diffusion constant in protein sequence space [13]. In this project, we attempt to address mechanism (ii) in the context of antibodies: What is the (width of the) histogram of phenotype values obtained by fixed sequence diversity in the non-scaffold part of the antibody in the context of different scaffolds?

Evolvability oftentimes goes hand in hand with modularity, pleiotropy, and autonomy (context-independence) [8, 16]. Modularity, which is the opposite of pleiotropy, and autonomy all reflect the idea that an object is internally organized and assembled from independent "building-blocks" and is considered a key-ingredient for evolvability: In a developmental context, modularity of organismic design is associated with few interference in adaptation for different functions [16] and the possibility for unconstrained changes of cell biological and developmental processes as a consequence of weak linkage, compartmentalization, redundancy conferring robustness, flexibility, and evolvability [8]. At the molecular scale, the observation of sectors in proteins [36, 10, 37], *i.e.* disjoint subunits encoding for different protein functions, predicts independence of mutations that affect different functions and, thus, evolvability and the possibility of improvement for any of these functions without compromizing the others. The efficiency of evolution in modular systems results from additive contributions of different functions (phenotypes) to total fitness and thus smooth, convex landscape with only a global maximum, which can be achieved by simply following the gradient. The opposite extreme are landscapes from fully connected subunits that are rugged and in which gradients are meaningless and (local) optima are several mutations away.

Within biology, evolvability is argued to be acute across the scales, from the molecular (biomolecules [32, 13]), over the cellular (gene-regulatory networks [31]), to the developmental and organismic scale [8, 16]. Its existence and relevance is also mirrored in other contexts, in which evolutionary algorithms and processes are successfully applied for optimization, such as material sciences (origami [38, 39]) and computer science [16]. As an example, the standard task of supervised learning consists in building a neural network capable of classifying given data (*e.g.* pictures of cats and dogs), which can be in principle achieved by evolutionary algorithms upon taking the error on a train dataset as (negative) fitness. In this computational context, evolvability echoes as the so-called representation problem: how to choose the degrees of freedom accessible to "genetic variation" in order to realize all from very well to very poorly performing networks. The effectiveness of evolutionary algorithms is indeed not guaranteed here. Consider as an example the task of improving a computer program by random changes in the code: If the mutations are chosen to be "point mutations" at the level of single strings, the change will almost always be detrimental. But recombining branches of parse trees can lead to improved performance [16]. The difference to biology is, however, that the choice of the evolutionary degrees of freedom is preset by nature in the form of DNA and its sequence.

Despite the number of theoretical insights and the apparent importance in biology, a direct observation of evolvability on an experimental system has not yet been reported. The observation and measurement of (differences in) selection potentials are the main goal of this project and the key result of this thesis.

1.4.2 Protein evolvability hand in hand with other properties

In the literature, several biophysical and structural properties of proteins have been shown or proposed to correlate with their evolvability: protein dynamics and conformational flexibility, as well as stability and polarity/modularity.

In general, protein dynamics contributes to protein function, such as catalytic activity [40] and allostery. Does it also contribute to evolvability? Here again, the antibody is presumably the model system that should allow to address this question: Antibody dynamics has been extensively studied in literature, especially comparatively between naïve and mature antibodies. A general finding is that antibody dynamics and binding mechanisms are modified upon maturation: In many cases, involving various rather invariant binding targets, a decrease in antibody flexibility, *i.e.* increase in rigidity, is observed. This finding conveys a picture of induced-fit binding mechanisms in naïve antibodies endowed with conformational isomerism *versus* lock-and-key binding mechanisms in mature antibodies which are specialized to specific recognition of their cognate targets [4]. More generally, the role of rigidity in molecular recognition has been outlined [41]. In fact, this conformational degeneracy may be a factor of evolvability as it allows for broad recognition spectra realized by few naïve scaffolds in the immune response [4].

There is a number of examples of specialization by rigidification in initially flexible antibodies upon maturation observed by various mathods [42, 3, 4, 43, 44, 45, 46]: (i) A good example of increased affinity, but lost cross-reactivity as a result of decreased antibody flexibility upon maturation is the antibody 7G12: As revealed by X-ray crystal structures [43], the naïve version of 7G12 can use its structural isomerism to nucleate around different target structures required for binding to epitopes on the unrelated molecules hapten and jeffamine. The mature 7G12 is the result of affinity maturation against hapten, which binds to hapten with increased affinity, but no longer binds to jeffamine, which implies higher binding specificity. X-ray structures show stabilization of the binding site structure towards complementarity with hapten, while excluding complementarity with jeffamine. (ii) The antibody 48G7, cognate to hapten and featuring 9 somatic mutations off the binding interface and up to 15 Å away from the binding interface, has $3 \cdot 10^4$ x higher affinity to hapten than its naïve ancestor. Comparing X-ray crystal structures of 48G7 and its naïve ancestor in complex with hapten shows that the presence of the somatic mutations stabilized the antibody to the target configuration that binds hapten [46]. Molecular dynamics simulations and computation of absolute free energies of the same antibody in complex with hapten in explicit solvent (water) draws identical conclusions by showing that structural fluctuations, as measured by root mean squared displacements around an average structure, are more restricted in the mature antibody, see figure 1.3(a) [45]. In addition, they reveal that maturation mediates improved binding through rearrangement of electrical charges and polar/electrostatic interactions in the antibody binding site, while leaving favorable nonpolar/hydrophobic and van-der-Waals interactions unaltered [45]. (iii) The antibody 4-4-20, cognate to fluorescein, as well as some intermediates on its maturation trajectory and somatic mutational reversals have been studied by various approaches and with similar conclusions, see figure 1.3(b): by three-pulse photon echo peak shift spectroscopy [4] (1 or 2 reverse mutations in V_L) which quantifies flexibility in terms of amplitudes and frequencies of motion in response to the onset of a constant forcing in the antibody binding site, by non-linear laser spectroscopy [3] (10 intermediates in V_H, while keeping V_L mutations fixed), and by molecular dynamics simulations [42, 3] (both V_L reversals and V_H intermediates). A possible mechanism behind rigidification upon maturation is the stabilization of the naïve paratope around its conformation in bound state, which increases the association rate k_+ by decreasing the entropic burden of finding the "correct" conformation [44]. However, this represents a kinetic rather than thermodynamic selection force. However, decrease in dissociation rate k_- simultaneously with increases in association rates have been observed on other systems [47].

The picture of an antibody converging in structure and affinity towards a given target does not hold indefinitely: (i) The prevalence and relevance of polyspecific antibodies in the repertoire (beyond the naïve ones) capable of recognizing several potentially unrelated antigens has been emphasized [48]. Here again, flexibility is proposed to be the main mechanism of antibody polyreactivity, but their function is less clear and speculative so far; they are not necessarily involved in immune response against one or several pathogens, but may be important for the control of autoimmunity and self tolerance. Usually, these polyspecific antibodies are of IgM isotype, have lower affinities, and are closer to naïve antibodies in terms of sequence identity, than fully mature antibodies, meaning that they have lower maturation level and less somatic mu-

Fig. 1.3: Flexibility, rigidity, polarity. **(a)** Root mean-square (RMS) fluctuations in mobile atoms of a hapten-cognate mature antibody and its naïve ancestor antibody in complex with hapten as measured from molecular dynamics simulations [45]. **(b)** RMS fluctuations depicted on the structures of the mature anti-fluorescein 4-4-20 antibody (right) and its naïve ancestor (left) in complex with fluorescein (yellow) [3]. **(c)** Different scenarios for the evolution of antibody affinity (to conserved pathogenic epitopes) and rigidity upon maturation to highly variable pathogens, depending on initial affinity values [54]. **(d)** Model of bnAbs stating rigidification upon maturation to prevent steric exclusion with highly variable pathogen surface structure [51]. **(e)** Fraction of active-site residues that are non-scaffolding (measuring polarity between active-site and scaffold) correlated with the number of catalytic activities (measuring innovability) in various enzymes (one point per enzyme) [55]. **(f)** The (normalized) number of contacts between active-site and scaffold residues (measuring modularity between active-site and scaffold) correlated with the number of catalytic activities (measuring innovability) for the same enzymes as in (e) [55].

tations. (ii) Computational, graph-based models of antibodies come to seemingly contradictory conclusions [49, 50]. In such graph-based models, nodes and the presence or absence of edges between pairs of nodes correspond to amino acids and conformationally constraining interactions between them, respectively. Using a distance-constraint model, it was shown that affinity maturation increases rigidity in V_H and flexibility in V_L, mediated through more or less constraining hydrogen-bond networks in the antibody (more bonds implies more spatial constraints and, thus, more restricted motion), as well as induces a more intricate entanglement of V_H and V_L [50]. However, it was pointed out that less flexibility is not necessarily concomitant with less mobility, which refers to translational motions of rigid subparts of the antibody. Another large-scale study on thousands of antibodies based on the degree-of-freedom counting of CDR3 residues using a pebble-game algorithm, concluded that the CDR3 of V_H shows no general trend of more or less flexibility upon maturation [49], but it is unclear if and how this result extends to the scaffold part of the antibody. (iii) The picture of increased rigidity upon maturation is mainly raised by studies that involve constant model targets such as haptens or fluorescein. However, *in vivo* targets such as HIV and influenza are themselves subject to evolution on similar timescales as the maturating antibody and may be very dissimilar in the ensemble and over time. Selection forces may differ between constant and co-evolving targets. But a very similar scenario as for constant targets was raised in the context of bnAbs against HIV, which features highly variable, potentially disordered epitopes [51], see figure 1.3(d): Using hydrogen/deuterium exchange with mass spectrometry to dissect contributions from paratope structure and dynamics, it was found that these bnAbs are associated with increased complementarity to buried, conserved epitopes of HIV and decreased interference with disordered surface structures that cover these conserved parts. To achieve this reduced effect of random steric exclusions, the likely importance of the FWR regions, outside the paratope, and its transition from disordered to stabilized structures upon maturation were described. However, other mechanisms of stabilizing effects of somatic mutations in bnAbs were described, including through modifications in the inter-domain (V_H-V_L) dynamics [52, 53]. A study of models of *in silico* affinity maturation for bnAb elicitation and molecular dynamics simulations of bnAb structures revealed different scenarios for the evolution of antibody flexibility upon affinity maturation [54], see figure 1.3(c): The initial, naïve affinities to conserved epitopes of the pathogen determine the initial effect of framework somatic mutations and the course of affinity maturation with regard to antibody dynamics; the weaker the initial affinity is, the more the initial effect of somatic mutation tends towards increased flexibility; only if the initial affinity is already strong enough, the flexibility would decrease from the beginning. To extend and account for this behaviour in the model of [51], we may speculate that the first somatic mutations in the case of weak initial affinity should act in a way to facilitate exploration of conformational space for better access to hidden, conserved epitopes. However, in later stages of affinity maturation the rigidity ultimately tends to increase irrespectively of the initial behaviour [54].

The concept of evolvability has been subdivided into the two necessary requirements of "robustness" to mutations and "innovability" [55, 56] which encode that the effect of mutations should mostly be non-deleterious and sometimes be beneficial regarding a new target function, respectively. The need for robustness can likely be regarded as a reformulation of the need for an excess

of thermodynamic stability [57, 11] that assures that the fold would not be compromised upon destabilizing, but possibly function-enhancing mutations. However, robustness also encodes for the requirement of not compromising already existing function upon trying to create a new one. As pointed out in [55], it may seem puzzling that both flexibility and stability/robustness [11] are required for evolvability. This apparent contradiction is resolved by the concept of polarity (or modularity) [55, 56], which hypothesizes the spatial separation of features and/or functions within the protein. To maximize evolvability by maximizing both stability and flexibility, both features may be accommodated into disjoint regions inside the protein (that are not necessarily continuous along the sequence), see figure 1.3(e),(f). Besides its actual appearance in theoretical models (see previous section 1.4.1), modularity/polarity has been observed in real-life systems: The the scaffold of the enzyme β-lactamase is highly stabilized and rigidified in some variants compared to others, while the active site has the same conformational diversity across variants [56]. High-polarity variants of β-lactamase feature a broader spectrum of non-zero activities and therefore higher evolvability. Separation of functions and properties into "sectors" has also be observed *e.g.* in trypsin, where catalytic activity, substrate specificity, and possibly stability (or mean lifetime) are encoded in 3 disjoint parts of the protein [37]. In the context of antibodies, polarity may be realized by the separation into FWRs and CDRs that account respectively for flexible binding sites capable of assuming various tasks and a scaffolding that assures stability of the overall fold (although a feedback of CDRs on stability of the scaffold has been reported in particular in the context of bnAbs). Indeed, a classification of somatic mutations in antibodies into binding-related and stability-related (or both) mutations, both leading the path to increased affinity: Somatic mutations in antibody binding sites that increase binding affinity can be destabilizing and followed by compensatory somatic mutations that repair the loss in stability but do not directly contribute to affinity [58]; affinity- and thermodynamic stability-related somatic mutations would have to fixate in presence of each other. Moreover, it was found that destabilizing effects and stability-rescuing mutations in regions far from the binding sites of bnAbs are required to achieve large neutralization breadth [52].

Chapter 2

The physics, information theory, and universality of binding

In this chapter, we are going to set up the theoretical basis for the study of selection, in particular in a context where the selective pressure is defined by binding. We will also bridge the gap between physical parameters (binding free energies) and experimentally observable quantities (enrichments). The theory of systems under selection for binding is rooted at the crossroads of kinetics and statistical mechanics (see section 2.1), as well as information theory (see section 2.3). The benefits of the use of binding are its conceptual completeness, yet relative mathematical straightforwardness (as compared to other phenotypes such as stability, allostery, ...). We will argue that selection enrichments are strongly constrained by universality: The central-limit theorem, extreme-value theory, and order statistics altogether provide predictions on selection coefficients (see section 2.2), regardless of microscopic details of binding mechanisms. Interpretations and implications of the remaining free parameter for the specificity of interactions (see section 2.3) and selection dynamics (see section 2.4) will be discussed.

2.1 Kinetics and statistical physics of selection

In a first place, we are going to expose how physics constrains enrichments, the quantities that determine selection dynamics and can be measured experimentally. This happens at two levels: First, we are going to show that selection probabilities at equilibrium obey Fermi-Dirac statistics involving the free energy of binding ΔG, thus reflecting an analogy with quantum physics. This result will be obtained by expressing the fraction of ligands engaged in binding at equilibrium using the solution to the kinetic equations of the binding reaction. Second, we will be in need for a prediction on ΔG itself, as the actual evolutionary degrees of freedom are the sequence x of a

ligand. A prediction of the mapping from sequence x to phenotype ΔG has to be made in the context of next-to absent knowledge about the underlying binding mechanisms and interactions of random ligands. But, as in many other contexts in- and outside biology, the solution consists in the use of a certain class of random models.

2.1.1 Kinetics of the binding reaction

Our focus here is the description of the interactions between two complementary classes of objects that may associate and dissociate. We will call these respectively the ligands and the targets. Let us for simplicity consider the case of a single type of ligands A and a single type of targets T in a first place. The kinetics and the equilibrium state of the binding reaction $A + T \rightleftharpoons AT$ will be discussed here.

The kinetic equations of the binding reaction between A and T are given by

$$\frac{\mathrm{d}}{\mathrm{d}t}[AT] = k_+[A][T] - k_-[AT], \tag{2.1}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}[A] = \frac{\mathrm{d}}{\mathrm{d}t}[T] = k_-[AT] - k_+[A][T], \tag{2.2}$$

where $[A]$, $[T]$ and $[AT]$ denote respectively the concentrations of the ligand, the target and the complex consisting of a single copy of both the ligand and the target. Moreover, k_+ (k_-) denotes the kinetic rate of association (dissociation) of the ligand-target complex. Kinetic equations assume the limit of large numbers of copies of all involved reactants and products, as well as a spatially uniform distribution of all species (well-mixed soup), so that the binding process becomes essentially deterministic. The (products of) concentrations measure the probability of encounter between one copy of each reactant at a given point in space, while k_+ (k_-) measures the probability of association (dissociation) given the event of such an encounter. The product of both then measures the probability of association (dissociation) of a complex at a given point in space. The equations (2.1) and (2.2) are not independent as a consequence of the conservation of matter that imposes $[T](t) + [AT](t) = [T]_{\mathrm{tot}}$ and $[A](t) + [AT](t) = [A]_{\mathrm{tot}}$, *i.e.* the total concentration of ligands and targets (bound and unbound combined) is each one constant through time. Here, $[A]_{\mathrm{tot}} = [A](t = 0)$ and $[T]_{\mathrm{tot}} = [T](t = 0)$ denote the initial condentrations of respectively the ligands and the targets. This reduces the system to a single non-linear differential equation for, say, $[AT]$,

$$\frac{\mathrm{d}}{\mathrm{d}t}[AT] = k_+ \left([AT] - [A]\right)\left([AT] - [T]\right) - k_-[AT]. \tag{2.3}$$

The solution to this equation for generic initial conditions is derived in appendix C.1 and reads

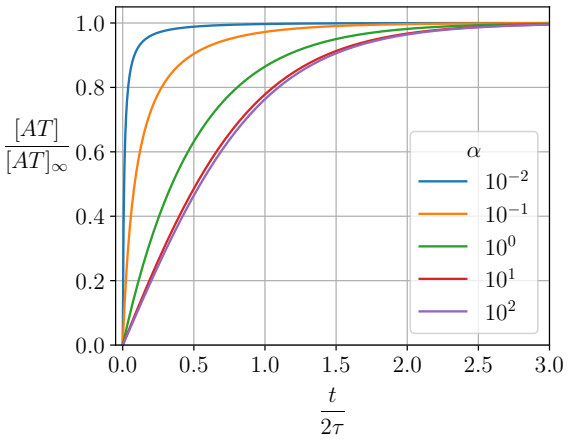


Fig. 2.1: Solution of the kinetic equations for the binding reaction $A+T \rightleftharpoons AT$: the concentration of ligand-target complexes $[AT]$ as a function of time t given by equation (2.5).

in the special case of $[AT](t=0)=0$

$$[AT](t) = \frac{2\gamma_0 \tanh\left(\frac{k_+ t\sqrt{-\Delta}}{2}\right)}{\sqrt{-\Delta} + \gamma_1 \tanh\left(\frac{k_+ t\sqrt{-\Delta}}{2}\right)}, \tag{2.4}$$

where $\Delta = 4\gamma_0 - \gamma_1^2$, $\gamma_0 = [A]_{\text{tot}}[T]_{\text{tot}}$, $\gamma_1 = K_{AT} + [A]_{\text{tot}} + [T]_{\text{tot}}$, and K_{AT} is the equilibrium constant discussed below. This solution can be reparametrized using the equilibrium complex concentration $[AT]_\infty = \frac{\gamma_1 - \sqrt{-\Delta}}{2}$, the time scale parameter $\tau = \frac{1}{k_+\sqrt{-\Delta}}$, and $\alpha = \frac{\gamma_1}{\sqrt{-\Delta}}$ as

$$\frac{[AT](t)}{[AT]_\infty} = \frac{(\alpha+1)\tanh\left(\frac{t}{2\tau}\right)}{\alpha + \tanh\left(\frac{t}{2\tau}\right)} \xrightarrow{t\to\infty} 1. \tag{2.5}$$

The equilibrium concentration of complex $[AT]_\infty$ corresponds to (one of the two) time-independent solutions to the quadratic equation $\frac{\mathrm{d}}{\mathrm{d}t}[AT] = 0$. A plot of the solution is shown is figure 2.1. After equilibration, the time derivatives vanish and the first equation gives

$$K_{AT} = \frac{k_-}{k_+} = \frac{[AT]_\infty}{[A]_\infty [T]_\infty}, \tag{2.6}$$

where K_{AT} is called the equilibrium constant or (binding) affinity. It is important to note that k_+ and k_-, and thus K_{AT} are a property of the ligand-target combination in play only, but independent of the experimental conditions (initial ligand and target concentrations, etc.); if one of the quantities on the right-hand site of equation (2.6) is perturbed by the experimenter, the remaining would adjust such that the value of K_{AT} remains unchanged.

The goal is to link the equilibrium constant K_{AT} (and in subsection 2.1.2 the binding free energy ΔG_{AT}) to quantities that are easily measurable through deep sequencing of selected populations. We define the enrichment s_{AT} as the binding probability of a ligand A to a target

T at equilibrium. This is equivalent to the probability to pass a round of selection if, during selection, only the bound ligands after sufficient incubation time are kept as "survivors". This probability will depend on the binding energy ΔG_{AT} and revealing the mapping $K_{AT} \mapsto s_{AT}$ (and $\Delta G_{AT} \mapsto s_{AT}$) is our objective. Assuming large numbers of particles (thermodynamic limit) and that sufficient time has passed for the binding reaction to equilibrate, s_{AT} can be equated to the fraction among all copies that is in bound state at equilibrium,

$$s_{AT} = \frac{[AT]_\infty}{[A]_{\text{tot}}} = \frac{[AT]_\infty}{[AT]_\infty + [A]_\infty} = \frac{1}{1 + \frac{[A]_\infty}{[AT]_\infty}} = \frac{1}{1 + \frac{K_{AT}}{[T]_\infty}}. \tag{2.7}$$

using equation (2.6) for the last equality. Indeed, this probability s_{AT} increases as the affinity increases (*i.e.* K_{AT} decreases), and as the target concentration is increased. To express s_{AT} in terms of thermodynamic potentials only, it remains to express $[AT]_\infty$ (or equivalently $[T]_\infty$) in terms of K_{AT}, as well as the initial parameters $[A]_{\text{tot}}$ and $[T]_{\text{tot}}$ which are controlled by the experimenter. In the generic case, this is not possible analytically, as $[AT]_\infty$ is the solution to a complicated systems of non-linear equations. Here, however, $[AT]_\infty$ can be computed explicitly,

$$\begin{aligned}
[AT]_\infty &= \frac{2\gamma_0}{\sqrt{-\Delta} + \gamma_1} = \frac{\gamma_1 - \sqrt{-\Delta}}{2} \\
&= \frac{1}{2}\left[[T]_{\text{tot}} + [A]_{\text{tot}} + K_{AT} - \left(([T]_{\text{tot}} - [A]_{\text{tot}} + K_{AT})^2 + 4K_{AT}[A]_{\text{tot}}\right)^{\frac{1}{2}}\right] \\
&\simeq \frac{[A]_{\text{tot}}}{1 + \frac{K_{AT}}{[T]_{\text{tot}}}}
\end{aligned} \tag{2.8}$$

where the approximation corresponds to first order in $\epsilon = [A]_{\text{tot}}/[T]_{\text{tot}} \ll 1$ (see appendix C.1). As expected, this result depends on k_+ and k_- only via their ratio $K_{AT} = \frac{k_-}{k_+}$. Equation (2.8) implies

$$\begin{aligned}
[T]_\infty &= [T]_{\text{tot}} - [AT]_\infty \\
&= \frac{1}{2}\left[[T]_{\text{tot}} - [A]_{\text{tot}} - K_{AT} + \left(([T]_{\text{tot}} - [A]_{\text{tot}} + K_{AT})^2 + 4K_{AT}[A]_{\text{tot}}\right)^{\frac{1}{2}}\right].
\end{aligned} \tag{2.9}$$

Equation (2.7) together with either equation (2.8) or (2.9) represents the solution to our problem of expressing s_{AT} as a function of equilibrium quantities and initial conditions. In experimental practice, an asymmetry between ligands and targets allows for a simplification of this result: When the number of targets exceeds that of the ligands, only few targets will be engaged in binding and the equilibrium concentration of free targets $[T]_\infty$ will be approximately the total concentration $[T]_{\text{tot}}$. In this case, inserting the first order expansion in $\epsilon = [A]_{\text{tot}}/[T]_{\text{tot}}$ of equation (2.8) instead of the exact expression into equation (2.7) yields

$$s_{AT} = \frac{1}{1 + \frac{K_{AT}}{[T]_\infty}} \simeq \frac{1}{1 + \frac{K_{AT}}{[T]_{\text{tot}}}}. \tag{2.10}$$

This result can also be directly obtained in the absence of an exact result for $[AT]_\infty$ by simply replacing $[T]_\infty \simeq [T]_{\text{tot}}$ in equation (2.7) and we are done. In the next subsection, we will obtain a

similar expression to equation (2.7) from a physics approach, involving the free energy of binding ΔG_{AT}.

2.1.2 Equilibrium binding obeys Fermi-Dirac statistics

From a physical viewpoint, the quantity of interest is not the equilibrium constant K_{AT}, but the free energy of binding ΔG_{AT}. It represents the difference in Gibbs free energy of the system between the bound and unbound state of the binding reaction. It systematically appears in studies where binding between two sets is a key phenotype, such as antibodies [59, 60, 61, 62] and transcription factors (DNA binding proteins reponsible for the regulation of DNA expression and with binding specificities to precise DNA sequences) [63, 64, 65, 66]. ΔG_{AT} is an equilibrium quantity, meaning that it entirely determines the state of the binding system at equilibrium, and is directly linked via

$$K_{AT} = \exp(\beta \Delta G_{AT}) \tag{2.11}$$

to the affinity. Here $\beta = (k_B T)^{-1}$ denotes the inverse temperature. The use of ΔG_{AT} requires the assumption that binding is let to happen for sufficiently long time for the binding reaction to have reached equilibrium. The equilibration time that defines "sufficiently long", however, is not fully determined by ΔG_{AT} or K_{AT}, but also depends through both k_+ and k_- on the "height" of the energetic barrier separating the bound and unbound states ΔG^*_{AT} (activation energy). This dependence is captured by the Arrhenius law, $k_+ = \omega \exp(-\beta \Delta G^*_{AT})$ and $k_- = \omega \exp(\beta(\Delta G_{AT} - \Delta G^*_{AT}))$. ω defines the fundamental time-scale of the system, which is found by Kramers' turnover problem which considers a random walk along the reaction coordinate under the effect of the energetic landscape between the bound and unbound state, and of thermal agitation and dissipation from and to orthogonal degrees of freedom (solvent, etc.) [67]. The goal of this subsection is to express the binding probability s_{AT} in terms of ΔG_{AT}.

To make the link between the kinetics and the statistical physics of the binding system governed by ΔG, consider a simplified system consisting of a single copy of the (monovalent) ligand bathed into a sea of targets and thus in contact with a thermal and chemical reservoir at (inverse) temperature β and chemical potential μ. Call $n \in \{0, 1\}$ the binary occupation number indicating whether a target is unbound ($n = 0$) or bound ($n = 1$) to the ligand. The grand-canonical partition function is $\Xi(\beta, \mu) = \sum_{n \in \{0,1\}} \mathrm{e}^{-n\beta(\Delta G - \mu)}$, and, hence, the binding probability reads

$$s_{AT} = P(n = 1) = \frac{\mathrm{e}^{-\beta(\Delta G_{AT} - \mu)}}{\Xi(\beta, \mu)} = \frac{1}{\mathrm{e}^{\beta(\Delta G_{AT} - \mu)} + 1}, \tag{2.12}$$

which is the Fermi-Dirac distribution. Figure 2.2 shows a plot of s_{AT} as a function of $\Delta G_{AT} - \mu$. This is the exact same computation as in the occupation number formalism for the quantum equilibrium statistics of fermionic systems that are constrained by the Pauli exclusion principle.

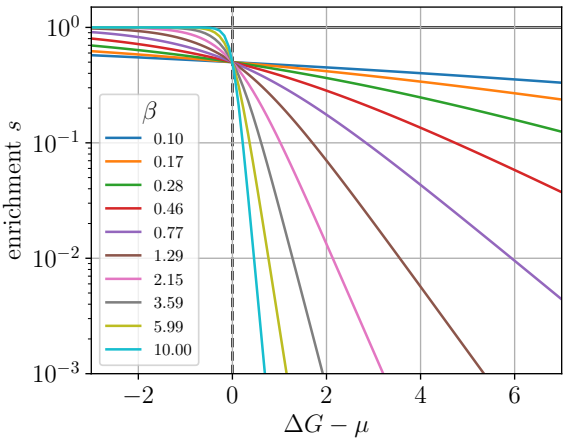


Fig. 2.2: Enrichment s as a function of binding free energy ΔG and chemical potential μ which represents target availability, given by the Fermi-Dirac statitics in equation (2.12). Various values of inverse temperature β. The saturation and Boltzmann regimes are visible to the left and right of the dashed vertical line, respectively.

In the theory of binding, the equivalent to the Pauli exclusion principle resides in the fact that at most a finite number of targets (in the case of monovalent ligands at most one) can bind to a given copy of the ligand. The bare result of equation (2.12) was discussed *e.g.* in [64, 65] and is also used *e.g.* in [68].

The equations (2.7) and (2.12) correspond to the same result for s_{AT} obtained from two different points of view, and both become identical upon setting $\Delta G = \ln(K_D/[T]^*)$ and $\mu = \ln([T]_\infty/[T]^*)$, where $[T]^*$ denotes an arbitrary reference concentration. This is coherent with equation (2.11), and μ becomes a true potential upon expressing $[T]_\infty$ as a function of other potentials (ΔG_{AT}) and initial conditions ($[A]_{\mathrm{tot}}$ and $[T]_{\mathrm{tot}}$) which has been done in equation (2.8). Hence, the chemical potential μ can be associated with the availability of target molecules and $\mathrm{e}^{\beta\mu} = [T]_\infty$ can be referred to as the associated fugacity. In practice, binding happens at constant temperature and β is typically set to unity or, equivalently, all energies are expressed in units of $\beta^{-1} = k_{\mathrm{B}}T$, but it may also be sometimes interpreted as the strength of the selective pressure (see section 2.4).

Under certain conditions that need to be specified, the Fermi-Dirac binding statistics can be approximated by a Boltzmann statistics. When the binding free energy ΔG exceeds the chemical potential μ in such a way that $\mathrm{e}^{\beta(\Delta G-\mu)} \gg 1$, the binding probability becomes small and the Fermi-Dirac statistics simplifies to the Boltzmann statistics,

$$s_{AT} \simeq \mathrm{e}^{-\beta(\Delta G_{AT}-\mu)}. \tag{2.13}$$

The Boltzmann regime is visible to the right of the dashed vertical line in figure 2.2. Note that in this approximation, s_{AT} is no longer strictly a probability; in the regime $\Delta G_{AT} \lesssim \mu$ where the

Boltzmann approximation does not hold, the value of s_{AT} can be larger than 1. In the context of binding, the requirement $e^{\beta(\Delta G-\mu)} \gg 1$ translates into a condition on the chemical potential μ and binding affinity K_{AT}, namely $[T]_\infty \ll K_{AT}$ or, equivalently, $[AT]_\infty \ll [A]_\infty$. Thus the Boltzmann limit corresponds to a limiting regime in which low binding probabilities are achieved either by low concentrations of targets or low binding affinities (*i.e.* large K_{AT}). The deviation from Boltzmann statistics at high target concentrations or strong affinities is due to saturation effects caused by limited availability of ligands (all ligands are bound with probability close to 1, irrespectively of their affinity). The Boltzmann regime is the one to be targeted in experiments to ensure that enrichments s_{AT} represent well the binding affinities K_{AT}.

This represents a similar realization of the Boltzmann limit as in quantum physics by decreasing the density of fermionic particles in a Fermi gas and thus decreasing the Fermi energy (which is identical to the chemical potential) [69], which defines the energy scale at which the quantum effects of the Pauli exclusion principle become important. The interpretation of the Boltzmann limit is again analog between both contexts: By diluting the targets in binding or the particles in a Fermi gas, the probability of each available state to be occupied by one particle is small (compared to 1), let alone the probability of two or more particles competing for the same state. The exclusion of multiple occupancy of each such state required by the Pauli principle becomes thus automatically satisfied without explicit imposition.

The bosonic equivalent to the fermionic statistics of monovalent ligands can be obtained by considering multivalent ligands instead. Multivalent ligands consist of ≥ 1 identical binding sites, allowing for the binding of several copies of the target. Multivalent ligands appear e.g in cellular signaling processes, where they are referred to as scaffold proteins [70, 71]. These scaffold proteins with $m \geq 1$ binding sites define networks of binding reactions involving $\sum_{j=0}^{m} \binom{m}{j} = 2^m$ species and $m2^m$ possible reactions between them if the m binding sites are distinguishable. So far, we have studied the special case of $m = 1$. Consider for our purpose multivalent ligands with $m \geq 1$ identical, independent and indistinguishable binding sites that can bind up to m targets at a time. Due to indistinguishability of binding sites, their positions on the ligand does not matter, thus defining only $\sum_{j=1}^{m} 1 = m$ different species and $2m$ binding and unbinding reactions. We will denote by AT^j the complex of a ligand with j targets (AT^0 is identical to A). In this case, the probability of a ligand to be bound at equilibrium (to at least one target) reads

$$s_{AT} = \frac{\sum_{j=1}^{m}[AT^j]_\infty}{[A]_\infty + \sum_{j=1}^{m}[AT^j]_\infty} = \frac{1}{1 + [A]_\infty \left(\sum_{j=1}^{m}[AT^j]_\infty\right)^{-1}}. \tag{2.14}$$

By making use of

$$K_{AT} = \frac{[AT^{j-1}]_\infty [T]_\infty}{[AT^j]_\infty}, \quad \forall j = 1, 2, \ldots, m, \tag{2.15}$$

we obtain

$$s_{AT} = \left(1 + \left(\sum_{j=1}^{m} \left(\frac{[T]_\infty}{K_{AT}}\right)^j\right)^{-1}\right)^{-1} = \frac{1}{1 + \left(\sum_{j=1}^{m} \mathrm{e}^{-j\beta(\Delta G - \mu)}\right)^{-1}}, \tag{2.16}$$

where ΔG and μ are defined as before. Denoting $z = \mathrm{e}^{-\beta(\Delta G - \mu)}$ and using the (incomplete) geometric series $\sum_{j=1}^{m} z^j = \frac{z(1-z^m)}{1-z}$, we can further simplify and find

$$s_{AT} = \frac{z\,(1 - z^m)}{1 - z^{m+1}}, \tag{2.17}$$

which converges towards the Boltzmann statistics $s_{AT} = z = \mathrm{e}^{-\beta(\Delta G - \mu)}$ as $m \to \infty$, provided that $z < 1$ and thus $\Delta G > \mu$ as required for the convergence of the geometric series. In the limit $m \to \infty$, a ligand can bind an arbitrary number of targets which represents the bosonic counterpart of the monovalent ligands obeying fermionic statistics. Surprisingly, the Boltzmann approximation for the enrichments s_{AT} is exact in this bosonic case, the reason being that the presence of infinitely many binding sites excludes competition of targets for ligands. Again, the same expression for s_{AT} as in equation (2.17) is obtained by considering a ligand in the grand-canonical ensemble with partition function $\Xi(\beta, \mu) = \sum_{j=0}^{m} \mathrm{e}^{-j\beta(\Delta G - \mu)} = \frac{1-z^{m+1}}{1-z}$,

$$s_{AT} = \frac{1}{\Xi(\beta, \mu)} \sum_{j=1}^{m} \mathrm{e}^{-j\beta(\Delta G - \mu)} = \frac{1 - z}{1 - z^{m+1}} \frac{z(1 - z^m)}{1 - z} = \frac{z(1 - z^m)}{1 - z^{m+1}}, \tag{2.18}$$

which is equivalent to equation (2.17). The assumption of indistinguishability of the m sites is, however, biologically irrelevant. Similar expressions to equation (2.17) can be obtained for distinguishable binding sites with biological relevance: For independent sites, one obtains simply the m-th power of the Fermi-Dirac statistics for a monovalent case, $s_{AT} = \left(\frac{z}{1+z}\right)^m$, which features an entropic barrier at $j = \frac{m}{2}$. For interacting sites [72], typical examples are i) sequential binding as for example in hemoglobin, *i.e.* binding to the $(j+1)$-th site requires j-th site to be in bound state (sequential model). Here, one obtains $s_{AT} = \frac{z^m(1-z)}{1-z^{m+1}}$, which is identical to a random walker with step probability z. ii) For all-or-none cooperativity between binding sites (symmetry model) [72], one obtains $s_{AT} = \frac{z^m}{1+z^m}$, leading to a switch (Hill function).

2.1.3 Conditions and implications for library selections

We now want to generalize the results of subsections 2.1.1 and 2.1.2 to libraries of many different ligands A and targets T. The binding energies ΔG_{AT} and affinities K_{AT} will be different for different combinations of ligands and targets. Ligands A differ in their amino acids sequences x, conferring them different structures and chemical properties, and thus various affinities for the targets. The same is true for targets T if they can be defined on a sequence space, such as DNA and protein targets. Ligand sequences may be beneficial or obstructive for binding depending for

instance on whether they encode for complementary versus unfitting structures, or carry opposite versus identical electrical charges with respect to the targets.

The theory developped in the context of a single ligand and a single target can be easily generalized to a diversity of ligands and targets. In the presence of $|A|$ different types of ligands A and $|T|$ different types of targets T, the reaction kinetics now consists of $|A||T|+|A|+|T|$ reaction equations for the $|A||T|$ possible complexes and for the ligands and targets. These are subject to $|A|+|T|$ conservation constraints for the total concentrations of ligands and targets. This yields $|A||T|$ independent equations of the form

$$\frac{\mathrm{d}[A_iT_j]}{\mathrm{d}t} = k_{+,ij}[A_i][T_j] - k_{-,ij}[A_iT_j], \quad i = 1,\dots,|A|, \quad j = 1,\dots,|T| \tag{2.19}$$

that are coupled *a priori* because of $[T_j] = [T_j]_{\mathrm{tot}} - \sum_{i=1}^{|A|}[A_iT_j]$. The conclusions at the level of the enrichments s_{AT} remain, however, mainly unaffected.

Consider first the case of a single target, but $|A|$ different ligands. This corresponds to the situation we will realize experimentally in the chapters 3 and 4. In this case, the computation of the enrichments s_{AT} in equations (2.7) and (2.17) remains unchanged. Only the chemical potential $\mu = \beta^{-1}\ln([T]_\infty)$ is modified as the target availability now depends on the equilibrium concentrations of all possible ligand-target complexes and thus introduces a coupling between the enrichments of different ligands. Thus, s_{A_iT} still follows a Fermi-Dirac statistics,

$$s_{A_iT} = \frac{1}{1+\mathrm{e}^{\beta\left(\Delta G_{A_iT}-\mu\right)}}, \tag{2.20}$$

with chemical potential

$$\mu = \frac{1}{\beta}\ln\left([T]_{\mathrm{tot}} - \sum_{i=1}^{|A|}[A_iT]_\infty\right) \simeq \frac{1}{\beta}\ln([T]_{\mathrm{tot}}) - \frac{\sum_{i=1}^{|A|}[A_iT]_\infty}{\beta[T]_{\mathrm{tot}}}, \tag{2.21}$$

where the approximation holds to first order if $\sum_{i=1}^{|A|}[A_iT]_{\mathrm{eq}} < \sum_{i=1}^{|A|}[A_i]_{\mathrm{tot}} \ll [T]_{\mathrm{tot}}$ using $\ln(1-\epsilon)\approx -\epsilon$. Again, to make μ a true potential, the $[A_iT]_\infty$ need to be expressed in terms of K_{A_iT}, $[T]_{\mathrm{tot}}$, $[A_i]_{\mathrm{tot}}$, $i = 1,\dots,|A|$ by solving a complicated system of coupled non-linear equations defined by $\frac{\mathrm{d}[A_iT]}{\mathrm{d}t} = 0$. However, this will no longer be possible in all generality in the case of many ligand types. According to equation (2.21), the coupling effect becomes neglectable if the total target concentration $[T]_{\mathrm{tot}}$ exceeds the final concentration of binding products.

Consider for the sake of completeness the case with $|T| \geq 1$ instead of a single target species in addition to the $|A|$ ligand types. In this case, a ligand can bind to any of the targets and the overall binding probability for ligand A reads

$$s_{A_iT} = \frac{\sum_{j=1}^{|T|}[A_iT_j]_\infty}{[A_i]_\infty + \sum_{j=1}^{|T|}[A_iT_j]_\infty} = \frac{1}{1+\left(\sum_{j=1}^{|T|}K_{AT_j}^{-1}[T_j]_\infty\right)^{-1}} = \frac{1}{1+\mathrm{e}^{\beta\left(\Delta\mathcal{G}_{A_i}-\mu\right)}}, \tag{2.22}$$

where $\mathrm{e}^{-\beta(\Delta\mathcal{G}_{A_i}-\mu)} = \sum_{j=1}^{|T|} \mathrm{e}^{-\beta(\Delta G_{A_iT_j}-\mu_j)}$ defines an effective free energy over the different targets. Thus, we still obtain Fermi-Dirac statistics, but involving $\Delta\mathcal{G}_{A_i}$ that summarizes the various $\Delta G_{A_iT_j}$.

Applying the Boltzmann approximation from equation (2.13) to all variants A will ensure that differences in binding affinities among ligand types translate into differences in enrichments s_{AT}. In order to achieve the overall validity of the Boltzmann limit for all possible interaction pairs, the choice of the target concentration $[T]_{\mathrm{tot}}$ in selection experiments must be adjusted to an intermediary regime that is flanked by two high and low target concentration limits featuring unwanted saturation effects, namely

$$\sum_{i=1}^{|A|}[A_i]_{\mathrm{tot}} \ll [T]_{\mathrm{tot}} \ll \min_{i\in\{1,\ldots,|A|\}} K_{A_iT}. \tag{2.23}$$

The origin of the second constraint was explained in subsection 2.1.2, where it was expressed as $[T]_{\mathrm{tot}} \ll K_{AT}$ in the context of a single ligand. It must be generalized to the above form to assure that best binders (represented by $\min_i K_{A_iT}$) are not in the saturation regime: It is required to prevent flattening effects resulting from competition of targets for limited amount of ligands, thus saturating the binding reaction with binding probabilities close to 1. The first constraint excludes the inverse scenario in which ligands compete for targets. Optimal selection conditions that assure enrichments represent, and represent only, intrinsic properties of the ligand (*i.e.* selection for differences in binding affinities and independently from one another) are achieved by setting the target concentration to its optimal value in between the two limits. We anticipate here that both constraints are satisfied in our phage display selection experiments: A population of $\sum_i[A_i]_{\mathrm{tot}} \simeq 10^{11}\,\mathrm{mL}^{-1}$ ligands are incubated with an excess of $[T]_{\mathrm{tot}} \simeq 10^{14}\,\mathrm{mL}^{-1}$ of targets, thus satisfying the first constraint in equation (2.23). Moreover, the selection yield ranges from about $\sum_i[A_iT]_\infty \simeq 10^5\,\mathrm{mL}^{-1}$ at the first selection round dominated by random binders to $\sum_i[A_iT]_\infty \simeq 10^7\text{–}10^8\,\mathrm{mL}^{-1}$ at the third round onwards which is dominated by good binders, down from initially $\sum_i[A_i]_{\mathrm{tot}} \simeq 10^{11}\,\mathrm{mL}^{-1}$. We can thus estimate that $[T]_{\mathrm{tot}}^{-1}\min_i K_{A_iT} \simeq \sum_i[A_i]_{\mathrm{tot}}/\sum_i[A_iT]_\infty \simeq 10^3$, which meets the second constraint in equation (2.23). The potential relevance of the Fermi-Dirac form of s_{AT} in practice has been pointed out by in the context of SELEX experiments on transcription factors [64].

2.1.4 Spin-glass models for biophysical interactions

We will continue to restrict to the case of a fixed target T and a diversity of ligands A (as this will be the case throughout most of the chapters 3 and 4). We may for this scenario adopt a modified notation in which we refer by $\Delta G(x)$ instead of ΔG_{AT} to the binding free energy of ligand A that has the sequence $x = (x_1, x_2, \ldots, x_L)$. Here, L denotes the length, *i.e.* the number of sequence positions, of sequence x. Each position x_i may take on an alphabet of q letters, in biology typically $q = 20$ for amino acids and $q = 4$ for nucleotides. The mapping $x \mapsto G(x)$ will be the object of

interest in this subsection.

The definition and study of selection potential must be done in light of some binding landscape that maps a sequence x to its selection probability $s(x)$, likely depending on some external parameter. With the results of subsections 2.1.1 to 2.1.3, we have reduced this problem to the study of $x \mapsto \Delta G(x)$ by expressing the selection probability $s(x)$ of a sequence x in terms of its free energy of binding $\Delta G(x)$,

$$s(x) = \frac{1}{1 + \mathrm{e}^{\beta(\Delta G(x) - \mu)}}, \tag{2.24}$$

where μ is a chemical potential accounting for target availability. Here, the task thus consists in defining a suitable class of binding landscapes. However, a bottom-up approach toward $x \mapsto \Delta G(x)$ *a priori* requires the knowledge about microscopic details of binding mechanisms of random ligands and the construction of a possibly complicated Hamiltonian reflecting these mechanisms. Such a detailed modeling of $x \to \Delta G(x)$ seems, however, tedious if not impossible due to a large number of potential binding interactions and our insufficient knowledge about the nature and relevance of these interactions. Rational design of ligands for given targets based on structural and chemical aspects of binding led to affinities higher than non-specific but far less than what is achieved through directed evolution [73, 74, 75]. Yet, this knowledge may be non-essential for an understanding of selection at the ensemble level if it is possible to define classes of random models that (statistically) reproduce features of the true landscapes. The idea and hope is that such random models should capture the statistical properties of the true landscapes. Therefore, we here discard a possibly complicated modeling of $x \mapsto \Delta G(x)$ in favor of a precise, more easily tractable class of random models and justify their likely applicability to our problem. Such an approach is historically reminiscent of the quantum description of atomic nuclei in which the use of random matrix theory may successfully replace the search for and study of complicated true Hamiltonians [76, 77]. However, the use of such statistical models in quantitative biology is not new either due to the omnipresence of untractable complexity, but increasing availability of biological data (in particular sequences and structures). They are now extensively used in protein evolution [34, 78, 79, 80, 81, 82] and data-driven approaches [83] in various contexts ranging from structural [84] or functional [36, 85, 37] decomposition of proteins or both [17], structural [86, 87, 88] and functional [89] prediction, over to binding specificities of transcription factors [63, 64, 65, 66] and signaling proteins.

In spite of the variety of contexts, the statistical models in use are usually shared across different applications and correspond to (combinations and generalizations to Potts spins of) mean-field p-spin glass models with Hamiltonian

$$\mathcal{H}(x) = \sum_{i_1 < i_2 < \cdots < i_p} J_{i_1, i_2, \ldots, i_p}(x_{i_1}, x_{i_2}, \ldots, x_{i_p}) \tag{2.25}$$

where the $J_{i_1, i_2, \ldots, i_p}(a_1, a_2, \ldots, a_p)$ are contributions to the total energy $\mathcal{H}(x)$ of a sequence (or configuration) x and encode for p-body interactions, *i.e.* interactions between *a priori* all subsets

of size $0 \leq p \leq L$ among the L positions (Potts spins). The use of mean-field may be justified by the fact that sites, which are far away along the sequence, may still be close in real space within a folded protein. In the forward study of such models, a typical because analytically tractable choice of these interactions is an independent, Gaussian disorder,

$$J_{i_1,i_2,\ldots,i_p}(a_1, a_2, \ldots, a_p) \overset{d}{=} \mathcal{N}\left(0, \frac{p!\sigma^2}{L^{p-1}}\right). \tag{2.26}$$

The choice of the variance $\frac{p!\sigma^2}{L^{p-1}}$ assures extensivity of $\mathcal{H}(x)$, *i.e.* proportionality with system size L. On the contrary, the inverse study of these models involves the inference of the interactions from experimental data, such as empirical correlation functions. Beyond proteins, these models are also successfully applied in fitness inference [90], neuroscience, regulatory network reconstruction, and outside biology *e.g.* in finance [83].

Upon setting $p = 1$ in equation (2.25), we obtain an additive model consisting of sites that contribute independently one from another to the sequence energy $\mathcal{H}(x)$

$$\mathcal{H}(x) = \sum_{i=1}^{L} J_i(x_1) \equiv \sum_{i=1}^{L} h_i(x_i), \tag{2.27}$$

where the $J_i(a) \equiv h_i(a)$ are called local field functions. By combining $p = 1$ with $p = 2$, we obtain the DCA Hamiltonian [84]

$$\mathcal{H}(x) = \sum_{i=1}^{L} h_i(x_i) + \sum_{i=1}^{L} \sum_{j=1}^{i-1} J_{ij}(x_i, x_j) \tag{2.28}$$

that also allows for interactions $J_{ij}(a, b)$ between pairs of sites i, j. Higher-order interactions are then easily obtained by adding more terms of the form of equation (2.25) with increasing p. In the limit $p \to \infty$ (after $L \to \infty$), equation (2.25) becomes a "double mean-field" model in which all Potts spins explicitly interact altogether. This case will be of central interest within our work and is discussed in more detail below. Another sub-class of equation (2.25) are the so-called NK models that have been used as fitness landscapes in studies of protein evolution and affinity maturation [34, 13, 91, 82]. Here, each Potts spin or site i contributes by $h_i(x_i, x_{i_1}, x_{i_2}, \ldots, x_{i_K})$ to the overall energy (or fitness) $\mathcal{H}(x)$,

$$\mathcal{H}(x) = \sum_{i=1}^{L} h_i(x_i, x_{i_1}, x_{i_2}, \ldots, x_{i_K}). \tag{2.29}$$

(The original definition of the NK model defines the fitness as the intensive equivalent of $\mathcal{H}(x)$, *i.e.* including a normalization by the system size, $\frac{1}{L}\mathcal{H}(x)$). The contribution of a site i depends on the state of K other sites, thus the Hamiltonian is again made up of p-body interactions with $p = K + 1$ in equation (2.25). The difference to equation (2.25) is that only a fixed set of K positions interacts with position i, while the more general model assumes interactions between i and all possible subsets of p positions. The connection between equations (2.25) and (2.29) is

formally achieved by setting $J_{i_1,\dots,i_p}(x_{i_1},\dots,x_{i_p}) \leftarrow h_i(x_i, x_{i_1},\dots,x_{i_K})$.

The general form of the model in equation (2.25) interpolates between a smooth and convex landscape for $p = 1$ with a single optimum and a perfectly uncorrelated and rugged landscape for $p \to \infty$ where the energies $\mathcal{H}(x)$ are uncorrelated even between neighboring sequences x. In the case of NK models, such a landscape contains on average $\frac{2^L}{L+1}$ local optima [82]. The complete decorrelation as $p \to \infty$ can be observed by computing the covariance between two sequences x and y,

$$\begin{aligned}
\langle \mathcal{H}(x)\mathcal{H}(y)\rangle &= \sum_{i_1<i_2<\cdots<i_p} \sum_{j_1<j_2<\cdots<j_p} \langle J_{i_1,i_2,\dots,i_p}(x_{i_1},x_{i_2},\dots,x_{i_p}) J_{j_1,j_2,\dots,j_p}(y_{j_1},y_{j_2},\dots,y_{j_p})\rangle \\
&= \sum_{i_1<i_2<\cdots<i_p} \sum_{j_1<j_2<\cdots<j_p} \underbrace{\langle J_{i_1,i_2,\dots,i_p}(x_{i_1},x_{i_2},\dots,x_{i_p}) J_{j_1,j_2,\dots,j_p}(y_{j_1},y_{j_2},\dots,y_{j_p})\rangle}_{=\frac{p!\sigma^2}{L^{p-1}}\prod_{k=1}^{L}\delta(x_{i_k},y_{j_k})} \prod_{k=1}^{p}\delta_{i_k,j_k} \\
&= \frac{p!\sigma^2}{L^{p-1}} \sum_{i_1<i_2<\cdots<i_p} \prod_{k=1}^{L}\delta(x_{i_k},y_{j_k}) \\
&= \frac{\sigma^2}{L^{p-1}} \sum_{i_1,i_2,\dots,i_p} \prod_{k=1}^{L}\delta(x_{i_k},y_{i_k}) \\
&= L\sigma^2 O(x,y)^p, \qquad O(x,y) = \frac{1}{L}\sum_{i=1}^{L}\delta(x_i,y_i),
\end{aligned} \tag{2.30}$$

where $O(x,y) \in [0,1]$ denotes the overlap between sequences x and y (normalized Hamming distance). Because of $O(x,y) = 1$ only if $x = y$ and $O(x,y) < 1$ otherwise, it follows for non-identical sequences x and y that $\langle \mathcal{H}(x)\mathcal{H}(y)\rangle \to 0$ as $p \to \infty$, and thus $\langle \mathcal{H}(x)\mathcal{H}(y)\rangle \to L\sigma^2\delta(x,y)$ as $p \to \infty$. This means that the landscape completely decorrelates and similar sequences do no longer have similar energy.

The cases of $p = 1$ and $p = 2$ and the Hamiltonians in equations (2.27) and (2.28) oftentimes appear in the context of *e.g.* i) protein structure and ii) binding specificities: i) A model of the form of equation (2.28), along with the Boltzmann distribution $P(x) = Z^{-1}\mathrm{e}^{-\beta\mathcal{H}(x)}$ with $Z = \sum_x \mathrm{e}^{-\beta\mathcal{H}(x)}$, are used to model the probability of occurrence $P(x)$ of a sequence x in alignments of homologous sequences or any other vector x of possibly correlated quantities (*e.g.* neural status, gene expression levels, stock values, etc.). One motivation is that equation (2.28) arises naturally as the model that maximizes entropy under the constraints of fixed first- and second-order correlation functions $\langle\delta(x_i,a)\rangle_{P(x)}$ and $\langle\delta(x_i,a)\delta(x_j,b)\rangle_{P(x)}$ that can be measured from data. It represents a Potts model because it is normalized such that $\sum_x P(x) = 1$. In practice, the correlation functions are empirically estimated from single- and two-point frequencies $f_i(a)$ and $f_{ij}(a,b)$ in the alignment, and the parameters $h_i(a)$ and $J_{ij}(a,b)$ are then inferred in such a way that the model statistics matches these frequencies. Here, the energy of a sequence $\mathcal{H}(x)$ has no immediate physical meaning, although it is predictive of protein thermal stability [92]. The model parameters $h_i(a)$ and in particular $J_{ij}(a,b)$ are of interest: Non-zero values of $J_{ij}(a,b)$

can oftentimes be associated with physical and evolutionary coupling of the sites i and j in the protein fold. ii) In the context of binding, models of the form of equations (2.28) are used to model the binding free energy landscape $\mathcal{H}(x) \equiv \Delta G(x)$. However, the motivation for these models and the interpretation of $P(x)$ are different to i). In addition, it is not a Potts model as $P(x)$ is related to $\mathcal{H}(x)$ by the Fermi-Dirac statistics (see section 2.1) and $\sum_x P(x) \neq 1$ in general. Equations (2.27) and (2.28) can be justified as being a Taylor-like expansion (cut-off at $p = 1$ or $p = 2$) of an arbitrary, but not too random ("$p \ll \infty$") true Hamiltonian $\mathcal{H}(x)$. Here, $h_i(a)$ can represent beneficial versus deleterious amino acids independently of the sequence context, while the meaning of $J_{ij}(a, b)$ is less clear. The couplings may be the result of a global non-linearity, such as the Fermi-Dirac form of $P(x)$, or to cooperative effects between sites [60]. In either context, the models are generally stopped at the second-order term due to the explosion in the number of parameters and the estimation of three- and higher-order correlation functions requiring unachievable amounts of data. The feasibility in principle of the inference of higher-order couplings has, however, been shown [93]. Variants of equation (2.28) may impose translational invariance, so that couplings are constrained to $J_{ij}(x_i, x_j) = J_{|i-j|}(x_i, x_j)$. This appears in particular in the context of transcription factors where the position of the target sequence along the DNA does not matter [63].

Within the scope of this thesis, we will consider statistical models of the form of equation (2.25) with $p = \infty$ and $p = 1$ (see chapter 4) for selections from antibody libraries. The $p = \infty$ resumes to the $\Delta G(x)$ being independent (and identically distributed) random variables, *i.e.* the binding energy $\Delta G(x)$ for each sequence x is drawn independently of all others from (a common) probability distribution $P(\Delta G)$. The class of distributions that should be used for $P(\Delta G)$ remains free so far. In section 2.3, we will graft conclusions from the $p = 1$ case onto the $p = \infty$ case in order to fix a class of distributions for $P(\Delta G)$: Using the central limit theorem, the result will be that a Gaussian distribution with mean μ and variance σ^2 should provide a reasonable approximation to the true distribution of binding energies,

$$\mathcal{H}(x) \equiv \Delta G(x) \overset{d}{=} \mathcal{N}(\mu, \sigma^2) \tag{2.31}$$

This Hamiltonian is identical to the one of the Derrida random energy model (up to a shift and rescale by μ and σ). The random energy model is one of the simplest spin glass models in the theory of disordered systems bearing a phase transition [94, 95]. This model is defined on L spins, where the energy $\mathcal{H}(x)$ of each configuration x is an iid Gaussian random variable with mean $\mathbb{E}[\mathcal{H}(x)] = \mu = 0$ and covariance $\langle \mathcal{H}(x)\mathcal{H}(y) \rangle = L\sigma^2\delta(x, y) = L\delta(x, y)$. Alternatively, if only the left tail of $P(\Delta G)$ or, equivalently, the right tail of $P(s)$ matter (as is typically the case in practice), extreme-value theory constrains the choice of the model.

2.2 Universality of selection statistics

In this chapter, we will provide arguments for similar selection phenomenologies in landscapes described by the class of random models presented in section 2.1.4, thus introducing the notion of universality in the context of selection. Universality is a key observation in statistical mechanics stating that multi-component systems may be insensitive to a majority of microscopic details of the underlying interactions, thus constraining their collective behaviour to few qualitatively different classes (universality classes) [96, 97]. As an example, the phenomenology of an Ising model close to a critical point (encoded by the critical exponents) is governed by few relevant interaction terms (or operators) in a potentially complicated Hamiltonian [96, 97]. Similarly, we can make statements about selection properties yet leaving aside all the complicated and even unknown details of the selection-driving mechanisms. We will see in chapter 4 that such coarse-grained models can be sufficient to analyze experimental selection data and to capture and dissociate qualitatively different phenomenologies.

2.2.1 The central-limit theorem predicts lognormality of enrichments

Our goal is to define a potentially relevant class of distributions $P(s)$ for the enrichments s in libraries of random ligands under selection for binding. Let us assume that binding between a fix target and ligands that differ in their sequences $x = (x_1, x_2, \ldots, x_L)$ is described by an additive binding free energy $\Delta G(x)$ with independent contributions from all sequence positions, *i.e.* the simplest form within the class of interaction models defined by equation (2.25) (with $p = 1$),

$$\mathcal{H}(x) \equiv \Delta G(x) = \sum_{i=1}^{L} h_i(x_i). \tag{2.32}$$

Here, the local field functions $h_i(a)$ represent the context-independent contributions to binding energy that results from position i carrying amino acid a and are instances of a position-specific random variable H_i with some probability distribution $P_i(h)$. If these variables H_i have finite first and second moments $\langle H_i \rangle$ and $\langle H_i^2 \rangle$, then the central-limit theorem (CLT) implies a Gaussian distribution for $\Delta G = \sum_{i=1}^{L} H_i$ with mean and variance

$$\mu = \sum_{i=1}^{L} \langle H_i \rangle = L \langle H \rangle, \tag{2.33}$$

$$\sigma^2 = \sum_{i=1}^{L} \left(\langle H_i^2 \rangle - \langle H_i \rangle^2 \right) = L \left(\langle H^2 \rangle - \langle H \rangle^2 \right) \tag{2.34}$$

for sufficiently large sequence length L. The second equalities in equations (2.33) and (2.34) hold if all L sites contribute equally to binding, *i.e.* if $H_1 \stackrel{d}{=} H_2 \stackrel{d}{=} \ldots \stackrel{d}{=} H_L \stackrel{d}{=} H$. The meaning of "sufficiently large" L depends on the distribution of energy contributions per site $P_i(h)$; if these

distributions are already "close" to a Gaussian distribution, their sum will be even more so with only few sites.

Importantly, the central limit theorem is robust to a certain amount of correlation between sites i that may be introduced either in the form of correlations between the H_i or by higher-order terms in the model for ΔG involving *e.g.* pairwise couplings $J_{ij}(a,b)$ ($p=2$). Thus, strict additivity of the binding mechanism is not required and the above result is expected to hold even in presence of weak correlation between sites. However, it is generally difficult to find quantitative criteria for the amount of correlation tolerated by the CLT. Reciprocally, the observation of Gaussian binding energies ΔG does not necessarily imply strict additivity of the underlying binding mechanism.

According to the results of section 2.1.3, the binding probability at equilibrium is given by $s(x) \simeq \mathrm{e}^{-\beta\Delta G(x)}$ in a regime of intermediate target concentrations, with $\beta = (k_B T)^{-1}$ the inverse temperature. Hence, if ΔG follows a Gaussian distribution with mean μ and variance σ^2 across ligand sequences, it follows that the enrichments s should obey a lognormal distribution with PDF

$$P(s) = \frac{1}{\sqrt{2\pi}\sigma s} \exp\left(-\frac{(\ln(s)-\mu)^2}{2\sigma^2}\right), \tag{2.35}$$

where the parameters μ and σ are redefined in units of $\beta^{-1} = k_B T$, $\mu = -\beta\sum_{i=1}^{L}\langle H_i\rangle$ and $\sigma^2 = \beta\sum_{i=1}^{L}\left(\langle H_i^2\rangle - \langle H_i\rangle^2\right)$.

Further support for lognormal distributions comes from their stability under iteration of the selection process. If a first selection with enrichment s_1 is followed by a second selection with enrichment s_2, then the lognormality of s_1 and s_2 implies lognormality of the overall enrichment $s = s_1 s_2$, with parameters $\mu = \mu_1 + \mu_2$ and $\sigma^2 = \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2$, where ρ is the correlation between the two selective pressures. This property is inherited from the Gaussian distribution which is a fix point under addition. Lognormality of enrichments in selection is a special case of the more general property of lognormal distributions as attractors of evolutionary dynamics [98]. However, this stability requires the limit of large populations: It has been shown that the disappearance-by-chance of rare (in particular good, but rare) sequences may lead to pathological behaviour at the population scale [99]. Finally, the relevance of lognormal distributions for binding affinities has already been shown in the literature [100, 101, 102, 103, 104, 105].

In practice, deviations from the lognormal distribution may occur if the assumptions for this result are not met: For instance, global non-linearities as for instance introduced by saturation effects may invalidate the lognormal model. Such non-linearities will find their way into ΔG where they may give rise to apparent pairwise couplings: If the non-linearity reads $f(\Delta G) = \Delta G + \alpha(\Delta G)^2$ to the lowest non-linear order, we obtain

$$\Delta G^{\mathrm{app}} = \sum_{i=1}^{L} h_i(x_i) + \alpha \sum_{i=1}^{L}\sum_{j=1}^{L} \underbrace{h_i(x_i)h_j(x_j)}_{=J_{ij}^{\mathrm{eff}}(x_i,x_j)}, \tag{2.36}$$

where $J_{ij}^{\text{eff}}(a,b)$ denotes an effective pair-wise coupling unrelated to the true ΔG. For example, in the case of the Fermi-Dirac statistics that relates s to ΔG as derived in section 2.1.2, we have $f(\Delta G) = \Delta G - \frac{1}{2}(\Delta G)^2$. But most importantly, a model for ΔG with additive binding energy contributions itself appears to be a strong assumption. Deviations from an additive model (that are not due to a global non-linearity) may be the result of the presence of several epitopes, or couplings between several sites due to cooperative and adverse effects between nearby amino acids upon binding. Finally, all potentially diverting factors mentioned here are in principle tractable perturbatively by extending the first-order model presented here, possibly at the expense of introducing additional parameter: Known non-linearities may be systematically accounted for by a simple change of variables: Taking again the Fermi-Dirac statistics, we find

$$P(s) = \frac{1}{\sqrt{2\pi}\sigma s(1-s)} \exp\left[-\frac{1}{2\sigma^2}\left(\ln\left(\frac{s}{1-s}\right) - \mu\right)^2\right], \tag{2.37}$$

which is supported on $s \in [0,1]$ and thus a "probability distribution of a probability". Alternatively, unknown global non-linearities may be fitted by splines, *i.e.* expanding the unknown true non-linearity into some set of nonlinear base functions [106], or by discretization of the non-linearity [93]. Beyond global non-linearities, true interactions between sequence positions may be accounted for by extending the binding model by a second-order term invoking pairwise interactions $J_{ij}(a,b)$ in addition to the local fields $h_i(a)$. However, we will show later that lognormal distributions can provide a reasonable fit to experimental enrichments in some cases (see chapter 4), and, most interestingly, that an additive model may capture the binding landscape of an antibody binding site surprisingly well in some cases (see chapter 4).

2.2.2 Mathematical constraints: extreme-value theory

A typical observation in biological and other contexts are power-law distributions of observables such as frequencies of occurrence [107], *e.g.* of antibody sequences in the immune repertoire [108], and so it happens to appear for enrichments in selection data [1]: When sorting a list of empirical enrichments in decreasing order such that $s_1 \geq s_2 \geq \cdots \geq s_N$ and plotting s_r against their rank $r \in \langle 1, 2, \ldots, N\rangle$, we sometimes observe power-law decrease (linearity in log-log scale). Such power-laws are oftentimes associated with (near-to) criticality of the underlying interactions of a system's constituents [109, 108], although it has been shown that inferred models are inherently likely to yield critical points in parameter space [110]. Moreover, such power-laws are seemingly inconsistent with lognormal distributions that we have motivated in the previous subsection. However, we will show here based on [111] that *truncated* data may indeed be consistent with both power-laws and lognormal distributions and other Gumbel-type distributions. In sequencing data-based approaches, truncation is a result of finite sequencing depth, meaning that the true diversity exceeds (by far) the sequencing budget.

We will continue to assume the case where enrichments are iid variables from a probability

density $P(s)$. Regardless of any prior statement on $P(s)$, such as the lognormality motivated in subsection 2.2.1, the shape of the tail of $P(s)$ is constrained by mathematics within the so-called *extreme-value theory* (EVT) [112, 113]. The tail of $P(s)$ is of particular interest in selection, because selected populations will be dominated by strong binders with high enrichment s. The conclusions of EVT have previously been applied to selection data within the group [1]. For any random variable S with PDF $P(s)$, the PDF of threshold-exceedance, *i.e.* of the probability of having $S \geq s$ conditioned to $S \geq s^*$ with $s \geq s^*$, converges in distribution to a generalized Pareto distribution $f_{\kappa,s^\star,\tau}(s) = \tau^{-1} f_\kappa\left((s - s^\star)/\tau\right)$ as $s^\star \to \infty$ [112], where

$$f_\kappa(x) = \begin{cases} (1+\kappa x)^{-\left(1+\frac{1}{\kappa}\right)} & \text{if} \quad \kappa \neq 0, \\ \mathrm{e}^{-x} & \text{if} \quad \kappa = 0. \end{cases} \tag{2.38}$$

The qualitative behaviour of these PDF is determined by the sign of the shape parameter κ, which in turn is determined by the shape of the tail of $P(s)$. It defines three universality classes called respectively the Weibull class ($\kappa > 0$), the Gumbel class ($\kappa = 0$) and the Fréchet class ($\kappa < 0$). The Weibull class comprises all infinitely supported distributions $P(s)$ decreasing as a power-law as $s \to +\infty$, $P(s) \sim Cs^{-\alpha}$ with some constant $C > 0$. The Fréchet class comprises all finitely supported distributions with $P(s) \sim (s_+ - s)^{-\alpha}$, where s_+ denotes a finite upper bound to S. Finally, all "intermediate" distributions with infinite support, but decreasing faster than any power-law fall into the Gumbel class, including lognormal distributions. The PDF in equation (2.38) in the case of $\kappa = 0$ is obtained by taking the analytical continuation of the case with $\kappa \neq 0$.

2.2.3 Order statistics and power-law mimicry: implications for finite data

Sequencing data from selection experiments gives access to a list of $N < q^L$ empirical enrichments $\{s_r\}_{r\in\langle 1,2,\ldots,N\rangle}$, typically the highest among all q^L enrichments. In subsection 2.2.1, we have argued for a lognormal distribution of these numbers. Under this assumption, extreme-value theory then suggests that top enrichments should asymptotically obey a generalized Pareto distribution with $\kappa = 0$. However, we will show here that the extremes of finite samples from lognormal distribuions may actually be consistent with a generalized Pareto distribution with non-zero shape parameter $\kappa \neq 0$, although lognormal distributions fall into the class with $\kappa = 0$ strictly. This phenomenon was described in [111] as *power-law mimicry*. To this aim, the question about the statistics of extremes already mentioned in subsection 2.2.1 needs to be slightly reformulated mathematically: Given a model for $P(s)$, such as the lognormal or the generalized Pareto distribution from subsections 2.2.1 and 2.2.2, what is the distribution of the r-th largest value $S_{N:r}$, $1 \leq r \leq N$, within a sample of size N taken from $P(s)$? This is the central question in *order statistics* [114]. The theory provides the scaling in N and distribution of $S_{N:r}$ which again displays universality, with the exact same three universality classes as in extreme-value theory (see subsection 2.2.2).

This constrains in particular $\mathbb{E}[S_{N:r}]$ which, when expressed as a function of r, predicts the shape of enrichment-rank plots $s_r(r)$ (for the highest order statistics $r \ll N$ at least). Thus, data from different models $P(s)$ from the same universality class may thus give rise to similar shape of $s_r(r)$. In particular, N values $s_1 \geq s_2 \geq \cdots \geq s_N$ drawn from Gumbel-type distributions $P(s)$, such as the lognormal distribution and the generalized Pareto distribution with $\kappa > 0$, give rise to an apparent power-law behaviour of $s_r(r)$,

$$\mathbb{E}[\ln S_{N:r}] \simeq a_N - b_N \ln(r). \tag{2.39}$$

Only the a_N and b_N depend on the precise choice of the (Gumbel-type) $P(s)$. Conversely, this also suggests that empirical data with power-law appearance may be consistent with several Gumbel-type models. Note that for analytical purposes, it is beneficial to study $\mathbb{E}[\ln S_{N:r}]$ rather than $\ln \mathbb{E}[S_{N:r}]$, as we will see below.

We will reproduce here the derivation of equation (2.39), which holds for Gumbel-type variables S and compute the coefficients a_N and b_N for lognormal distributions with parameters μ and σ, as well as for generalized Parato distributions with shape parameter $\kappa > 0$. Consider a Gumbel-type variable S and its logarithm $Y = \ln(S)$. If S is a lognormal (generalized Pareto with $\kappa > 0$) variable, then Y is a Gaussian (exponential) variable,

$$\begin{aligned} S \overset{d}{=} \ln \mathcal{N}(\mu, \sigma) \quad &\rightarrow \quad Y = \ln(S) \overset{d}{=} \mathcal{N}(\mu, \sigma) \\ S \overset{d}{=} \text{GenPareto}(\kappa, \tau, s^*) \quad &\rightarrow \quad Y = \ln(S) \overset{d}{=} \text{Exp}(\kappa^{-1}, \tau\kappa^{-1}, s^*). \end{aligned} \tag{2.40}$$

This can be seen by performing a simple change of variables or, alternatively, by simply replacing $s \leftarrow \mathrm{e}^y$ in the lognormal and generalized Pareto CDFs. The CDFs and PDFs of the various random variables involved here are

$$\mathcal{N}(\mu, \sigma): \qquad F(y|\mu, \sigma) = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{y-\mu}{\sqrt{2}\sigma}\right), \qquad P(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \tag{2.41}$$

$$\ln \mathcal{N}(\mu, \sigma): \quad F(s|\mu, \sigma) = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{\ln(s)-\mu}{\sqrt{2}\sigma}\right), \quad P(s|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma s} \exp\left(-\frac{(\ln(s)-\mu)^2}{2\sigma^2}\right) \tag{2.42}$$

$$\text{Exp}(\alpha, \epsilon, s^*): \qquad F(y|\alpha, \epsilon, y^*) = 1 - \epsilon^\alpha \mathrm{e}^{-\alpha y}, \qquad P(y|\alpha, \epsilon, y^*) = \alpha \epsilon^\alpha \mathrm{e}^{-\alpha y} \tag{2.43}$$

$$\text{GenPareto}(\alpha, \epsilon, s^*): \qquad F(s|\alpha, \epsilon, y^*) = 1 - \left(\frac{\epsilon}{s}\right)^\alpha, \qquad P(s|\alpha, \epsilon, y^*) = \alpha \epsilon^\alpha s^{-1-\alpha}. \tag{2.44}$$

Let us denote by $F_{N:r}(y)$ the CDF of the r-th order statistic $Y_{N:r}$ in a sample of size N. $F_{N:r}(y)$ is given by the probability that $Y_{N:r}$ is smaller than or equal to y, *i.e.* the probability at most $r-1$ among N sample values are larger than y,

$$F_{N:r}(y) = \mathbb{P}[Y_{N:r} \leq y] = \sum_{k=0}^{r-1} \binom{N}{k} F(y)^{N-k} \left(1 - F(y)\right)^k, \tag{2.45}$$

where $F(y)$ is the single sample CDF given by equations (2.41) or (2.43). For all fixed r and $y < \infty$, we have that $F_{N:r}(y) \to 0$ as $N \to \infty$. In order to allow for a non-trivial limit as $N \to \infty$, the rescaled variable $\tilde{Y}_{N:r} = \frac{Y_{N:r} - a_N}{b_N}$ with CDF $\tilde{F}_{N:r}(y) = F_{N:r}(a_N + b_N y)$ should be considered instead. With a suitable choice of the coefficients a_N and b_N, and some function $\gamma(y)$ such that

$$F(a_N + b_N y) = 1 - \frac{\gamma(y)}{N} + \mathcal{O}\left(\frac{1}{N^2}\right), \tag{2.46}$$

we obtain for the first order statistic CDF

$$\begin{aligned} \lim_{N\to\infty} \tilde{F}_{N:1}(y) &= \lim_{N\to\infty} F_{N:1}(a_N + b_N y) = \lim_{N\to\infty} F(a_N + b_N y)^N \\ &= \lim_{N\to\infty} \left(1 - \frac{\gamma(y)}{N} + \mathcal{O}\left(\frac{1}{N^2}\right)\right)^N = \mathrm{e}^{-\gamma(x)}, \end{aligned} \tag{2.47}$$

where we have used in the second equality the fact that $F(y)^N$ is the probability that all N values are $\leq y$. The coefficients a_N and b_N reveal the scaling of $Y_{N:r}$ in N, $Y_{N:r} = a_N + b_N \tilde{Y}_{N:r}$, where $\tilde{Y}_{N:r}$ is a random variable of order 1 encoding the dependence of $Y_{N:r}$ in r. In general, *i.e.* beyond the case case of Gumbel-type distributions, they can be determined by solving

$$F(a_N) = 1 - \frac{1}{N}, \quad b_N = \frac{1}{N a_N}. \tag{2.48}$$

In the $N \to \infty$ limit, $F_{N:1}(y)$ converges towards $\mathrm{e}^{-\gamma(y)}$ and $Y_{N:1}$ necessarily converges in distribution to one of only three classes of probability distributions,

$$\text{Gumbel:} \quad \tilde{F}_{N:1}(y) = \mathrm{e}^{-\mathrm{e}^{-y}}, \quad \tilde{P}_{N:1}(y) = \mathrm{e}^{-y-\mathrm{e}^{-y}}, \quad y \in (-\infty, \infty) \tag{2.49}$$

$$\text{Weibull:} \quad \tilde{F}_{N:1}(y) = \mathrm{e}^{-(-y)^\alpha}, \quad \tilde{P}_{N:1}(y) = \alpha(-y)^{\alpha-1}\mathrm{e}^{-(-y)^\alpha}, \quad y \in (-\infty, 0] \tag{2.50}$$

$$\text{Fréchet:} \quad \tilde{F}_{N:1}(y) = \mathrm{e}^{-y^{-\alpha}}, \quad \tilde{P}_{N:1}(y) = \alpha x^{-(\alpha+1)}\mathrm{e}^{-y^{-\alpha}}, \quad y \in [0, +\infty). \tag{2.51}$$

These are the same universality classes as in EVT encountered in subsection 2.2.2. Finally, it can be shown that the CDF of the r-th order statistic $\tilde{F}_{N:r}(y)$ can be expressed in terms of the one for the first order statistic $\tilde{F}_{N:1}(y)$ through [114]

$$F_{N:r}(y) = F_{N:1}(y) \sum_{k=0}^{r-1} \frac{(-\ln F_{N:1}(y))^k}{k!} = \frac{1}{(r-1)!} \int_{-\ln F_{N:1}(y)}^{\infty} \mathrm{e}^{-\zeta} \zeta^{r-1} \mathrm{d}\zeta. \tag{2.52}$$

In the case of exponential Y with CDF $F(y)$ given in equation (2.43), the scaling coefficients a_N and b_N can be easily found using equation (2.46) through

$$1 - \left(\frac{\tau}{\kappa}\right)^{\frac{1}{\kappa}} \mathrm{e}^{-\frac{a_N + b_N y}{\kappa}} = 1 - \frac{\gamma(y)}{N}, \tag{2.53}$$

which is solved by $a_N = \kappa \ln(N) + \ln\left(\frac{\tau}{\kappa}\right)$, $b_N = C\kappa$, $\gamma(y) = \mathrm{e}^{-Cy}$, where C is an arbitrary constant that we conveniently set to $C = 1$. This confirms that we do indeed find the limiting CDF of the

Gumbel class in equation (2.49). In the case of the Gaussian distribution for Y with CDF given in equation (2.43), the computation is more complicated and yields [115]

$$a_N = \mu + \left(\sqrt{2\ln N} - \frac{\ln\ln N + \ln(4\pi)}{2\sqrt{2\ln N}}\right)\sigma, \qquad b_N = \frac{\sigma}{\sqrt{2\ln N}}. \tag{2.54}$$

In order to reveal the dependence of $\mathbb{E}[\ln S_{N:r}] = \mathbb{E}[Y_{N:r}] = a_N + b_N\mathbb{E}[\tilde{Y}_{N:r}]$ on r, it remains to compute $\mathbb{E}[\tilde{Y}_{N:r}]$ for Gumbel-type distributions. Inserting the first-order statistic CDF in equation (2.49) into equation (2.52) yields

$$\tilde{F}_{N:r}(y) = \mathrm{e}^{-\mathrm{e}^{-y}} \sum_{k=0}^{r-1} \frac{\mathrm{e}^{-ky}}{k!}. \tag{2.55}$$

The probability distribution function is then obtained by taking the derivative with respect to y,

$$\begin{aligned}\tilde{P}_{N:r}(y) = \frac{\mathrm{d}\tilde{F}_{N:r}}{\mathrm{d}y}(y) &= \mathrm{e}^{-\mathrm{e}^{-y}}\left(\sum_{k=0}^{r-1}\frac{\mathrm{e}^{-(k+1)y}}{k!} - \sum_{k=1}^{r-1}\frac{\mathrm{e}^{-ky}}{(k-1)!}\right)\\ &= \mathrm{e}^{-\mathrm{e}^{-y}}\left(\sum_{k=0}^{r-1}\frac{\mathrm{e}^{-(k+1)y}}{k!} - \sum_{k=0}^{r-2}\frac{\mathrm{e}^{-(k+1)y}}{k!}\right) = \frac{1}{(r-1)!}\mathrm{e}^{-ry-\mathrm{e}^{-y}}.\end{aligned} \tag{2.56}$$

Finally, the expectation $\mathbb{E}[\tilde{Y}_{N:r}]$ reads

$$\begin{aligned}\mathbb{E}[\tilde{Y}_{N:r}] &= \frac{1}{\Gamma(r)}\int_{-\infty}^{\infty} y\mathrm{e}^{-ry-\mathrm{e}^{-y}}\mathrm{d}y = \frac{-1}{\Gamma(r)}\frac{\partial}{\partial r}\underbrace{\int_{-\infty}^{\infty}\mathrm{e}^{-ry-\mathrm{e}^{-y}}\mathrm{d}y}_{=\int_0^\infty \mathrm{e}^{-y}y^{r-1}\mathrm{d}y} = \frac{-1}{\Gamma(r)}\frac{\partial\Gamma}{\partial r}(r)\\ &= -\psi(r) = \gamma - H_{r-1},\end{aligned} \tag{2.57}$$

where $\psi(\cdot) = (\ln\Gamma)'(\cdot)$ is the Digamma function, $\gamma = \lim_{n\to\infty}\left(-\ln(n) + \sum_{r=1}^{n}\frac{1}{r}\right) \simeq 0.577$ is the Euler-Mascheroni constant, and $H_r = \sum_{k=1}^{r}\frac{1}{k}$ is the r-th harmonic number with the convention $H_0 = 0$. Using $H_r = \ln(r) + \gamma + \frac{1}{2r} - \frac{1}{12r^2} + \mathcal{O}(r^{-3})$ and thus $H_{r-1} = \ln(r) + \gamma - \frac{1}{2r} + \mathcal{O}(r^{-2})$, we have $\mathbb{E}[\tilde{Y}_{n:r}] = -\ln(r) + \mathcal{O}(r^{-1})$. Thus, $\mathbb{E}[\ln S_{N:r}] \simeq a_N - b_N\ln(r)$ which is the result stated in equation (2.39).

Taking together the results for the enrichment-rank $s_r(r)$ dependence of Gumbel-type enrichments and the scaling coefficients a_N and b_N in the particular cases of lognormal and generalized Pareto enrichments, we find that

$$\mathbb{E}[\ln S_{N:r}] \simeq \mu + \sqrt{2\ln N}\sigma - \frac{\sigma}{\sqrt{2\ln N}}\ln(r) \qquad \text{(lognormal)}, \tag{2.58}$$

$$\mathbb{E}[\ln S_{N:r}] \simeq \kappa\ln(N) + \ln\left(\frac{\tau}{\kappa}\right) - \kappa\ln(r) \qquad \text{(generalized Pareto)}. \tag{2.59}$$

Thus, both lead to apparent power-law behaviour of $s_r(r)$ with exponent $-b_N$, *i.e.* affine behaviour in log-log scale with slope $-b_N$. The relevance of this finding is supported by a simple numerical

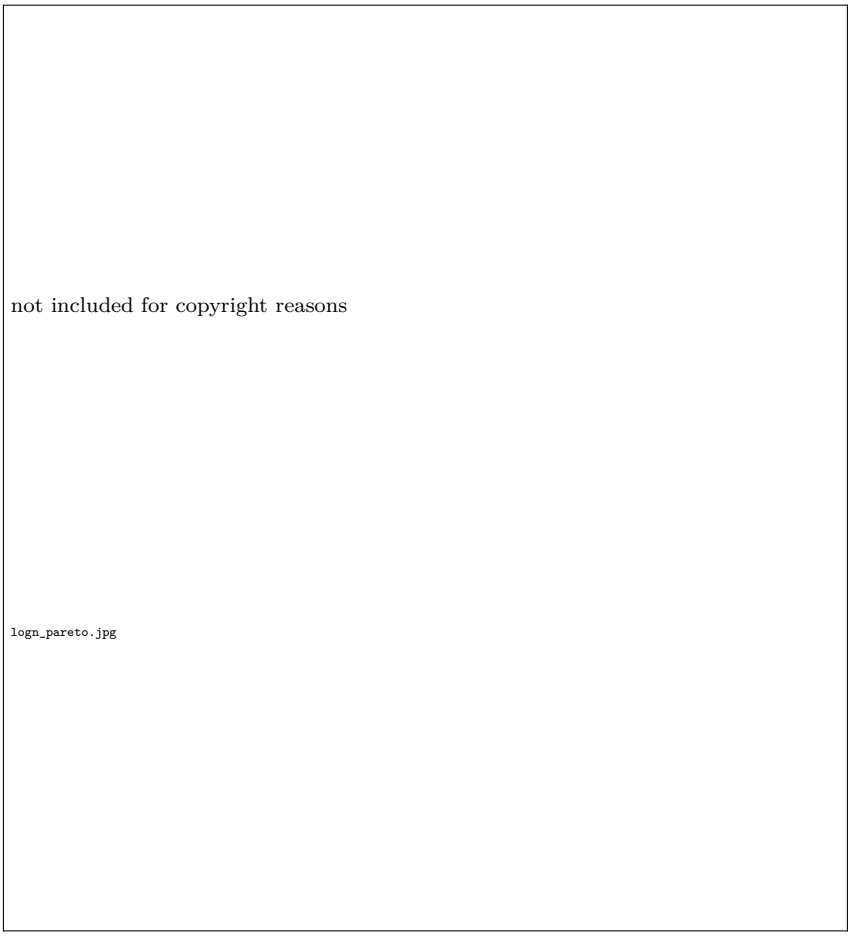


Fig. 2.3: Examples of power-law mimicry. Taken from [111]. Samples s_r from a Pareto and from a lognormal distribution ordered in decreasing order, $s_1 \geq s_2 \geq \ldots$, and plotted against their rank r. The data are truncated to show only the top 10^3 among $10^3 - 10^6$ points (see legends). At high truncation, both datasets become indistinguishable and display apparent power-law behaviour.

experiment shown in figure 2.3. This result suggests that empirical data of extremes displaying such power-law behaviour may be consistent with both the lognormal and generalized Pareto assumption with $\kappa > 0$ (and other Gumbel-type distributions), and thus that a true power-law distribution of the data is not a necessary conclusion from such an observation. Moreover, the slopes in equations (2.58) and (2.59) become identical for lognormal and generalized Pareto samples if the parameters are chosen such that

$$\kappa = \frac{\sigma}{\sqrt{2 \ln N}}. \tag{2.60}$$

In particular, truly lognormal data with some value of σ may appear to be consistent with an apparent $\kappa_N = \frac{\sigma}{\sqrt{2 \ln N}}$ as a finite-size effect for $N < \infty$. This apparent κ_N decreases to 0 very slowly, emphasizing its potential relevance for real data; observing the mathematically exact value of $\kappa = 0$ would require astronomical data size N. Note, however, that equation (2.39)

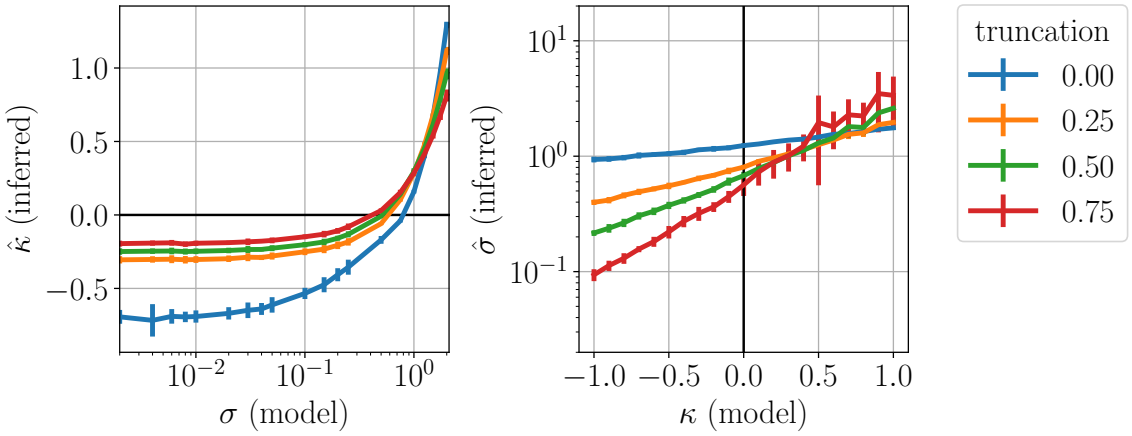


Fig. 2.4: Interdependence of inferred generalized Pareto distribution parameter κ and lognormal distribution parameter σ for finite dataset size N. **Left** $\hat{\kappa}$ as a function of σ for $N = 10^4$ obtained from fitting a generalized Pareto distribution with parameters κ and τ by MLE to the largest among N iid lognormal numbers with $\mu = 0$ and various values of σ and several truncation levels (see legend, *e.g.* a truncation level of 0.75 means largest 25 % among N values were kept). Solid curve and error bars represent respectively mean and standard deviation over 25 independent realizations of the numerical experiment. **Right** Reverse situation: $\hat{\sigma}$ as a function of κ for $N = 500$ obtained from fitting a lognormal distribution with parameters σ and μ by MLE to the largest among N iid numbers from a generalized Pareto distribution with $\tau = 0.115$, $s^* = 0.001$ and various values of κ and several truncation levels.

itself requires that N be large enough. In a numerical experiment, we fitted by MLE generalized Pareto distributions to truncated lognormal data and *vice versa*. Figure 2.4 shows the relationship between the inferred κ_N and σ obtained from such numerical simulations when fixing $N = 10^4$ and $\mu = 0$. As predicted by equation (2.60), truncated lognormal data may be fitted best by a shape parameter $\kappa > 0$. However, we also observe that for a given value of N, it breaks down when σ is below some threshold value $\sigma^\star$. Numerically, we find $\sigma^\star \simeq 0.5$. In such cases, the data may even appear to arise from a bounded distribution with $\kappa_N < 0$, which is not captured by the above prediction from order statistics that assumes the $N \to \infty$ limit.

2.3 Information theory of selection: a definition of specificity

In this section, we will study an information-theoretic interpretation of the parameter σ in the lognormal model for the distribution of enrichments $P(s)$, as well as its implications. Using a definition of specificity inspired from information theory (subsection 2.3.1), we will show that σ quantifies the specificity of interactions between two classes of objects, as well as the information content of selection based on these interactions (subsection 2.3.2). We will show that, as a consequence, σ also constrains the emergence of sequence motifs under selection and the "area under

the curve" of sequence logos drawn from selection data (subsection 2.3.3). Sequence logos are a commonly used representation of sequence specificities in the literature [116, 117]. These results generalize the idea to define specificity as the amount of information encoded in interactions [118].

2.3.1 Relative entropies for model testing

The problem of quantifying specificity arises when two classes of objects or properties A and T, with respectively $|A|$ and $|T|$ variants on both sides, may interact with one another and associate. Nonspecificity corresponds to the case of equally likely association between members of both classes and random pair formation, while specificity of interactions is at play in case of preferential association of few A variants with few T variants. Mathematically, one can start defining specificity in terms of the probability $P_1(A_i, T_j)$ that a randomly picked pair from a population of associated pairs consists of A_i linked to T_j. The problem is already acute in the binary case of a single A and a single T: How does a property differ between the two objects to identify each of them as either A or T, or, from a statistical point of view, how to discriminate the two objects based on their properties? Beyond, this simple case, the problem generalizes to the one-to-many case and the many-to-many case. Our goal is to identify characterizations that, in such cases, involve fewer numbers of parameters than the number of possible pairwise comparisons. Below, we are going to motivate the use of the Kullback-Leibler divergence $D(P^1\|P^0)$ as a measure of specificity with respect to a null hypothesis represented by $P_0(A_i, T_j)$, *e.g.* the expectation from associations at random. In practice, $P_1(A_i, T_j)$ is unknown and empirical observation of N instances $(A^1, T^1), \ldots, (A^N, T^N)$ provides an empirical measurement of the probability $P_1(A_i, T_j)$ that A_i is associated with T_j.

In this definition of the problem, the question of specificity of interactions between A_i and T_j can be translated into a hypothesis testing problem: Given a set of N interactions $(A^1, T^1), \ldots, (A^N, T^N)$ sampled from the true distribution of interactions P_1, can the null hypothesis of unspecific (random) interactions defined by the distribution P_0 be excluded given the data? The theoretical framework to answer this class of problems comes from asymptotic inference: The central quantity here is the relative entropy $D(P^1\|P^0)$, also known as the Kullback-Leibler divergence [119], which quantifies how samples drawn from a true (typically unknown) distribution P_1 are consistent with a hypothesized distribution P_0 and which is defined by

$$D(P_1\|P_0) = \sum_{i,j=1,1}^{|A|,|T|} P_1(A_i, T_j) \ln \frac{P_1(A_i, T_j)}{P_0(A_i, T_j)} = \left\langle \ln \frac{P_1}{P_0} \right\rangle_{P_1}, \tag{2.61}$$

where $\langle\cdot\rangle_{P_1}$ denotes the average taken with respect to P_1. It measures the distance of two distributions, though not in the mathematical sense: It satisfies $D(P_1\|P_0) \geq 0$, $D(P_1\|P_0) = 0$ if $P_0 = P_1$ in the sense of distributions, but $D(P_1\|P_0) \neq D(P_0\|P_1)$. (The symmetrized quantity $d(P_0, P_1) = D(P_1\|P_0) + D(P_0\|P_1)$ may be used as a true distance between P_0 and P_1.) The positivity of $D(P_1\|P_0)$ can be confirmed by applying Jenssen's inequality $\langle \ln f(x)\rangle \geq \ln\langle f(x)\rangle$ to the function

$f(x) = P_0(x)/P_1(x)$ and the average with respect to P_1, $\langle\cdot\rangle_{P_1}$: $\ln\langle P_0/P_1\rangle_{P_1} = \ln\langle 1\rangle_{P_0} = \ln 1 = 0$ and $\langle\ln(P_0/P_1)\rangle_{P_1} = -D(P_1\|P_0)$, and thus $D(P_1\|P_0) \geq 0$. The relative entropy appears naturally by considering the posterior probability $P(P_0|y)$ given the data $y = \left(y^1, y^2, \ldots, y^N\right)$, $y^i = (A^i, T^j)$ in the limit of large sample size N,

$$\begin{aligned} P(P_0|y) &= \frac{P(y|P_0)}{Z(y)} = \frac{1}{Z(y)}\prod_{i=1}^{N} P_0(y^i) = \frac{1}{Z(y)}\exp\left[\sum_{i=1}^{N}\ln P_0(y^i)\right] \\ &\simeq \frac{1}{Z(y)}\exp[N\langle\ln P_0\rangle_{P_1}] = \frac{1}{Z(y)}\exp\left[N\left(\left\langle\ln\frac{P_0}{P_1}\right\rangle_{P_1} + \langle\ln P_1\rangle_{P_1}\right)\right] \\ &= \frac{1}{Z(y)}\exp[-N\left(D(P_1\|P_0) + S[P_1]\right)], \end{aligned} \tag{2.62}$$

where Bayes' theorem with a uniform prior on different models is used in the first line, and the CLT is used to go to the second line. $S[P_1] = -\langle\ln(P_1(s))\rangle_{P_1}$ denotes the standard entropy of P_1. Thus, the probability of the data y under a model P_0 different from P_1 decreases exponentially with sample size N, and the sample size N required to discriminate and exclude P_0 in favor of P_1 scales as $N \sim D(P_1\|P_0)^{-1}$ [119]. The emergence of a simple scalar measure is thus rooted in the CLT and therefore relevant to large N.

In the context of specificity, when P_0 defines a null model of interactions, $D(P_1\|P_0)$ measures to what extent the true interactions divert from unspecificity, *i.e.* how specific they are. In practice, only a finite number N of observations can be made and specificity cannot be sensed as long as $N \lesssim D(P_1\|P_0)^{-1}$, and a conclusion will be made in favor of unspecificity. Thus, if P_0 and P_1 are very different, *i.e.* P_1 encodes for highly specific interactions, few observations will be needed to conclude on the specificity of interactions and the presence of selection.

With the choice of $P_0(A_i, T_j) = P_1(A_i)P_1(T_j)$ where $P_1(A_i) = \sum_{j=1}^{|T|} P_1(A_i, T_j)$ and $P_1(T_j) = \sum_{i=1}^{|A|} P_1(A_i, T_j)$ are the marginal distributions of A_i and T_j under P_1, $D(P^1\|P^0)$ corresponds to the mutual information

$$I(A_i; T_j) = \sum_{i,j=1,1}^{|A|,|T|} P_1(A_i, T_j)\ln\left(\frac{P_1(A_i, T_j)}{P_1(A_i)P_1(T_j)}\right) \tag{2.63}$$

between the random variables A_i and T_j [119]. This choice of P_0, however, generally does not reflect the expectation from random associations as we shall see in subsection 2.3.2. The relevant measure of specificity is therefore not captured by a mutual information in general, but by the more general relative entropy $D(P^1\|P^0)$. A previous study proposed the mutual information as a measure of specificity [118]. It is justified, however, only within the special model considered in [118] where, because of the overall symmetry of the interactions between the $|A| = M$ locks A and $|T| = M$ keys T, $P_1(A_i) \simeq P_1(T_j) \simeq 1/M$, and therefore $P_0(A_i, T_j) = 1/M^2 \simeq P_1(A)P_1(T)$.

2.3.2 Information theory of binding interactions

Now consider again that the two sets of objects A and T are ligands (*e.g.* antibodies) and targets, respectively, and that the mechanism behind association of these is equilibrium binding characterized by K_{AT} (see section 2.1). In the case of a single target ($|T| = 1$) and many ligands, the probability $P_1(A_i, T)$ reads

$$P_1(A_i, T) = \frac{K_{A_iT}[A_i]_{\text{tot}}}{\sum_{k=1}^{|A|} K_{A_kT}[A_k]_{\text{tot}}}, \tag{2.64}$$

This expression is easily generalized to the case of many targets,

$$P_1(A_i, T_j) = \frac{K_{A_iT_j}[A_i]_{\text{tot}}[T_j]_{\text{tot}}}{\sum_{k,m=1}^{|A|,|T|} K_{A_kT_m}[A_k]_{\text{tot}}[T_m]_{\text{tot}}}. \tag{2.65}$$

Here, $[\cdot]_{\text{tot}}$ denotes total concentrations. This reflects the fact that associations are seen with high probability if the binding partners are strongly binding or simply are present with high frequency in the soup. If all concentrations are equal, this simplifies to

$$P_1(A_i, T) = \frac{K_{A_iT}}{\sum_{k=1}^{|A|} K_{A_kT}}, \qquad P_1(A_i, T_j) = \frac{K_{A_iT_j}}{\sum_{k,m=1,1}^{|A|,|T|} K_{A_kT_m}}. \tag{2.66}$$

The unspecific case with association probability $P_0(A_i, T_j)$ corresponds to the case of identical equilibrium constants $K_{A_iT_j} \equiv K$, $\forall i = 1, \ldots, |A|; j = 1, \ldots, |T|$, thus

$$P_0(A_i, T) = \frac{[A_i]_{\text{tot}}[T]_{\text{tot}}}{\sum_{k=1}^{|A|}[A_k]_{\text{tot}}[T]_{\text{tot}}}, \qquad P_0(A_i, T_j) = \frac{[A_i]_{\text{tot}}[T_j]_{\text{tot}}}{\sum_{k,m=1,1}^{|A|,|T|}[A_k]_{\text{tot}}[T_m]_{\text{tot}}}, \tag{2.67}$$

and in the case of equal concentrations simply $P_0(A_i, T_j) = (|A||T|)^{-1}$. Note that in the case of unequal concentrations, P_0 does not factorize *a priori*, $P_0(A_i, T_j) \neq P_0(A_i)P_0(T_j)$ with $P_0(A_i) = \sum_{j=1}^{|T|} P_0(A_i, T_j)$ and $P_0(T_j) = \sum_{i=1}^{|A|} P_0(A_i, T_j)$.

Let us compute the specificity, defined as $D(P_1\|P_0)$ in equation (2.61), of the binding system defined characterized by the probabilities P_1 and P_0 in equations (2.65) and (2.67). To this aim, we denote by $\langle\cdot\rangle_0$ and $\langle\cdot\rangle_1$ the average taken with respect to P_0 and P_1. Thus, for some observable $\mathcal{O}_{A_iT_j}$ that depends on the ligand-target combination,

$$\langle\mathcal{O}_{A_iT_j}\rangle_0 = \frac{1}{Z_0}\sum_{i,j=1,1}^{|A||T|}[A_i][T_j]\mathcal{O}_{A_iT_j}, \qquad \langle\mathcal{O}_{A_iT_j}\rangle_1 = \frac{1}{Z_1}\sum_{i,j=1,1}^{|A||T|}K_{A_iT_j}[A_i][T_j]\mathcal{O}_{A_iT_j}, \tag{2.68}$$

with the normalization constants

$$Z_0 = \sum_{i,j=1,1}^{|A||T|}[A_i][T_j], \qquad Z_1 = \sum_{i,j=1,1}^{|A||T|}K_{A_iT_j}[A_i][T_j]. \tag{2.69}$$

With these notations, it follows that $\langle \mathcal{O}_{A_iT_j} \rangle_1 = Z_0/Z_1 \langle K_{A_iT_j} \mathcal{O}_{A_iT_j} \rangle_0$, and in particular $\langle K_{A_iT_j} \rangle_0 = Z_1/Z_0 \langle 1 \rangle_1 = Z_1/Z_0$. Besides, we have $P_0(A_i, T_j) = Z_0^{-1}[A_i][T_j]$ and $P_1(A_i, T_j) = Z_1^{-1} K_{A_i,T_j}[A_i][T_j]$ and thus $\frac{P_1(A_i,T_j)}{P_0(A_i,T_j)} = \frac{Z_0}{Z_1} K_{A_iT_j} = \frac{K_{A_iT_j}}{\langle K_{A_iT_j} \rangle_0}$. Finally, we thus find

$$\begin{aligned} D(P_1 \| P_0) &= \left\langle \ln\left(\frac{P_1(A_iT_j)}{P_0(A_iT_j)} \right) \right\rangle_1 = \left\langle \ln\left(\frac{K_{A_iT_j}}{\langle K_{A_iT_j} \rangle_0} \right) \right\rangle_1 \\ &= \frac{Z_0}{Z_1} \left\langle K_{A_iT_j} \ln\left(\frac{K_{A_iT_j}}{\langle K_{A_iT_j} \rangle_0} \right) \right\rangle_0 = \left\langle \frac{K_{A_iT_j}}{\langle K_{A_iT_j} \rangle_0} \ln\left(\frac{K_{A_iT_j}}{\langle K_{A_iT_j} \rangle_0} \right) \right\rangle_0 . \end{aligned} \tag{2.70}$$

In the next subsection, we consider the case of a single target and many ligands with a lognormal distribution of binding affinities. It should be noted that the result for $D(P_1 \| P_0)$ in equation (2.70) is invariant under rescale of the binding affinities, $K_{AT} \leftarrow \lambda K_{AT}$, showing that the overall scale of binding strength is irrelevant in the problem of specificity; only differences in affinity among ligands and targets matter. In a similar approach, the relative entropy $D(P_1 \| P_0)$ has been related to the change in Malthusian fitness [120].

2.3.3 The case of lognormal interactions

The binding affinity K_{AT} is related to the binding free energy ΔG_{AT} by $K_{AT} = \mathrm{e}^{\beta \Delta G}$ (see section 2.1). At least for a single target T, we have argued in section 2.2 that the ΔG_{AT} should follow a Gaussian distribution, and the K_{AT} thus a lognormal distribution. In the intermediate regime for target concentrations (see section 2.1), this directly translates into a lognormal distribution $P(s)$ for the selection coefficients/enrichments s_{AT}. If further assuming that the ligands and targets are equi-concentrated, *i.e.* no initial bias in frequencies $[A_i]_{\mathrm{tot}}$, the average over P_0 becomes identical to the average over $P(s)$, $\langle \cdot \rangle_0 = \langle \cdot \rangle_{P(s)} \equiv \langle \cdot \rangle$. Equation (2.70) then becomes

$$D(P_1 \| P_0) = \left\langle \frac{K}{\langle K \rangle} \ln \frac{K}{\langle K \rangle} \right\rangle = \left\langle \frac{s}{\langle s \rangle} \ln \frac{s}{\langle s \rangle} \right\rangle = \frac{\langle s \ln s \rangle}{\langle s \rangle} - \ln \langle s \rangle . \tag{2.71}$$

These averages are most conveniently computed as Gaussian averages involving ΔG. Thus, assuming that $\Delta G \overset{d}{=} \mathcal{N}(-\mu, \sigma^2)$, we obtain

$$\langle s \rangle = \langle \mathrm{e}^{-\beta \Delta G} \rangle_{\mathcal{N}} = \exp\left(\beta \mu + \frac{\beta^2 \sigma^2}{2} \right), \tag{2.72}$$

$$\langle s \ln s \rangle = -\beta \langle \Delta G \mathrm{e}^{-\beta \Delta G} \rangle_{\mathcal{N}} = \beta \frac{\partial}{\partial \beta} \langle \mathrm{e}^{-\beta \Delta G} \rangle_{\mathcal{N}} = \exp\left(\beta \mu + \frac{\beta^2 \sigma^2}{2} \right) \left(\beta \mu + \beta^2 \sigma^2 \right), \tag{2.73}$$

where the third equality in equation (2.73) uses equation (2.72). Note that we have here defined μ and σ^2 as the mean and variance for the Gaussian distribution for ΔG, as opposed to section 2.3 where they were defined as the parameters of the lognormal distribution of s. The consequence is that the (inverse) temperature β enters into these results. This is meaningful because the binding affinities K_{AT} also depend on temperature, $K_{AT} = \mathrm{e}^{\beta \Delta G_{AT}}$, and increasing β increases differences in binding affinity across ligands. However, we will show in section that β may also be

re-interpreted as the number of selection rounds: Repeating the selection t times at temperature β has the same effect as performing a single selection step at temperature $t\beta$. In fact, enrichments are potentiated over several selection rounds, *i.e.* after $t \geq 0$ selection rounds, the total enrichment is s^t. Indeed, repeating the above computation with generic t yields

$$\langle s^t \rangle = \langle \mathrm{e}^{-\beta t \Delta G} \rangle_{\mathcal{N}} = \exp\left(\beta t \mu + \frac{\beta^2 t^2 \sigma^2}{2} \right), \tag{2.74}$$

showing exchangability of t and β. For the moment, we may simply set $\beta = 1$. Inserting equations (2.72) and (2.73) into equation (2.71), we thus obtain

$$D(P_1 \| P_0) = \frac{\sigma^2}{2}, \tag{2.75}$$

for lognormal binding affinities, irrespectively of μ. Again, this reflects the fact that specificity quantifies only relative differences in binding free energies between different ligands. The parameter σ thus encodes for the specificity in a library of ligands with lognormally distributed binding affinities. As expected, the value of the specificity $D(P_1 \| P_0)$ vanishes in the perfectly unspecific case of random ligand-target assemblies, which is realized within the lognormal family by $\sigma = 0$. where all $K_{A_i T_j}$ are equal, *i.e.* $P(K) = \delta(K + \mu)$ and $P(s) = \delta(s - \mathrm{e}^{\beta\mu})$.

To be precise, the result in equation (2.75) quantifies the specificity of the target T in light of a diversity of ligands A_i, $i = 1, \ldots, |A|$. It does not relate to the specificity of a given ligand A_i which would have to be defined with respect to a diversity of targets T. However, the results obtained here should be easily generalizable if the argument for Gaussian binding energies also applies to targets (for instance, if targets are defined on a sequence sequence alike the ligands).

Note that the result in equation (2.75) also remains unchanged upon iteration of the selection step: After $(t-1)$ selection rounds, the initially equal frequencies of ligands are changed to $[A_i T] = s_{A_i T}^{t-1} / \sum_{k=1}^{|A|} s_{A_k T}^{t-1}$. Taking into account this bias in frequencies, the average over P_0 thus becomes $\langle \mathcal{O} \rangle_0 = \langle s^{t-1} \mathcal{O} \rangle / \langle s^{t-1} \rangle$ or, equivalently,

$$\langle \mathcal{O} \rangle_0 = \frac{\int_0^\infty \mathrm{d}s P(s) s^{t-1} \mathcal{O}(s)}{\int_0^\infty \mathrm{d}s P(s) s^{t-1}}. \tag{2.76}$$

For simplicity, set again $\beta = 1$. Then, using equation (2.74), we obtain

$$\langle s \rangle_0 = \frac{\int_0^\infty \mathrm{d}s P(s) s^t}{\int_0^\infty \mathrm{d}s P(s) s^{t-1}} = \exp\left(\mu + \left(t - \frac{1}{2} \right) \sigma^2 \right) \tag{2.77}$$

$$\langle s \ln s \rangle_0 = \exp\left(\mu + \left(t - \frac{1}{2} \right) \sigma^2 \right) \left(\mu + t\sigma^2 \right), \tag{2.78}$$

and thus again $D(P_1 \| P_0) = \left\langle \frac{s}{\langle s \rangle} \ln \frac{s}{\langle s \rangle} \right\rangle = \frac{\sigma^2}{2}$, independently of t.

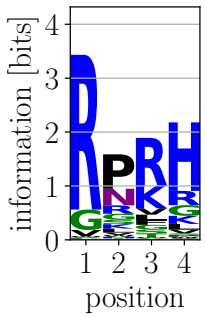


Fig. 2.5: Example of a sequence logo. Starting from a PWM $f_{1,i}(a)$, it shows on each sequence position i a stack of height equal to the relative entropy (or "information") $D(f_{1,i}\|f_{0,i})$ given in equation (2.79), in which each letter a occupies a height of $f_{1,i}(a)D(f_{1,i}\|f_{0,i})$.

2.3.4 Implications for sequence motifs and logos

The equivalence between σ and specificity of interactions also has implications for selection at the sequence level: We are going to show here that σ constrains the "area under the curve" of sequence motifs (or logos) [116, 117] which quantify the information content of underlying interactions and the "goodness" of certain sequences over others. Let us denote by L the length of a sequence and by q the size of the alphabet ($q = 20$ for amino acids). Such sequence motifs take as input position-specific letter frequencies, or position weight matrices, $f_{1,i}(a)$ and $f_{0,i}(a)$ (the probability of observing letter a at position i under a null model $f_{0,i}(a)$) and assign to each position $i = 1, \ldots, L$ a stack of height

$$D(f_{1,i}\|f_{0,i}) = \sum_{a=1}^{q} f_{1,i}(a) \ln\left(\frac{f_{1,i}(a)}{f_{0,i}(a)}\right), \tag{2.79}$$

which is identical to the relative entropy between $f_{1,i}(a)$ and $f_{0,i}(a)$ on position i. Thus, a sequence motif appears the larger the more $f_{1,i}(a)$ deviates from the null model $f_{0,i}(a)$, *i.e.* the more the frequencies $f_{1,i}(a)$ deviate from the expectation at random. In particular, it vanishes when $f_{1,i}(a) = f_{0,i}(a)$, $\forall a = 1, \ldots, q$. In practice, the $f_{1,i}(a)$ are estimated empirically from sequence data and $f_{0,i}(a)$ is typically the uniform distribution over the alphabet, $f_{0,i}(a) = \frac{1}{q}$, reflecting irrelevance of amino acids in the unspecific case. The area under the curve

$$D(f_1\|f_0) = \sum_{i=1}^{L} D(f_{1,i}\|f_{0,i}) \tag{2.80}$$

then sums the contributions from all sites to the overall information content of interactions across the sequence. Thus, sequence logo representations inplicitly assume independence of the L sites. An example of a sequence logo is shown in figure 2.5.

Consider a fixed target T and let $x = (x_1, x_2, \ldots, x_L)$ denote the sequences of ligands A that

have a length of L positions, each x_i taking on the alphabet of size q. Our goal is to define relevant sequence motifs in the context of selection. This means that we define frequencies $f_{1,i}(a)$ and $f_{0,i}(a)$ from the probabilities to observe x associated with T, $P_1(x) \propto K(x)[x]$ and $P_0(x) \propto [x]$, rather than from actual frequencies in a population $[x]$ and $\frac{1}{q^L}$. In order to construct PWMs from $P_1(x)$ and $P_0(x)$, we need to factorize them into position-wise contributions $f_{1,i}(a)$ and $f_{0,i}(a)$ such that

$$P_1(x) = \prod_{i=1}^{L} f_{1,i}(x_i), \qquad P_0(x) = \prod_{i=1}^{L} f_{0,i}(x_i). \tag{2.81}$$

The $f_{\cdot,i}(a)$ are chosen normalized on each position, *i.e.* $\sum_{a=1}^{q} f_{\cdot,i}(a) = 1$, $\forall i = 1, \ldots, L$. The inverse transformation of equation (2.81) reads

$$f_{\cdot,i}(a) = \sum_{x} P_{\cdot}(x)\delta(x_i, a). \tag{2.82}$$

In our case, these factorizations amount to assume additive models for both binding free energies $\Delta G(x)$ of the form

$$K(x) = \mathrm{e}^{\beta G(x)} = \exp\left(\sum_{i=1}^{L} h_{1,i}(x_i)\right), \tag{2.83}$$

and for concentrations $[x]$,

$$[x] = \exp\left(\sum_{i=1}^{L} h_{0,i}(x_i)\right). \tag{2.84}$$

The uniform distribution over sequences, $P_0(x) = \frac{1}{q^L}$, is realized by $f_{0,i}(a) = \frac{1}{q}$ and any constant $h_{0,i}(a) = c$. Given the expressions for $P_1(x)$ and $P_0(x)$ in equations (2.65) and (2.67), the amino acid frequencies $f_{\cdot,i}(a)$ and local fields $h_{\cdot,i}(a)$ are related by

$$\begin{aligned} \ln f_{1,i}(a) &= h_i(a) - \frac{1}{L}\ln\left(\sum_{y} K(y)[y]_{\mathrm{tot}}\right) \\ &= h_{1,i}(a) + h_{0,i}(a) - \frac{1}{L}\sum_{j=1}^{L}\ln\left(\sum_{b_j=1}^{q} \mathrm{e}^{h_{j,i}(b_j)+h_{0,j}(b_j)}\right), \end{aligned} \tag{2.85}$$

$$\ln f_{0,i}(a) = h_{0,i}(a) - \frac{1}{L}\ln\left(\sum_{y}[y]_{\mathrm{tot}}\right) = h_{0,i}(a) - \frac{1}{L}\sum_{j=1}^{L}\ln\left(\sum_{b_j=1}^{q}\mathrm{e}^{h_{0,j}(b_j)}\right) \tag{2.86}$$

where the sums in the first equalities run over all q^L possible sequences. We can compute the specificity $D(P_1\|P_0)$ of interactions under such additive models starting from equation (2.61),

$$D(P_1\|P_0) = \sum_{x} P_1(x)\ln\left(\frac{P_1(x)}{P_0(x)}\right) = \sum_{x_1=1}^{q}\cdots\sum_{x_L=1}^{q}\prod_{i=1}^{L} f_{1,i}(x_i)\ln\left(\prod_{j=1}^{L}\frac{f_{1,j}(x_j)}{f_{0,j}(x_j)}\right)$$

$$
\begin{aligned}
&= \sum_{x_1=1}^{q} \cdots \sum_{x_L=1}^{q} \sum_{j=1}^{L} f_{1,j}(x_j) \ln\left(\frac{f_{1,j}(x_j)}{f_{0,j}(x_j)}\right) \prod_{\substack{i=1 \\ i\neq j}}^{L} f_{1,i}(x_i) \\
&= \sum_{j=1}^{L} \left(\sum_{x_j=1}^{q} f_{1,j}(x_j) \ln\left(\frac{f_{1,j}(x_j)}{f_{0,j}(x_j)}\right) \right) \underbrace{\left(\sum_{x_1=1}^{q} \cdots \sum_{x_{j-1}=1}^{q} \sum_{x_{j+1}=1}^{q} \cdots \sum_{x_L=1}^{q} \prod_{\substack{i=1 \\ i\neq j}}^{L} f_{1,i}(x_i) \right)}_{=1} \\
&= \sum_{j=1}^{L} \sum_{x_j=1}^{q} f_{1,j}(x_j) \ln\left(\frac{f_{1,j}(x_j)}{f_{0,j}(x_j)}\right). \qquad (2.87)
\end{aligned}
$$

By definition in equation (2.80), this corresponds to the size of a sequence logo for a PWM defined by $f_{1,i}(a)$, taking $f_{0,i}(a)$ as null model. Under the assumption of a uniform null model P_0, $f_{0,i}(a) = \frac{1}{q}$, we thus have

$$
\begin{aligned}
D(P_1\|P_0) &= \sum_{j=1}^{L} \sum_{x_j=1}^{q} f_{1,j}(x_j) \ln\left(\frac{f_{1,j}(x_j)}{f_{0,j}(x_j)}\right) = \sum_{j=1}^{L} \sum_{x_j=1}^{q} f_{1,j}(x_j) \ln(f_{1,j}(x_j)q) \\
&= L\ln(q) + \sum_{j=1}^{L} \sum_{x_j=1}^{q} f_{1,j}(x_j) \ln(f_{1,j}(x_j)) = S_{\max} - S[f_1], \qquad (2.88)
\end{aligned}
$$

where $S[f_1] = -\sum_{j=1}^{L} \sum_{x_j=1}^{q} f_{1,j}(x_j) \ln(f_{1,j}(x_j))$ denotes the standard entropy of $f_{1,i}(a)$ and $S_{\max} = \max_{f_1} S[f_1] = L\ln(q)$ is its maximum value that occurs if f_1 is itself the uniform distribution. Upon comparing equations (2.75) and (2.88), we find the self-consistent relation

$$
\frac{\sigma^2}{2} = S_{\max} - S[f_1], \qquad (2.89)
$$

which provides a direct link between σ and the area under the curve of a sequence logo. Intuitively, high specificity and thus high σ indeed means low entropy of P_1 and large sequence logos. This relation is also consistent in the unspecific case where $\sigma = 0$ and the entropy is maximal, $S[f_1] = S_{\max}$. At the other extreme, however, this result necessarily breaks down at some point because σ can be in principle arbitrarily large, whereas the right-hand side has an upper bound of $S_{\max} = L\ln(q)$. This is because in the computation of $D(P_1\|P_0)$ in subsection 2.3.3, we took the full distribution $P(s)$ which does not account for the finiteness of sequence space and neglects that $P(s)$ is thus not sampled above a certain threshold.

In practice, the frequencies may be estimated in line with equation (2.82) from a list of empirical (and unnormalized) enrichments $s(x)$ for the sequences x by

$$
f_{1,i}(a) = \frac{\sum_x s(x)\delta(x_i, a)}{\sum_{b=1}^{q} \sum_x s(x)\delta(x_i, b)} = \frac{\sum_x s(x)\delta(x_i, a)}{\sum_x s(x)}. \qquad (2.90)
$$

The result in equation (2.89) will, however, be difficult to observe in real data where enrichments are available only for small subset among all q^L sequences. Leaving out unobserved sequences in the computation of $S[f_1]$ resumes to assuming their enrichment be zero (although it is simply

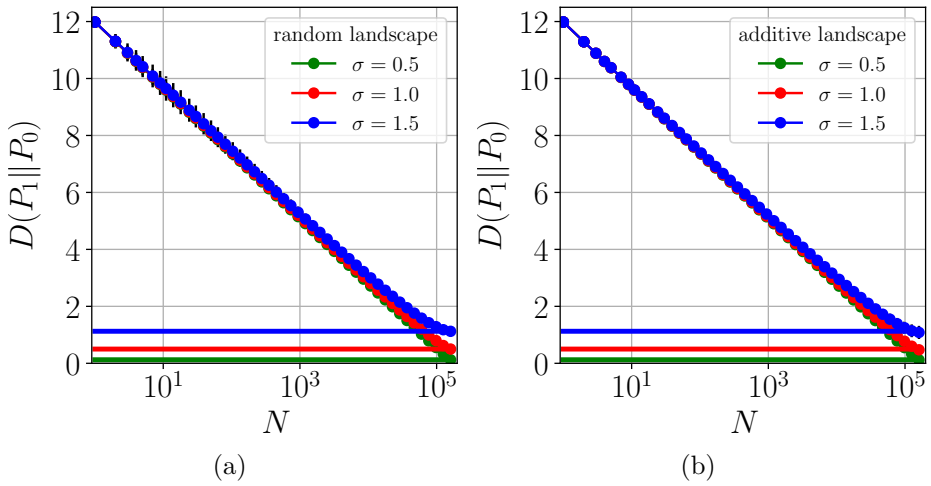


Fig. 2.6: Over-estimation of PWM entropy $D(P_1\|P_0)$ for incomplete sets of enrichments. **(a)** Apparent $D(P_1\|P_0)$ computed from the largest $N \leq q^L$ among $q^L = 1.6 \cdot 10^5$ iid enrichments as a function of N. The q^L enrichments are drawn from lognormal distributions with $\mu = 0$ and several values of σ (see legend). The horizontal lines represent the predicted values of $D(P_1\|P_0) = \sigma^2/2$ which is indeed reached for the complete set of enrichments $N = q^L$. **(b)** Same as (a) but enrichments are drawn from an additive model, $s = \exp\left(-\sum_{i=1}^{L} h_i(x_i)\right)$ with iid $h_i(a) = \mathcal{N}(0, \sigma/\sqrt{L})$.

unknown) and thus systematically underestimating the entropy of f_1 (*i.e.* overestimating the area under the curve of a sequence logo). This effect is shown in figure 2.6 on simulated data: If enrichments are available for all q^L sequences, the expected value of $\sigma^2/2$ is achieved, both for $D_{p=\infty}(P_1\|P_0)$ in the case of a random landscape and for $D_{p=1}(P_1\|P_0)$ in the case of an additive landscape. If less than q^L enrichments are available, the sequence logo overestimates the specificity, the more so as data availability decreases. This finite-size effect may be corrected for, at least partly, upon inferring an additive model from selection data and using the model enrichments rather than empirical enrichments (see chapter 4). Finally, this result relies on the uniform distribution over the ligand sequences ($f_{0,i}(a) = \frac{1}{q}$, $\forall i = 1, \ldots, L; a = 1, \ldots, q$) and the additivity of ligand positions to the binding free energy, and deviations from these assumptions in true data may lead to deviations from equation (2.89).

2.4 Dynamics of selection: evolutionary time as a temperature

As yet, we have justified that, under certain conditions, binding energies and enrichments under selection for binding should follow respectively a Gaussian and a lognormal distribution with parameters μ and σ in a library of ligands with random binding positions. Within the framework

of information theory, we associated the parameter σ to the level of binding specificity in such libraries. In addition to these results, we are going to show in this section that σ also determines the fate of such libraries in competition with other libraries and thus their selection potential. We will derive the time-dependence of frequencies respectively of sequences within a library (subsection 2.4.1) and of libraries within a mix of libraries (subsection 2.4.2). The exact solution that is obtained in the special case of lognormal enrichments will be discussed (subsection 2.4.3). Combining all these results, σ is identified as a key quantity in selection with various equivalent interpretations and implications: It quantifies binding specificities and the dispersal of enrichments, but is at the same time also a measure of selection potential.

2.4.1 Recursion for sequence frequencies and Fisher's equation

We here compute the evolution of the frequencies of variants in a population that repeatedly undergoes selection and amplification. By time, we here mean the discrete time defined by the number of selection rounds t. At time t, such a population of variants x is determined by $\{N_t(x)\}_{x\in\ell}$, the list of numbers of copies of each variant x inside a library ℓ. If we assume the thermodynamic limit in which $N = \sum_{x\in\ell} N_t(x) \to \infty$, we can introduce a kind of continuum limit in which the use of frequencies $f_t(x) = N_t(x)/N$ instead of $N_t(x)$ is meaningful. In what follows, this assumption implies that no variant x ever disappears. The goal is to determine $f_t(x)$ at any selection round t, given the initial condition at round $t = 0$, $f_0(x)$, and the enrichments $s(x)$. The continuous-time equivalent of this discrete-time problem has already been discussed in chapter 1.

In the limit of large N_t, the selection is deterministic and the dynamics is governed by the recursion

$$f_{t+1}(x) = \lambda_t s(x) f_t(x) \tag{2.91}$$

stating that frequencies $f_t(x)$ are updated proportionally to enrichments $s(x)$ as a consequence of selection. The prefactor λ_t assures proper normalization of the new frequencies $f_{t+1}(x)$, $\sum_{x\in\ell} f_t(x) = 1$, $\forall t \geq 0$, and can be interpreted as an amplification factor required to recover the initial population size after selection. It is thus given by $\lambda_t = \left(\sum_{x\in\ell} s(x) f_t(x)\right)^{-1}$ and the recursion for $f_t(x)$ becomes

$$f_{t+1}(x) = \frac{s(x) f_t(x)}{\sum_{y\in\ell} s(y) f_t(y)}. \tag{2.92}$$

The solution is obtained by simply reinserting the recursion into itself T times, giving

$$f_t(x) = \frac{s(x)^T f_{t-T}(x)}{\sum_{y\in\ell} s(y)^T f_{t-T}(y)}, \tag{2.93}$$

for $t \geq T$, and, for $T = t$,

$$f_t(x) = \frac{s(x)^t f_0(x)}{\sum_{y\in\ell} s(y)^t f_0(y)}, \tag{2.94}$$

thus giving $f_t(x)$ only as a function of the enrichments $s(x)$ and the initial frequencies $f_0(x)$.

In the Boltzmann limit where the enrichments $s(x)$ of ligands directly relate to the binding free energies $\Delta G(x)$, $s(x) = \mathrm{e}^{-\beta\Delta G(x)}$, equation (2.94) can be rewritten as

$$f_t(x,\beta) = \frac{1}{Z(t,\beta)} f_0(x)\mathrm{e}^{-t\beta\Delta G(x)}, \qquad Z(t,\beta) = \sum_{x\in\ell} f_0(x)\mathrm{e}^{-t\beta\Delta G(x)}. \tag{2.95}$$

The frequencies $f_t(x)$ can also be interpreted as the probabilities that a randomly picked ligand in the population at selection round t has sequence identity x and binding energy $\Delta G(x)$. Thus, equation (2.95) allows for an analogy with a system of discrete energy levels $\Delta G(x)$, $x \in \ell$, in the canonical ensemble and in contact with a thermal reservoir at (inverse) temperature $t\beta$: The probability to find such a system in state x with energy $\Delta G(x)$ is given by $f(x,\beta) = \frac{1}{Z(\beta)}\mathrm{e}^{-\beta\Delta G(x)}$ with the canonical partition function $Z(\beta) = \sum_{x\in\ell} \mathrm{e}^{-\beta\Delta G(x)}$ [69]. The analogy is completed upon interpreting $Z(t,\beta)$ in equation (2.95) as the partition function of the selection problem, with additional parameter t and an *a priori* biasing field given by $f_0(x)$. Interestingly, the selection round t can be absorbed into the temperature, $Z(t,\beta) = Z(t\beta)$, showing that t consecutive rounds of selection at physical temperature β have the same effect as a single round of selection at a temperature of $t\beta$. In particular, when the physical temperature β is kept constant throughout the selection process, the selection round t plays itself the role of a temperature: Repeatedly selecting from a population with a given diversity (*i.e.* no mutations) is equivalent to cooling down the system and to eventually approach zero temperature, *i.e.* $\beta \to \infty$, as t goes to infinity. Zero temperature here means a complete takeover of the population by the variant with the highest enrichment $\max_{x\in\ell} s(x)$ among all variants x as $t \to \infty$ or, equivalently, the probability of a randomly picked individual having the lowest energy $\min_{x\in\ell} \Delta G(x)$ is one. This is equivalent to the general observation in statistical mechanics that, at $T = 0$, a particle resides in the ground state $\min_{x\in\ell} \Delta G(x)$ with probability one.

Moreover, we can define thermodynamic quantities in complete analogy to the equilibrium statistical mechanics of other systems, such as the ensemble average and variance in energy. For the following discussion, we keep β constant and express all $\Delta G(x)$ in units of β, formally setting $\beta = 1$. In our case, this translates into the ensemble and population-averaged binding energy

$$\langle \Delta G\rangle_{\mathrm{pop}}(t) = \frac{1}{Z(t)} \sum_{x\in\ell} f_0(x)\mathrm{e}^{-t\Delta G(x)}\Delta G(x) = -\frac{\partial}{\partial t}\ln Z(t), \tag{2.96}$$

where $\langle\cdot\rangle_{\mathrm{pop}}$ denotes population- and ensemble average. In addition, we find for the second

derivative of $\ln Z(t)$ that it equals the variance over the population of the binding energy,

$$\frac{\partial^2}{\partial t^2}\ln Z(t) = \frac{\sum_{x\in\ell} f_0(x)\mathrm{e}^{-t\Delta G(x)}\Delta G(x)^2}{\sum_{x\in\ell} f_0(x)\mathrm{e}^{-t\Delta G(x)}} - \left(\frac{\sum_{x\in\ell} f_0(x)\mathrm{e}^{-t\Delta G(x)}\Delta G(x)}{\sum_{x\in\ell} f_0(x)\mathrm{e}^{-t\Delta G(x)}\Delta G(x)^2}\right)^2$$
$$= \langle\Delta G^2\rangle_{\mathrm{pop}}(t) - \langle\Delta G\rangle_{\mathrm{pop}}(t)^2 = \mathrm{var}(\Delta G)_{\mathrm{pop}}(t). \tag{2.97}$$

Combining the results in equations (2.96) and (2.97), we can relate the population average and variance of ΔG in a similar way as in Fisher's equation [121],

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle\Delta G\rangle_{\mathrm{pop}} = -\mathrm{var}(\Delta G)_{\mathrm{pop}}. \tag{2.98}$$

Both equations are formally identical if we define the (Malthusian) fitness of ligands under selection for binding by (minus) their binding free energy ΔG. The additional sign appears as a consequence of enrichments and binding energies being inversely related to each other, $s(x) = \mathrm{e}^{-\beta\Delta G(x)}$.

2.4.2 Renormalization to library frequencies

We are now seeking to generalize the result of subsection 2.4.1 to the case of several ligand libraries ℓ in competition with one another during selection. Each library is itself composed of many ligands, but is characterized by a different model for the distribution of binding energies ΔG. We define by $f_t(\ell, x)$ the frequency of sequence x in the context of library ℓ in the total population, which are normalized such that $\sum_\ell \sum_{x\in\ell} f_t(\ell, x) = 1, \forall t \geq 0$. We also define the coarse-grained frequencies $f_t(\ell)$ that define the frequency of library ℓ in the library mix, with normalization $\sum_\ell f_t(\ell) = 1, \forall t \geq 0$. They are obtained from the $f_t(\ell, x)$ by summing over all of a library's constituent sequences x,

$$f_t(\ell) = \sum_{x\in\ell} f_t(\ell, x). \tag{2.99}$$

In order to compute $f_t(\ell)$ as a function of enrichments and initial frequencies, we need the result of equation (2.94) which generalizes to

$$f_t(\ell, x) = \frac{s(\ell, x)^t f_0(\ell, x)}{\sum_{\ell'}\sum_{y\in\ell'} s(\ell', y)^t f_0(\ell', y)}. \tag{2.100}$$

Inserting into the definition of $f_t(\ell)$ in equation (2.99) yields

$$f_t(\ell) = \frac{\sum_{x\in\ell} s(\ell, x)^t f_0(\ell, x)}{\sum_{\ell'}\sum_{y\in\ell'} s(\ell', y)^t f_0(\ell', y)}. \tag{2.101}$$

This result can be simplified under the additional assumptions that i) sequences are uniformly represented in the initial population across and inside libraries, *i.e.* $f_0(\ell, x) = \frac{1}{|\ell|q^L}$ for all (ℓ, x), ii) enrichments s follow a library-specific distribution $P_\ell(s)$. Then, by denoting the average with

respect to $P_\ell(s)$ as $\langle\cdot\rangle_\ell$, we obtain

$$f_t(\ell) = \frac{\langle s^t\rangle_\ell}{\sum_{\ell'}\langle s^t\rangle_{\ell'}}. \tag{2.102}$$

With all these assumptions, $f_t(\ell)$ thus involves the t-th moments of the enrichment distributions $P_\ell(s)$.

2.4.3 Exact solution for lognormal interactions and implications

We now want to study the prediction for library frequencies $f_t(\ell)$ under the additional assumption of lognormal distributions $P_\ell(s)$ of the enrichments,

$$P_\ell(s) = \frac{1}{\sqrt{2\pi}\sigma_\ell s}\exp\left(-\frac{(\ln(s)-\mu_\ell)^2}{2\sigma_\ell^2}\right) \tag{2.103}$$

with the library-dependent parameters μ_ℓ and σ_ℓ. In this case, the t-th moments are explicitly known (see section 2.3.3) and read

$$\langle s^t\rangle_\ell = \int_0^\infty \mathrm{d}s P_\ell(s) s^t = \exp\left(t\mu_\ell + \frac{t^2\sigma_\ell^2}{2}\right). \tag{2.104}$$

Using equation (2.102), the library frequencies are then given by

$$f_t(\ell) = \left(\sum_{\ell'} \mathrm{e}^{t(\mu_{\ell'}-\mu_\ell)+\frac{t^2}{2}\left(\sigma_{\ell'}^2-\sigma_\ell^2\right)}\right)^{-1} = \left(1+\sum_{\ell'\neq\ell} \mathrm{e}^{t(\mu_{\ell'}-\mu_\ell)+\frac{t^2}{2}\left(\sigma_{\ell'}^2-\sigma_\ell^2\right)}\right)^{-1}. \tag{2.105}$$

For the following discussion, we rewrite equation (2.105) as $f_t(\ell) = \left(1+\sum_{\ell'\neq\ell}\mathrm{e}^{t g_{\ell,\ell'}(t)}\right)^{-1}$, with $g_{\ell,\ell'}(t) = \mu_{\ell'} - \mu_\ell + \frac{t}{2}\left(\sigma_{\ell'}^2 - \sigma_\ell^2\right)$. We shall outline two main consequences of this result: i) Let us have a look at small t: If we analytically continue $f_t(\ell)$ to real $t \geq 0$ and take the derivative with respect to t, we obtain

$$\frac{\partial f_t(\ell)}{\partial t} = -f_t(\ell)^2 \sum_{\ell'} g_{\ell,\ell'}(t)\,\mathrm{e}^{t g_{\ell,\ell'}(t)}. \tag{2.106}$$

It follows that

$$\left.\frac{\partial f_t(\ell)}{\partial t}\right|_{t=0} = \frac{1}{|\ell|}\left(\mu_\ell - \langle\mu\rangle\right) \tag{2.107}$$

and thus $\mathrm{sgn}\left(\left.\frac{\partial f_t(\ell)}{\partial t}\right|_{t=0}\right) = \mathrm{sgn}(\mu_\ell - \langle\mu\rangle)$, where $\langle\mu\rangle = \frac{1}{|\ell|}\sum_\ell \mu_\ell$ denotes the average μ across libraries. Hence, the parameter μ defines the behaviour of the library mix in the early stages of selection: The frequency $f_t(\ell)$ of a library ℓ increases if its μ_ℓ exceeds the average $\langle\mu\rangle$ across competing libraries and decreases otherwise. ii) Without loss of generality, if we let ℓ be the library

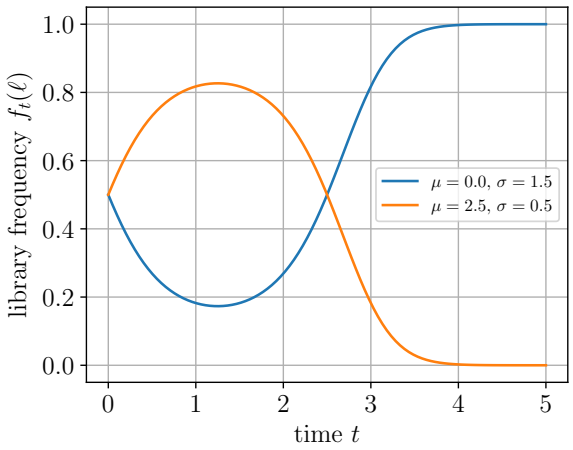


Fig. 2.7: Time dependence of library frequencies in an initially uniform mix of two libraries with lognormal enrichments, given by equation (2.105) (analytical continuation to non-integer t). The parameters of the two libraries are respectively $\mu = 0$, $\sigma = 1.5$ and $\mu = 2.5$, $\sigma = 0.5$. The library with larger σ wins the competition albeit a smaller μ and initial decrease in frequency.

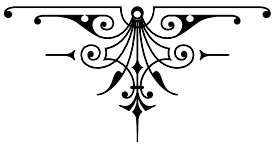
with highest variance σ_ℓ^2 among all libraries in the mix, *i.e.* $\sigma_\ell > \sigma_{\ell'}$, $\forall \ell' \neq \ell$, then $g_{\ell,\ell'}(t) \to -\infty$ as $t \to \infty$ because of $\sigma_{\ell'}^2 < \sigma_\ell^2$, $\forall \ell' \neq \ell$ and, hence, $f_t(\ell) \to 1$ as $t \to \infty$, meaning ℓ will eventually take over the mix. Remarkably, this result holds completely independently of the means μ_ℓ. Thus, the fate of the competition at large $t \to \infty$ is controlled by σ only and the library which maximizes σ dominates, even if its enrichments are in mean low compared to other competing libraries.

In summary, in a library mix with lognormally distributed enrichments, the parameters μ and σ take effect in different limits of the selection process: The short-term response to selection ($t \to 0$) is dominated by the mean binding energies μ, whereas the long-term response ($t \to \infty$) is encoded in the variance σ^2. In particular, this leads to a non-trivial prediction for a library with low μ but high σ (compared to competing libraries): It is expected to first decrease in frequency in the first round(s) of selection, followed by a catch-up and takeover of the mix in the long-term limit. The reason for this behaviour is that such libraries consist mostly of the worst binders across all libraries, as well as of few variants that on the contrary are the best binders across all libraries. A majority of variants of this library will thus be selected out and removed during the first round(s), before few variants of this library impose in the later rounds. An example of such a case is shown in figure 2.7.

However, it should be noted that these conclusions reside on a number of assumptions that have been made throughout this section: First, they rely on the assumption of initially uniform frequencies, which is difficult to achieve in practice. Second, they require the limits of infinite population size ($N \to \infty$), followed by the limit of infinite diversity (so that the use of the full distribution $P_\ell(s)$ is justified). In practice, stochastic finite-size effects may be important: In finite populations, stochasticity of selection may lead to elimination by chance of low-frequency variants, irrespectively of their binding capacity. In addition, for finite diversity, the distributions

$P_\ell(s)$ are not sampled beyond some threshold enrichment. Thus, it may happen by chance that a library features the global maximum in enrichment albeit not maximizing σ.

A finite-size ($N < \infty$) analysis of this selection dynamics will no longer be deterministic and can possibly be carried out analytically in terms of survival probabilities and take-over probabilities within the framework of branching processes.



Chapter 3

Choice and design of antibody libraries and binding targets, strategies for *in vitro* selection

We will here expose the experimental basis and strategies followed for our study and comparison of selection potentials. Unlike former approaches to evolvability (see chapter 1), these are rooted in the standard repertoire of molecular biology, using notably phage display and biopanning for quantitative selection experiments, as well as deep sequencing of antibody libraries. As a reminder, our goal is to test antibodies with previous maturation against HIV for their susceptibility to new selective pressures unrelated to HIV. The question is about the factors that confer large enrichment values to an antibody, where "large" can be defined in different ways. This is a first step towards the more general question about the impact of the presence or absence of past maturation on the initiation of a new maturation trajectory. To this purpose, we here propose to study synthetic V_H libraries built on the basis of three natural antibodies, two of which are matured *in vivo* to different degrees as part of the immune response against HIV in human, starting from the third antibody which is a naïve one (section 3.1). The V_H libraries were built on the basis of these three template antibodies by introducing variation at the level of the highly variable antigen binding site (section 3.1). These libraries are expressed by phage display, a standard technique that allows to physically link together phenotype and genotype of variants (section 3.2), and selected for their binding capacity to different target molecules: We choose different protein and DNA molecules, each one unrelated to HIV, as binding targets for these V_H libraries (section 3.3). The protocol for selection by phage display biopanning is outlined and we propose to perform selections within libraries, between libraries, and from subsampled libraries. These different selection schemes reflect the idea that enrichments may be considered "large" relative to sequences within the same library, *i.e.* with same scaffold and different CDR3, or to sequences from other libraries, *i.e.* with different scaffolds (section 3.4). As a reminder, we denote by "scaffold" the germline-encoded

part of naïve V_H which comprises notably the FWRs, as well as CDR1 and 2 (see also B). The computation of frequencies and enrichments of variants based on high-throughput sequencing of the libraries will be described (sections 3.4 and 3.5). In summary, our approach described here can be regarded as an *in vitro* equivalent to the initiation of affinity maturation *in vivo*: Randomized CDR3 mimick random junctions between V and D segments in primary repertoire formation (see section 1.3.1). Antibody display on phage is the correspondence to their display on the B cell surface [122]. Selection for randomly chosen targets translates into selection for binding to antigens newly encountered by the organism. The cloning of V_H libraries and setup of phage display in the group's lab were performed as part of another PhD project [1]. Detailed experimental protocols of our selection experiments can be found in appendix A.

3.1 Combinatorial libraries of synthetic, human-based V_H segments with different maturation levels and randomized CDR3

In this section, we will briefly motivate the use of V_H fragments instead of complete antibodies for our selection experiments (subsection 3.1.1). Then, we explain our choice of three V_H with different degrees of maturation against HIV based on [123] as templates for antibody library construction (subsection 3.1.2). These three template antibodies correspond to a fully matured broadly neutralizing antibody, a naïve germline(-reversed) antibody, and an antibody with intermediate maturation which we here, and henceforth, refer to as respectively BnAb, Germ, and Lmtd. These are evolutionarily related, as Germ is the common ancestor of both Lmtd and BnAb, though Lmtd and BnAb are located on different maturation trajectories. Finally, we explain the design and construction of synthetic, recombinant V_H libraries which were performed as part of a former PhD project within the group [1] (subsection 3.1.3): The library design starts from a template V_H and introduces diversity by complete sequence randomization at the level of four consecutive CDR3 residues. Our libraries thus represent libraries of random antibody binding pockets operating in the context of fixed antibody scaffolds with different maturation levels that consist of the framework regions FWR1, 2, 3, 4, as well as CDR1, 2.

3.1.1 V_H domains as model system: advantages and shortcomings

Within this project, we will be working with a strongly reduced version of the complete antibody structure discussed in chapter 1, as is done in most antibody-based therapeutic and diagnostic contexts. These reductions are necessary to accommodate large antibodies to feasible sizes for display techniques. The two largest among the commonly used reduced formats are the antigen-binding fragment (Fab) followed by the single-chain variable fragment (scFv). The former comprises vari-

able and constant domains of the heavy and light chains (V_H, V_L, C_H, and C_L), while the latter consists only of a single V_H domain fused covalently to a single V_L domain via a synthetic, highly flexible glycine-serine linker [124]. The scFv imitates its natural counterpart, the V_H-V_L heterodimer called variable fragment (Fv), which was long considered as the minimal building block necessary for binding. It comprises all parts of the antibody directly involved in binding: As addressed in section 1, the regions most crucial for binding specificities are the complementarity-determining regions located on the V_H and V_L, and in particular the CDR3 of V_H which bears extraordinary sequence diversity [21] and where sequence diversity is most likely to yield functional variants [22]. It has been shown that scFv are indeed sufficient to retain affinities and specificities of the underlying full antibody and to yield antibodies with specificities to nearly any therapeutically relevant antigen [125, 126, 122]. Here, we will use an even more reduced format that consists of the V_H domain only, see figure 3.1. Such standalone V_H domains were also shown to be functional [127, 128] albeit being oftentimes associated with reduced solubility (*i.e.* tendency to form aggregates) and stability due to exposed hydrophobic residues otherwise buried inside the V_H-V_L interface and missing contacts with V_L residues [129, 130]. In addition, V_H-like domains unpaired to any V_L and with extended, stabilizing CDR3, referred to as V_HH, also appear naturally in camels [131, 132]. In our case, we will be able to conclude the viability of our V_H domains from the enrichment of CDR3 sequences upon selection that confer to the V_H domains the capacity to bind (see chapter 4).

The use of V_H domains is also beneficial in combination with high-throughput sequencing methods where the sequence length is oftentimes a key limitation. Sequencing reads can not be arbitrarily long and the upper bound is a strong constraint. As an example that is acute to our project, consider Illumina MiSeq sequencing which provides paired-end sequencing reads of up to 350 bp in length (including barcodes), while the length of a single V_H or V_L domain is $\simeq 100$ aa, *i.e.* $\simeq 300$ bp. Illumina MiSeq sequencing thus allows in principle for a paired-end sequencing readout of a complete V_H or V_L domain, whereas a fusion of both including the glycine-serine linker would be too long to be read at once. More generally, limitations in sequencing length are relevant for the sequencing of entire genes with typical lengths of 1 kb. Within our work on V_H domains where only the CDR3 sequence is highly variable, sequence length does not represent a strong constraint and we can design our sequencing amplicons in way that a single sequencing read provides all the information about the V_H sequence identity. More involved sequencing strategies are required when the sequence of interest exceeds the readable length: In such cases, the total sequence can be devided into several shorter, overlapping reading windows that can be sequenced each one separately. To recover the complete sequence, reads from different windows then need to be associated *a posteriori* using the sequence overlaps. But this is a non-trivial task when sequences are similar, *i.e.* only a few mutations away from each other. Another strategy relies on barcoding: Long and similar sequences (mutants of a gene) are tagged by shorter random sequences (barcodes). If one ensures that each barcode represents a single mutant sequence, the problem is reduced to sequencing the short barcodes. However, the non-trivial step consists in establishing the mapping from barcode sequence to mutant sequence.

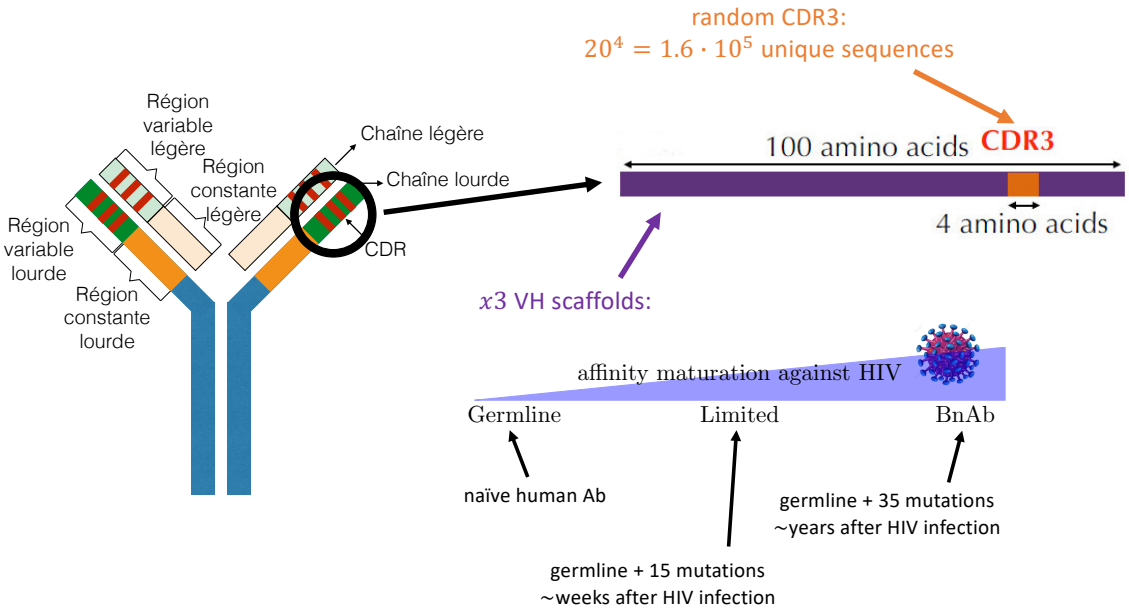


Fig. 3.1: **Left** Only the variable part of the heavy chain (V_H) of the full antibody is kept for our antibody library design, leaving aside notably its adjacent light chain. **Upper right** In our library design, the V_H is subdivided into a scaffold region, comprising FWR1, 2, 3, 4 and CDR1, 2, that is kept constant within a library. Diversity is introduced by complete sequence randomization at the level of 4 consecutive residues in CDR3, a region directly involved in antibody binding. A library thus consists of $20^4 = 1.6 \cdot 10^5$ different unique variants that display various CDR3 in the context of a fixed scaffold. **Lower right** Three such libraries are considered and systematically compared regarding their selective potentials within this project: These are built on three scaffolds with varying degrees of maturation based on the V genes of three natural human antibodies evolved to different degrees as part of the immune response against HIV: a germline(-reversed) antibody with no previous maturation (Germline), a matured antibody with limited neutralization spectrum of HIV strains (Limited), and an extensively matured antibody with broad neutralization spectrum against HIV (bnAb). Germline is a common ancestor of Limited and BnAb which are respectively 15 and 35 somatic mutations away from Germline. Figure assembled using a drawing of the full antibody and the library design from [133].

3.1.2 Choice of template V segments with different maturation levels for library construction

Comparing selection potentials at different degrees of maturation by our means requires the knowledge of antibody sequences that differ in their degrees of maturation, but are otherwise identical. Upon affinity maturation, an antibody gradually accumulates mutations in its framework regions, the so-called somatic mutations. Therefore, the maturation level of an antibody is encoded in the set of its somatic mutations; a trajectory of somatic mutations defines the maturation trajectory of an antibody. In practice, determination of the immune repertoire by deep sequencing is possible and used for the purpose of vaccine development [134, 24], the understanding of physiochemical properties of antibodies [135], statistical properties and constraints of repertoires [108], inference

of repertoire dynamics [136], and prediction of antibody-antigen interactions [89]. However, the reconstruction of single maturation trajectories from *e.g.* the sequencing of *in vivo* immune repertoires is a difficult task and only few such trajectories are available in the literature, notably [137]. This is because differences between any two given antibody sequences are not necessarily the result of somatic mutations mapping one to the other: Simultaneous challenges to an organism's immune system lead to antibodies on unrelated maturation trajectories carrying unrelated somatic mutations. In addition, the combinatorial use of different V, D, and J genes for initial immune repertoire formation also leads to differences between antibody sequences. A similar tracking of evolution trajectories by sequencing has already been performed in the context of asexual cell populations [138].

For the purpose of this project, we use information from [123] to define a trio of antibodies that are evolutionarily related through affinity maturation against HIV and we will compare with regard to their selection potentials (see also figure 3.1). This paper studies the effect of the reversal of somatic mutations from antibodies matured *in vivo* against HIV on their HIV neutralization breadth and associates matured antibody sequences with their reversed-to-germline counterparts obtained by reverting somatic mutations. Both antibodies with limited degrees of maturation and limited neutralization spectra against HIV strains, as well as deeply matured antibodies with broad neutralization of HIV strains, referred to as *broadly neutralizing antibodies* (bnAb), are considered. These bnAb appear only in few HIV patients typically after years of infection and immune response [139, 140]. They are generalists capable of neutralizing various strains of the HI virus by targeting hidden, conserved parts of HIV spike proteins that are otherwise highly variable [140]. In the literature, bnAbs are considered strong candidates for HIV immune therapy [141, 142, 143], in spite of their unlikely appearance and induction through vaccination which is explained by difficult access of hidden conserved epitopes [143] and two strong attractors that divert bnAb generation to respectively specialists with low neutralization breadth and frustrates that result from too different and contradictory selective pressures from different HIV strains [25, 6]. From the sequences provided in the paper, we inferred trios consisting of i) a bnAb, ii) an antibody with limited maturation, and iii) a germline-reversed antibody that is shared between i) and ii). The limited and bnAb were associated based on the similarity of their respective germline-reversed: The germline-reversed of any limited was associated with the closest among the germline-reversed of the bnAbs based on pairwise alignments. However, minor sequence differences between the associated germline-reversed could still arise from different V gene usage, in which case they would not represent a single germline. Excluding all pairs with non-identical V gene (as found by comparison with the set of 51 human V genes), only three pairs of bnAb and limited with common germline-reversed sequences are found, namely (`10-1074`, `2-491`), (`4E10`, `17b`), and (`PGT128`, `6-187`). For (`10-1074`, `2-491`) and (`4E10`, `17b`), the paper concluded that the matured antibodies do not significantly lose their HIV neutralization breadth upon reversal-to-germline, suggesting that the difference between mature and germline may not be relevant in these cases. Such a behaviour is also observed elsewhere [144]. The choice is thus made in favor of the `PGT128` [145] and `6-187` antibodies, and their common germline origin, `IGHV4-39`. Note that the germline-reversed of Lmtd and BnAb differ at position 83 in figure 3.2 (`Q` in germline-reversed

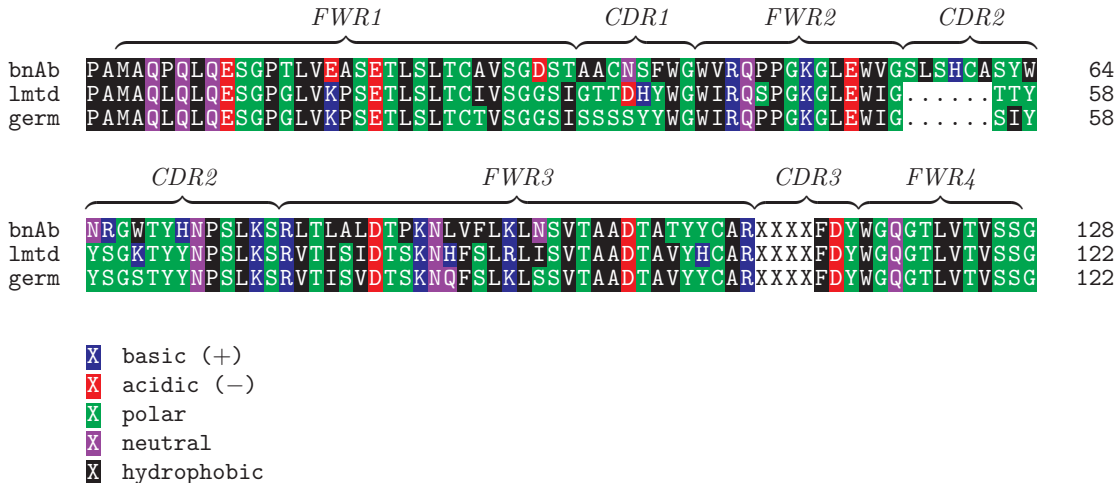


Fig. 3.2: V_H scaffold sequences. Alignment (using ClustalW 2.1) of the three human HIV-specific antibody V_H scaffolds of interest. The four randomized positions in the CDR3 are indicated by `XXXX`.

of BnAb, `H` in germline-reversed of Lmtd; `IGHV4-39` contains `Q` at this position), but which is not explained by different V genes. Also note that the CDR1 and CDR2 of the germline `IGHV4-39` have been used to define Germ.

In what follows, we will refer to these three sequences by the less cryptic names “BnAb”, “Limited” (or “Lmtd”), and “Germline” (or “Germ”). An alignment between their amino acid sequences is shown in figure 3.2. In summary, both Lmtd and BnAb were derived from Germ by affinity maturation *in vivo*, but were isolated from different patients [145, 146], meaning they are located on distinct maturation trajectories [123]. Lmtd features an intermediate level of maturation against HIV (15 somatic mutations in the V region compared to Germ), while BnAb is the result of extensive maturation against HIV (37 somatic mutations in the V region compared to Germ including insertions). They have respectively limited and broad spectra of neutralization of HIV strains [123, 147]. Lmtd and BnAb are 41 mutations apart from one another. A notable difference is the elongated CDR2 of BnAb that has six additional residues compared to Lmtd and Germ. Note that there are changes at a few residues that were necessary to accomodate for restriction sites as explained in subsection 3.1.3.

It should be emphasized that this restricted choice of three antibodies only provide a first step towards a more systematic study of selection potentials and maturation trajectories in the future: First, while tracking affinity maturation *in vivo* is hard, maturation trajectories under controlled conditions may be obtained by mimicking the process of affinity maturation *in vitro*. The benefits are well-defined and controlled selective pressures, as well as the possibility to sequence the simulated immune repertoire at every discrete “maturation step” that consists of one random mutagenesis along the antibody sequence followed by selection and thus to record the appearance of somatic mutations over evolutionary time. In particular, this will also allow to study more

than three evolutionary time points on a maturation trajectory as is done here. However, the study of many maturation trajectories in parallel requires parallelization and automation of the experimental protocol and efforts in this direction are being made within the group [148].

3.1.3 Library design and construction: mimicking the initial step of maturation

Starting from the V genes of the three antibodies Germ, Lmtd, and BnAb identified in subsection 3.1.2, Boyer *et al.* [1] constructed V_H libraries by introducing diversity at four consecutive among the seven CDR3 positions and grafting a common FWR4 sequence downstream, see figures 3.1 and 3.2. This step is akin to initial repertoire formation in the immune system in which antibody sequences with random CDR3 are created by imprecise joining of a V_H segment (encoding for FWR1, 2, 3 and CDR 1, 2) and a D_HJ_H seqment (encoding in particular for FWR4): random nucleotides are added at this junction to yield the CDR3 of the newly created antibody, which is thus highly variable across realizations of the VDJ recombination. The CDR3 encodes for much of the chemical diversity of the primary immune repertoire [126] and can be sufficient to define the binding specificity of antibodies [21].

The choice of a small number of random positions for our libraries, encoding for only $20^4 = 1.6 \cdot 10^5$ unique variants per library, is mainly grounded in limitations in sequencing depth: In order to compute meaningful frequencies and enrichments of antibody sequences from sequencing data, many sequences must be counted sufficiently many times (see chapter 3.4.3. At constant sequencing budget, the average number of counts per sequence is increased by decreasing the number of unique variants. For our libraries, we can conclude *a posteriori* that such a small diversity does contain binding sequences, again from the presence of selection for binding at the sequence level (see chapter 4). This is in contrast to recombinant antibody libraries used in other, *e.g.* therapeutical, contexts which oftentimes have (much) larger (typically $> 10^8$) sequence diversity across all CDRs of V_H and V_L and thus chemical and conformational diversity [126]. Popular examples are the Tomlinson and Griffin antibody repertoires [150, 151]. However, statistical analysis in combination with high-throughput sequencing is *hitherto* rarely performed in these contexts; rather, randomly picked sequences in the selection output are tested for their secretion and true binding capacity by ELISA and possibly sequenced.

During phage display (see section 3.2), these V_H libraries will be expressed in fusion with pIII phage surface protein in TG1 cells, a display strain of *E. coli*. To this purpose, the DNA coding for our V_H must be purchased as synthetic genes and cloned into a phagemid (*i.e.* plasmid or circular dsDNA with a phage origin of replication) carrying a phage display vector with all the genetic ingredients for display of the V_H on phage. The phagemid that we use, (a modified version of) pIT2, is presented in section 3.2. TG1 cells transformed with the pIT2-V_H phagemid represent the atarting point for the phage display and selection experiments and will here be referred to as

Fig. 3.3: Classical cloning procedure. Illustrations taken, adapted, and assembled from [149]. The goal is to insert a target gene or DNA sequence into a display vector, here in particular cloning of a V_H sequence into pIT2 display vector or cloning of a target CDR3 sequence into pIT2 with the V_H sequence already present. **Top** A display vector (on a circular plasmid) containing an unrelated sequence in the insert region is used as a template: In order to remove the "old" sequence, the plasmid is digested (cut) at restriction sites flanking the insert region using restriction enzymes. The linearized template containing the display cassette is kept and the "old" insert region sequence is removed by gel purification. The restriction enzymes for cloning of a gene into pIT2 are NcoI and NotI, the ones for cloning of CDR3 into a V_H are BssHII and XhoI. **Bottom** The insert region containing the target sequence flanked by the same restriction sites is purchased as annealed (ds) DNA oligo and disgested and purified in the same way as the display vector. **Right** The linear template and the insert are ligated and the circular display vector containing the target gene or DNA sequence is obtained. The ligation product is transformed into cells which are then cultured in selective growth medium containing the antibiotic for which the display vector contains a resistance gene. The cell strain used for transformation of ligation products is not necessarily the display strain used later for phage display, but can be a cloning strain optimized for transformation efficiency. To check for the correct sequence in the insert region, plasmid DNA is extracted from the cells and is Sanger sequenced. If a cloning strain was used, the plasmid still needs to be transformed into the display strain.

2289 "library cells".

2290 The cloning is schematized in figure 3.3 and the protocol is provided in A.2. They proceeded as
2291 follows: First, in order to obtain gene sequences optimized for TG1 codon usage, the V_H amino acid
2292 sequences were back-translated to nucleotide sequences by taking for each amino acid the codon
2293 that maximizes *E. coli* codon usage. Then, the obtained gene sequences were slightly modified
2294 in order to also accomodate for restriction sites at few relevant positions in the gene. The DNA
2295 can be specifically cut at these restriction sites by digestion with restriction enzymes. The final

```
                FWR1
               ------------------------------------------------------------------
                   |||||| <- NcoI
bnAb           CCGGCCATGGCGCAGCCGCAGCTGCAGGAATCTGGTCCGACCCTGGTTGAAGCGTCTGAAACCCTGTCTC  70
lmtd           CCGGCCATGGCGCAGCTGCAGCTGCAGGAATCTGGTCCGGGTCTGGTTAAACCGTCTGAAACCCTGTCTC  70
germ           CCGGCCATGGCGCAGCTGCAGCTGCAGGAATCTGGTCCGGGTCTGGTTAAACCGTCTGAAACCCTGTCTC  70

                FWR1                        ------------------------ FWR2
               ------------------------------ CDR1                   -----------------
bnAb           TGACCTGCGCGGTTTCTGGTGACTCTACTGCGGCGTGCAACTCTTTCTGGGGTTGGGTTCGTCAGCCGCC 140
lmtd           TGACCTGCATCGTTTCTGGTGGTTCTATCGGTACCACCGACCACTACTGGGGTTGGATCCGTCAGTCTCC 140
germ           TGACCTGCACCGTTTCTGGTGGTTCTATCTCTTCTTCTTCTTACTACTGGGGTTGGATCCGTCAGCCGCC 140

                FWR2                    -------------------------------------------
               ------------------------- CDR2
bnAb           GGGTAAAGGTCTAGAATGGGTTGGTTCTCTGTCTCACTGCGCGTCTTACTGGAACCGTGGTTGGACCTAC 210
lmtd           GGGTAAAGGTCTAGAATGGATCGGT..................ACCACCTACTACTCTGGTAAAACCTAC 192
germ           GGGTAAAGGTCTAGAATGGATCGGT..................TCTATCTACTACTCTGGTTCTACCTAC 192

               ---------------------- FWR3
                CDR2                 ---------------------------------------------------
bnAb           CACAACCCGTCTCTTAAGTCTCGTCTGACCCTGGCGCTGGACACCCCGAAAAACCTGGTTTTCCTGAAAC 280
lmtd           TACAACCCGTCTCTTAAGTCTCGTGTTACCATCTCTATCGACACCTCTAAAAACCACTTCTCTCTGCGTC 262
germ           TACAACCCGTCTCTTAAGTCTCGTGTTACCATCTCTGTTGACACCTCTAAAAACCAGTTCTCTCTGAAAC 262

                FWR3                                          ----------------------
               ------------------------------------------------ CDR3                --
                                        BssHII -> ||||||
bnAb           TGAACTCTGTTACTGCGGCGGACACCGCGACCTACTACTGTGCGCGCTCTTCTTCTTCTTTCGACTACTG 350
lmtd           TGATCTCTGTTACTGCGGCGGACACTGCGGTTTACCACTGTGCGCGCTCTTCTTCTTCTTTCGACTACTG 332
germ           TGTCTTCTGTTACTGCGGCGGACACTGCGGTTTACTACTGTGCGCGCTCTTCTTCTTCTTTCGACTACTG 332

                FWR4                                 -----------------------------------
               ---------------------------------- glycine linker
                              XhoI -> ||||||
bnAb           GGGTCAGGGTACCCTGGTTACCGTCTCGAGCGGTGGAGGCGGTTCAGGCGGAGGTGGCAGCGGCGGTGGC 420
lmtd           GGGTCAGGGTACCCTGGTTACCGTCTCGAGCGGTGGAGGCGGTTCAGGCGGAGGTGGCAGCGGCGGTGGC 402
germ           GGGTCAGGGTACCCTGGTTACCGTCTCGAGCGGTGGAGGCGGTTCAGGCGGAGGTGGCAGCGGCGGTGGC 402

               ------

                                  NotI -> ||||||||
bnAb           GGGTCGACGGACATCCAGATGACCCAGGCGGCCGCA 456
lmtd           GGGTCGACGGACATCCAGATGACCCAGGCGGCCGCA 438
germ           GGGTCGACGGACATCCAGATGACCCAGGCGGCCGCA 438
```

Fig. 3.4: Synthetic genes coding for the V_H sequences shown in figure 3.2. The CDR3 contains the placeholder sequence `TCTTCTTCTTCT`. Restriction sites are indicated.

gene sequences are shown in figure 3.4. In particular, the CDR3 is flanked up- and downstream by respectively the BssHII and XhoI restriction sites, allowing to cut and replace the CDR3 sequence. The whole V_H gene is flanked by the NotI and NcoI restriction sites, allowing for inserting of the gene into a phagemid containing the same restriction sites, such as pIT2. The placement of such restriction sites requires modification of the DNA sequence under the constraint of leaving the amino acid sequence unchanged. Sometimes, however, this is not possible and slight modifications of amino acid sequences are unavoidable. In our case, an `N` had to be changed into a `K` in the CDR2 of Lmtd (position 70 in figure 3.2). The final antibody gene sequences with placeholder CDR3 sequence `TCTTCTTCTTCT` (coding for amino acid sequence `SSSS`) were purchased as synthetic

genes from Genewiz (South Plainfield, NJ, USA) and cloned into the pIT2 phage display vector using the NcoI and NotI restriction enzymes. Then, random CDR3s purchased as degenerate oligonucleotides flanked by restriction sites from Eurogentec (Angers, France) were cloned into the pIT2-V_H vectors to replace the placeholder CDR3 using the restriction sites BssHII and XhoI. The final products are libraries of V_H in the form of library cells, *i.e.* cloned into a phage display vector (pIT2) and transformed into a bacterial display strain (TG1). Sequencing of the libraries reveiled that the distribution of CDR3 sequences is not uniform but contains biases [1], likely due to differences in transformation efficiencies between sequences.

3.2 Phage display: physically linking genotype and phenotype

We briefly explain the concept and benefits of protein display (subsection 3.2.1), which is a widely used technique in molecular biology and in therapeutic and diagnostic contexts. We then focus on phage display where the protein of interest is displayed on a filamentous bacteriophage as display platform. We remind the mechanistics of the particular variant of phage display realized by the (modified) pIT2 phagemid that we use here and that fuses the protein of interest to the pIII phage surface protein: How are displaying phage obtained starting from library cells? (subsection 3.2.2) Finally, the experimental protocol we use for the phage display of our antibody libraries is discussed (subsection 3.2.3). Our variant of the phage display protocol was set up as part of a previous PhD project and was presented in a former publication [1].

3.2.1 The concept and variants of protein display

Protein display in its various flavours allows to physically link together genotype and phenotype of a suitable protein of interest: The expressed protein is to carry its own gene in ways that differ between display techniques. The power of this genotype-phenotype link resides in the possibility to access and read out by sequencing the genetic information of certain phenotypes of interest, which themselves can be enriched by screening and selection from libraries of many random proteins simultaneously at display. Display and selectivity of foreign proteins on the surface of filamentous phage (phage display) was first described in 1985 [152] and has since then found its way into all of molecular biology, antibody-based therapeutics, and diagnostics. Directed evolution, which makes extensive use of these display techniques was worth the Nobel prize for Chemistry in 2018. Many alternative display platforms have been developped, using *e.g.* mRNA (mRNA display) [153], yeast (yeast display), ribosomes (ribosome display) instead of phage [126]. But there are also variants within phage display that differ mainly in the positioning of the protein of interest on the phage particle (the protein of interest may also be fused to pVI, pVII, or pIX surface proteins instead of pIII, see subsection 3.2.2) and in the number of times it is presented [154].

In clinical contexts, the main goal typically consists in identifying (human) antibodies to any given antigen or therapeutically relevant target [125, 126]. The solution to this problem is obtained through the display and enrichment by binding affinity of recombinant antibody libraries containing many random candidate sequences. The output of this procedure, phage (or other platforms) carrying antibodies of the desired binding specificity and their genes, can then be used for downstream analysis and applications, such as gene read-out by sequencing, cloning of the winner sequences, and binding essays (ELISA). However, while this approach solves the general problem in principle, selection protocols tend to be fine-tuned in not very straightforward ways, depending on the target of interest. The advantage of this *in vitro* methods over immunization of *e.g.* mice with these targets are numerous: First, the selective pressure can be defined controlled *in vitro*, while the selective pressures *in vivo* can not precisely known and may interfer with other challenges to the immune system. Second, antibodies identified from *e.g.* mice need to be "humanized" which is a non-trivial task.

To accomplish the goal of identifying functional sequences, it is typically sufficient to pick at random few sequences from the selected library. In combination with high-throughput sequencing of selected libraries, large numbers of binding sequences can be identified [155, 156]. However, such simple analysis largely underestimate the potential of display techniques and sequencing for the study of more fundamental questions around evolution and in particular immunity by statistical modeling and analysis of such sequencing data using the tools presented in chapter 2. These allow to measure relevant quantities and observables on populations under selection (or evolution), such as frequencies and enrichments (and mutation rates). Increasing availability of high-throughput sequencing techniques (see section 3.5) comes with increasing opportunities to sequence at larger scales and with increased depth such libraries and repertoires under selection or directed evolution: For instance, the process of affinity maturation can be mimicked *in vitro* by applying random mutagenesis and selection in controlled conditions to displayed antibody libraries and repertoires such as in [157]. The tracking of such evolution by deep sequencing at various time points would allow to trace the dynamics and maturation trajectories of its consituent sequences. Such an approach represents an *in vitro* equivalent to previous studies on *in vivo* immune repertoires [136, 89, 158, 108]. *In vitro* affinity maturation of antibodies are already being performed [125] but without statistical analysis of the generated *in vitro* repertoires or even without high-throughput sequencing [159].

3.2.2 Phagemid architecture for phage display

In this subsection, we will briefly revisit phage display from the mechanistic viewpoint, starting from helper phage and expression strain cells carrying a phage display vector and going towards secreted phage displaying protein or peptide of interest.

A popular choice for the phage system used for phage display is the M13 filamentous bacte-

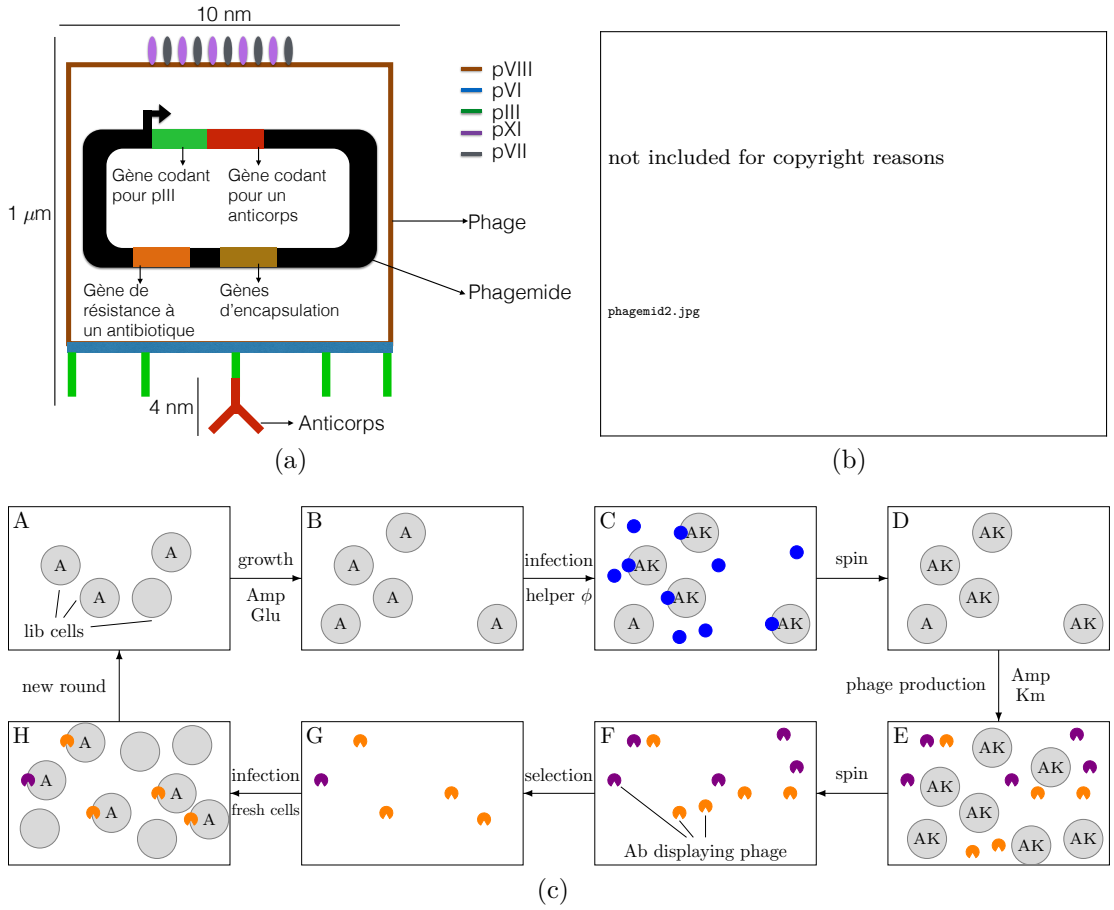


Fig. 3.5: Phage display. **(a)** Schema of an antibody-displaying M13 filamentous phage (not true-to-scale). Illustration taken from [133]. The phage capsid is assembled from pIII, pVI, pVIII, pIX surface proteins. In our phage display, at most one among the 5 copies of the pIII is actually a pIII-antibody fusion protein presenting the antibody in a physically accessible manner. Instead of the M13 phage genome, the displaying phage encapsulates the phagemid which codes for the pIII-antibody fusion. **(b)** The phage display vector or phagemid vector. Illustration taken from [162]. The phagemid backbone bears an *E. coli* origin of replication (*colE1 ori*), an M13 phage origin of replication (M13 *ori*), as well as a resistance gene against the antibiotic ampicillin (Amp^r). The display cassette consists of an inducible *lac* promoter (P*lac*), a signal peptide for the transport of the pIII-antibody fusion to the cell periplasm (SS), as well as the pIII-antibody fusion (M13 gene 3, V_H[-V_L] flanked by the restriction sites NcoI and NotI). These are linked through tag sequences (in our case: PolyHis tag and myc tag) and an amber stop codon `TAG` partially suppressed and coded as glutamine `Q` by display strain cells. **(c)** Schema of the phage display workflow in combination with selection (discussed later). Big gray circles, display strain cells. Small blue circles, helper phage. Small orange (violet) circles, displaying phage with a highly (lowly) selective antibody. The letter A (K) indicates the presence of the phagemid (phage genome) that comes with resistance against ampicillin (kanamycin) antibiotic. Library cells containing the phagemid (A.) are grown in selective growth medium containing ampicillin and glucose (B.) and infected with helper phage that inject the phage genome into the cells (C.). After removal by centrifugation of excess helper phage (D.), cell growth and phage production in selective growth medium containing both ampicillin and kanamycin (E.). The supernatant containing the displaying phage is kept and the cells removed by centrifugation and filtering (F.). A selection step that enriches good over bad antibody variants can be performed (G.) and the remaining displaying phage used to infect (an excess of) fresh display strain cells, thus injecting them the phagemids of the selected phage (H.) With these new library cells, another iteration of library phage display and selection can be started (A.).

riophage whose capsid is depicted in figure 3.5(a). Its genome contains nine genes coding for 11 phage proteins, labelled pI through pXI, 6 of which are involved in phage replication while the remaining 5, namely pIII, pVI, pVII, pVIII and pIX, are the constituent proteins of the phage capsid. The genome is encapsulated as ssDNA in a phage capsid that is assembled from $\simeq$ 2700 copies of pVIII that form the lateral surface of the phage particle, as well as 5 copies of each the pIII and pVI that form one base surface and a few copies of each pVII and pIX that form the other base surface. The pIII protein plays a particular role as it is responsible for infectivity of the phage particle and infects a cell by docking to an *E. coli* F' pilus displayed on the cell surface, leading to a chronical infection of the cell accompanied by production and release of new phage particles. Moreover, only The M13 replication cycle is described in [160] and modelled in [161]. The pIII, pVI, pVII and pIX have been used for phage display [154]; the most popular choice, however, is pIII. The phage retains its infectivity when the wild-type pIII is replaced by a pIII-protein of interest fusion.

A vector map of pIT2, a standard phage display vector that we use here, is shown in figure 3.5(b). It encodes for the pIII-antibody fusion and a number of genetic ingredients required for the production and release of displaying phage. In this variant of phage display, at most one of the 5 copies of pIII on a phage particle is a pIII-antibody fusion; the remaining ones are wild-type pIII. In the case of antibodies, owing to their size and in order to control the copy number at the surface of M13, phagemids and helper phage are used.

The phagemid backbone (see lower part in figure 3.5(b)) consists of an ampicillin resistance gene (Amp^{r} or amp), an M13 phage origin of replication (M13 *ori*), and an *E. coli* origin of replication (*colE1 ori*). The phagemid thus confers resistance against the antibiotic ampicillin to the cell that carries it. In a cell culture, cells can be selected for the presence of the phagemid by adding ampicillin to the growth medium: cells with the phagemid continue to grow while those without do not grow or die. Note that in absence of ampicillin, the phagemid is likely a burden to the cell and cells without phagemid would then have a selective advantage over cells with phagemid. *colE1 ori* allows for replication of the phagemid by the cell's replication mechanism, while M13 *ori* enables replication by the M13 replication mechanism. Replication by the cell is necessary for cell division (so that all daughter cells obtain the phagemid), while replication by the phage is required for encapsulation in new new phage particles.

The expression cassette (or display cassette, see upper part in figure 3.5(b)) notably contains the gene of the protein of interest (cloned in between the NcoI and NotI restriction sites) and gIII (or M13 gene 3), the gene for pIII surface protein. These genes are bridged by tag sequences (Tag) that can be targeted by primary antibodies in ELISA (here: a PolyHis tag and a myc tag), as well as an amber stop codon `TAG` (amber). In partial amber codon suppressor strains, such as TG1, the amber codon allows for expression of both the antibody alone (for when the translation is stopped at the amber codon) or of the pIII-antibody fusion protein required for phage display (for when the amber codon is read through). These strains feature ribosomes that happen to mistranslate amber codons as glutamine (`Q`) rather than stop (in about 1/3 of the

cases). Upstream of the pIII-antibody fusion gene is the *lac* promoter (P*lac*) and the sequence for a signal peptide (SS or pelB leader) that triggers the export of the (pIII-)antibody (fusion) to the cell periplasm where phage particles are assembled. Expression of the whole construct which is initiated by RNA polymerase binding to the *lac* promoter can be regulated by adding glucose or IPTG (isopropyl-β-D-thiogalactopyranoside) into the cell growth medium: Glucose represses expression by turning off the *lac* operon, while IPTG as a lactose analog induces expression by turning the *lac* operon on. (The cells preferentially metabolize glucose over lactose.)

The helper phage M13KO7 is a modified version of wild-type M13 phage and is required to initiate the production of displaying phage in library cells. Instead of the wild-type M13 phage genome, it contains a genome with all genes of the M13 phage, but the M13 origin of replication being replaced by both a high-copy plasmid origin of replication and a resistance gene against the antibiotic kanamycin. After infection of library cells with helper phage, all proteins of the phage are expressed inside the cell, and the phagemid is replicated via M13 *ori* more efficiently than the helper phage genome with the corrupted origin of replication. New phage particles are produced in the cell periplasm: Phage capsids are assembled by taking together expressed pIII-antibody fusions and all other phage surface proteins. Importantly, the phagemid carrying the protein of interest gene is more efficiently encapsulated into the new phage particles than the helper phage genome.

As a result, M13 phage particles that display at their surface, fused to pIII surface protein, the antibody of interest and that contain inside their capsids the phagemid with the genetic information of the antibody are produced and secreted into the cell growth medium. In our design, at most 1 among the 5 pIII copies is fused to the antibody in most phage particles, while the remaining ones are wild-type pIII. In parallel, non-displaying phage, *i.e.* phage particles with the phagemid but only wild-type pIII, and, to a lesser extent, displaying phage particles with the helper phage genome rather than the phagemid are also produced and secreted as side products. Non-displaying phage are expected to be removed upon selection for binding due to the absence of the antibody. However, displaying phage carrying the helper phage genome can pass the selection, but they are a minority (typically $\simeq 100\,\mathrm{x}$ times less frequent than displaying phage carrying the phagemid as a result of the helper phage *ori* being $\simeq 100\,\mathrm{x}$ less efficient than M13 *ori*).

3.2.3 Production of displaying phage

The protocol for the displaying phage production that we follow in our experiments is provided in section A.4. The workflow is schematized in figure 3.5(c) and goes as follows: We start an overnight liquid culture of library cells with ampicillin and glucose (to select for the presence of the phagemid and suppress its expression; expression of pIII-antibody is not yet needed at this stage). The quantity of glycerol stock needed to start this liquid culture had to be chosen carefully: The criterion is that all of the $q^L \simeq 10^5$ unique sequences be sufficiently oversampled,

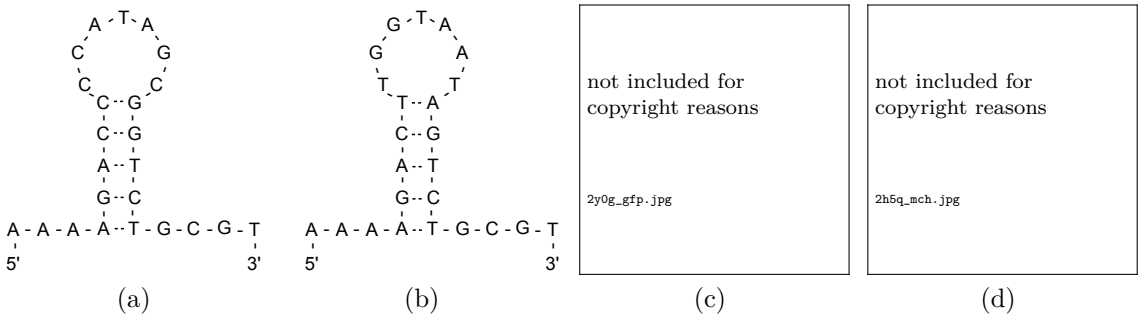
say on average $5 \cdot 10^2$ times. Thus, the liquid culture had to be started with initially $5 \cdot 10^7$ cells, which for a typical bacterial density in glycerol stocks of $OD_{600} = 100$ or, equivalently, $8 \cdot 10^{10}\,mL^{-1}$ means that a volume of $\frac{5 \cdot 10^7}{8 \cdot 10^{10}\,mL^{-1}} \simeq 0.6 \cdot 10^{-3}\,mL = 0.6\,\mu L$ of glycerol stock had to be added. The following day, we started a fresh liquid culture with glucose and ampicillin by diluting the overnight culture 100 x. As soon as the exponential growth phase was reached, *i.e.* at a bacterial density of $OD_{600} \simeq 0.4$, we added an excess of helper phage to the culture. During a 30 min incubation, the infection of the library cells is let to happen. Then, we centrifuged the culture and resuspended the pellet in fresh growth medium containing the antibiotics ampicillin and kanamycin. Thus, we select for the co-presence of the phagemid (ampicillin resistance) and the helper phage genome (kanamycin resistance) in the cells. In addition, the absence of glucose turns on the lac promoter for the expression of pIII-antibody fusions. We incubate the culture for 7 h in order to let the production, assembly and secretion of up to $10^{11} - 10^{12}$ displaying phage happen. By the end of the incubation, we separate the supernatant containing displaying phage from the cells by high-speed centrifugation (11,000 g) and filtering through 0.22 µm filters in order remove cell debris, and store it at 4 °C for selection experiments on the following day and only on the following day. These displaying phage can then be used for selections and to subsequently to infect fresh TG1 cells to obtain library cells representing the selected library. Beyond 24 h from the phage production step, displayed antibodies may be denatured (unfolded) and no longer functional. We therefore avoided the use of displaying phage older than 24 h for selections.

3.3 Choice and handling of target molecules for binding

We here present our choice of target molecules for the binding of antibodies, which includes two DNA hairpin targets denoted henceforth DNA1 and DNA2, as well as two protein targets, eGFP and mCherry, denoted henceforth prot1 and prot2 (subsection 3.3.1). In our selection experiments, the selective pressure will be defined by the binding affinity of antibodies to bind to these targets. The production of the protein targets, as well as their immobilization on magnetic beads is summarized (subsection 3.3.2). The immobilization is required in order to hold and separate these target molecules and displaying phage particles bound to them from unbound ones by applying a magnetic field.

3.3.1 Choice and production of target molecules

The task is to define suitable targets for the binding of antibodies. Past studies have shown that antibodies with affinity to almost any target can be identified from recombinant antibody libraries, including metallic gold [125, 126]. Seemingly, the only constraint for the design of such binding targets is the existence of a well-defined structure.



(a) (b) (c) (d)

Fig. 3.6: Target molecules for binding. ssDNA targets: **(a)** DNA1 (alternative name: Noire), **(b)** DNA2 (alternative name: Bleue). The DNA targets have identical stem sequence, but different loop sequences. At ambient temperature, they display hairpin structures. Protein targets: **(c)** prot1 (eGFP enhanced green fluorescent protein, PDB accession: `2Y0G`) [163], **(d)** prot2 (mCherry red fluorescent protein, PDB accession: `2H5G`) [164]. The protein targets have similar structures, but little sequence similarity.

The target molecules that we propose for our study of selection potentials are two single-stranded DNA targets, as well as two protein targets, shown in figure 3.6. The two ssDNA targets in figure 3.6(a), (b), that we denote by respectively "DNA1" and "DNA2" (or alternatively "Noire" and "Bleue", the French words for "black" and "blue"), are 24 nt in length and display hairpin structures at ambient temperature. This type of structure is the result of the presence of two complementary subsequences of length 5 nt each along the ssDNA sequence. These associate to form the stem of the DNA hairpins. The 7 nucleotides in between these stem sequences form the loop of the hairpin. The stem sequence is shared between DNA1 and DNA2, whereas the loop sequences are different. Generally, the antigenicity of DNA is well established. The two protein targets in figure 3.6(c), (d) are derived by directed evolution from two distinct natural fluorescent proteins taken from two phylogenetically distant species: They correspond to the enhanced green fluorescent protein (eGFP) which is derived from the wild-type GFP produced in jellyfish *Aequorea victoria* [163] and the red fluorescent protein (mCherry) which is derived from a wild-type produced in the *Discosoma* (mushroom anemone) genus [164]. They correspond to PDB accessions `2Y0G` and `2H5G`, respectively. In what follows, these proteins will be referred to as simply prot1 and prot2. These proteins have similar structure (see figure 3.6(c), (d)), but have little sequence identity (not shown here).

For the purpose of attaching the targets to streptavidin-coated magnetic beads, extensions to these target molecules are necessary: The 5′ ends of the DNA targets are fused to a biotin, the natural binding partner of streptavidin. The DNA targets are thus immobilized on the beads via streptavidin-biotin binding, one of the strongest known non-covalent interactions in nature. The protein targets are fused to a SBP tag sequence that also binds to streptavidin. It should be emphasized that the effective targets seen by the antibodies are not the molecules shown in figure 3.6 alone, but the whole complex formed of the magnetic bead, the streptavidin, and the defined target molecules. Such a complex is shown in figure 3.7(a).

The biotinylated DNA hairpin targets (DNA1 and DNA2) were purchased from IDT (Leuven, Belgium), diluted in deionized and filtered (MilliQ) water, and stored at −20 °C. The protein targets (prot1 and prot2) in fusion with a SBP tag were, however, expressed by ourselves using the corresponding genes kindly provided by Sandrine Moutel (Institut Curie, Paris, France) and the following expression protocol (details in section A.5): We transformed the genes to T7 Express *E. Coli* cells by electroporation (see section A.2) and plated them on selective growth medium containing ampicillin. Then, we diluted overnight liquid cultures from colonies of transformants 100 x in 200 mL of growth medium with ampicillin and induced the fluorescent protein expression at a bacterial density of $OD_{600} = 0.5$ with 300 µM Isopropyl-β-D-1-thiogalactopyranoside (IPTG, Sigma-Aldrich, Saint-Louis, MO, USA) final. We incubated the induced cell cultures overnight at 30 °C, shaking. The following day, the cell cultures had acquired clearly visible red or green tinting, confirming successful expression of the fluorescent proteins. We harvested the fluorescent proteins, *i.e.* extracted them from the cells, by threefold flash freezing of the cell cultures in liquid nitrogen and quick thawing in a water bath at 42 °C, which leads to cell lysis. We further incubated the lysates with 50 µg mL^{-1} lysozyme final and 2.5 U.mL^{-1} DNase I final at 30 °C for 15 min and subsequently centrifugated at 15,000 g and 4 °C for 30 min. Cell lysates contain fluorescent protein, as well as other (degraded) cellular components such as cells' proteins, DNA, and membrane; during the following immobilization step (see next subsection 3.3.2), predominantly fluorescent protein will bind to streptavidin-coated magnetic beads via their SBP tag. Finally, we aliquoted the fluorescent proteins into protein low-bind tubes (Protein LoBind, Eppendorf, Hamburg, Germany), flash froze them in liquid nitrogen and stored them at −80 °C until use.

3.3.2 Immobilization on magnetic beads

The targets must be immobilized on a controllable substrate. There are various platforms for target presentation [126] and for our selection experiments, we opt for immobilization on streptavidin-coated megnetic beads (Dynabeads(R) M-280 Streptavidin) that we purchased from Invitrogen Life Technologies (Carlsbad, CA, USA). The detailed protocol for target immobilization can be found in section A.5). We perform the binding of target molecules to streptavidin-coated magnetic beads 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. First, we washed the beads in 500 µL of 1 x PBS to remove any unwanted objects. The liquid was removed while holding the magnetic beads back in the tube by applying a magnetic field. In order to bind the DNA targets [protein targets], we first resuspended the beads in 90 µL 1 x PBS [50 µL 1 x PBS] and added 10 µL of DNA at 400 µM [50 µL of protein]. For negative control selections, we added the same volume of MilliQ water instead of targets. The binding was let to happen by incubation at ambient temperature on a rocker for 15 min. Finally, we removed all unbound targets by, first, removing all liquid from the beads and, second, a threefold washing of the beads using Bw1x buffer (1 M NaCl, 5 mM Trizma at pH = 7.4, 0.5 mM EDTA) [1 x PBS with 0.1 % Tween20] as washing solution: addition of 500 µL of washing solution, vortexing, and removal

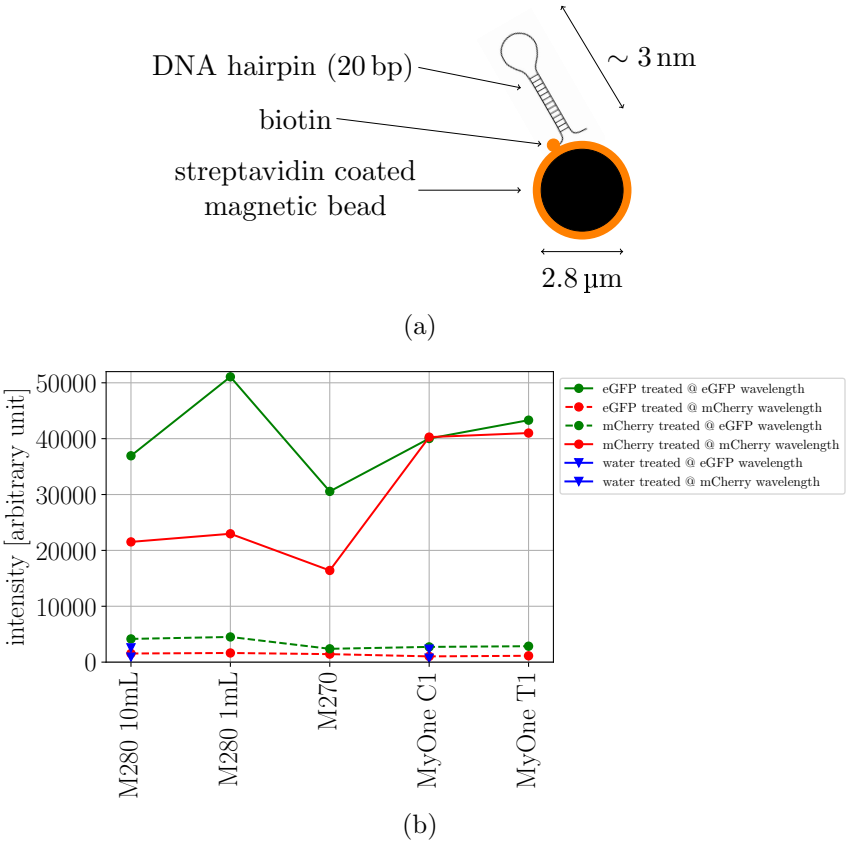


Fig. 3.7: **(a)** Schema of a biotin-tagged DNA hairpin target immobilized on a streptavidin-coated magnetic bead via streptavidin-biotin binding (not true to scale; several copies of the target per bead in real). **(b)** Fluorescence measurements of fluorescent protein treated magnetic beads. prot1 (eGFP) or prot2 (mCherry) targets were immobilized on different sorts of streptavidin-coated magnetic beads (x-axis) by myc tag to streptavidin binding. Fluorescence intensity at green and red wavelengths were measured (y-axis). Negative controls using water treated beads were performed (measurement of background fluorescence and auto-fluorescence of the beads).

of all liquid from the beads using a magnetic field. The presence of NaCl [the surfactant Tween20] in the washing solution is meant to screen unspecific electrostatic [hydrophobic] interactions that may involve and unintentionally hold back unbound target molecules. Finally, the beads were resuspended in 50 µL washing buffer and stored at 4 °C for use in selection experiments on the following day.

A simple calculation shows that the quantity of magnetic beads used in one selection has a capacity of 10^{14} binding sites and thus allows for the uptake and presentation of 10^{14} target molecules. This number of targets present during the selection exceeds that of possible binders, $10^{11} - 10^{12}$ in phage display, thus making competition between binders for targets unlikely. This justifies the Boltzmann approximation made in the computation of enrichments s in chapter 2.

For the fluorescent protein targets, we confirmed successful immobilization by fluorescence measurements on protein-treated beads after washing, see figure 3.7(b). We prepared different kinds of protein-treated beads, Dynabeads(R) M-280 Streptavidin and others, according to the protocol described above and diluted all 50 µL in 1 x PBS to have 1 mL final. After vortexing, we pipetted 10 µL from the dilution into a black 96-well plate and further diluted by adding another 90 µL of 1 x PBS (200-fold dilution final) in order to cover the ground of the well. Measurement of the intensity at red and green wavelengths then yielded the data reported in figure 3.7(b). Measurements on water-treated beads, as well as crossed measurements (eGFP-treated beads at red wavelength and mCherry-treated beads at green wavelength) were performed as negative controls. As expected in the case of successful immobilization, the eGFP- (mCherry-) treated beads show high signal at green (red) wavelengths compared to the signals obtained when the wavelengths are exchanged and for water-treated beads at both wavelengths. Measurements were carried out on a fluorescence plate reader (Spark, Tecan, Männedorf, Switzerland) equipped with monochromaters for excitation and emission that we set optimally to match the excitation/emission spectra of either eGFP or mCherry. For DNA targets, such a check of successful immobilization is not possible, but their presence during selection can be confirmed *a posteriori* (see chapter 4).

3.4 The selection step and strategies for library screens by phage display

Given the phage displayed V_H libraries from section 3.2 and immobilized targets from section 3.3, we can now proceed with the selection step. We here present the protocol for selection by biopinning which is as previously published in [1] (subsection 3.4.1). For the later analysis of these selection experiments, the goal is to compute frequencies and enrichments of sequences in selected and unselected libraries (subsection 3.4.2). Different selection schemes that provide complementary information about enrichments are possible for our three V_H libraries: Selections from each library separately focus on differences in enrichments between sequences of the same library, while selecting from a mix of all libraries also allows to compare sequences across libraries in absolute terms (subsection 3.4.3). In addition, subsampling the full libraries allows to estimate enrichments for less sequences but with higher precision (subsection 3.4.4).

3.4.1 Protocol for and effect of selection on a diverse population

We here present our protocol for the selection step by biopinning which simply consists in incubating the antibody-displaying phage together with target-coated magnetic beads the same tube, followed by a washing step leaving behind only the magnetic beads and everything bound to them, including the phages with affinity for the targets. The selection step is schematized in figure 3.8. The effect of enriching good binders over bad or nonspecific binders is put to the numbers in

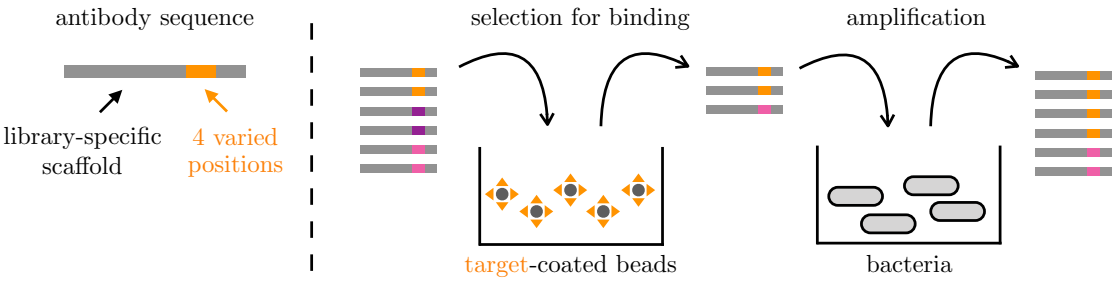


Fig. 3.8: Principle of our antibody selections. **Left** Antibody library design: A scaffold sequence (in gray, comprising the framework regions, as well as CDR1 and CDR2) is kept constant and defines the library, while four positions of the CDR3 (in orange) are randomized. **Right** Selection for binding of a population of antibodies with different sequences (here: different colors) displayed on phage by incubation with target-coated beads: Antibodies bind or not to targets according to their binding affinity. Amplification to original population size by replication in fresh bacteria. During this procedure, good binders (here: in orange) are enriched, while poor binders (here: in violet) are depleted.

figure 3.9(b). The detailed selection protocol is provided in section A.6.

As a first step, we adjust the culture supernatant containing the displaying phage to pH = 7.0 by adding Na_2HPO_4 and NaH_2PO_4 and centrifuge to collect possible cell debris leftovers at the bottom of the tube. Prior to the actual, positive selection, we perform a null selection by incubation of displaying phage with water-treated beads (this step was defined in a previous work [1], but as binding probabilities are globally low compared to 1, the effect of negative selection should be close to absent): We remove the washing solution from the naked beads and add 1 mL of supernatant containing $\simeq 10^{11}$ displaying phage particles. We incubate the tube, shaking (to avoid sedimentation of the beads), at ambient temperature for 90 min. We then proceed likewise using target-treated beads: We remove the washing solution from the target-treated beads and transfer the supernatant from the water-treated beads to the target-treated beads. Again, we incubate shaking at ambient temperature for 90 min to let the binding reaction between the displaying phage and the targets happen. By the end of the incubation, we pour away the supernatant containing the unbound phage and subject the beads to a 10-fold washing step with the goal of diluting and discarding all dead volumes and "stuck" but unbound phage: We add 10 mL of PBS with 0.1 % Tween20 surfactant to the beads, let the beads traverse the liquid by relocating the magnetic field, pour away the liquid and repeat the same procedure for another 9 times. Finally, we incubat the beads in 1 mL of 1.4 % triethylamine (TEA, Sigma-Aldrich, Saint-Louis, MO, USA) to elute the bound phage from the beads and transfer the eluted phage to 330 mM Trizma (2-Amino-2-(hydroxymethyl)-1,3-propanediol or Tris(hydroxy-methyl)aminomethane) final. Throughout the entire procedure, whenever we add any liquid to the beads or pour it away, the beads are hold back by applying a magnetic field.

The eluted phage represent the selected population of "survivors" and contain the genes of the V_H segments that allowed them to bind to the targets. We use these phage particles to infect a

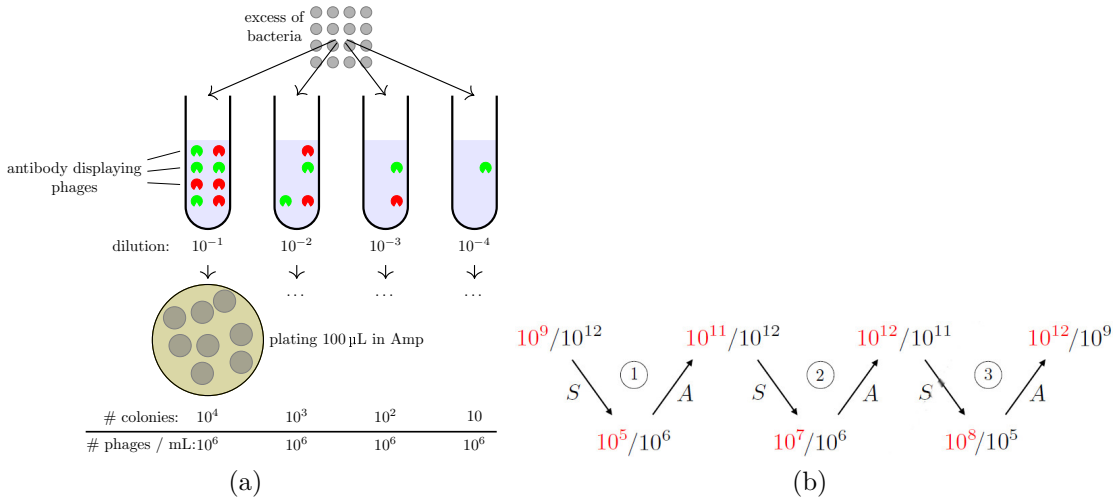


Fig. 3.9: Selection yield. **(a)** Estimation of selection yield. The number of displaying phage (among the $\simeq 10^{11}$ present) bound to targets during the selection step is estimated by performing serial dilutions of the phage elution, infection with an excess of cells and plating on selective growth medium containing ampicillin (thus selecting for the presence of the phagemid acquired upon infection). Each colony on the plate is the result of an infection event and thus of a selected phage. The number of colonies provides an estimation of the number of selected phage, *e.g.* 100 colonies from 100 µL of a 10^3-fold dilution resumes to $100 \cdot 10^3 = 10^5/100\,\mu\text{L} = 10^6\,\text{mL}^{-1}$, *i.e.* 10^6 selected phage, as a volume of 1 mL of displaying phage is used for a selection. **(b)** Increase of selection yield with the number of selection rounds [133]. The number of good binders (red numbers, binding probability 10^{-4}) and bad/unspecific binders (black numbers, binding probability 10^{-6}) is multiplied by their respective binding probabilities upon selection (S) and by a common factor to recover a population size of 10^{12} upon amplification (A). While the initial library is dominated by bad binders (10^9 good and 10^{12} bad), good binders are enriched upon several selection rounds to eventually become dominant (10^{12} good and 10^9 bad at round $t = 3$).

fresh exponential liquid culture of TG1 *E. coli* cells. This infection step is performed in a way to ensure that every phage gives rise to one infected cell, *i.e.* the number of cells exceeds the number of phage (otherwise a competition of phage for cells may lead to distortion of frequencies in the selected library). Then, we plate the cell culture on a large Petri dish containing selective growth medium with ampicillin and glucose and incubate overnight at 37 °C for cell growth and phagemid replication by the cells (amplification). The ampicillin allows only infected cells to grow, while the glucose suppresses expression of pIII-antibody fusion which is needless at this step. In parallel, we also plate serial dilutions of the infected cell culture, as well as serial dilutions of the input library separately on ampicillin (with glucose) and kanamycin. Counting the colonies on ampicillin allows to estimate the number of phage carrying a phagemid in the selection input and output, and thus the selection yield (how many among the initially present phage particles survived the selection step?), see figure 3.9(a). Counting the colonies on kanamycin allows to estimate the number of phage particles carrying the helper phage genome instead of the phagemid, which is typically $\simeq 100$ x lower. On the following day, the plates were covered by cell colonies that are indistinguishably many in platings of the undiluted cell culture (cell carpet), but distinguishable

and countable in platings of the serial dilutions. Each colony is the result of one bound phage during the selection step. We scrape the cell colonies in the large Petri dish are scraped into 25 % glycerol in growth medium and mix well by brief vortexing. The glycerol stock of selected library cells are then stored at −80 °C. At this stage, we are technically back to the beginning of the experiment, except that the newly obtained library cells represent the selected antibody library and display the shift in frequencies due to selection. These are used to initiate a new round of selection by phage displaying the selected antibody library. More generally, the selection can be iterated as many times as necessary. Also, the library cells at several selction rounds are the input for sequencing of the selected libraries (see section 3.5.2).

3.4.2 Empirical enrichments as proxy for binding affinity

A statistical analysis of these selection experiments will be performed in light of the theoretical aspects presented in chapter 2. We here propose to study empirically distributions of enrichments $P(s)$ (see chapter 4) and the inference of simple biophysical models that provide predictions for the mapping from sequence to binding energy $x \mapsto \Delta G(x)$ (see chapter 4). Computing enrichments and inferring biophysical parameters require the measurement of sequencing counts and frequencies in the initial libraries and after (several rounds of) selection. This can be achieved through high-throughput sequencing of the initial and selected libraries which allows to determine sequence identities of a large number of randomly picked individuals in the libraries. Upon counting the number of times a sequence x appears at selection round t in the sequencing data of size N_t, $n_t(x)$ such that $\sum_x n_t(x) = N_t$, we may compute (relative) enrichments according to equation (2.91) from

$$s(x) = \lambda_t^{-1} \frac{f_{t+1}(x)}{f_t(x)} = \lambda_t^{-1} \frac{n_{t+1}(x)}{n_t(x)}, \tag{3.1}$$

where the sequence-independent amplification factor λ_t is unknown but irrelevant for the study of differences between sequences x. (An order of magnitude for λ_t can be estimated from the ratio of the total number of phages in the selection input over output, but this method is rather imprecise.) Within a single selection experiment, it can be set to an arbitrary value as long as only differences in enrichments matter. (Note that amplification biases may amplify some sequences x more than others, but this effect can be absorbed into $s(x)$.) Note however that the lack of λ_t does not allow for the comparison of enrichments between different selection experiments. In another and more involved approach based on multi-type branching processes [63, 68, 138, 64, 61], one writes down a model for $P\big(\{n_t(x), n_{t+1}(x)\}_x | J_{i_1,\dots,i_p}(a_1,\dots,a_p)\big)$, where the $J_{i_1,\dots,i_p}(a_1,\dots,a_p)$ are parameters of a biophysical model for ΔG of the form of equation (2.25). Such a model can then be used to infer the values of $J_{i_1,\dots,i_p}(a_1,\dots,a_p)$ given the data for $\{n_t(x), n_{t+1}(x)\}_x$, see section 4.5.

However, such analysis will be limited by finite sequencing depth: Current sequencing techniques allow to sequence $N = \sum_x n_t(x) = 10^5 - 10^7$ individuals as opposed to population sizes of

$N_{\text{pop}} = 10^{11} - 10^{12}$ in phage display. To sequence our selection experiments, we will use Illumina MiSeq 2×250 bp paired-end sequencing that provides sequencing reads for $10^5 - 10^6$ individuals per sample, *i.e.* per library-target combination and per selection round. (A complete Illumina Miseq 2×250 bp paired-end run guarantees 10^7 sequencing reads but these need to be dispensed over several samples that are sequenced simultaneously.) Poorly represented (but non-absent) sequences may therefore not be observed in sequencing data albeit being present with low frequencies in the libraries. In our experiments, sequences with frequencies lower than $\frac{N}{N_{\text{pop}}} \simeq 10^{-6} - 10^{-5}$ will typically remain unseen. This may affect bad sequences that are depleted below this observation threshold, but also good but rare sequences that are not yet enriched above this threshold. Conversely, sequences x with a few counts, $n_t(x) \lesssim 10$, may be observed by chance and counts may not represent frequencies well. In summary, counts represent frequencies well only if the sequence count numbers are sufficiently high. Here, we compute empirical enrichments from equation (3.1) only if $n_t(x) \geq n_{\text{thr}}$ and $n_{t+1}(x) \geq n_{\text{thr}}$ to ensure they are meaningful. We use $n_{\text{thr}} = 10$ for all full-library selections [1] and $n_{\text{thr}} = 100$ for mini library selections (see subsection 3.4.4).

3.4.3 Isolate versus library mix selections

Two different selection scenarios are conceivable: i) Each of the three libraries Germ, Lmtd, and BnAb is subjected to selection separately and independently from each other. ii) All three libraries are pooled together in equal proportions and are subjected to the same selection altogether. We can expect these two selection strategies to reveal complementary aspects of comparing enrichments across all antibody variants involved: Selecting according to scheme i) is governed by differences in binding affinities and enrichments between variants of the same library, *i.e.* between antibodies with identical scaffold but different CDR3 sequences. On the contrary, selection trajectories according to scheme ii) would additionally be determined by global differences in enrichments between scaffolds in a way that, in the case of lognormally distributed enrichments, is encoded in equation (2.105).

We shall recapitulate these ideas in terms of the theory developped in chapter 2 that led to the assumption in which each library ℓ is characterized by a lognormal distribution of enrichments with parameters μ_ℓ and σ_ℓ encoding roughly for the global level of binding energy and the variance in binding energies among members of the same library, respectively. In scenario i), only the variance σ_ℓ within library ℓ, which represents the variance of binding energies between CDR3 sequences given the scaffold, determines the evolution of frequencies according to Fisher's equation (2.98). The parameter μ_ℓ should not matter for selection in this case as it only encodes for the overall binding capacity of a scaffold, irrespectively of the CDR3 sequence. In the scenario ii), however, differences in both differences in μ_ℓ and σ_ℓ among libraries should matter in a particular way that is encoded in equation (2.105) and that was discussed in section 2.4.3.

In summary, while selections according to scenario i) are expected to yield better resolution

of enrichment differences between sequences of the same library, selections using scenario ii) put emphasis on differences across libraries. Selections from a mix of libraries are required to learn differences between the libraries, as relative enrichments computed from equation (3.1) are not comparable between experiments. Within this project, we realize both selection strategies: On the one hand, we generated selection trajectories against all four targets independently starting from a uniform mix of all three libraries Germ, Lmtd, and BnAb. On the other hand, we also generated trajectories against the DNA1 and DNA2 targets for all three libraries separately. Note that the number of experiments required is multiplied by three when selecting the three libraries separately rather than in mix. In practice, the uniform mix of libraries is achieved by simply pooling together the culture supernatants of the three phage-displayed libraries (see section 3.2) in equal volumes before the first selection round (this assumes equal concentrations of displaying phage in all three phage-displayed libraries, but this is more or less true).

3.4.4 Trade-off between diversity and degeneracy: mini libraries

One goal is to infer and compare the values of μ_ℓ and σ_ℓ of the different libraries under the lognormal model using enrichment data from selection experiments. While this is feasible from lists of empirical enrichments determined according to equation (3.1) from mixed selections using a MLE approach (see chapter 4), a more direct measurement of these parameters can be achieved by yet another selection strategy that uses "mini libraries". These mini libraries represent subsamples of the full CDR3 diversity with drastically reduced number of unique variants ($\simeq 10$ instead of $\simeq 10^5$). The advantage of using mini libraries is twofold: i) If such mini libraries are designed in a way to contain few strongly binding as well as few randomly chosen sequences, which represent respectively the extremes and the mode (most likely value) of the enrichment distribution, the measurement of their enrichments can immediately lead to the values of μ_ℓ and σ_ℓ. This uses the fact that the maximum and mode of q^L lognormal enrichments are given by

$$\max(s) \simeq \exp\left(\mu + \sqrt{2\ln(q^L)}\sigma\right), \qquad \mathrm{mode}(s) = \exp\left(\mu - \sigma^2\right). \tag{3.2}$$

Note that the mean $\langle s\rangle = \exp\left(\mu + \frac{\sigma^2}{2}\right)$ is very different from the mode in case of large σ, as a consequence of the strong skew of the lognormal distribution in this case and the mean being dominated by large outliers. ii) At constant sequencing budget, a decrease in diversity implies an increase in degeneracy, meaning that more sequencing counts will be recorded per variant when fewer unique variants are present in the library. As a consequence, frequencies and enrichments can be computed with higher accuracy in mini libraries compared to full libraries. To put this argument into the numbers, consider that sequencing allows to read out the identities of $\simeq 10^6$ individuals. With a diversity of $\simeq 10^5$, we thus have on average $\simeq 10$ sequencing counts per variant. While good binders are enriched far beyond 10 counts upon selection, bad binders are depleted below 10 counts and are not taken into account for the computation of enrichments. If the diversity is, however, only of $\simeq 10$ different sequences, then we have on average $\simeq 10^5$ sequencing

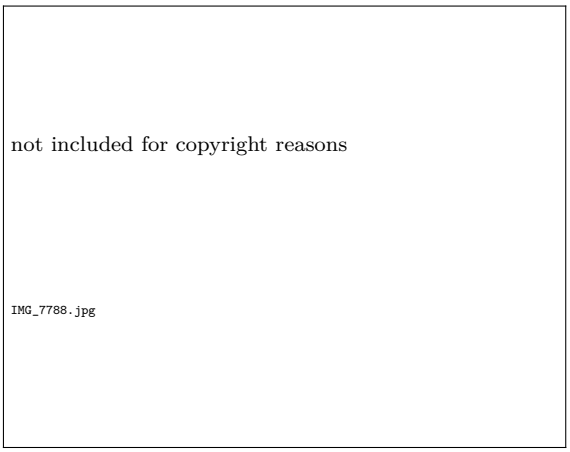


Fig. 3.10: Comparison of commonly used high-throughput sequencing technologies.

counts per sequence. Thus, enrichments can be computed even for strongly depleted sequences and the overall increase in sequence counts improves the accuracy of empirical enrichments.

In practice, such mini libraries can be obtained by cloning of few strongly binding sequences and pooling them together, along with a few randomly picked sequences from the initial full libraries. Strongly binding CDR3 sequences need to identified from preceding selections using the full libraries. Once the relevant CDR3 sequences are defined, they can be purchased as dsDNA and cloned into the pIT2-V_H phage display vector similarly to the cloning of the full libraries described in section 2.1.3. Rather than cloning the full V_H into pIT2 again, the pIT2 already containing the V_H gene and an arbitrary CDR3 sequence is used as template. Digestion is performed using the restriction sites BssHII and XhoI. The detailed cloning protocol is given in section A.2.

3.5 (High-throughput) Sequencing: measurement of frequencies and enrichments

To sequence our libraries, we opt for Illumina Miseq 2×250 bp paired-end sequencing that guarantees for $\geq 10^7$ paired forward and reverse reads, each 250 bp in length, of a sequencing amplicon (subsection 3.5.1). We describe our sequencing amplicon design and the preparatory PCR reactions under the premise that we seek to identify the scaffold and CDR3 sequences from the sequencing data (subsection 3.5.2). The preprocessing pipeline that we use to "clean" the raw data and to read out V_H sequence identities is also described (subsection 3.5.3).

3.5.1 A comment on sequencing methods used in this project

A very good review of "next-generation" sequencing (NGS) technologies can be found in [165]. In general, these methods allow to read out the sequences of large numbers of copies of a given DNA construct. They rely on a variety of different conceptual basis and have been and are still subject to continuous improvement and innovation, allowing for ever-increasingly deep and large-scale sequencing. However, they are not exchangable with one another in general as they apply to different regions of a "phase diagram" spanned notably by the number of sequencing reads, the sequence length, and the sequencing error level as relevant parameters: Figure 3.10 compares different commonly used high-throughput sequencing platforms with respect to these and other parameters. A few trade-offs between these parameters exist: i) High-throughput methods allow for many reads at once, but are also prone to various levels of sequencing errors. This is in contrast to *e.g.* Sanger sequencing, a more conventional, first-generation sequencing technique, which allows for a single read at a time but is (next to) error-free. ii) NextSeq and HiSeq provide more but shorter sequencing reads than MiSeq sequencing. iii) Pacbio and Nanopore provide very long but much more erroneous sequencing reads than MiSeq, HiSeq and NextSeq. The genetic construct that one wishes to sequence, as well as the information that one wishes to obtain thus determine the most useful among these sequencing methods.

Within this project, we used Sanger sequencing and Illumina MiSeq $2 \times 250\,\mathrm{bp}$ paired-end sequencing; the workflow for these first- and second-generation techniques is comparatively described in [165] and figure 1 therein. To sequence our V_H libraries, we opted for Illumina MiSeq $2 \times 250\,\mathrm{bp}$ paired-end sequencing that provides $\geq 10^7$ (per run) paired forward and reverse reads, each 250 bp in length, of a given DNA amplicon. A V_H gene already exceeds 250 bp and can thus not be read without fragmentation from a single such sequencing read (if one wishes the paired forward and reverse reads to be fully overlapping). But, as discussed below, the budget of 250 bp is sufficient to read at once the CDR3 and a sufficiently long part of the scaffold in order to identify the scaffold identity of a read. It allows us to obtain $10^5 - 10^6$ sequencing reads per sample (given that several samples are sequenced within one run of $\geq 10^7$ reads in total) in order to estimate frequencies and enrichment of variants. Sanger sequencing was used mainly on two purposes: First, it was used to check and confirm V_H sequences in the phagemid. Most importantly, we re-cloned single V_H variants from our full libraries into the pIT2 plasmid in order to construct mini libraries (see section 3.4). Transformants after the cloning were Sanger sequenced to check for the correctness of the cloned V_H sequence (which is far from obvious, see chapter A). Second, Sanger sequencing was typically performed on few (typically $\simeq 10$) colonies prior to high-throughput sequencing of a complete library as a plausibility check and a way to exclude to exclude failure of the selection experiment that created the library. For instance, in selections with a mix of the three V_H libraries, 10 reads allowed to obtain a rough estimate of the frequencies of the three libraries within the mix. This practice allows to check the success of the selection experiment and to draw predictions for the outcome of the Illumina sequencing. The Illumina sequencing was used to count sequences and compute frequencies from initial and selected libraries and library mixes.

The primers used here for Sanger sequencing were either M13-rev which is located upstream of the V_H, or pHEN which is located right upstream of the amber stop codon and extends against the V_H reading frame.

3.5.2 Amplicon design and preparatory PCR reactions for Illumina MiSeq sequencing

The setup of the sequencing amplicon for Illumina MiSeq 2×250 bp paired-end sequencing and its generation through two subsequent PCR reactions are shown in figure 3.11. The sequence of the amplicon for our case which is 316 bp in length is given in figure E.1(b). To generate the amplicon, the region of interest must be amplified from the plasmid by PCR and completed in two steps (PCR1 and PCR2) by random barcodes (represented by `NNNNN`) that allow to discriminate between neighboring clusters that differ in the realization of these 5 random positions, sample-specific DNA barcode sequences (index P5 and index P7) encoding library, target and selection round, the primer sequences used for annealing during the sequencing procedure (read 1 and read 2), as well as the adapter sequences for the Illumina sequencing platform (adapter f and adapter r) at both ends, see figure 3.11(a).

The amplicon design for the MiSeq sequencing of our V_H libraries must account for two essential requirements: i) In order to uniquely identify V_H sequences, the sequencing data must provide the information about both the scaffold identity, which takes only on the three possible values Germ, Lmtd, and BnAb, and the CDR3 identity, which is determined by its sequence of 12 bp. However, it is not possible to sequence the full V_H when we want the forward and reverse reads to be (fully) overlapping: Excluding the space required for primer and barcode sequences, this sequencing method only allows to define a region of around 200 bp in length along the V_H, but the full V_H (from CDR1 to FWR4) is > 400 bp in length (see figure 3.1). ii) The primer sequences used for the preparatory PCR reactions must be common to all three different V_H scaffold sequences (otherwise sequences from either of the three libraries would not be PCR amplified and be projected out in the sequencing data). Thus, the challenge consists in defining a region along the V_H gene that satifies the length and primer sequence constraints, yet provides sufficient information about V_H sequence identities. A region that satisfies all these requirements and that we use for the sequencing of our selection experiments is shown in figure 3.12 and is 128 bp in length (170 bp including the primer sequences): It consists of essentially FWR3, CDR3, and FWR4 and is flanked both up- and downstream by sufficiently long sequences in CDR2 and FWR4 that are conserved across the three V_H scaffolds and can thus be used as PCR primers: `GCTCGAGACGGTAACCAGG` as forward primer (F1) and `ACAACCCGTCTCTTAAGTCTCGT` as reverse primer (R1). Note that the forward primer is opposite to the V_H reading frame; this choice was made to maximize the reading quality of CDR3 (as the reading quality is highest on forward reads and on the first nucleotides along a sequence read). The Hamming distances between the library-specific FWR3 sequences are $d_H(\text{Germ}, \text{Lmtd}) = 10$ nt, $d_H(\text{Lmtd}, \text{Bnab}) = 25$ nt and $d_H(\text{Germ}, \text{Bnab}) = 22$ nt, sufficient to discriminate between them

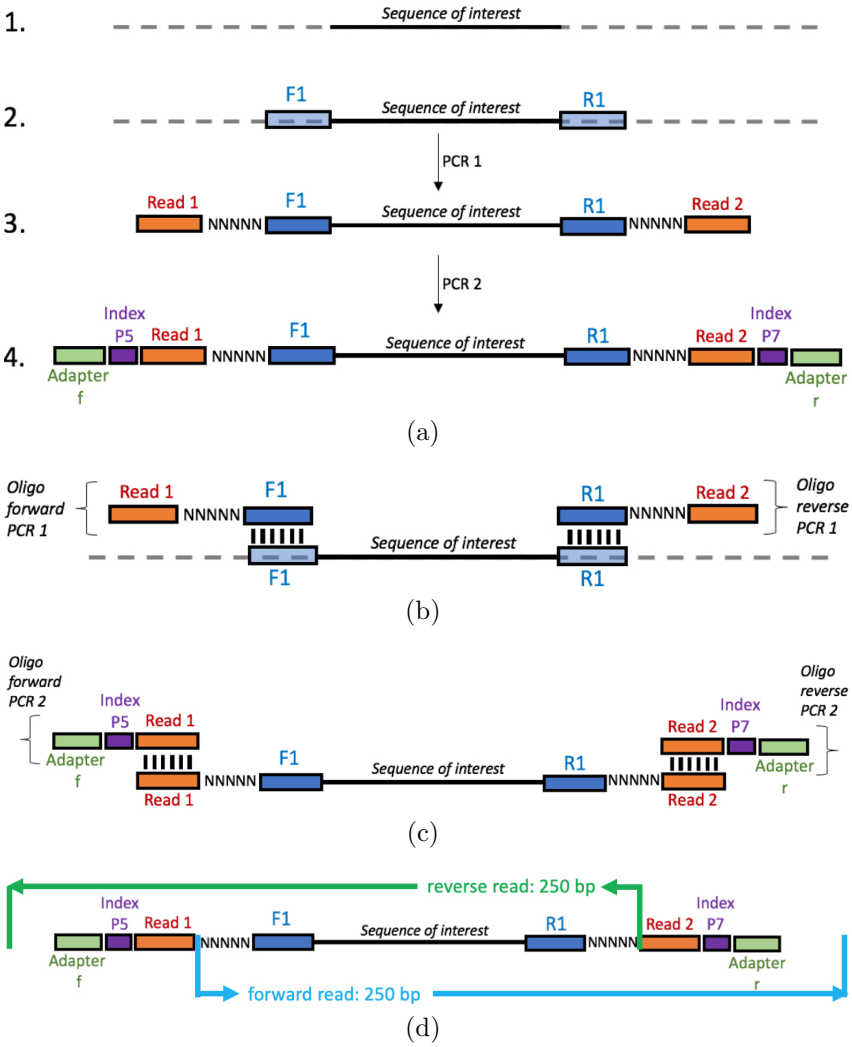


Fig. 3.11: The amplicon for Illumina MiSeq sequencing. Illustrations by Mégane Matysiak, adapted. **(a)** 1. A region of interest must be defined along the DNA sequence, under the respect of certain criteria specified in the main text (notably sequence length). 2. Primer sequences flanking the region of interest must be defined (here labelled F1 and R1), again under the respect of contraints given in the main text. 3. A first PCR reaction (PCR1) amplifies the region of interest from the plasmid while adding cluster barcode sequences 5 bp in length (`NNNNN`) and primer sequences for a second PCR reaction and the sequencing procedure later (Read1 and Read2). 4. A second PCR reaction (PCR2) amplifies the product of the first while adding sample-specific barcodes (P5 and P7 indices), as well as adapter sequences (adapter f and r), thus yielding the final sequencing amplicon. The adapter are used to immobilize the amplicon during the sequencing procedure by annealing to a surface that displays the complementary adapter sequences. **(b)** The forward and reverse degenerate (on the 5 random positions) oligo required for PCR1. **(c)** The forward and reverse oligo required for PCR2, one forward (or reverse) oligo per P5 (or P7) barcode sequence. **(d)** Content of the forward and reverse paired-end sequencing reads.

```
                CDR2                          primer rev PCR1     FWR3
                ------------------------------------------------
                                                                  --------------------
                                              ---------------------->
germ            GTTCTATCTACTACTCTGGTTCTACCTACTACAACCCGTCTCTTAAGTCTCGTGTTACCATCTCTGTTGA
lmtd            GTACCACCTACTACTCTGGTAAAACCTACTACAACCCGTCTCTTAAGTCTCGTGTTACCATCTCTATCGA
bnAb            CGTCTTACTGGAACCGTGGTTGGACCTACCACAACCCGTCTCTTAAGTCTCGTCTGACCCTGGCGCTGGA
germ <-> lmtd     - - -             ---                                              - -
germ <-> bnAb   --   --  ---  --     --      -                       - -   - -- -- -
lmtd <-> bnAb   --- ---  ---  --    ---      -                       - -   - -- -- -

                FWR3
                ------------------------------------------------------------------------
germ            CACCTCTAAAAACCAGTTCTCTCTGAAACTGTCTTCTGTTACTGCGGCGGACACTGCGGTTTACTACTGT
lmtd            CACCTCTAAAAACCACTTCTCTCTGCGTCTGATCTCTGTTACTGCGGCGGACACTGCGGTTTACCACTGT
bnAb            CACCCCGAAAAACCTGGTTTTCCTGAAACTGAACTCTGTTACTGCGGCGGACACCGCGACCTACTACTGT
germ <-> lmtd                -         ---   ---                                 -
germ <-> bnAb      - -       - - - --         ---                    -   ---
lmtd <-> bnAb      - -       --- - --   ---    -                     -   ---   -

                FWR3  CDR3                FWR4          primer fwd PCR1
                ------                    ---------------------------------
                      ---------------------             <------------------
germ            GCGCGCNNNNNNNNNNNNTTCGACTACTGGGGTCAGGGTACCCTGGTTACCGTCTCGAGCGGTGGAGGCG
lmtd            GCGCGCNNNNNNNNNNNNTTCGACTACTGGGGTCAGGGTACCCTGGTTACCGTCTCGAGCGGTGGAGGCG
bnAb            GCGCGCNNNNNNNNNNNNTTCGACTACTGGGGTCAGGGTACCCTGGTTACCGTCTCGAGCGGTGGAGGCG

germ            GTTCAGGCGGAGGTGGCTCTGGCGGTAGTGCACAGGTCCAACTGCAGGAGCTCGATATCAAACGGGCGGC
lmtd            GTTCAGGCGGAGGTGGCTCTGGCGGTAGTGCACAGGTCCAACTGCAGGAGCTCGATATCAAACGGGCGGC
bnAb            GTTCAGGCGGAGGTGGCTCTGGCGGTAGTGCACAGGTCCAACTGCAGGAGCTCGATATCAAACGGGCGGC
```

Fig. 3.12: The region of the V_H sequences targeted by PCR reactions and Illumina sequencing, comprising essentially FWR3, CDR3, and FWR4. This region is flanked on both sides by the forward and reverse primer sequences (indicated by arrows) for the first of two preparatory PCR reactions (PCR1). These primer sequences are common to all three scaffolds. Note that the primer which defines the forward direction of the Illumina sequencing is opposite to the V_H reading frame. The region of interest including the primer sequences is 170 bp in length. Mutations in DNA sequence between the three scaffolds used later for scaffold identification are indicated below.

even in the presence of sequencing errors.

The detailed protocol for amplicon generation is provided in section A.7. In short, we first defrost glycerol stocks of library cells at relevant selection cycles and extracted the plasmids using purification kits (Macherey-Nagel, Düren, Germany). No liquid cultures are performed prior to plasmid extraction to avoid potential additional biases in frequencies that may arise from growing an overnight culture beforehand. The resulting plasmids as well as primers of the form shown in figure 3.11(b)) were used as input for the first Illumina sequencing preparation PCR (PCR1): This adds the random cluster barcode sequences and the read 1 and read 2 sequences. The sequence of the product of PCR1 is 247 bp in length and is shown in figure E.1(a). The read 1 and read 2 sequences are targeted as primer sequence for the second PCR reaction (PCR2): We amplify the product of PCR using primers of the form shown in figure 3.11(c) which adds the P5 and P7 indices, as well as the forward and reverse adapter sequences. The P5 and P7 indices encode for the sample from which the amplicon originates: For each library, target and selection round, we use a different combination of P5 and P7 indices. A list of P5 and P7 barcode combinations used is provided in table D.3. The product of PCR2 is the final amplicon with sequence shown in figure E.1(b) which is 316 bp in length.

After each PCR reaction, the PCR products are gel-purified before going ahead: We subject the PCR products to electrophoresis on agarose gels, in parallel to a ladder sample containing DNA of referenced sizes. The PCR products are treated with a DNA intercalating dye that is excited and emits in the visible part of the spectrum (to avoid UV-induced DNA damage). Under blue light, the signal of the amplicon was checked for the expected size by locating the bands (fluorescence signals) of the PCR products with respect to the ladder. Examples of gel electrophoresis of PCR1 and PCR2 are shown in figure 3.13(a) and (b), respectively. Sometimes, weak bands or smear is observed off the expected size. We excise the main bands at the expected amplicon sizes from the gel and purify the PCR products by agarose gel purification kits (Macherey-Nagel, Düren, Germany). The purified product of PCR1 was used as template for PCR2. The purified products of PCR2 were mixed together to obtain the final amplicon mix. To know in which volumic proportions the samples have to be mixed, the DNA concentrations of all samples are measured using a Qubit fluorometer. The concentrations and projected number of sequencing reads (and amplicon sizes if they differ between samples) then determine how much of each sample has to be added to the final amplicon mix. These calculations are conveniently performed using an excel file; an example is shown in figure E.2. A final check for correct amplicon size is performed by running the amplicon mix on a TapeStation (a kind of high precision electrophoresis; from Agilent Technologies). An example of a fluorescence profile as a function of DNA size, with a peak at the expected amplicon size is shown in figure 3.13(c), (d).

An independent quality control, as well as the MiSeq sequencing itself, as well as the demultiplexing, *i.e.* assigning sequencing reads to samples based on P5 and P7 barcodes, were performed at I2BC, Gif-sur-Yvette, France. For this purpose, the amplicon mix, as well as a table of sample names and their respective P5 and P7 barcodes had to be handed to the the sequencing platform

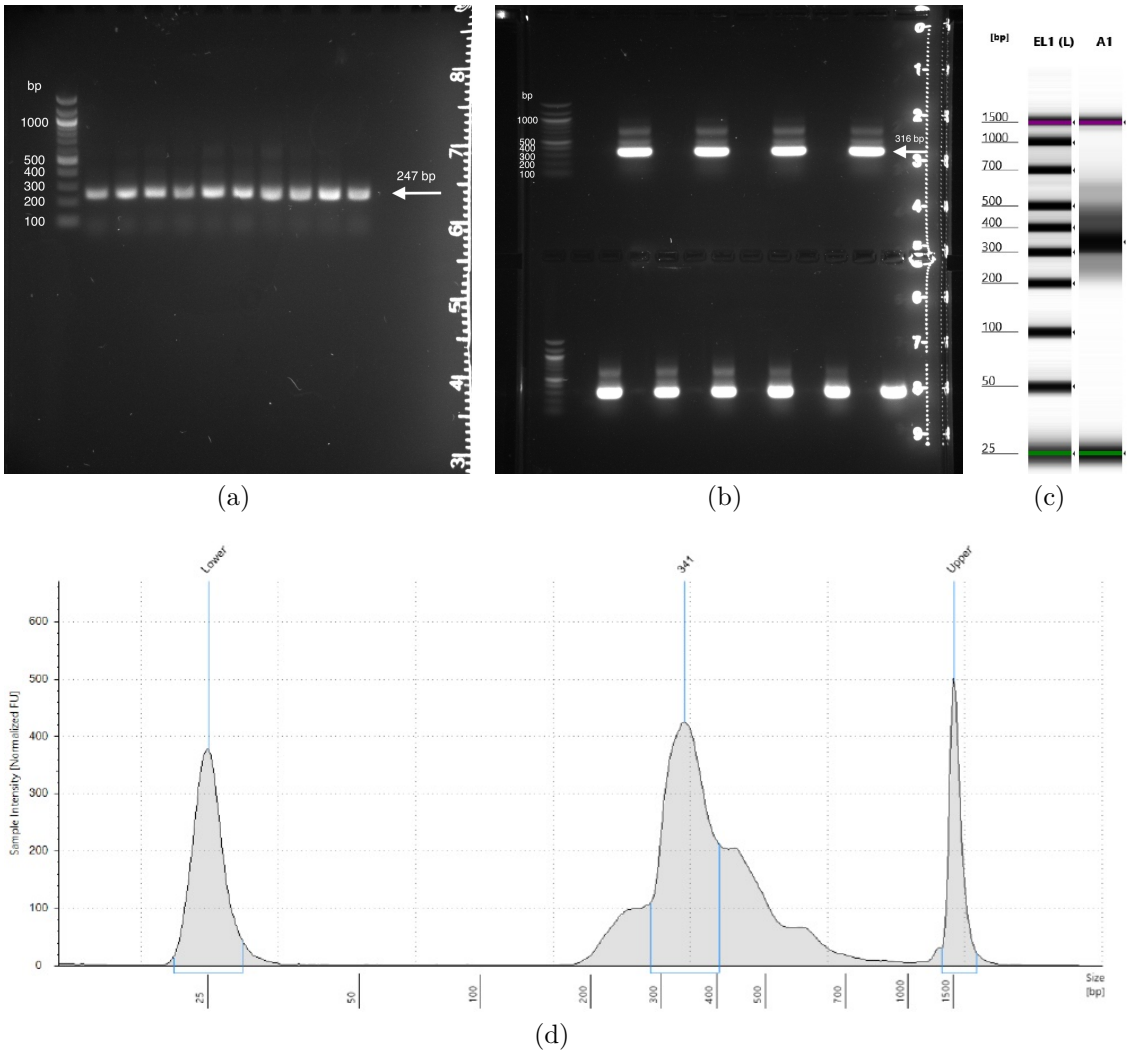


Fig. 3.13: Preparation of libraries for Illumina MiSeq sequencing. **(a)** Agarose gel of the products of the first PCR reaction targeting the DNA region of interest on the plasmid and adding random barcodes. Different lanes contain various samples corresponding to different library-target combinations and selection rounds. The leftmost lanes contain a 1 kb ladder sample with dsDNA of known sizes of 100, 200, 300, ... bp. All samples show strong fluorescence signal at the expected size of the amplicon of ≈ 250 bp. Unspecific signal around this size is removed by gel purification. **(b)** Agarose gel of the products of the second PCR reaction adding sample-specific barcodes and the Illumina adapter sequences to obtain the final amplicon. The strongest bands appear at the expected size of ≈ 320 bp. Once again, unspecific bands are eliminated by gel purification. **(c)** Tape Station (high-precision gel) run of the final gel-purified amplicon mix in which all samples are pooled together. Left lane: ladder, right lane: amplicon mix along with upper and lower markers at 25 bp and 1500 bp. **(d)** Fluorescence intensity profile of the right lane in (c). The signal peaks at 341 bp which is within 10 % of the expected 316 bp, a typical error for gels. The samples shown here correspond to selections of the library mix against the DNA1 and DNA2 targets, as well as the first replica of the mini-library selections against the DNA1 and DNA2 targets.

staff.

3.5.3 Sequencing data preprocessing and availability

The Illumina sequencing yields for each sample (i.e., each library, target and selection round) between 10^5 and $5 \cdot 10^6$ sequencing clusters and paired reads. The sequencing data is organized into two files per sample containing respectively the forward and reverse reads for all clusters, which in our case are entirely overlapping, also see figure 3.11(d). The forward and reverse data thus provide two (more or less) independent readings of the same sequences and are expected to be identical up to reverse-complementarity and sequencing errors. In addition, each identified nucleotide is tagged by a quality read Q that encodes for the confidence on the correctness of the given nucleotide according to the Illumina machinery. Q is given as an ASCII character comprised between `!`, `?`, `#`, ..., `G`, `H`, `I` that represents an integer between (including) 33 and 73 and translates into an error probability p_{err} via

$$p_{\mathrm{err}} = 10^{-\frac{Q-33}{10}}. \tag{3.3}$$

For the following preprocessing of the sequencing data, we have first reverse-complemented all the forward reads (which are themselves opposite to the V_H reading frame) so as to have both forward and reverse reads in V_H reading direction. The python code used for the proprocessing is provided as supplementary material in section F.1.

For the purpose of sequencing data analysis, we first subjected the raw data to a "cleaning" step that notably selects for the presence and correct length of the V_H region of interest and its primer and restriction site sequences, as well as for sufficiently high average quality read Q. Each cluster was accepted or discarded based on the following simple procedure: First, both the forward and reverse reads were screened for the presence of the primer sequences F1 and R1 (up to 4 nt mismatch accepted for each) and cut to keep only the part between the primers (including the primers) which corresponds to the region of interest containing FWR3, CDR3, and FWR4. Either one the two reads was discarded if the primer search was unsuccessful; the whole cluster was discarded if primer search was unsuccessful on both reads. Second, we checked if the forward and/or reverse sequence fragments of the remaining cluster had the expected length of 170 nt. 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 Q. 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 (positions 108-116 and 129-137 in reverse direction, repectively; up to 4 nt mismatch accepted for each). The cluster was discarded if either one of these two criteria was not fulfilled. The output of this procedure are

sequences of length 170 bp (one per cluster) and their associated quality reads which are written to files named `<sample name>_cleaned.txt`. The typical yield of this procedure is $95-99\,\%$, *i.e.* $< 5\,\%$ of the clusters are discarded as they fail to any one of the above criteria. Exceptions where the yield is lower occur when the Germline library is selected (alone or in mixture with others) against the DNA2 and protein targets where contaminant CDR3 sequences with respectively 7 aa (sequence: `RGGGRRF`) and 3 aa (sequences: `GPA` and `GPM`) rather than 4 aa appear and which are projected out by the region length requirements.

In a second preprocessing step, we determined the sequence identities for all of the remaining clusters. This task consists in identifying (i) the scaffold (Germ, Lmtd, or BnAb) and (ii) the CDR3 sequence identity. Task (i) was performed by computing the Hamming distance of the library-specific FWR3 (of length 116 nt) to all three scaffold reference sequences. The read was assigned to the nearest scaffold if the Hamming distance to the nearest scaffold was ≤ 7 nt *and* the difference in Hamming distances between the nearest and next-nearest scaffolds was ≥ 3 nt. For task (ii), the CDR3 sequences were simply extracted from the read in the case of samples from full-library experiments. For the mini-library selections with reduced CDR3 diversity a similar procedure as for FWR3 was applied to CDR3: the measured CDR3 sequence was assigned to the nearest among $\simeq 20$ reference CDR3 sequences if the Hamming distance was ≤ 3 nt and the difference in Hamming distances between nearest and next-nearest was ≥ 1 nt. After assessment of the V_H sequence identity of all clusters in a dataset, the CDR3 sequences were translated into amino acids and the number of occurrences of each V_H sequence identity (determined by its scaffold and CDR3 sequence identities in either nucleotides or amino acids) was counted. The results were stored in files named `<sample name>_counted_nt.txt` and `<sample name>_counted_aa.txt`. For the mixed full-library selections, these final data files contain three columns: 1) scaffold identity ('`germ`' for Germline, '`lmtd`' for Limited, '`bnAb`' for BnAb, '`????`' if scaffold inference failed), 2) CDR3 identity given by the sequence of either 4 amino acids or 12 nucleotides ('`????`' is given if the CDR3 readout failed), 3) the number of counts of this sequence in the sequencing data of the corresponding sample. The sequence identities are sorted in decreasing order with respect to their number of occurrences in column 3).

Finally, we checked that the results are unaffected by the choice of the various parameters in the preprocessing described here. In total, we analyzed in this way sequencing data from three full Illumina MiSeq 2×250 bp sequencing runs with a combined total of $\simeq 4.2 \cdot 10^7$ sequencing reads distributed over 45 samples that were generated throughout the PhD project. Note that the procedure described here only applies to this new sequencing data. Data from previous selection experiments, some of which were reported in [1], was obtained on the basis of a different amplicon design and slightly different preprocessing. This data was reused in its preprocessed form for our analysis. Their amplicon design included all of the V_H gene but non-overlapping paired-end sequencing reads and made use of the same forward primer (F1) but a different reverse primer sequence (R1) that is located upstream of FWR1. See [133] for more details on their data analysis.

We deposited the raw high-throughput sequencing data from selection experiments performed

during this PhD project, as well as from previously reported experiments [1] to the NCBI Sequence Read Archive with respective SRA accessions PRJNA592656 ($\simeq$ 57 GB in size) and PRJNA600801 ($\geq$ 1.5 GB in size). The preprocessed high-throughput sequencing data from all (new and previous) selection experiments ($\simeq$ 220 MB in size) is available through a shared Dropbox folder.

Chapter 4

Inference of selection potentials from high-throughput sequencing of V_H libraries

The matter of this chapter is the analysis of selections of the V_H libraries in light of the biophysical models discussed in chapter 2. From selections that combine the three V_H libraries and four targets defined in chapter 3, this leads to the following main conclusion: The parameter σ, *a priori* different for different antibody-target combinations, is independent of the target and thus a property of the V_H library. Moreover, σ appears to decrease with increasing degree of maturation of the underlying V_H scaffold of a library, suggesting that the Germ library, which is based on a naïve human antibody, maximizes selection potential.

First, we will provide a brief summary of the experiments performed as a part of this project, notably selections of the three libraries Germ, Lmtd, and BnAb each one separately and altogether in a mixture against each of the four targets. General features of selection trajectories are discussed, including optimality and reproducibility of enrichments, as well as unspecific binding, and amplification biases (section 4.1). Given histograms of empirical enrichments from sequencing data, we ask for lognormal and generalized Pareto distributions that best describe these empirical enrichment distributions $P(s)$. The model parameters are inferred by a simple MLE approach and the quality of fit is assessed graphically by quantile-quantile plots and probability-probability plots (section 4.2). The ensemble of model parameters for the various experiments is plotted and interpreted, which leads to the aforementioned conclusions. These simple models are confronted with mini library measurements to check for consistency and with library frequencies, *i.e.* observables that were not used for model inference, to assess their predictive power (section 4.3). Finally, we plot sequence logos based on frequencies and enrichments which reveal target specificities, and perform "crossed" selections on mini libraries to also reveal antibody specificities (section 4.4).

Beyond the study of enrichments, we also propose extensions to our simplistic (but sufficient for the purposes of this project) biophysical model inference from high-throughput sequencing data of selected combinatorial libraries (section 4.5). These should reveal more insights into our model system and its underlying interactions using the same data presented here again and are likely to be required for the modeling of any selection beyond unimodal, additive binding.

In the main text of this chapter, we will show examples of plots that have been generated for a number of different selection experiments. A more complete enumeration of these figures for various datasets are provided as supplementary figures in chapter E.

4.1 Selection trajectories

In this introductory section, we will provide a brief summary of the entirety of selection experiments performed as part of the PhD project and that are the basis of most of the results presented in later sections (subsection 4.1.1). Then, we will showcase at a single representative example the effect of the selection step on a library and at the level of sequencing data before and after selection and motivate the computation of empirical enrichments from sequencing counts (subsection 4.1.2. These empirical enrichments computed from sequencing data are reproducible and thus meaningful (subsection 4.1.3), and we verify by phage ELISA that differences in enrichments are related to differences in binding affinities (subsection 4.1.4). Finally, we show the existence of non-zero minimal enrichments as a signature of non-specific binding which defines a lower bound to the binding probability (subsection 4.1.5). This minimal enrichment differs between libraries but not inside libraries, suggesting that the level of unspecific affinity to non-cognate targets may be determined by the scaffold and its degree of maturation.

4.1.1 Summary of selection experiments performed

We performed and sequenced selection experiments involving the three V_H libraries Germ, Lmtd, and BnAb (see section 3.1) and the four binding targets DNA1, DNA2, prot1, and prot2 (see section 3.3) in three different selection contexts (i) to (iii) (see section 3.4): (i) We selected from a uniform mix of the three libraries (which we call Mix3) independently against all four targets. The selections against the protein targets were performed in two replicates. The two replicates are used to conclude reproducibilty on empirical enrichments. (ii) Moreover, each of the three libraries was also selected separately against the two DNA targets. (iii) Finally, we identified, cloned, and pooled the most enriched sequences from these separate selections to construct two mini libraries containing $\simeq$ 10 top DNA1- and $\simeq$ 10 top DNA2-specific sequences, respectively. In addition, we prepare a third mini library by pooling $\simeq$ 10 randomly picked sequences from the full libraries. We then pool the the top mini libraries with the random mini library and

selected against the corresponding cognate DNA target in order to directly measure differences in enrichments between top and random sequences from all three libraries. We also pool the two top mini libraries and select independently against both DNA1 and DNA2 to learn about specificities and cross-reacticity of the top enriched sequences. In control experiments, we selected these mini libraries against naked beads (without targets) and against nothing (no beads, no targets). As in most other phage display contexts, we perform three to four rounds of selection for full libraries, while a single selection round is found to be sufficient for the mini libraries. Using full libraries, the first 1 to 2 selection rounds are typically still dominated in frequency by bad binders. The full library selections (i) and (ii) are used to infer parameters in models for $P(s)$, the distribution of enrichments in libraries, that were discussed in chapter 2. The direct measurement of the mode and maximum of $P(s)$ by mini library also allows for a consistency check.

We also consider previously published selection data [1] for a re-analysis under the lognormal model for $P(s)$ and comparison with the new data. This includes in particular selection experiments in which i) the three libraries Germ, Lmtd, and BnAb were selected in the context of 20 other V_H libraries (Mix24) of the same design but using other natural V_H scaffolds from various species (nurse shark, frog, *etc.*), ii) those 20 were selected together but in absence of the three libraries studied here (Mix21).

4.1.2 Characteristics of selection trajectories and optimality of inference

The effect of iterated selection on a library is showcased on two examples of selection trajectories in figure 4.1 which compares the number of counts of CDR3 sequences at round t with those at round $t+1$ for $t = 0, 1, 2(, 3)$. These examples correspond to Germ selected alone against DNA1 and the Germ part of Mix3 selected against DNA1, where 3 and 4 selection rounds have been performed, respectively. Each dot thus represents a single CDR3 sequence in the context of the Germ scaffold. The presence of selection is clearly visible from this representation: As an example, the selection is particularly strong in the first selection round of Germ alone. There are sequences x with $n_0(x) < 10$ but $n_1(x) > 10^4$ counts in sequencing datasets of approximately equal size at $t = 0, 1$, meaning that they were strongly enriched. By the end of the selection trajectory at round $t = 3$, the library consists of essentially sequences with the maximum selection probability and there is no effect of further selection (sequences are no longer enriched over others as all points are concentrated along the diagonal $y = x$). These findings are in line with Fisher's fundamental theorem of selection which relates the strength of selection response to the (population) variance of enrichments and predicts that selection stops as there is no more variance in enrichment values (*i.e.* diversity). The same selection effect is observed for the Germ part in Mix3, but the intra-selection selection response seems delayed, likely as a result of simultaneous selection at the inter-library level between Germ and the other libraries.

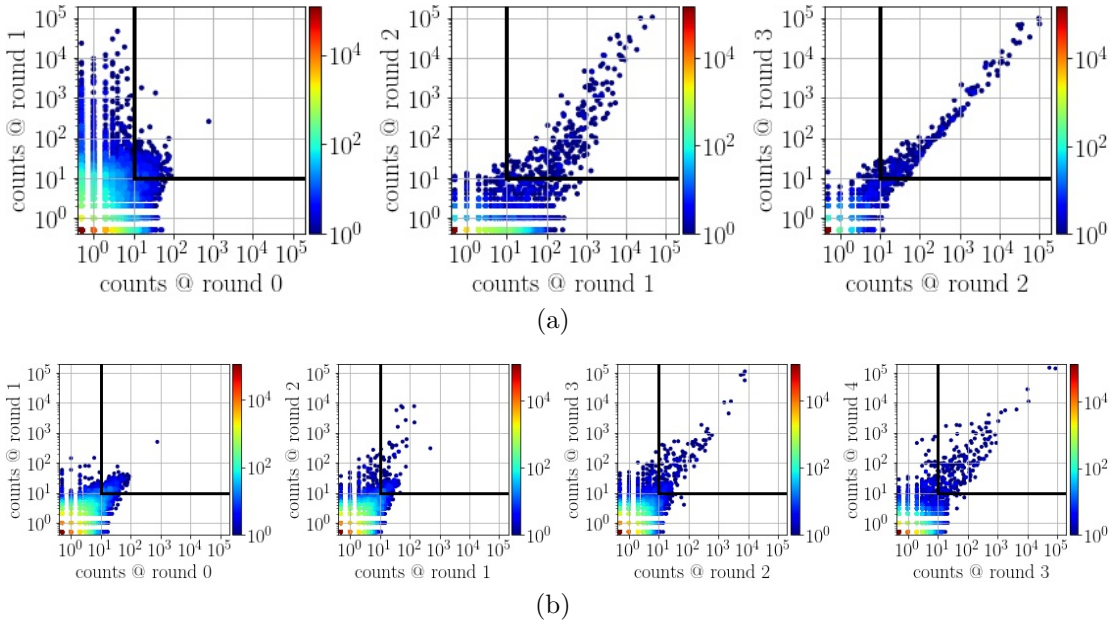


Fig. 4.1: The effect of selection on a library, as shown by directly comparing frequencies of sequences at consecutive rounds of selection: For all $q^L = 1.6 \cdot 10^5$ sequences, the number of counts at round $t+1$ is plotted against the number of counts at round t. Each point represents one unique sequence. The color encodes the density of points. The window in which enrichments can be computed ($n_{t+1}(x) \geq 10$ and $n_t(x) \geq 10$) is delimited by bold black lines. The example shown is the Germline library selected against the DNA1 target, **(a)** alone, **(b)** in a uniform mix with Limited and BnAb. A complete enumeration of count plots from all experiments can be found in E.

This has implications for the computation of empirical enrichments from sequencing counts. On the one hand, we have argued in section 3.4 that enrichments computed from sequence counts as $s(x) = \lambda_t^{-1} n_{t+1}(x)/n_t(x)$ are meaningful and non-random only if the count numbers are sufficiently high. The window in which enrichments can be computed confidently ($n_{t+1}(x) \geq 10$ and $n_t(x) \geq 10$) is highlighted in figure 4.1 by bold black lines. In early selection rounds, most sequences of the specific signal are outside this window, as they are not sufficiently enriched yet. On the other hand, non-optimal binders become increasingly depleted as selection progresses and enrichments computed in late selection rounds thus do no longer represent well the diversity of enrichment values of the initial libraries. As a consequence, there exists an optimal selection round for the computation of empirical enrichments, which in our case is located between rounds $t = 1$ and $t + 1 = 2$ or between $t = 2$ and $t + 1 = 3$. Note that this is a limitation due to finite sequencing depth: Increasing sequencing depth would lift sequences with low or no counts to values ≥ 10, thus allowing to compute more enrichment values. Future perspectives of this approach may consider improved sequencing strategies.

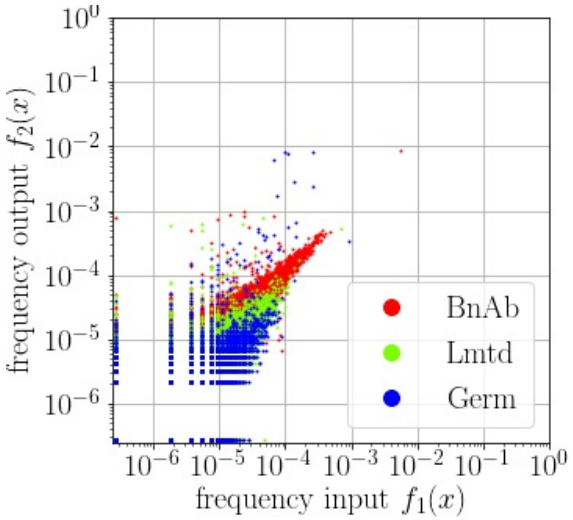


Fig. 4.2: Direct comparison of the level of unspecific binding across the libraries. The frequencies between rounds $t = 1$ and $t+1 = 2$ of selection are compared for the three libraries selected in mix against the DNA1 target. All three exhibit bulks of sequences concentrating alongs lines of slope one, thus representing large numbers of unspecifically (irrespectively of CDR3 sequence) binding sequences. (Unity slope, *i.e.* $\ln(f_2) = \ln(f_1) + \text{const}$, implies $s \propto f_2/f_1 = \text{const}$.) They differ in the value of the unspecific binding probability (as indicated by their vertical stacking). This unspecific binding probability appears to increase with the maturation level of the library.

4.1.3 Library-dependent levels of unspecificity

The overall selection signal that we obtain is typically a superposition of specific (CDR3 sequence-dependent) and unspecific (CDR3 sequence-independent) signal. The figures 4.1 and 4.2 reveal the existence of a non-zero minimal enrichment value that no sequence can go below: Its signature is the accumulation of many sequences around a line of slope one in plots of $n_{t+1}(x)$ against $n_t(x)$ (in log-log scale). This implies that in addition to binding, there is one (or are several) alternative strategies to survive the selection protocol which do(es) not depend on CDR3 sequence. The most likely explanation is the presence of one (or several) unspecific binding modes allowing the antibody to bind without using its CDR3. Note that unspecific binding occurs not necessarily only to the binding target but my involve the whole complex formed by the streptavidin-coated bead and the target, as well as the selection tube made of polypropylene. From negative control selections of mini libraries in the case of beads with targets (polypropylene+, bead+, target+) against beads without targets (polypropylene+, bead+, target−), as well as without beads and without targets (polypropylene+, bead−, target−), we conclude that specific binding occurs against the target and that unspecific binding also occurs against the streptavidin of the beads but not against polypropylene. (In what concerns the specific binding mode, the relevance of the target over the bead also follows from target-dependent CDR3 pattern, *i.e.* different CDR3 sequences are selected for different targets.)

The presence of unspecific binding modifies the expression for the enrichment of sequence x, $s(x)$, to include a CDR3 sequence-independent unspecific binding energy ΔG_{us},

$$s(x) = \frac{\mathrm{e}^{-\beta\Delta G(x)} + \mathrm{e}^{-\beta\Delta G_{\mathrm{us}}}}{1 + \mathrm{e}^{-\beta\Delta G(x)} + \mathrm{e}^{-\beta\Delta G_{\mathrm{us}}}} \tag{4.1}$$

It sets a lower, non-zero bound for the enrichment given by

$$s_{\mathrm{us}} = \frac{\mathrm{e}^{-\beta\Delta G_{\mathrm{us}}}}{1 + \mathrm{e}^{-\beta\Delta G_{\mathrm{us}}}} = \frac{1}{1 + \mathrm{e}^{\beta\Delta G_{\mathrm{us}}}}, \tag{4.2}$$

independently of sequence x.

Figure 4.2 compares the level of unspecific binding between the three scaffolds; in comparison to figure 4.1, it shows the data for sequences from all three libraries (instead of only Germ) from a Mix3 selection against DNA1, which allows for a direct comparison. The bulks of nonspecific sequences of the three libraries are non-coincident which implies non-identical levels of unspecific binding across the libraries. This suggests that this unspecific mode is a property of the scaffold and that the scaffold itself (which also includes CDR1, 2) engages in binding rather than mediating the binding through CDR3. If this were true, it would relate to the "stickiness" of the scaffold, *i.e.* an affinity of the scaffold for a random target. Interestingly, the level of unspecific binding increases with the degree of maturation of the scaffold; Germ features lowest while BnAb has the highest unspecific binding. However, the comparison of only three libraries does not allow yet to draw a general conclusion in this regard. More points on a maturation trajectory should be studied in order to provide evidence.

4.1.4 Empirical enrichments are reproducible, target-dependent, and related to binding affinity

From the sequencing counts at consecutive rounds of selection, we compute enrichments for all sequences (ℓ, x) (ℓ denotes scaffold identity: Germ, Lmtd, or BnAb; x denotes CDR3 identity) with $n_t(\ell, x) \geq 10$ and $n_{t+1}(\ell, x) \geq 10$ according to

$$s(\ell, x) = \lambda_t^{-1} \frac{n_{t+1}(\ell, x)}{n_t(\ell, x)}. \tag{4.3}$$

For all other sequences with lower sequencing counts, we do not compute enrichments (which does not mean they have vanishing enrichment!) due to the relevance of sampling stochasticity. Two choices of the arbitrary factor λ_t coexist throughout this work. Choosing λ_t such that $\sum_\ell \sum_x s(\ell, x) = 1$, where the sums run over all sequences (ℓ, x) for which an enrichment $s(\ell, x)$ could be computed, is in line with [1] and inspired by the idea of $s(\ell, x)$ representing the probability of a randomly picked clone in the selected library having sequence identity (ℓ, x) rather than the binding probability of (ℓ, x) (see section 2.3). Note however that the choice of λ_t is irrelevant and

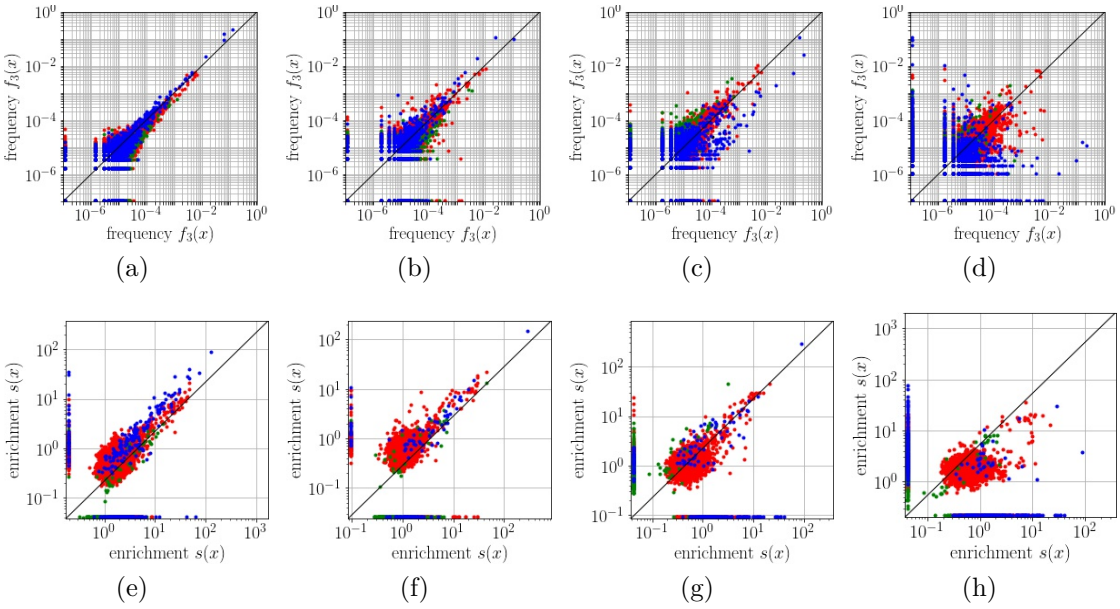


Fig. 4.3: Reproducibility of selection experiments and target specificity of selected libraries. In all plots, each point represents one CDR3 sequence, colors encode the library, **blue** Germline, **green** Limited, **red** BnAb. **(a)** Comparison of frequencies at selection round $t = 3$, $f_3(x) = n_3(x)/\sum_y n_3(y)$ of two independent replicate selections of the Mix3 library mix against the prot1 target. It shows the same sequences are selected in both replicates (points are concentrated on the diagonal) and thus reproducibility of the experiment. **(b)** Same as (a) for the prot2 target. **(c)** Comparison of frequencies $f_3(x)$ between selections of the Mix3 library mix against the prot1 (x-axis) and prot2 (y-axis) targets at selection round $t = 3$. This shows very similar sequences are selected for the protein targets, but still more difference than between replicates ((a) and (b)). **(d)** Comparison of frequencies $f_3(x)$ between selections of the Mix3 library mix against the prot1 (x-axis) and DNA1 (y-axis) targets at selection round $t = 3$ showing different sequences are selected (points far away from the diagonal to either side) and thus target specific selection response. **(e,f,g,h)** Enrichments $s(x) \propto f_3(x)/f_2(x) \propto n_3(x)/n_2(x)$ are compared for the same experiments as in (a,b,c,d). Sequences where no enrichments could be calculated in either of the two experiments are drawn to the left or lower edge of the panel.

affects in no way the results of this chapter, as long as the Boltzmann limit for s is true and the theory of sections 2.2 through 2.4 is thus invariant under rescale of s.

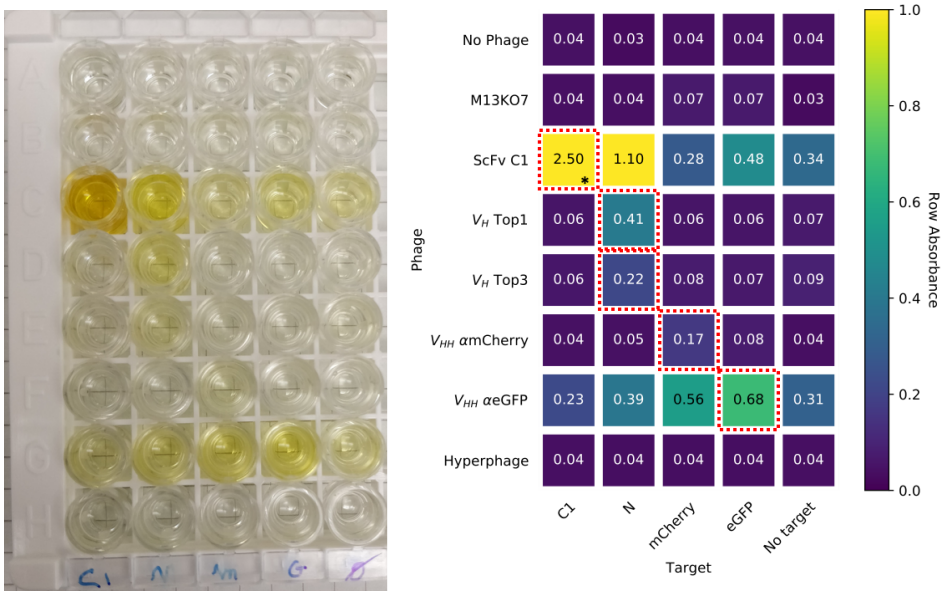
In figure 4.3, we compare frequencies and enrichments computed from replicate experiments using the same target and between experiments using different targets on scatter plots. These plots allow for two main observations and conclusions: (i) The plots show obvious correlation between replicate experiments. Thus, the selection experiments, as well as the enrichment values are reproducible and meaningful in a sense that they represent properties of the underlying antibody sequences. (ii) Comparing selections with different targets, notably prot1 and DNA1, shows weaker correlation, in particular for sequences from the Germ library. Note the sequences enriched to high frequencies for one target that are depleted to (apparent) zero frequency for the other. Thus, frequencies after several rounds of selection and enrichment values are target-dependent,

suggesting that enrichments are indeed related to different interaction mechanisms of the antibody with different binding targets. The case of prot1 versus prot2 shows evident correlation, likely as a consequence of structural similarity that allows for similar binding mechanisms in both cases. However, there are a few sequences that are relatively strongly enriched in all selections with any target; these can be explained by amplification biases unrelated to binding that favor certain sequences over others as a result of differences in replication efficiency (see next subsection 4.1.5).

We check that high enrichment is indeed related to binding affinity to the target molecule. From selections of the Germ library (alone) against DNA1, we have identified the sequences `RKKH` and `RTKH` as the ones with respectively strongest and third-strongest enrichment, whence their parallel names V_H Top1 and V_H Top3. The enrichment of `RKKH` is roughly twice the one of `RTKH`. We questioned if these antibodies were enriched as a result of affinity for the DNA1 target and tested their binding capacity to DNA1 by phage ELISA. ELISA is a binding essay in which the presence of affinity of a candidate antibody for a given target is tested: The target is presented on a surface and incubated with the antibody in question. After washing and removal of all unbound antibodies, as well as downstream treatment with secondary antibodies targeting the tag sequences between antibody and pIII (see section 3.2) and fused to the HRP enzyme (horseradish peroxidase) catalyzing a fluorescent reaction, the binding of the antibody in question to the target is revealed or ruled out by the turning-on or silence of the blue fluorescence upon adding a TMB substrate of HRP. In phage ELISA, the antibody is not presented in solution but displayed on the M13 phage. The binding essay was carried out by Guillaume Villain and the result is shown in figure 4.4 and in particular the rows $1-5$ therein; after adding hydrochloric acid which turns the blue into a yellow fluorescence, the absorbances of the samples at a blue-to-violet (complementary to yellow) wavelength were measured. Row 3 corresponds to a positive control with an antibody called ScFv C1 that strongly binds to another DNA target C1. In rows 4 and 5, `RKKH` and `RTKH` are tested against various targets including C1, DNA1, prot1 and prot2. Negative controls containing non-displaying phage (row 2) and nothing (row 1), as well as no targets (last column) were also performed. As expected, the positive control shows strong signal against C1, but also cross-specific signal for the DNA1 target and strong background fluorescence unrelated to binding (see negative control with no target). While the negative controls are all off, there is significant signal for both against only the DNA1 target. This finding essentially shows that both Germ antibodies bind specifically to DNA1, which is a rationale for their strong enrichment during the selection. Interestingly, the fluoresence signal is twice as strong for `RKKH` as for `RTKH`, possibly reflecting the difference by a factor of 2 of their enrichments. More generally, enrichments can indeed be calibrated on binding affinities measured from binding essays [59].

4.1.5 Orthogonality of binding and amplification biases

The previous subsection showed that strong enrichments are linked to binding. In another negative control experiment, we aim at studying shifts in frequencies and (contributions to) enrichments



(a) (b)

Fig. 4.4: Phage ELISA showing specific binding to their targets of top clones selected by phage display. **(a)** Binding assay in a 96-well plate revealed with anti-pVIII phage surface protein antibody. **(b)** Absorbance at 450 nm wavelength (blue to violet, complementary to the yellow fluorescence color) of the plate in (a) measured by a plate-reader. The lines correspond to different antibodies displayed on phages, the columns correspond to different targets. The first two and the last line are negative controls with no phage and non-displaying phage. The last column is a negative control with no target. Within each row (*i.e.* each antibody), the signal is constant across non-cognate and no targets which represents the background absorbance level; the signals to cognate targets are higher (probably significantly higher) than these background levels. The experiment was performed and the plot generated by Guillaume Villain [148].

that may be unrelated to binding. We perform selections without the selection for binding step but otherwise unmodified protocol, *i.e.* phage production from library cells, followed by immediate reinfection of fresh cells and growth in a solid cell culture/on a plate. This protocol involves amplification by replication of the phagemid carrying the V_H gene using the phage's replication mechanism (phage production) and the cells' replication mechanism (in the solid cell culture). To assure that we measure the amplification bias of enriched sequences, we perform the control experiment on a library mix (Mix3) selected for binding twice rather than on the initial libraries.

Figure 4.5 compares enrichments computed from experiments with and without selection for binding to DNA1. Enrichments have been computed on amino acid sequences or nucleotide sequences, as binding is a phenotype of the antibody while replication biases occur at the genotypic level. The conclusions are, however, identical in both representations: Some sequences are equally enriched with and without selection for binding, showing that these are enriched purely as a result of amplification biases. On the other hand, there are sequences more strongly enriched with selection for binding while being among the worst performing sequences without this selection

step. In consequence, binding and amplification bias are orthogonal in sequence space. This effect is particularly obvious in the case of DNA targets that select for positively charged amino acids in the CDR3 (K, R, H, see section 4.4) which appear to be the least easily replicated sequences. A closer look on the well-amplified sequences shows a small effect for V- and L-rich sequences, but a dominant effect comes from outliers (see upper right corners in panels of figure 4.5) with isolated CDR3 sequences. This suggests that these strong biases may not be related to the antibody sequence itself, but possibly elsewhere on the phagemid and outside the reading window, such as a mutation in the regulatory network of the M13 phage [166].

Note that figure 4.5 provides evidence that there is selection for binding in the BnAb library as well, even though the effect is weaker than in Germ and comparable to amplification bias: There are sequences in the upper left corners of panels in figure 4.5 that, in addition, display the same sequence patterning as enriched Germ sequences, *e.g.* K-, R-, and H-rich sequences.

4.2 Parameter inference from truncated enrichment data

The parameters of the most simple models for the distribution of enrichments $P(s)$, the generalized Pareto and the lognormal distributions, are inferred by maximum-likelihood estimation, taking into account the fact that $P(s)$ is not uniformly sampled. This requires conditioning $P(s)$ to lower threshold enrichments (subsection 4.2.1) which are themselves inferred by threshold scanning (subsection 4.2.2). Finally, the quality of fit of these simple models is assessed graphically by QQ and PP plots (subsection 4.2.3). The python code associated with this inference procedure is provided as supplementary material in section F.2.

4.2.1 Threshold-conditioned maximum-likelihood estimators

Given a list of enrichments $s(x)$ computed from sequencing data of selection experiments according to equation (4.3) and a model for the distribution of enrichments $P(s)$, we seek to infer and later compare the model parameters of $P(s)$. Two possible candidates for $P(s)$ justified in chapter 2 are the generalized Pareto distribution

$$P(s|s \geq s^*) = \frac{1}{\tau}\left(1 + \kappa\frac{s - s^*}{\tau}\right)^{-1-\frac{1}{\kappa}} \tag{4.4}$$

and the lognormal distribution

$$P(s) = \frac{1}{\sqrt{2\pi}\sigma s}\exp\left(-\frac{(\ln(s) - \mu)^2}{2\sigma^2}\right). \tag{4.5}$$

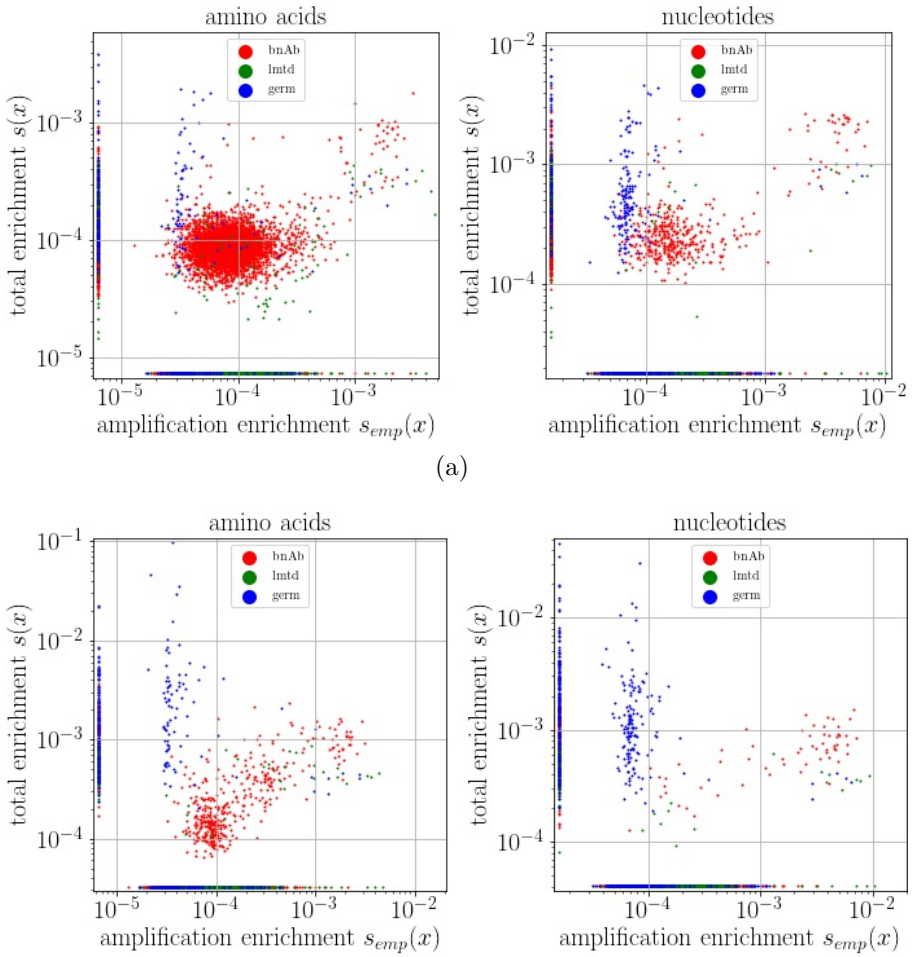


Fig. 4.5: Orthogonality of binding and amplification bias. Enrichments due to amplification s_{ampl} are computed as count ratios from before/after amplification. Comparison of total enrichments $s(x)$ (comprising binding and amplification bias) and amplification enrichments $s_{\text{ampl}}(x)$ (comprising only amplification bias). Total enrichments computed between selection rounds **(a)** $t = 2$, $t+1 = 3$, **(b)** $t = 3$, $t+1 = 4$ of the library mix (Mix3) selection against the DNA1 target are used. **Left** Enrichments are computed for amino acid sequences. **Right** Enrichments are computed for nucleotide sequences. Similar plots using total enrichments at rounds $t = 1$, $t + 1 = 2$ and $t = 3$, $t + 1 = 4$ are shown in figure E.24.

The generalized Pareto distribution from EVT is a model for the tail of $P(s)$ only, whereas the lognormal model is a global prediction for the full distribution $P(s)$.

Possibly suitable parameter values can be inferred by a standard maximum likelihood estimation (MLE) which maximizes the log-likelihood function $\mathcal{L} = \prod_x P(s_x)$ with respect to the parameters of $P(s)$. In these and the following computations, the products and sums run again only over sequences x for which an empirical enrichment value $s(x)$ is available. The challenge, however, consists in the fact that $P(s)$ is not uniformly sampled in finite sequencing data as only

high values of s can be typically observed. While this justifies the use of the generalized Pareto model for the inference precisely in the form of equation (4.4) and with log-likelihood function

$$-\mathcal{L}(\kappa, \tau, s^*) = \ln(\tau) + \left(1 + \frac{1}{\kappa}\right) \ln\left(1 + \kappa \frac{s - s^*}{\tau}\right), \tag{4.6}$$

further considerations are needed in the case of the lognormal model in equation (4.5). Its parameters must be inferred from truncated data representing only the tail of the full distribution. The inverse situation of Gaussian data where only the bulk of the distribution is sampled and its tail(s) are truncated and projected out is discussed in the literature [167]. For analytical and numerical purposes, it is beneficial to consider log-transformed enrichments $y_i = \ln(s_i)$, which should obey a Gaussian distribution with the same parameters μ and σ. We opt for a similar threshold exceedance approach as is done in the generic case to asymptotically obtain the EVT model: If restricting to values y_i larger than a given threshold y^*, the probability density $P(Y = y|Y \geq y^*)$ of observing y given that $y \geq y^*$ is

$$P(Y = y|Y \geq y^*) = \frac{P(Y = y)}{\mathbb{P}[Y \geq y^*]} = \frac{1}{1 - F(y^*)} \frac{\mathrm{d}}{\mathrm{d}y} F(y) = \sqrt{\frac{2}{\pi}} \frac{\mathrm{e}^{-\frac{(y-\mu)^2}{2\sigma^2}}}{\sigma \left[1 - \mathrm{erf}\left(\frac{y^* - \mu}{\sqrt{2}\sigma}\right)\right]}, \tag{4.7}$$

where $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{e}^{-\xi^2} \mathrm{d}\xi$ is the Gauss error function and the last equality invokes the cumulative distribution function of the Gaussian distribution $F(y)$ with

$$F(y) = \mathbb{P}[Y \leq y] = \int_{-\infty}^{y} P(y)\mathrm{d}y = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{y - \mu}{\sqrt{2}\sigma}\right)\right]. \tag{4.8}$$

The log-likelihood function $\mathcal{L}(\mu, \sigma, y^*)$ then verifies

$$\begin{aligned} -\frac{1}{N}\mathcal{L}(\mu, \sigma, y^*) &= -\frac{1}{N} \sum_{i=1}^{N} \ln P(Y = y_i|Y \geq y^*) \\ &= \ln(\sigma) + \ln\left[1 - \mathrm{erf}\left(\frac{y^* - \mu}{\sqrt{2}\sigma}\right)\right] + \frac{1}{2N\sigma^2} \sum_{i=1}^{N} (y_i - \mu)^2, \end{aligned} \tag{4.9}$$

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 $\hat{\mu}(y^*)$. In the limit $y^* \to -\infty$, we recover the log-likelihood function of the Gaussian distribution,

$$\lim_{y^* \to -\infty} \left(-\frac{1}{N}\right) \mathcal{L}(\mu, \sigma, y^*) = \ln(\sigma) + \frac{1}{2N\sigma^2} \sum_{i=1}^{N} (y_i - \mu)^2, \tag{4.10}$$

again up to irrelevant additive constants. In practice, equations (4.6) and (4.9) are optimized with respect to κ, τ or μ, σ given the value of s^* or y^*, respectively. The values of s^* and y^* are themselves fixed in a different way; the two constraints for y^* (s^*) are as follows: (i) Both $\mu(y^*)$ and $\sigma(y^*)$ ($\kappa(s^*)$) must be constant as functions of y^* (s^*) within uncertainty bars for all $y \geq y^*$ ($s \geq s^*$), *i.e.* the enrichments with $y \geq y^*$ ($s \geq s^*$) are described by unique values of μ and σ

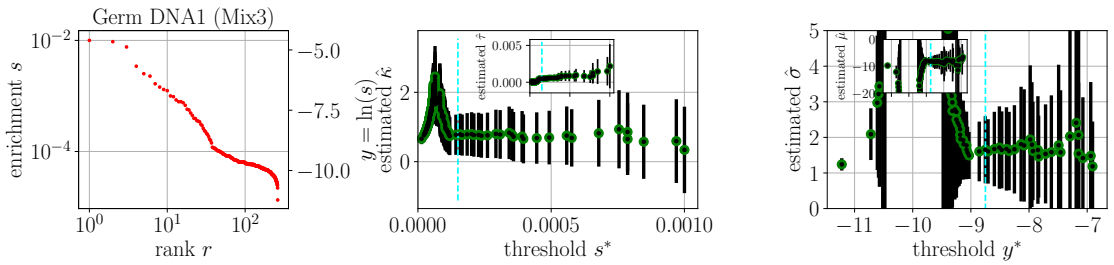


Fig. 4.6: Example of threshold scan plots showing the values of model parameters as functions of truncation values for enrichment data from the Germline library selected in Mix3 against the DNA1 target. **Left** Enrichments s and log-enrichments $y = \ln(s)$ sorted in decreasing order as a function of their rank r. **Center** ML genralized Pareto model parameter $\hat{\kappa}$ ($\hat{\tau}$ in inset) as a function of s^*. **Right** ML lognormal model parameter $\hat{\sigma}$ ($\hat{\mu}$ in inset) as a function of y^*. Error bars show 90 % confidence intervals (1.96 x the standard deviation) estimated from the Fisher information matrix and the Cramér-Rao bound. The vertical dashed cyan lines indicate the chosen values of s^* and y^* used for the inference. More examples are shown in figures E.26, E.27, E.28, E.29.

(κ). (ii) As discussed, the specific binding to the target to which the lognormal model applies is superposed by unspecific binding. Thus, y^* must be chosen to exclude unspecific enrichment values (with s such that $\ln(s) < y^*$) and to include only specific enrichments (with s such that $\ln(s) \geq y^*$). In the generalized Pareto case, the enrichments with $s < s^*$ are declared as being insufficiently far in the tale of the distribution. The lognormal distribution provides a prediction for the complete distribution of enrichments and, thus, the argument here is that enrichments with $\ln(s) < y^*$ are dominated by unspecific binding; in the absence of unspecificity, we should have in principle $y^* = \min_x \ln(s(x))$. The different interpretations of s^* and y^* imply that they are not necessarily directly related to each other, *i.e.* $y^* \neq \ln(s^*)$ in general. In practice, both thresholds can be identified in a similar way by threshold scanning. In previous work [1], only condition (i) was considered to define s^*; some cases of previously published values of s^* do not account for the presence of unspecificity and need to be modified to also satisfy criterion (ii). Finally, lower bounds on the uncertainties of the model parameter values are estimated using the Cramér-Rao bound for unbiased estimators, $\mathrm{var}_\theta(\theta_i) \geq \left(\mathcal{I}(\theta)^{-1}\right)_{ii} \geq \left(\mathcal{I}(\theta)_{ii}\right)^{-1}$ with $\theta = (\kappa, \tau)$ or $\theta = (\mu, \sigma)$ and $\mathcal{I}(\theta)$ is the Fisher information matrix, $\mathcal{I}(\theta)_{ij} = \frac{\partial^2 \mathcal{L}}{\partial\theta_i \partial\theta_j}$.

Note that the generalized Pareto model has larger representational power than the lognormal model: It represents the limiting case for all three universality classes (corresponding respectively to $\kappa > 0$, $\kappa = 0$, and $\kappa < 0$) and can therefore be expected to fit a wide range of qualitatively different data, while the lognormal distribution falls into the class of $\kappa = 0$ (asymptotically for finite data, though very slowly converging, see section 2.3). In addition to providing a prediction for the bulk of $P(s)$, the lognormal assumption is in this respect also more constraining to the shape of the tail. As found by numerical experiments (not shown), equation (4.9) is not even guaranteed to have a global maximum at finite parameter values when the data is seemingly inconsistent with the lognormal assumption.

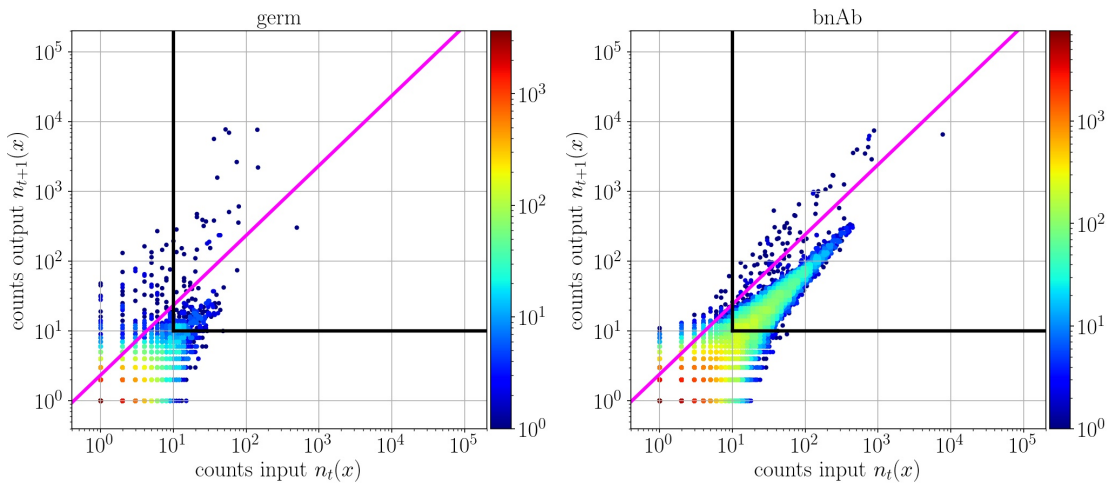


Fig. 4.7: Examples for the choice of the threshold enrichment s^* for model inference. The selection in- and output counts $n_t(x)$, $n_{t+1}(x)$ are plotted one against another along with the window defined by the upper triangle between the black and pink lines containing the sequences taken into account for model inference. The **black** and **pink** lines represent respectively the count threshold for reliable enrichment computation ($n_t(x) \geq 10$, $n_{t+1}(x) \geq 10$) and the choice of s^* (the line is parametrized by $n_{t+1}(x)/n_t(x) = s^*$). **(a)** Germline library in Mix3 against the DNA1 target, $t = 1$, $t + 1 = 2$, **(b)** BnAb library in Mix3 against the DNA1 target, $t = 2$, $t + 1 = 3$. More examples from [1] in figure E.25.

4.2.2 Threshold scanning

The goal is to identify threshold parameter y^* (s^*) such that for any $y \geq y^*$ ($s > s^*$) both $\hat{\sigma}(y)$ and $\hat{\mu}(y)$ ($\hat{\kappa}(s)$) are nearly constant within their uncertainty intervals. Such values are sufficient to satisfy the criteria (i) and (ii) in the previous subsection 4.2.1. In figure 4.6, we show the enrichment-rank plot $s_r(r)$, as well as the inferred model parameters as a function of the corresponding threshold parameter for enrichment data from selection rounds 1 and 2 of the Germ part of Mix3 selected against DNA1 (same example as in figure 4.1(b), second panel). The presence of unspecific binding creates a plateau in $s_r(r)$ and renders $s_r(r)$ non-convex, which is inconsistent with the assumption of lognormality. We observe three regimes: (i) For small s^* and y^*, the unspecific enrichments are included which leads to non-contant $\hat{\kappa}(s^*)$ and the failure of inference of μ and σ. (ii) For large s^* and y^*, too few data points lead to large uncertainty intervals on the model parameters. (iii) In between (i) and (ii), we can identify optimal threshold values $\hat{s}^*$ and $\hat{y}^*$ (vertical cyan dashed lines in figure 4.6) with minimal uncertainty bars and such that the inferred parameter values remain unchanged for $y > \hat{y}^*$ ($s > \hat{s}^*$) within confidence intervals.

The choice of y^* is again plotted together with the sequence count data in figure 4.7 (figure 4.7(a) again in the same example as in figures 4.1(b), second panel, and 4.6). In a plot of $n_{t+1}(x)$ versus $n_t(x)$, the condition $s > \exp(y^*)$ translates into $n_{t+1} > \exp(y^*)n_t$; the curve

$n_{t+1} = \exp(y^*)n_t$ is shown in magenta in figure 4.7. Indeed, the bulk of unspecific sequences below the threshold curve are excluded in the inference which uses only sequences x with $s(x) > \exp(y^*)$.

Note that this explicit exclusion of the nonspecific binding mode is required only for enrichment data from the first 1 to 2 selection rounds of a selection trajectory. In later rounds, non-specific sequences become increasingly depleted from the libraries and the sequencing data is dominated by specific binders. The gradual disappearance of unspecific sequences from sequencing data (at constant sequencing depth) with increasing selection round can be observed in the examples shown in figure 4.1.

4.2.3 Graphical assessment of quality of fit

In figure 4.8, we show examples of enrichment histograms (empirical enrichment distributions $P(s)$) together with model distributions for $P(s)$ inferred as described in the previous subsections. These examples correspond to enrichment data from rounds $t = 2$ and $t + 1 = 3$ of the Germ, Lmtd, and BnAb parts of Mix3 selected against DNA1. These show rough agreement between histograms and the simple model distributions, especially for Germ. Note, however, that our goal does not consist in representing local details of the true $P(s)$, which the simple models considered here are likely unable to account for, and we do not claim that the lognormal distribution is an accurate model for $P(s)$. Rather, our hope is to be able to capture global aspects of $P(s)$ such as its first few moments (which will be the focus of the next section 4.3).

A way to graphically assess the quality of fit are quantile-quantile (QQ) and probability-probability (PP) plots [112]. PP plots compare the empirical cumulative distribution function (CDF) which is given at y_i by $F(y_i|y^*) = \frac{i}{N+1}$ with the model CDF which for the lognormal model reads

$$z = F(y|y^*) = \mathbb{P}[Y \geq y | Y \geq y^*] = \frac{\operatorname{erf}\left(\frac{y-\mu}{\sqrt{2}\sigma}\right) - \operatorname{erf}\left(\frac{y^*-\mu}{\sqrt{2}\sigma}\right)}{1 - \operatorname{erf}\left(\frac{y^*-\mu}{\sqrt{2}\sigma}\right)} \tag{4.11}$$

and for the generalized Pareto model

$$z = F(s|s^*) = 1 - \left(1 + \kappa\frac{s - s^*}{\tau}\right)^{-\frac{1}{\kappa}}. \tag{4.12}$$

Both are estimates for the fraction of enrichment data points y_i that satisfy $y^* \leq y_i \leq y$. When the model exactly reproduces the data, *i.e.* $F(y_i) \equiv \frac{i}{N+1}$, the PP plot coincides with the diagonal $y = x$. QQ plots compare the data y_i itself with the i-th quantile of the model which is given by inverse distribution function $y = F^{-1}(z|y^*)$ with

$$y = F^{-1}(z|y^*) = \mu + \sqrt{2}\sigma \operatorname{erf}^{-1}\left[\left(1 - \operatorname{erf}\left(\frac{y^* - \mu}{\sqrt{2}\sigma}\right)\right) y + \operatorname{erf}\left(\frac{y^* - \mu}{\sqrt{2}\sigma}\right)\right]. \tag{4.13}$$

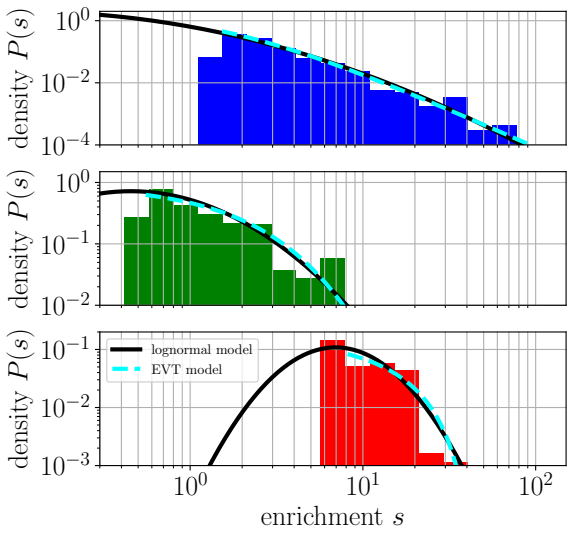


Fig. 4.8: Example of enrichment histograms plotted with the fitted generalized Pareto and lognormal models. The histogram of enrichment values $s(x) \geq \max(s^*, \exp(y^*))$ is plotted for all three libraries of the Mix3 selections against the DNA1 target at selection round $t = 2$, $t + 1 = 3$, **top, blue** Germline library, **center, green** Limited, **bottom, red** BnAb. The inferred models for $P(s)$ with the parameters from figure 4.10 are shown, **black solid** lognormal $P(s)$, **cyan dashed** generalized Pareto $P(s|s \geq s^*)$. Similar figures for the other targets and separate selections are shown in figures E.30 and E.31.

in the lognormal case and

$$s = F^{-1}(z|s^*) = s^* + \frac{\tau}{\kappa}\left((1-z)^{-\kappa} - 1\right) \tag{4.14}$$

in the generalized Pareto case, taking $z = \frac{i}{N+1}$. These provide estimates of the i-th data point (sorted) and, as for PP plots, plotting one against another yields the diagonal $y = x$ in case of perfect agreement between data and model. Both plots contain identical information, strictly, but in complementary representations, namely in probability space and data space. These spaces are mapped non-linearly into each other by equations (4.11) to (4.14). As a consequence, good agreement in one representation does not necessarily imply good agreement in the other representation. (To see this, take the example of a single outlier in data that is otherwise perfectly reproduced by the model. This changes few in PP, but generates a point far off $y = x$ in QQ.)

The QQ and PP plots for the examples of figure 4.8 are shown in figure 4.9, which shows satisfactory agreement of the Germ enrichments with a lognormal distribution and rough consistency of the BnAb data with a lognormal distribution. Deviations from lognormality in the case of BnAb may also arise from contributions of amplification biases to overall enrichments which are not neglectable here. However, mean and variance of enrichments should be correctly captured by the lognormal model even in this case. In section 4.5, we will study the validity of the independent CDR3-site assumption, which is the basis of the lognormal distribution for $P(s)$, especially for the

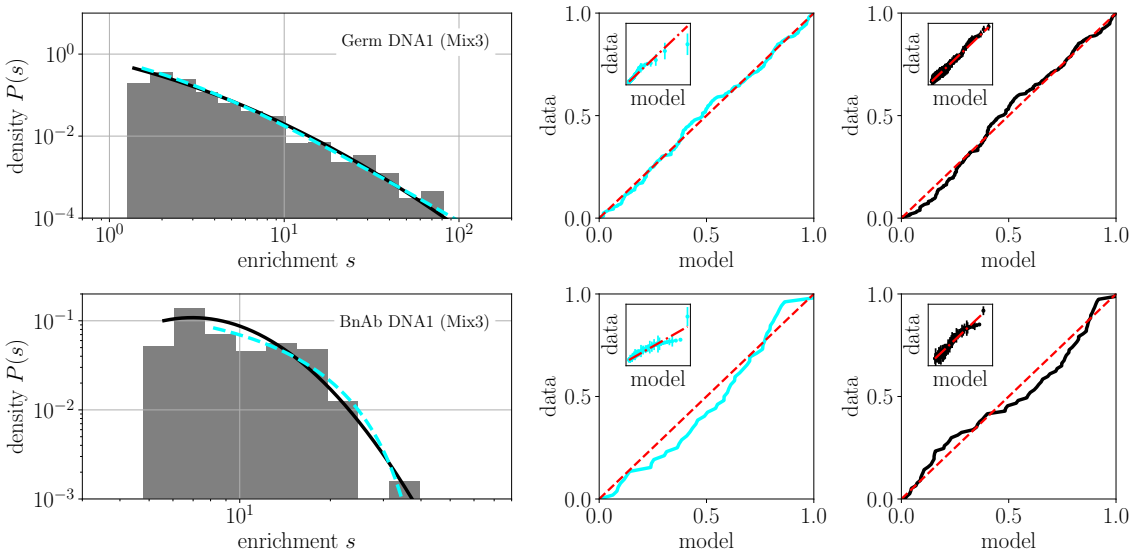


Fig. 4.9: Example of quality of fit assessment for the generalized Pareto and lognormal distributions. Germline (**top**) and BnAb (**bottom**) library selected in Mix3 against the DNA1 target, at selection round $t = 2$, $t + 1 = 3$. **Left** Histograms of enrichment values $s(x) \geq \max(s^*, \exp(y^*))$ are plotted along with the inferred model probability densities, **black solid** lognormal $P(s)$, **cyan dashed** generalized Pareto $P(s|s \geq s^*)$. **Center** PP plot and QQ plot (inset) in cyan for the generalized Pareto distribution comparing respectively the model and empirical cumulative distribution functions, and the model and empirical enrichments. **Right** PP plot and QQ plot (inset) in black for the lognormal distribution. **Red dashed** and **red dash-dotted** lines represent the expected plots in case of perfect agreement between model and data. More such plots for various experiments reported here are shown in figures E.32 to E.38.

Germ library.

4.3 Hierarchies in and between libraries are maturation-dependent, target-independent, and share a common origin

Analysis of the inferred models for $P(s)$ for the set of experiments summarized in section 4.1 leads to the following conclusions: The three libraries differ notably in their values of σ in a way seemingly correlated with the degree of maturation but irrespective of the target. The Germ (BnAb) library systematically features the maximum (minimum) σ among the three libraries, with the Lmtd library having intermediate values of σ. This reveals strong intra-library hierarchies that confer high affinity for the target to few CDR3 sequences in the Germ library (subsection 4.3.2). Different values of σ translate into different apparent shape parameters κ, showing that κ also captures intra-library hierarchies as previously suggested, but has less predictive power (subsec-

tion 4.3.1). In combination with the theory in section 2.4, the inferred values of σ and μ reproduce correctly the observed non-trivial selection dynamics in a mix of the three libraries: The Germ library, which maximizes σ and minimizes μ, eventually takes over the mix after an initial drop in frequency. This is at the basis of a hierarchy at the inter-library level which is also governed by σ and thus shares a common origin with the intra-library hierarchies (subsection 4.3.3). We directly measure maximum and mode of $P(s)$ by mini library selections with few top enriched and few random sequences and find that they are consistent with the values of σ found from full-library selections (subsection 4.3.4).

4.3.1 Parameters and intra-library hierarchies are scaffold-dependent

Lognormal and generalized Pareto models are fitted as described in section 4.2 to data from the selections summarized in section 4.1.1, typically using enrichments computed between selection rounds $t = 2$, $t+1 = 3$ and $t = 3$, $t+1 = 4$, sometimes also between $t = 1$, $t+1 = 2$. An (almost) complete listing of inferred model parameters can be found in table D.4. Earlier in the selection trajectory, the libraries are dominated by unspecific binders and too few specific enrichments are available to perform a meaningful inference. The quality of these fits is systematically challenged by QQ and PP plots, see figures E.32 to E.38. Generally, inferred models are consistent between replicates (see figure 4.10(a)), between selections of a library alone or in mixture with the other (see figure 4.10(a)), and between different rounds of the same selection trajectory (see table D.4). This is expected as enrichments, unlike frequencies, are fully determined by binding affinity and therefore independent of time t (as a reminder, "time" t here means selection round).

The parameters μ, σ of lognormal models for $P(s)$ inferred notably from Mix3 selections against all 4 targets are shown in figure 4.10(a). In Mix3 selections, the inferred values of μ can be directly compared between co-selected libraries, while this is not possible when they are selected separately due to the missing multiplicative constant λ in equation (4.3) that differs between experiments. In this case, we profitably used information from minimal library selections for calibration of μ, see subsection 4.3.4. Enrichments are here normalized such that $\mu_{\mathrm{Germ}} = 0$ as a reference value, which corresponds to a particular choice of λ.

Strikingly, the various inferred models clusterize in the (μ, σ)-plane based on scaffold identity of the underlying library; Germ, Lmtd, and BnAb libraries are all located in different regions of this plane. This has two major implications: (i) The distribution of CDR3 enrichments $P(s)$ is determined by the scaffold that displays the CDR3 sequence, but is independent of the binding target. (ii) This implies a hierarchy between libraries that holds irrespectively of the antigenic context. From figure 4.10, we observe that this hierarchy is defined by $\sigma_{\mathrm{Germ}} \geq \sigma_{\mathrm{Lmtd}} \geq \sigma_{\mathrm{BnAb}}$ but $\mu_{\mathrm{BnAb}} \geq \mu_{\mathrm{Lmtd}} \geq \mu_{\mathrm{BnAb}}$. Again, the comparison on only 3 scaffolds and 4 targets is an inherent weakness of the approach, but the above result is already significant in the biological nomenclature: The probability of observing the same hierarchy by chance (p-value) if the scaffold

had no effect, all 4 targets were independent, and each outcome were thus equally likely, is $(3!)^{-4} \simeq 7 \cdot 10^{-4} \ll 5 \cdot 10^{-2}$.

The parameter σ is associated with differences between CDR3 sequences of the same library, see section 2.3: Small σ indicates that all sequences are equally likely selected and no sequence is enriched over others upon selection, while large σ implies large differences between sequences of a library and a strong effect of selection that enriches top binding sequences over many bad sequences. Note that phenomenologically, the difference between a lognormal distribution with $\sigma \simeq 0.5$ and $\sigma \simeq 1.5$ is notable: The top enrichments among $q^L \simeq 10^5$ sequences (as in our libraries), which scales as $\exp(\mu + \sqrt{2\ln(q^L)}\sigma)$, are respectively 10 x and 10^4 x larger than a random enrichment characterized by the mode $\exp(\mu - \sigma^2)$. In figure 4.10, the more the scaffold is matured, the smaller is σ, suggesting that maturation is a key determinant of σ with the unmatured Germ scaffold allowing for a large diversity in terms of CDR3 enrichments and thus for efficient selection of strongly binding CDR3 sequences. It is characterized by strong intra-library hierarchies that favor few CDR3 sequences over most others, whereas BnAb has weak such hierarchies. These differences in enrichment spread are already visible in histograms as in figure 4.8. σ also has implications for specificities and selected sequence motifs, see section 4.4.

The parameter μ has opposite dependence on maturation degree compared to σ: It increases with the maturation level of the scaffold. These differences in μ are likely related to library-dependent unspecific binding strengths that were discussed in section 4.1.3, as they reproduce the same hierarchy of the three scaffolds in terms of unspecific binding strength shown in figure 4.2. In the theory, the lognormal model does not account for unspecific binding. In practice, however, it may be difficult to distinguish between lognormal numbers shifted by a constant (see equation (4.1)) and purely lognormal numbers with an increased μ from extreme values that sample only the tail of the distribution. But importantly, values for μ are consistently inferred from enrichments in early selection rounds where nonspecificity is important and later selection rounds where the library is already depleted of unspecific binders. To fit lognormal distributions to enrichments from early selection rounds, we exclude purely unspecific enrichments by the choice of enrichment thresholds (see section 4.2), but we do not subtract unspecific contributions from all remaining enrichments with a specific component.

While the existence of few strongly binding sequences in the Germ library appears to hold irrespectively of the target, their precise CDR3 sequences differ between targets, see section 4.4.

4.3.2 Relation between lognormal and generalized Pareto models

In figure 4.10(b), we compare σ of lognormal models with the shape parameter κ of generalized Pareto models fitted to identical enrichment datasets, along with the expectation from numerical experiments already discussed in section 2.2 and shown in figure 2.4. Here, we show the same

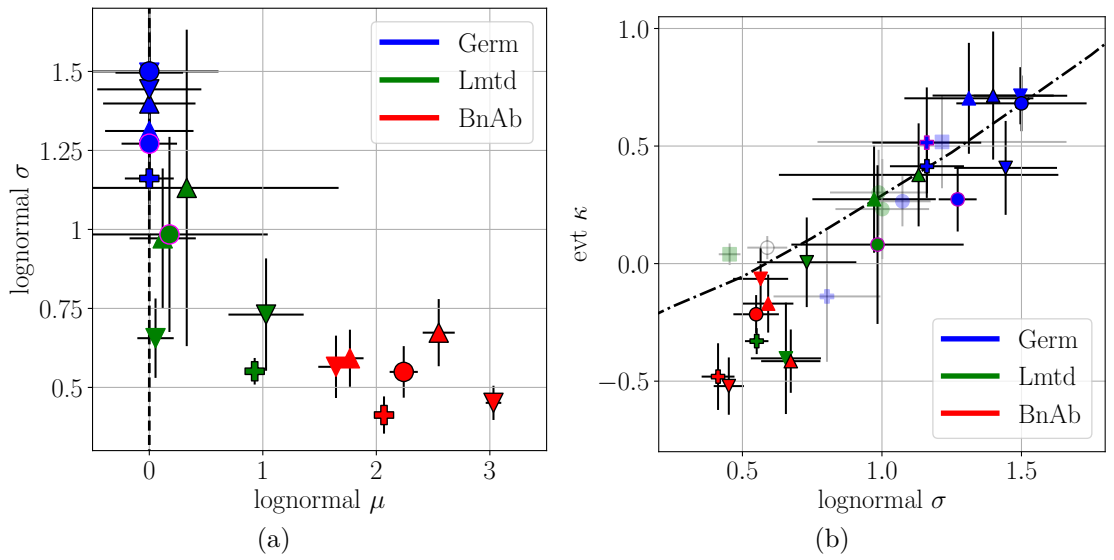


Fig. 4.10: Inferred EVT and lognormal model parameters κ, σ, μ. Different colors encode different libraries as indicated in the **legend**. Different symbols encode different targets, **circle** DNA1, **cross** DNA2, **triangle down** prot1, **triangle up** prot2. **Black encircled** and **white encircled** points are from mixed selections (different replica), **pink encircled** points are from separate selections. The error bars correspond to a single standard deviation around the maximum likelihood estimate as given by the Cramér-Rao bound. **(a)** The lognormal model parameters $\hat{\sigma}$ and $\hat{\mu}$ inferred from the library mix selections are plotted. **(b)** The EVT parameter $\hat{\kappa}$ is plotted against $\hat{\sigma}$. The behaviour is compared to the apparent $\hat{\kappa}$ as a function of σ (**dash-dotted** line) as found from a numerical experiment in which truncated iid lognormal numbers with given σ were fitted to a generalized Pareto distribution. A more complete version of (b) including values from previous experiments [1] is shown in figure E.39.

experiments as in subsection 4.3.1, as well as previously published ones [1], see also figure E.39 for a more complete version of figure 4.10(b). We observe that optimal values of κ are oftentimes non-zero. Moreover, QQ and PP plots reveal that both generally fit the data equally well. This is seemingly inconsistent with the fact that lognormal distributions fall into the Gumbel class with $\kappa = 0$ for any values of μ, σ [113]. In practice, this holds only asymptotically in the double limit $N \to \infty$, followed by $s^* \to \infty$ (where N is the number of samples from the lognormal distribution and s^* is the cut-off defined in section 4.2), which is not achieved due to finite sequence space and finite sequencing depth. However, relaxing the double-limit first to finite $s^* < \infty$ allows for negative $\kappa < 0$ according to numerical simulations; relaxing also to finite $N < \infty$ allows for positive $\kappa > 0$ predicted by power-law mimicry to be $\kappa = \sigma/(2 \ln N)^{1/2}$ [111].

Taken together, these findings provide a simple explanation for previous observations of all three classes $\kappa > 0$, $\kappa = 0$, and $\kappa < 0$ in selection data [1] in terms of σ in combination with finite-size effects. Figure 4.10(b) shows that σ maps one-to-one to κ, showing that κ is a valid measure of intra-library hierarchies as is σ, though it does not feature the same convenient physical and information-theoretical interpretations as σ (see chapter 2). Moreover, the predictive power of the

associated generalized Pareto distribution is certainly questionable as it provides a prediction on only the tail of $P(s)$, but bulk properties may be important, too, as pointed out in section 2.4.

4.3.3 Implications for evolutionary dynamics, model validation, and inter-library hierarchies

The observed pattern of lognormal model parameters μ and σ and their dependence on the maturation level of the underlying scaffold gives rise to a highly non-trivial selection dynamics: According to the findings of section 2.4, the library that maximizes μ initally increases in frequency in a uniform library mix, whereas the library that maximizes σ eventually invades the mix and wins the selection. Since we have in particular $\mu_{\text{Germ}} < \mu_{\text{BnAb}}$ but $\sigma_{\text{Germ}} > \sigma_{\text{BnAb}}$, we expect BnAb to grow before being ultimately taken over by sequences of the Germ library.

Figure 4.11 compares predicted library frequencies according to equation (2.105) with measured library frequencies from a Mix3 selection against DNA1. The discussed qualitative features of $f_t(\ell)$ are indeed observed experimentally. Deviations between theory and experiment can be explained by the numerous assumptions that led to the expression in equation (2.105) for $f_t(\ell)$, but that are not met in practice, notably the uniform distribution of CDR3 sequences in the initial libraries and the lognormal distribution being itself an approximation to the true enrichment distribution $P(s)$. Note that this prediction of frequencies from lognormal models provides a validation of the inferred lognormal distributions; these observables have not been used to establish the fits but do qualitative reproduce the observations.

The outcome of selection is thus the library that maximizes σ; this library is typically the most frequent one by the end of a selection trajectory. This finding generalizes as selections against the other target molecules show similar behaviour, see figure E.41. Thus, there exists an intrinsic hierarchy at the inter-library level encoded in the scaffolds that determines the winner library of a selection. Curiously, this inter-library hierarchy is determined by the set of σs of the competing libraries and thus by the same parameters that determine the intra-library hierarchies. As a consequence, the number of unique CDR3 sequences from the winning library still present at the end of the selection trajectory are few (compared to the initial diversity of the library), see also section 4.4; not only all other libraries but also most sequences within the winning library are selected out to give place to a few strongly binding sequences.

The selection against DNA2 is particular in the following sense: We observed strong enrichment of an antibody with Germ scaffold and a CDR3 of length 7 aa instead of 4 aa and amino acid sequence `RGGGRRF`. This is a contaminant sequence that was likely present in the purchased degenerate oligonucleotides used for library construction, see section 3.1, and carried over during library cloning and selections. While the presence of this sequence can be simply ignored for the matter of computating enrichments for all other sequences, it cannot for the computation of

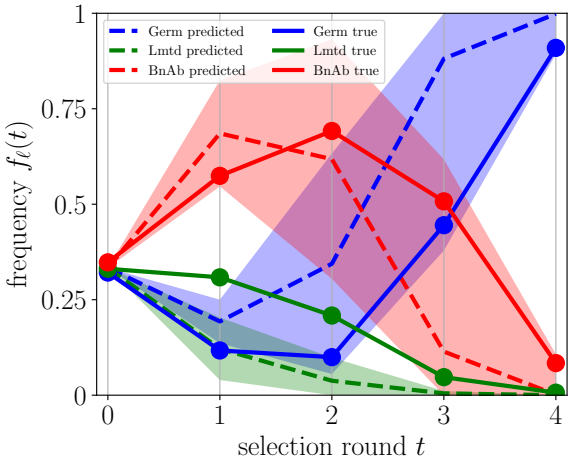


Fig. 4.11: Example of observed versus predicted selection dynamics. For the Mix3 selection against the DNA1 target, the frequencies for all three libraries (see legend) within the mix $f_t(\ell)$ is shown as a function of the selection round t. The observation (**solid**) is compared to the prediction of the lognormal model (**dashed**, shaded area corresponding to 68 % confidence intervals in the parameters μ and σ) under the assumption of initially (at $t = 0$) uniform distribution of sequences within the libraries. The same plot for Mix3 selections against target molecules are shown in figure E.41.

library frequencies. Therefore, an analysis of library frequencies as for the other three targets is not possible for DNA2.

4.3.4 Mini library selections and consistency

We seek to access the extremes and modes (most often enrichments) of enrichment distributions $P(s)$ more directly. To this goal, we construct mini libraries and identify relevant CDR3 sequences from independent selections of the three full libraries Germ, Lmtd, and BnAb against the DNA1 and DNA2 targets, and re-clone them into the corresponding pIT2-V_H phagemids. These comprise sequences among the most enriched to either target to build the mini libraries "top DNA1" and "top DNA2", as well as a few randomly picked sequences from the initial libraries to build a mini library "random". The top and random sequences are supposed to represent respectively the maximum and the mode of the enrichment distribution $P(s)$ of the full libraries and can provide a more direct estimate of σ and μ following equation (3.2). In particular, the (relative) modes can be used to calibrate selections in which libraries are selected independently one from another. The CDR3 sequences used in our mini libraries are summarized in table 4.1.

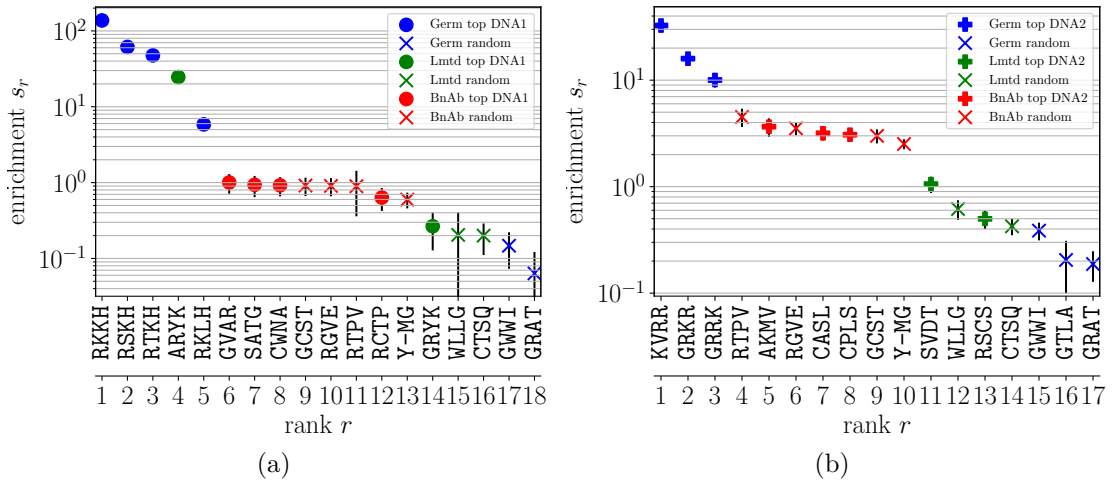
In figures 4.12 and 4.13, we show high-precision enrichments computed from sequencing counts before and after a single round of selection of mini library mixes of top DNA1 and random (top DNA2 and random) against DNA1 (DNA2). In addition, two controls were performed by

library	top DNA1 CDR3	rank r		top DNA2 CDR3	rank r		random CDR3
Germ	RKKH	1	3	KVRR	4	4	GLRS
	RSKH	2	10	KVRQ	5	7	GRAT
	RTKH	5	9	GRKR	11	1	GTLA
	RKLH	62	8	GRRR	18	8	GWWI
	RSSH	170	13	GRRK	19	3	
Lmtd	ARYH	2	2	SVDT	1	5	CTSQ
	ARYK	3	3	WAWA	2	6	GAGP
	GSHK	19	8	RSCS	3	2	GLLP
	ARHK	nb	1	EGGR	12	3	GRQL
	GRYK	nb	7	YRIE	8380	4	WLLG
BnAb	SATG	4658	18	CPLS	6	6	GCST
	VFFS	4785	14	CTVV	3151	15	GRTK
	GVAR	5635	3	FRWQ	8968	8	RGVE
	CWNA	6170	13	AKMV	nb	5	RTPV
	RCTP	7967	15	CASL	nb	2	Y*MG

Tab. 4.1: Single V_H sequences re-cloned into the pIT2 phagemid for the construction of three mini libraries: "top DNA1", "top DNA2", and "random" comprising respectively sequences selected against DNA1, DNA2, or randomly picked from the initial, unselected libraries. Successfully cloned sequences are indicated in black and are pooled together to obtain the mini libraries. Their ranks r according to enrichments s in separate selections of Germ, Lmtd, and BnAb against DNA1 or DNA2 between rounds $t = 1$, $t + 1 = 2$, as well as between $t = 2$, $t + 1 = 3$ are indicated. If no enrichment could be computed, and thus no ranking for the corresponding sequence is available, "nb" is given instead. The * in Y*MG is encoded by an amber stop codon which is sometimes expressed as Q in partial amber codon suppressor cell strains.

selecting the mix of top DNA1 and random against naked beads and in a void tube. First, these measurements are consistent with the results of the previous subsections: The top Germ sequences are $10^2 - 10^3$ x enriched over random Germ sequences, which is consistent with $\sigma \simeq 1.0 - 1.5$ in figure 4.10(a), while the differences between top and random BnAb sequences are minor, which is consistent with small σ. The random BnAb sequences are $\simeq 10$ x more enriched than random Germ sequences, consistent with a difference in μ of $\Delta\mu \simeq 2$ in figure 4.10(a).

Finally, it should be noted that these high-precision measurements of enrichments are no longer limited by sequencing depth but by the reproducibility of the selection experiment: Significant differences in BnAb enrichments (relative to Germ and Lmtd sequences) between replicates of the mini library selections, resulted from the use of different batches of magnetic beads, see figure E.40. Here, we are thus limited by the reproducibility of the target.



(a) (b)

Fig. 4.12: Mini library selections against DNA targets revealing target specificities. High-precision enrichments from libraries with around 20 different sequences are plotted in decreasing order and the CDR3 sequences are indicated. **(a)** DNA1-specific and random clones from all three libraries selected against DNA1, **(b)** DNA2-specific and random clones against DNA2. Error bars are 20 x enlarged. Reproducibility of and consistency between mini library selections is shown in figure E.40. Selections against beads in absence of targets is shown in figure 4.13.

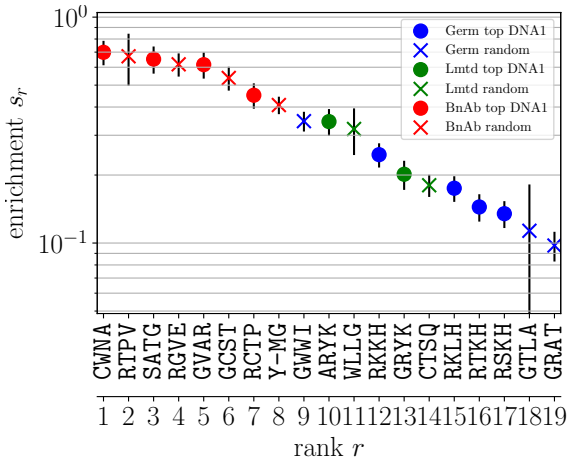


Fig. 4.13: Unspecific binding to magnetic beads. DNA1-specific and random clones from all three libraries (the same as in figure 4.12(a)) selected against magnetic beads in absence of targets. Error bars are 20 x enlarged.

4.4 CDR3 sequence motifs and binding specificities

The analysis of the previous section actually ignored an essential part of the information provided by the sequencing of the libraries: It considered only sequence counts and enrichment values, but ignored the CDR3 identities behind these enrichments. We here show sequence logos based on frequencies at each selection round (subsection 4.4.1), as well as based on enrichments which provide in principle a measure of the information content of the selection process as discussed in section 2.3 (subsection 4.4.2). We comment on the difference between two specificities that are captured by this sequence motif approach, namely the specificity of a target in light of a variety of ligands and, *vice versa*, the specificity of a ligand (CDR3) in light of a variety of targets (subsection 4.4.3). Finally, we confirm by mini library selections the specificities of Germ sequences enriched against the DNA targets, showing that they are indeed able to discriminate between DNA1 and DNA2 (subsection 4.4.4).

4.4.1 Emergence of target-specific CDR3 patterns

We plot sequence logos based on frequencies $f_{t,i}(a)$ that measure the frequency of amino acid a on CDR3 site i at selection round t and are estimated from sequencing counts $n_t(x)$ as

$$f_{t,i}(a) = \sum_x f_t(x)\delta(x_i, a) = \frac{\sum_x n_t(x)\delta(x_i, a)}{\sum_x n_t(x)}. \tag{4.15}$$

For each position i, a bar of total height

$$H_{t,i} = \sum_{a=1}^{q} f_{t,i}(a) \ln \frac{f_{t,i}(a)}{g_i(a)}, \tag{4.16}$$

with $q = 20$ the size of the alphabet and $g_i(a) = \frac{1}{q}$ independently of a, is divided into letters with heights $H_{t,i}(a)$ proportional to $f_{t,i}(a)$, *i.e.* $H_{t,i}(a) = f_{t,i}(a)H_{t,i}$. In figures 4.14 (separate selections against DNA targets), 4.15 (Mix3 selections against DNA targets), and 4.16 (Mix3 selections against protein targets), we plot such sequence logos as a function of selection round t. This illustrates how certain sequence motifs are enriched over others, which appears to happen particularly efficiently in the case of the Germ library and, sometimes, the Lmtd library. This is consistent with the observation of strong inter-library hierarchies within these libraries. The sequence logos based on $f_{0,i}(a)$ represent the initial bias in amino acid use in the initial libraries and before any selection; these are non-uniform due to differences in cloning efficiencies between CDR3 sequences (see section 3.1).

Inconveniently, these logos based on frequencies depend on selection round t, as well as on the realization of initial bias in CDR3 sequences $f_0(x)$ in the libraries. These are therefore unsuitable to represent binding properties of the underlying sequence diversity. In section 2.3, we motivated

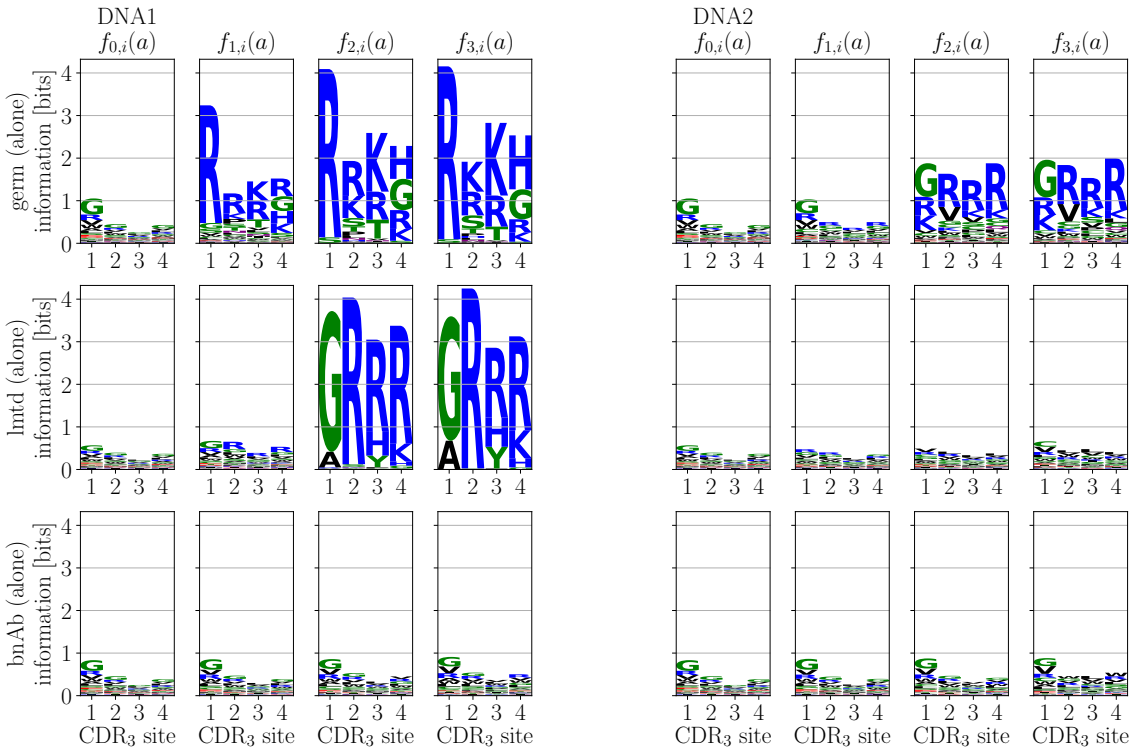


Fig. 4.14: Sequence logos based on amino acid frequencies $f_{t,i}(a)$. Data from all rounds t of separate selections against the DNA targets is shown. The total height at position i is given by $H_{t,i} = S_{i,max} - S[f_{t,i}(a)] = \ln(q) + \sum_a f_{t,i}(a)\ln(f_{t,i}(a))$, where $f_{t,i}(a) = \sum_x f_t(x)\delta(x_i, a)$ is the PWM, thus representing the information content (negative relative entropy) at position i. The larger the logo is, the more different the PWM is from the uniform distribution of amino acids. The height of letter (amino acid) a at position i is proportional to its frequency, $H_{t,i}(a) = f_{t,i}(a)H_{t,i}$, thus highlighting enriched amino acids. The logos for library mix selections are shown in figures 4.15 and 4.16. The logos for previously reported experiments [1] are shown in figures E.42 and E.43.

based on information-theoretic considerations the use of (in principle) selection round-independent enrichments instead of frequencies to construct time- and bias-independent and PWMs.

4.4.2 Enrichment sequence logos and the curse of finiteness of data

Instead of using frequencies $f_t(x)$, we construct time-independent PWMs $f_{t,i}(a)$ from enrichments $s(x)$ as

$$f_{t,i}(a) = \frac{\sum_x s(x)\delta(x_i, a)}{\sum_x s(x)}. \tag{4.17}$$

Formally, this is identical to setting $g_i(a) = f_{t-1,i}(a)$ in equation (4.16). In theory, such PWMs eliminate time-dependence and the effect of initial biases in CDR3 sequences. We used enrichments

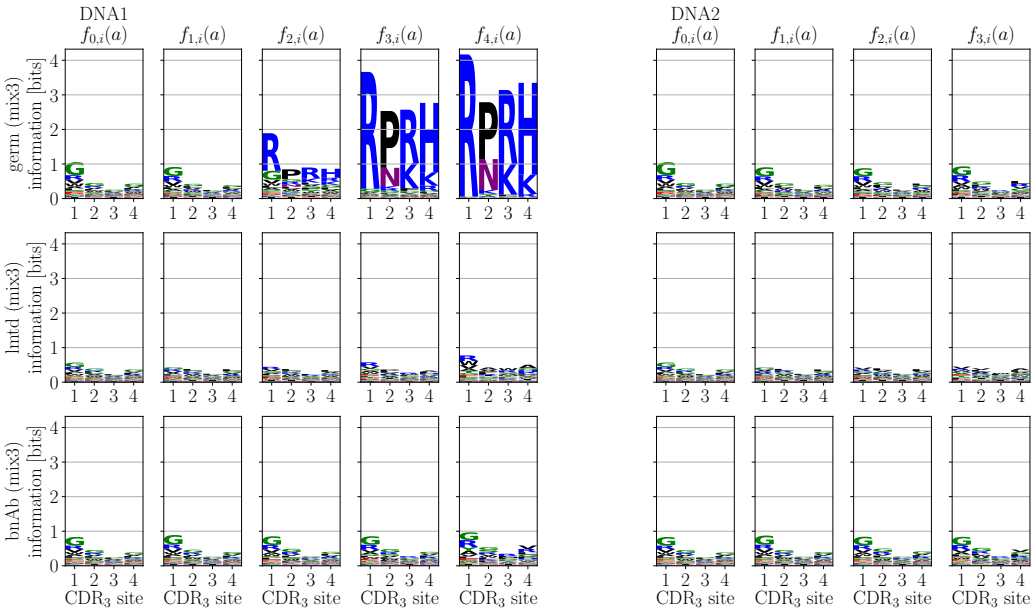


Fig. 4.15: Sequence logos based on amino acid frequencies $f_{t,i}(a)$. Similar to figure 4.14. Data from all rounds t of library mix (Mix3) selections against the DNA targets is shown.

$s(x)$ computed between selection rounds $t = 1$, $t+1 = 2$ for separate selections and $t = 2$, $t+1 = 3$ for Mix3 selections. In practice, however, this constancy over selection rounds is hardly observed in sequencing data due to finite sequencing depth which implies that empirical enrichments $s(x)$ are available only for a small subset of all q^L sequences x, and due to unspecific binding which superposes the specific contributions to enrichment in weakly binding sequences. In figures E.44, E.45, and E.46, we show enrichments sequence logos in which the effect of unspecific binding is removed by taking only sequences x with enrichments $s(x)$ such that $s(x) \geq \max(s^*, \exp(y^*))$. The same enrichments sequence logos without this correction are shown in figures E.49, E.50, and E.51. Taking only a subset of sequences into account systematically overestimates sequence logos, as found in section 2.3 and in figure 2.6. This is also observed here as, for instance, the area under the curve of enrichment sequence logos for the Germ library largely exceed the theoretical value from equation (2.89) of $\frac{\sigma^2}{2} \simeq 1.1$ for $\sigma_{\text{Germ}} \simeq 1.5$. The main conclusion of these logos that we present in the following subsection 4.4.3 remains, however, unaffected by these finite-size effects.

4.4.3 Target specificity and antibody specificity

In figure 4.17, we summarize the enrichment sequence logos obtained between selection rounds $t = 2$, $t + 1 = 3$ for Mix3 selections and between $t = 1$, $t + 1 = 2$ for separate selections. These logos give insight into two orthogonal specificities, namely the specificity of (i) the target and (ii) the antibody. In section 2.3, we defined and studied the overall specificity of interactions in the general case of many ligands interacting with many targets, before restricting to the case of

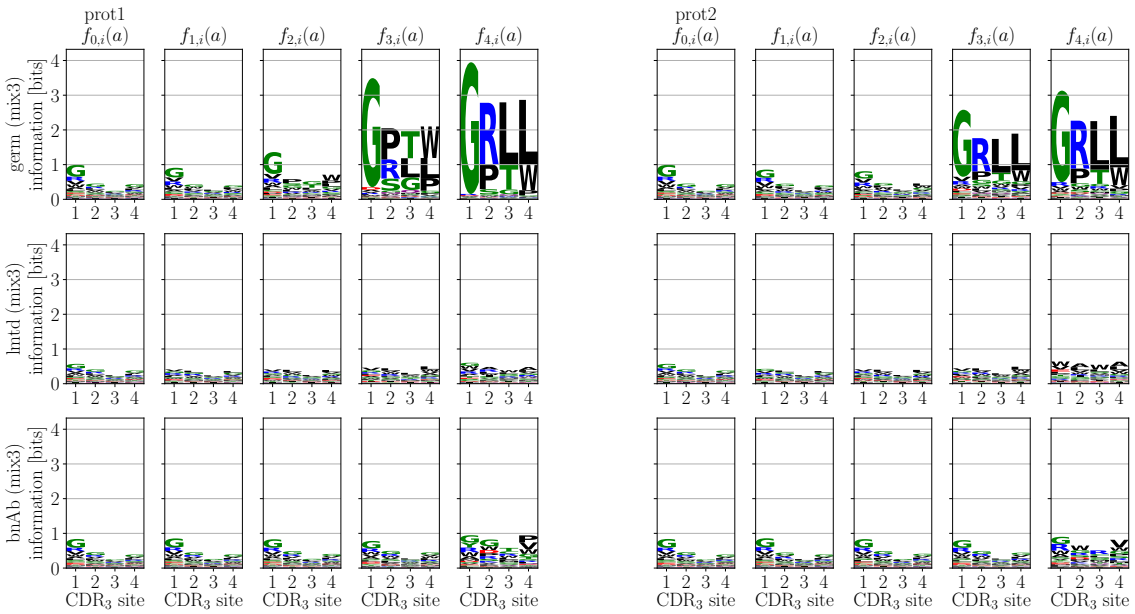


Fig. 4.16: Sequence logos based on amino acid frequencies $f_{t,i}(a)$. Similar to figure 4.14. Data from all rounds t of library mix (Mix3) selections against the protein targets is shown.

a library of ligands and a single target and deriving equation (2.89) which constrains the area under the curve of enrichment sequence logos in terms of σ. But this result does not constrain *which* sequences contribute to this area under the curve. These two pieces of information reflect respectively the marginals of the overall specificity of binding, *i.e.* the specificities of respectively the target and the ligand, but only the former one is predicted by the theory when considering a single target.

Importantly, when selecting a library of antibodies against a fixed target, as we do in our phage display biopanning experiments, we screen for the specificity of the *target* (rather than for the one of the antibody) in the context of a diversity of potential ligands. Inversely, the specificity of antibodies is defined in light of a set of several possible targets. A way to assess the specificity of antibodies by means of the same tools is to select a library independently against different targets. This resumes to either comparing sequence logos between targets or to direct measurements of specificities by crossed mini library selections. In subsection 4.4.4, we test for the specificity of Germ sequences that are enriched in selections against DNA1 and DNA2 by selecting them in mixture against either DNA1 or DNA2. More generally, the specificity of DNA-binding proteins can be assessed by SELEX experiments [168, 169, 103] that "reverse" the respective roles of ligands and targets compared to phage display biopanning experiments: Instead of selecting a library of ligands against a given DNA target, a library of DNA targets is selected against a given ligand. In the literature, such SELEX-based mathods are extensively used to measure specificities of transcription factors [63, 170, 64].

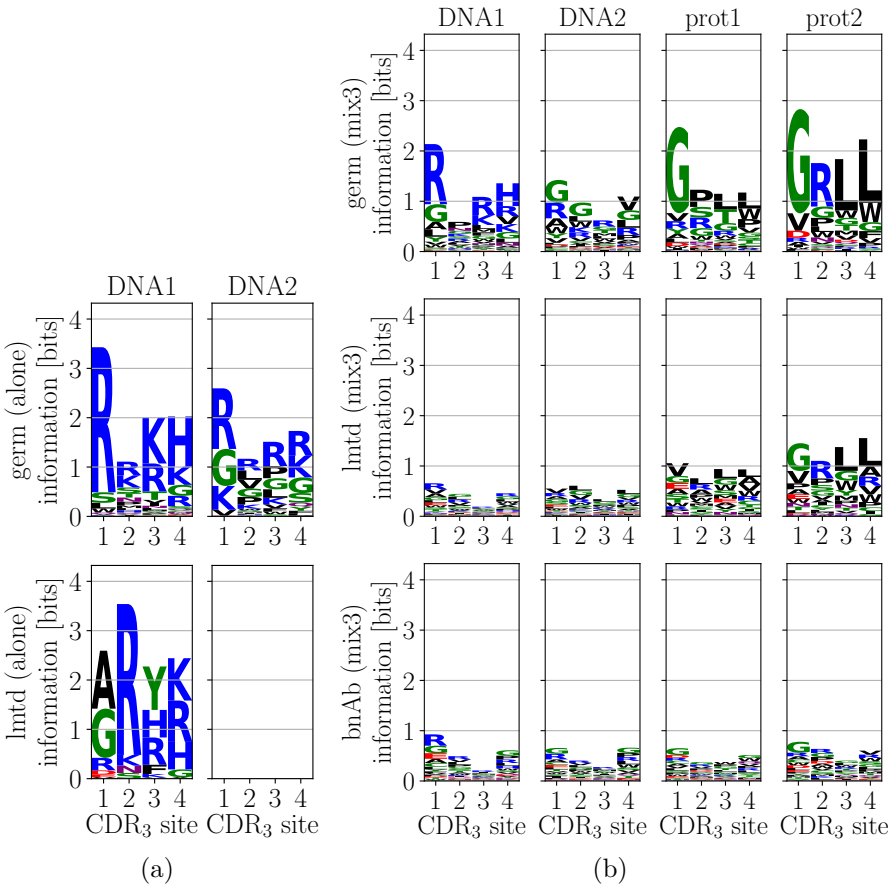


Fig. 4.17: Sequence logos based on enrichments $s(x)$. PWMs $\tilde{s}_i(a)$ are constructed from $s(x)$ according to $\tilde{s}_i(a) = \left(\sum_x s(x)\right)^{-1} \sum_x s(x)\delta(x_i, a)$, using only values $s(x) \geq s^*$ in order to exclude non-specific enrichments. Logo is empty if there is no specific signal. **(a)** Separate selections against the DNA targets. Here, enrichments $s(x)$ computed at round $t = 1$, $t + 1 = 2$, *i.e.* $s(x) \propto f_2(x)/f_1(x)$. **(b)** Library mix (Mix3) selections against DNA and protein targets. Enrichments computed at round $t = 2$, $t + 1 = 3$ are used, *i.e.* $s(x) \propto f_3(x)/f_2(x)$. In addition, logos for the same experiments but using enrichments computed at other rounds t are shown in figures E.44, E.45, and E.46. Logos from previously reported data [1] are shown in figures E.47 and E.48. Logos using all values of $s(x)$ (including those with $s(x) < s^*$) are shown in figures E.49, E.50, E.51, E.52, and E.53.

Although the areas under the curve are not predicted well by equation (2.89) in combination with the inferred values of σ, the sequence logos nonetheless seem to reproduce the same hierarchy found in terms of σ: In figure 4.17, the logos of the Germ library are systematically larger than those of the BnAb library. Remarkably, the Lmtd library behaves either like Germ or BnAb, depending on the target: A clear motif emerges when Lmtd is selected (alone) against DNA1, while it does not when selected against DNA2. Similarly, a Lmtd motif seems to appear when selected against prot2, but not against prot1. Moreover, the amino acids represented in the motifs are different between different targets and are consistent with the nature of the targets: The CDR3 sequences most enriched in both the Germ and Lmtd libraries in selections against

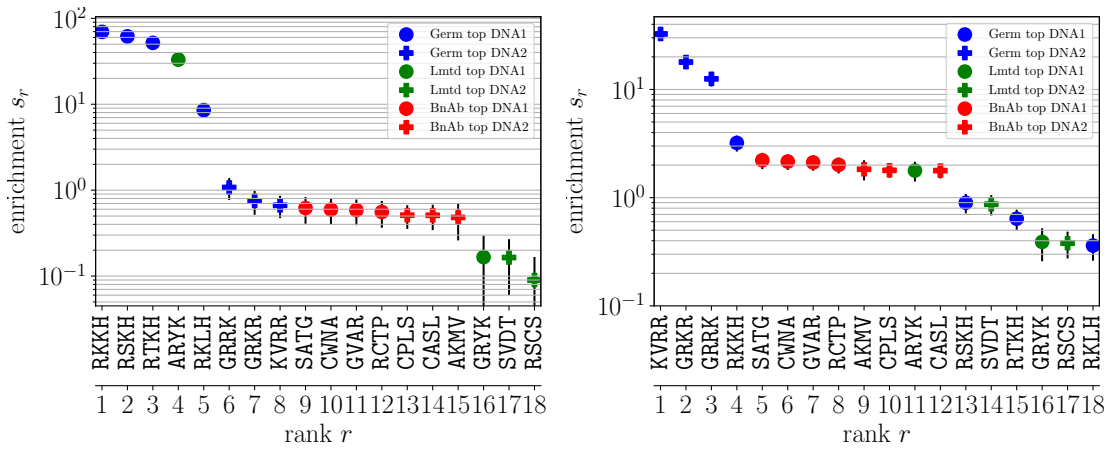


Fig. 4.18: Cross-selections of mini libraries against DNA targets revealing CDR3 sequence specificities. High-precision enrichments from mini libraries with around 20 different sequences are plotted in decreasing order and the CDR3 sequences are indicated. DNA1- and DNA2-specific clones against **(a)** DNA1 and **(b)** DNA2. Error bars are 20 x enlarged. Consistency with the other mini library selections in figure 4.12 is shown in figure E.40.

the negatively charged DNA targets are rich in positively charged amino acids (`K`, `R`, `H`, letters in blue). Selections against the protein targets, which are close homologs and thus structurally similar, are dominated by similar CDR3 sequences in these libraries. Note that the same CDR3 sequence patterns may also be enriched in BnAb though much weaker than in Germ and similarly strongly as the amplification bias, see figure 4.5. These different sequence logos are in line with the discussion of target-specific selection responses already discussed in section 4.1.4.

A few final remarks: (i) Large sequence logos are also observed in the winner libraries of former Mix24 and Mix21 selections, see figures E.42 and E.43, consistent with the finding that inter- and intra-library hierarchies are connected. (ii) Moreover, the conclusion about the chemical properties of Germ and Lmtd sequences enriched against DNA targets extends to third DNA target, DNA3, that was previously studied, see figure E.42. (iii) There is a significant difference in sequence motif between the Germ library selected alone or in Mix3 against DNA1, notably on position $i = 2$, though the chemical properties of the selected CDR3 are overall the same. This is likely due to stochasticity in the initial libraries where these sequences are rare.

4.4.4 Cross-selections with mini libraries

The selections against DNA1 and DNA2 resulted in different CDR3 sequences being strongly enriched, with consensus sequence `RKKH` against DNA1 *versus* `GRRR` against DNA2 in the Germ library and `GRRR` against DNA1 versus no strongly enriched sequence against DNA2 in the Lmtd

library (see figure 4.17). To test for the specificity of sequences in the top DNA1 and top DNA2 mini libraries, which are defined in table 4.1, we pool them together and perform a single round of selection independently against DNA1 and DNA2. The enrichments are shown in figure 4.18: The Germ top sequences are 10^2 x more enriched against their cognate DNA target than the top sequences against the other DNA target, which shows their specific binding to the respective DNA target. The same holds for the highly enriched Lmtd sequence with CDR3 `ARYK` which is strongly enriched against DNA2. This shows that the electric charges, which are common to DNA1- and DNA2-specific sequences are not sufficient to explain these specificities to either of the DNA targets. Moreover, this result confirms that both DNA targets can be regarded as independent as different binding mechanisms are used by the antibody to presumably target epitopes that differ between these two DNA targets. Curiously, the consensus sequence `GRRR` is shared between Germ against DNA1, Lmtd against DNA2, and also Germ against DNA3 (see figures E.47 and E.52), which could be explained by binding to the common stem sequence which is shared between all three DNA targets.

4.5 Beyond enrichments: inference of more detailed biophysical models

Our very abecedarian analysis of the selection data turned out sufficient for major conclusions on selection potentials in antibody libraries. However, a re-analysis of the same and future sequencing data under a more efficient use of the provided information as well as a refined modeling of both the biophysics of selection steps and the stochasticity of sampling steps should allow for more detailed insights into the interactions that drive selection and evolution in our model system (subsection 4.5.1). Skipping the mathematical details, we here motivate that the inference of biophysical models, such as those presented in chapter 2, can occur in the framework of multi-type branching processes (subsection 4.5.2). Additionally and importantly, a more careful modeling should also allow to deconvolute several selection-related and -unrelated factors of biasing in sequence frequencies, such as several binding modes, unspecific binding, cooperative/adverse effects, and amplification biases (subsection 4.5.3). We recently implemented the learning of biophysical models from our sequencing data in python and showcase here a result obtained for the data from the Germline library selected (in Mix3) against the DNA1 target assuming a binding model with one specific, additive and one unspecific binding mode (subsection 4.5.4).

4.5.1 Shortcomings of empirical enrichments

Our approach based on empirical enrichments comes with a number of inconvenients: (i) Empirical enrichments are simply inconclusive for rare and unseen sequences. Computing enrichments as the after-to-before selection ratios in count numbers, $s_{\text{emp}}(x) \propto \frac{n_{t+1}(x)}{n_t(x)}$, is meaningless for unseen

sequences with $n_t(x) = 0$ and/or $n_{t+1}(x) = 0$ and dominated by sampling noise rather then selection for low-count sequences. Here, we sequenced up to 10^6 individuals in populations of up to 10^{12} individuals. Sequences with $n_t(x) < 10^{-6} \cdot 10^{12} = 10^6$ copies in the population are thus typically not observed in the data despite being present in the library and taking part in selection; sequences with $n_t(x) < 10^7$ copies are observed only a few times, typically $n_t(x) < 10$ times, and have to be excluded from the empirical enrichment analysis. However, low-count sequences typically represent a significant part of the sequence space and do certainly provide useful information beyond sampling noise (see figure 4.1(a), rounds $t = 0$, $t + 1 = 1$) that could be profitably integrated into an alternative sequencing data analysis. (ii) The random-energy assumption ($p = L$ with L the number of sequence positions and p defined in section 2.1.4), which is the basis for empirical enrichments, discards any non-random structure that may exist in the actual binding energy landscape $x \mapsto \Delta G(x)$ (remember the positively charged amino acids in the CDR3 selected against the negatively charged DNA targets). However, the sequencing data does contain information about the shape of these landscapes, as it does provide not only the mere histogram of sequence counts, but also the connection between sequence count and sequence identity. Points (i) and (ii) imply that a significant amount of information in our sequencing data has remained unused so far. (iii) Empirical enrichments *a priori* are blind to the mechanisms that have generated them ("any model") while prior knowledge about these mechanisms, such as the physics of binding and stochasticity of sampling, could be profitably used to constrain the inference to certain relevant model spaces and to possibly dissect several selection pressures simultaneously at play and subtract selection-unrelated biasing. For instance, our empirical enrichments were found to be a superposition of binding and amplification biases. (iv) This blindness of empirical enrichments also implies the absence of predictive power of the approach towards unseen sequences. At the other extreme, statistical models with one-point, two-point, *etc.* interactions ($p = 1, 2, \dots$) have been shown to be generative in other contexts such as contact and structure prediction [87, 88].

4.5.2 Biophysical models and multi-species branching processes

The modeling of the experimental evolutionary process occurs at two levels: A model for (i) the binding landscape (or fitness landscape in general), $x \mapsto \Delta G(x)$, as well as for (ii) how true enrichments, or survival/offspring number probabilities upon selection and amplification, $s(x) \simeq \exp(-\beta \Delta G)$ translate into sequencing counts, $s(x) \mapsto n_t(x)$ for all t. Models for (i) have been discussed in section 2.1.4, while (ii) is stochastic in nature due to finiteness of population size and sequencing depth and conveniently captured by multi-type branching processes [63, 68, 138, 61, 171]. Denote by $N_t(x)$ and $n_t(x)$ the number of copies of sequence x at round t in respectively the full population and the sequenced sample. The conditional probabilities $\mathbb{P}[N_{t+1}(x)|N_t(x), s(x)]$ and $\mathbb{P}[n_t(x)|N_t(x), \phi]$ define the selection and sampling steps. They represent respectively the probability of having $N_{t+1}(x)$ copies of x after selection (and amplification) given $N_t(x)$ copies before selection and selection probability $s(x)$, and the probability of seeing $n_t(x)$ times sequence x

in the sample (sequencing data) at selection round t given $N_t(x)$ copies of x in the full population and CDR3 sequence-independent sampling probability $\phi \simeq \frac{10^6}{10^{12}} = 10^{-6}$. With a suitable choice of these distributions, typically $N_{t+1} \overset{d}{=} \mathrm{Bin}(N_t, s)$ for selection and $n_t \overset{d}{=} \mathrm{Bin}(N_t, \phi)$ for sampling, a likelihood function can be derived and used for maximum-likelihood estimation of model parameters, see subsection 4.5.4. The system can be reformulated as a hidden Markov model where the time series of $N_t(x)$ and $n_t(x)$ represent the hidden states and observed variables, respectively, and with transition and emission probabilities involving $s(x)$ and ϕ, respectively [172].

4.5.3 Dissecting binding and non-binding modes, epitope inference

The unimodal, additive ($p = 1$) binding model with

$$s(x) = \mathrm{e}^{-\beta\Delta G(x)}, \qquad \beta\Delta G(x) = \sum_{i=1}^{L} h_i(x_i) \tag{4.18}$$

in the Boltzmann limit as the simplest case can be extended to take into account other factors of selection (possibly) present in the system [63, 64, 66]: (i) In the presence of one or several additional binding modes with different amino acid preferences, *i.e.* other local field functions, the enrichment becomes

$$s(x) = \sum_{k=1}^{K} \mathrm{e}^{-\beta\Delta G_k(x)} \tag{4.19}$$

with K the number of binding modes, each of which could again be assumed additive in the simplest case. The co-presence of several binding modes may be mediated by several non-identical epitopes on the binding target. (ii) Unspecific binding, as observed in our data, can be accounted for by introducing a binding mode with CDR3 sequence-independent binding energy ΔG_{us},

$$s(x) = \mathrm{e}^{-\beta\Delta G(x)} + \mathrm{e}^{-\beta\Delta G_{\mathrm{us}}(x)}. \tag{4.20}$$

(iii) An apparent second mode of selection may also be linked to factors unrelated to binding, such as amplification bias. Compared to a second binding mode, an amplification mode is multiplicative and we thus have a total enrichment of the form $s(x)a(x)$, where $s(x)$ is the binding enrichment and $a(x)$ the amplification enrichment. However, upon performing a Taylor expansion using $s(x) = s_0 + s_1(x)$ and $a(x) = a_0 + a_1(x)$ with s_0 and a_0 representing global levels of binding and amplification, and small sequence-dependent perturbations $s_1(x)$, $a_1(x)$, we obtain $s(x)a(x) = s_0a_0 + s_0a_1(x) + a_0s_1(x)$ to first order, which may be indistinguishable from a model with unspecific binding and two sequence-dependent binding modes in practice upon formally taking $s_0a_0 = \mathrm{e}^{-\beta\Delta G_{\mathrm{us}}}$, $s_0a_1(x) = \mathrm{e}^{-\beta\Delta G_a(x)}$, and $a_0s_1(x) = \mathrm{e}^{-\beta\Delta G_s(x)}$. (iv) To go beyond the additive model

for $\Delta G(x)$, the next-order term accounts for pair-wise interactions ($p = 2$) between sites,

$$\beta\Delta G(x) = \sum_{i=1}^{L} h_i(x_i) + \sum_{i=1}^{L}\sum_{j=1}^{i-1} J_{ij}(x_i, x_j). \tag{4.21}$$

The interpretation of these couplings J_{ij} is, however, not straightforward; they could stem from cooperative and adverse effects between residues or from perturbative effects coming from the superposition of several binding and/or non-binding modes or from a global non-linearity, such as the Fermi-Dirac statistics (see section 2.1), due to ligand or target saturation effects.

4.5.4 Biophysical model inference for Germline against DNA1

We exemplify the approach on the Germline part of the Mix3 selections against the DNA1 target, taking the sequencing counts at selections rounds $t = 1$ and $t+1 = 2$, $\{n_1(x)\}$ and $\{n_2(x)\}$ for all CDR3 sequences x, as input for the binding model inference. This data was fitted to a binding model which comprises one unspecific and one specific, additive binding mode,

$$s(x) = \mathrm{e}^{-\beta\Delta G(x)} + \mathrm{e}^{-\beta\Delta G_{\mathrm{us}}}, \qquad \beta\Delta G(x) = \sum_{i=1}^{L} h_i(x_i) \tag{4.22}$$

where ΔG_{us} is an unspecific, *i.e.* CDR3 sequence-independent, binding energy and where we assumed the validity of the Boltzmann regime. The relevance of unspecific binding in this data results from the presence of many sequences at a minimal, non-zero enrichment $s_{\mathrm{us}} = \mathrm{e}^{-\beta\Delta G_{\mathrm{us}}}$ which is conveyed by an accumulation of points around a line parametrized by $n_2 = s_{\mathrm{us}} n_1$ in the n_2-n_1 plane, see figure 4.19(a). The model parameter $h_i(a)$ and ΔG_{us} are taken to be the ones that maximize the log-likelihood function $\mathcal{L}\left(s(x)|\{n_t(x), n_{t+1}(x)\}\right) = \sum_x \ln \mathbb{P}[n_{t+1}(x)|n_t(x), s(x), \phi]$ with

$$\begin{aligned}
\mathbb{P}[n_{t+1}|n_t, s, \phi] &= \sum_{N_t, N_{t+1}=0}^{\infty} \frac{\mathbb{P}[n_t|N_t, \phi]\mathbb{P}[N_t]}{\mathbb{P}[n_t]} \mathbb{P}[N_{t+1}|N_t, s]\mathbb{P}[n_{t+1}|N_{t+1}, \phi] \\
&\simeq \frac{\phi^{n_t+1}(s\phi)^{n_{t+1}}}{n_t! n_{t+1}!} \int_0^{\infty} \mathrm{e}^{-(1+s)\phi\xi} \xi^{n_t+n_{t+1}} \mathrm{d}\xi \\
&= \binom{n_t + n_{t+1}}{n_{t+1}} \left(\frac{s}{1+s}\right)^{n_{t+1}} \left(\frac{1}{1+s}\right)^{n_t+1},
\end{aligned} \tag{4.23}$$

independently of ϕ, which assumes deterministic selection, *i.e.* $\mathbb{P}[N_{t+1}|N_t, s] = \delta(N_{t+1} - sN_t)$ with the amplification factor λ absorbed into s, a Poisson distribution for the sampling step (as a limiting case of the binomial distribution), $\mathbb{P}[n_t|N_t, \phi] = \mathrm{e}^{-N_t\phi}\frac{(N_t\phi)^{n_t}}{n_t!}$, and a uniform prior $\mathbb{P}[N_t]$ which implies $\mathbb{P}[n_t] = \phi^{-1}$. Thus, we have up to terms independent of s

$$\mathcal{L}\left(s(x)|\{n_t(x), n_{t+1}(x)\}\right) = \sum_x n_{t+1}(x)\ln(s(x)) - (n_t(x) + n_{t+1} + 1)\ln(1 + s(x)). \tag{4.24}$$

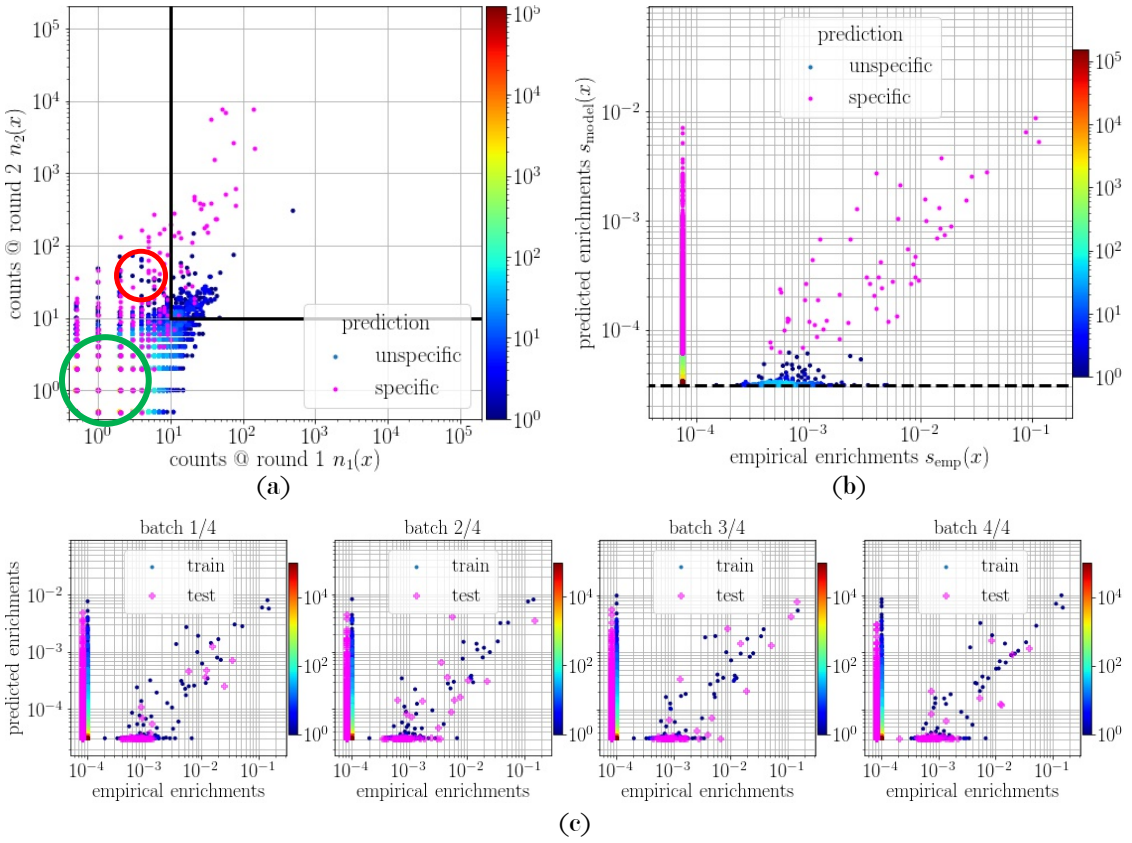


Fig. 4.19: Biophysical model inference beyond the random-energy model. For the Germline part of Mix3 at rounds $t = 1$ and $t + 1 = 2$ against DNA1, a binding model with one additive binding mode, $\beta\Delta G(x) = \sum_{i=1}^{4} h_i(x_i)$, and one unspecific binding mode, $\beta\Delta G_{\mathrm{us}} = h_{\mathrm{us}}$, was inferred from the sequencing counts $\{n_1(x), n_2(x)\}$ by maximum-likelihood estimation using equation (4.24). **(a)** Sequencing counts at round $t + 1 = 2$, $n_2(x)$, plotted against those at round $t = 1$, $n_1(x)$; one point per CDR3 sequence. Color bar between blue and red indicates density of points. Sequences x classified as specific binders to DNA1, as *per* the condition $\Delta G(x) < \Delta G_{\mathrm{us}}$, are colored in pink. **(b)** Model enrichments $s_{\mathrm{model}}(x) = \exp\left(-\sum_{i=1}^{4} h_i(x_i)\right) + \exp(-h_{\mathrm{us}})$ plotted against meaningful (*i.e.* $n_1(x) \geq 10$ and $n_2(x) \geq 10$) empirical enrichments $s_{\mathrm{emp}}(x) \propto n_2(x)/n_1(x)$. Sequences where no empirical enrichment is available are drawn to the left of the plot. Same classification into specific *versus* unspecific sequences as in (a). Dashed black line indicates minimal model enrichment equal to $\exp(-h_{\mathrm{us}})$. **(c)** Cross-validation of the inferred binding model. Sequence space is randomly partitioned into 4 batches; 1 batch is used for testing while the 3 others are used for training; this is repeated for all 4 possible choices of the test batch. The plots compare the model enrichments $s_{\mathrm{model}}(x)$ with empirical enrichments $s_{\mathrm{emp}}(x)$; the test batch is indicated in each plot and the sequences therein are colored in pink.

In figure 4.19(b), we show *preliminary* results: We compare model enrichments computed as $s_{\mathrm{model}}(x) = \exp\left(-\sum_{i=1}^{4} h_i(x_i)\right)$ with empirical enrichments computed as $s_{\mathrm{emp}}(x) \propto \frac{n_2(x)}{n_1(x)}$ (as long as $n_1(x) \geq 10$ and $n_2(x) \geq 10$ for a given sequence x). The correlation between s_{model} and s_{emp} for high-enrichment sequences suggests that an additive binding model explains enrichments in the Germline CDR3 and may provide an *a posteriori* support for the relevance of the central-

limit theorem to the distribution of enrichments $P(s)$. The slope of the correlation is close to 1 which suggests that empirical enrichments and the inferred binding model find similar values for σ. Such a model can be validated in several ways: Classifying sequences into specific *versus* unspecific binders according to whether $\Delta G(x) < \Delta G_{\mathrm{us}}$ (specific binder) or $\Delta G(x) > \Delta G_{\mathrm{us}}$ (unspecific binder), correctly identifies the bulk of sequences at low enrichment as unspecific binders while asserting that most sequences above the bulk are specific binders, see figure 4.19(a), (b). Sequences well above the bulk of minimal enrichment but classified as unspecific (red circle in figure 4.19(a)) are well-amplified sequences in the amplification bias which is not taken into account here. On the contrary, sequences poorly represented but predicted to be specific binders (green circle in figure 4.19(a)) are likely underrepresented in the initial library. Furthermore, a cross-validation is performed by randomly partitioning the sequence space of all 20^4 CDR3 sequences into 4 batches of approximately equal size and only the sequences in any 3 of them are used as training set for the maximum-likelihood estimation of model parameter, while the sequences in the remaining batch are used for the prediction of enrichments. In figure 4.19(c), these predicted enrichments are compared with empirical enrichments: high-enrichment sequences in the test set are predicted by the inferred models.

Chapter 5

Conclusion and perspectives

As usual, the work presented in this manuscript raises more questions than it answers. In this wrap-up chapter, we summarize again the answers and main contributions that this project was able to provide (section 5.1), and discuss possible directions for future research, experimentally and theoretically, that are motivated by this work (section 5.2).

5.1 Definition and measurement of selection potential, implications for evolvability

In this project, we performed quantitative selection experiments to study selection potentials of antibodies, a model system where evolvability is presumably a key evolutionary property (see chapter 1), which is amenable to controlled, experimental evolution and quantifiable through high-throughput sequencing (see chapter 3), and which is suitable for mathematical and modeling purposes (see chapters 2 and 4). Our notion of selection potential, which is captured by a single scalar parameter σ in the case of unimodal binding, encodes for the scale in which sequence diversity translates into phenotypic diversity and thus for the efficacy of selection to a new selective pressure. We describe a procedure to infer σ from *in vitro* selections and high-throughput sequencing of libraries of variants and thus provide answers to the following question: How to read evolvability from the sequence? (See subsection 5.1.1) Then, what determines evolvability? Upon measuring the selection potentials of several antibody libraries built around different antibody scaffolds for several binding targets, we identified the maturation degree of the scaffold, *i.e.* its amount of previous maturation towards another, unrelated binding target, as a key determinant of selection potentials. (See subsection 5.1.2) Beyond that, we may also ask: How does evolvability depend on the maturation degree? What other biophysical and/or structural properties does evolvability thus correlate with, and in which way? Our work also provides a preliminary

answer to the first question: Within the lineage of HIV-specific antibody scaffolds we studied, the selection potential tends to decrease as its maturation degree increases. (See subsection 5.1.3.) The second question, as well as the generality of the preliminary result to the first question will be targeted in future work, as discussed in the next section 5.2.

5.1.1 Reading selection potentials from the sequence

Evolvability defines the ability of an object to efficiently respond to and quickly yield improvement with regard to a new selective pressure. It is required for the success of any evolutionary optimization procedure with regard to a given target property or feature, starting from some initial condition with certain evolutionary degrees of freedom. Evolvability is oftentimes taken for granted and appears as a rather peculiar property or phenotype [15], but its potential relevance and selectivity has been repeatedly demonstrated [9, 10, 31, 11, 33, 12, 13]. We introduced the notion of "selection potential" to specifically refer to the amplitude of the initial response to selection and ability to enrich high-fitness mutants, *i.e.* the susceptibility to selection at the beginning of an evolutionary trajectory. Evolvability itself is usually defined with respect to the end-product of such a trajectory, but selection potential can be expected to favor evolvability. Our experiments targeted the selection potential as a component of evolvability, but our approach can be easily generalized to study evolvability, as discussed in the next chapter 5.2.

The major contribution of this work resides in the definition and inference of a quantity directly related to evolvability in an experimental model system. As yet, evolvability was only studied on the basis of mathematical and computational model systems or emerged as a side-product in studies of other properties (see chapter 1). None of these works resulted in suggestions and protocols to experimentally assess and measure evolvability. Our work should have implications in any context which relies on the optimization through (Darwinian) evolutionary procedures and where high evolvability of the initial guess is thus crucial for success, such as in the adaptive immune response, directed evolution of proteins, derivation of clinically relevant biomolecules, but possibly also for the training of neural and elastic networks. Being able to measure evolvability should open the doors towards the understanding and control of evolvability in the future.

Our model system consists of antibody libraries built from fixed scaffold sequences and randomized binding site sequences, which are selected for binding affinity to various targets. It is defined on a sequence space and governed by selection for unimodal binding at thermodynamic equilibrium. This is particularly convenient as the mapping (evolutionary degrees of freedom) $\mapsto$ (property/phenotype) $\mapsto$ (selection coefficient) can be modeled under the use of physics and universality arguments; the evolutionary degrees of freedom are the residues (or Potts spins) of a sequence x, the relevant phenotype is the binding free energy $\Delta G(x)$, and the selection coefficient s is the probability to be in bound state at thermodynamic equilibrium. The first mapping, $x \mapsto \Delta G(x)$, can be represented by a class of random models and we show that the second map-

ping, $\Delta G \mapsto s$, reads $s = \exp(-\beta\Delta G)$ in a regime of intermediate target concentrations. The central-limit theorem predicts the distribution of binding energies in a library of randomized binding site sequences to be close to Gaussian, and we denote the mean and variance of $(-\beta\Delta G)$ in this case by μ and σ^2. This argument notably relies on the assumption of close-to non-interacting binding site residues. In section 4.5, we confirm that an additive binding model can indeed capture the observed enrichments reasonably well. The selection coefficients $s = \exp(-\beta\Delta G)$ should then obey a lognormal distribution with the same parameters. The parameter μ sets the scale of binding affinities within libraries while σ is associated with the phenotypic diversity. The parameter σ can be equated to the selection potential of the system based on its alternative, information-theoretic interpretation as the interaction specificity between ligands and targets and its implications for selection dynamics within and between libraries of ligands: It determines the outcome of competitions between libraries with different values of σ and relates inter- and intra-library hierarchies between sequences. Moreover, σ relates to the time-derivative of the population fitness via Fisher's fundamental theorem of natural selection and thus to the efficacy of selection.

Lognormal model distributions for $P(s)$ fit the observed histograms of enrichments s reasonably well, as revealed by PP and QQ plots. We suggest the use of σ preferentially to another parameter, κ, previously proposed to quantify phenotypical diversity based on extreme-value theory [1]. We here show that both approaches fit our selection data equally well, but σ has the advantage of immediate physical and information-theoretic interpretations, its applicability to the full distribution of enrichments $P(s)$ rather than only the tail, as well as its predictive power: Beyond satisfactory quality of fit, lognormal distributions for $P(s)$ are good predictors of selection dynamics in competitive selections of library mixtures as we demonstrated by validation on the observed time series of library frequencies.

While σ is a parameter particular to our model selective pressure (unimodal binding at equilibrium), we think that the concept can be readily generalized to other, more complicated selective pressures and model systems.

5.1.2 The degree of maturation determines selection potentials

We performed and studied selections for equilibrium binding to 4 different target molecules of 3 antibody libraries built from different antibody scaffolds with no (naïve), intermediate, and profound (bnAb) maturation degree against HIV, respectively, and identical sequence diversity at the CDR3 (antibody binding site). These scaffolds are evolutionarily related, as the naïve scaffold is the common ancestor of both the intermediate and profoundly matured scaffolds, although the 2 mature scaffolds are on different branches of the phylogeny. Despite identical sequence diversity, which consists of 4 CDR3 residues completely randomized to all 20 amino acids each, the libraries show vastly different phenotypic diversities represented by significant different values of σ. As predicted by the theory (see chapter 2) and confirmed in the experiments (see chapter 4), this has

strong implications for their interaction specificities and behaviour under selection both within and between libraries, that is, for their selection potential. Curiously, the values of σ are largely unaffected by the precise HIV-unrelated binding target in use: Here, we used DNA and fluorescent proteins as binding targets, all unrelated to HIV and also unrelated among them. These findings suggest that the selection potential is fully determined by, and thus a property of, the antibody library and the underlying antibody scaffold. As the antibody scaffolds used here differ only by their maturation status (against HIV) as encoded in the quantity of fixed somatic mutations (0 for the naïve, 15 for the intermediate, 34 for the fully mature scaffold), these differences in selection potentials can be traced back to the maturation degree.

5.1.3 How do selection potentials depend on maturation degree?

The way in which selection potentials, as represented by σ, depend on the maturation degree is remarkable: For all binding targets we used, the selection potential decreases as the maturation degree of the library increases; the library based on a naïve scaffold systematically dominates the mature ones and the library based on a deeply matured bnAb scaffold systematically shows the least selection potential. As mentioned above, this hierarchy holds irrespectively of the binding target in use. Interestingly, the same CDR3 sequence patterns are selected in all 3 libraries for a given target and can sometimes be explained by the nature of the target, such as positively charged amino acids in the CDR3 being presumably selected for their electrostatic interactions with negatively charged DNA binding targets. Yet, the same beneficial CDR3 pattern features higher affinity and selectivity when appearing in the context of the naïve scaffold than in the context of the deeply matured scaffold which fails to provide high affinity and selectivity. This context-dependance is oftentimes termed “epistasis” [68, 106, 173] and has already been introduced in the context of antibodies [60]. Here, epistasis occurs between the CDR3 and the scaffold, but not necessarily among CDR3 residues as shown in section 4.5. A suited CDR3 sequence for a new binding target requires the naïve scaffold and, inversely, the naïve scaffold requires a suited CDR3 sequences, as most naïve antibodies were actually found to perform very poorly. The library based on the intermediate maturation degree features intermediate values of σ and displays an intermediate selection behaviour in the following way: Depending on the target, it is sometimes selectable and behaves more closely to the naïve one, sometimes it is not selectable and behaves more closely to the deeply matured one.

The picture raised by our measurements of selection potentials fits into a picture abundantly evoked in the literature and obtained from various viewpoints and approaches (see chapter 1): Upon specializing towards high affinity to its cognate target (here: some epitope on HIV), a naïve antibody loses its selection potential for other, non-cognate targets in the course of its affinity maturation. The naïve antibody itself, which is assembled from inheritable, germline-encoded gene fragments, may be evolved through generations towards high evolvability and high selection potentials to a variety of different binding targets. However, alternative but similar scenarios of

changes to the antibody upon affinity maturation have been proposed as well [54, 49, 51, 50, 174] and may be revealed upon screening an increased number of antibodies for their selection potentials compared to what was done here.

The significance of this particular hierarchy between the 3 libraries consistently found against 4 binding targets can be captured by a simple p-value, $(3!)^{-4} \simeq 8 \cdot 10^{-4}$. To more systematically address the question of how selection potential depends on maturation degree, the approach needs to be scaled up on several accounts: The selection potential should be measured for more antibody scaffolds by choosing scaffolds and building antibody libraries from (i) more affinity maturation trajectories (only a single trajectory here) and (ii) more timepoints per affinity maturation trajectory (only 3 timepoints here). Moreover, (iii) the selection potentials of these libraries should be determined for more binding targets (only 4 targets here). Such a scale-up of the selection potential assessment the will likely require a change in the experimental setup, as discussed in the following section 5.2.

5.2 Evolvability: what's next?

Combining library-based screening techniques with high-throughput sequencing is not completely new in general but has not yet been used to define and measure selection potentials and evolvability. While we show that our approach is able to capture differences in selection potentials and evolvability and to identify its major determinants, the quantity of data is currently insufficient to read out precise dependencies of evolvability on other protein properties. However, we think that this is simply a matter of scale-up and extension of the experiments reported here, as well as of the depth of the data analysis and thus does not call our new approach into question. We propose the next steps to be as follows: (i) To systematically assess selection potentials of many antibodies from several maturation trajectories and several timepoints per trajectory reequires a speed-up and parallelization of selection trajectories and will likely require a change in the selection protocol (subsection 5.2.1). (ii) A better control of the maturation status can be achieved by studying *in vitro* matured rather than *in vivo* matured antibodies, as was done here. Details of selective pressures during *in vivo* affinity maturation are generally unknown and may vary especially across patients. The choice of antibodies could thus be based on more informative *in vitro* affinity maturation trajectories with controlled evolutionary circumstances and sufficient sampling along evolutionary time (subsection 5.2.2). (iii) Our approach allows to probe selection potentials of antibody libraries and the undelying antibody scaffolds. By measuring the biophysical properties of the same antibodies, dependencies and interdependencies between evolvability and other protein properties such as binding specificity and stability could be systematically studied. Some of these opportunities have recently been initiated by other students within the group (subsection 5.2.3). (iv) Finally, we should also seek to thoretically understand the basis of evolvability, as well as its evolution and connections with other properties. Spin-glass models and elastic networks are intuitive candidates that have already been used in the literature and are relevant beyond the

context of biomolecules (subsection 5.2.4).

5.2.1 Improving and scaling up the assessment of selection potentials

The currently used display and selection protocols are costly in terms of time and manual labor which limits the number of antibodies that can be tested for their selection potentials. A single cycle of library display and selection takes at least 3 continuous days packed with laboratory work (without preparatory tasks and sequencing), *i.e.* at least 2 weeks for a selection trajectory with 4 rounds of selection as was done here. At most 4 selections can be performed in parallel. The scaling in time and amount of repetitive work is inadequate if one wants to perform a large-scale assessment of selection potentials: Take as an example the study of 10 maturation trajectories (instead of 1 here) at 10 time points each (instead of 3 here) against 10 different target molecules (instead of 4 here), which would require $10^3/4 \cdot 2 = 500$ weeks of continuous work which already far exceeds the PhD time scale of three years, *i.e.* 150 weeks.

Efforts to condense the duration of the experiment by parallelization and automation without sacrifice of controllability of the experimental evolution have been and are being made within the group, in parallel with the introduction of a mutagenesis step to extend the *in vitro* selection to *in vitro* maturation [175, 148] (see next subsection 5.2.2). In fact, more rapid laboratory evolution do exist but are essentially *in vivo*, which is detrimental for the controllability of the experimental evolution. One example is phage-assisted continuous evolution (PACE) [176] in which the evolutionary steps happen continuously and hands-off rather than iteratively with significant hands-on parts. In addition, the protocol is currently being modified to replace phage display by ribosome display. This allows to perform the selection step entirely *in vitro* and thus for an even improved control of selection conditions; as yet, the library display step occurs *in vivo* in an *E. coli* expression strain (see chapter A), which is probably responsable for the significant amplification biases we observed.

Finally, the precision of the selection potential inference can also be improved in two ways: (i) Increasing the sequencing depth will provide more sequencing reads per sequence and thus allow for the computation of empirical enrichments for an increased number of sequences. For instance, switching from the Illumina MiSeq sequencing used here to the Illumina HiSeq sequencing can increase the number of sequencing reads by a factor of 100 with no sacrifice in the sequencing error level, yet to the expense of shorter sequencing reads which would thus require a change in the sequencing strategy and a redefinition of the "region of interest" along the antibody gene. (ii) The overly simplistic analysis of sequencing data via empirical enrichments can itself be discarded in favor of a more involved, but more accurate and stable modeling of the both the binding landscape and the selection and sampling processes (section 4.5).

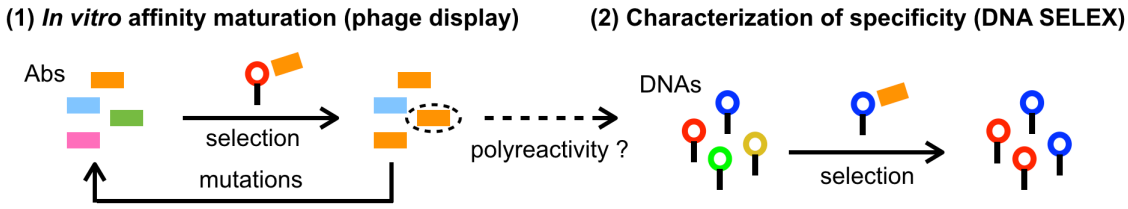


Fig. 5.1: Principle of controlled affinity maturation and SELEX experiments for DNA-binding antibodies. Taken from [177]. **(1)** Controlled affinity maturation by repeated cycles of selection and mutation of the antibody lineage: A selection step for binding to the DNA target, *e.g. in vitro* by phage display and biopanning as was done here, enriches strongly binding antibody sequences (orange rectangle) over others, while a subsequent mutagenesis step, *e.g. in vitro* by error-prone PCR of antibody genes, introduces fresh, random sequence diversity. **(2)** The recognition spectrum of any antibody of interest (orange rectangle) can be defined and assessed by a SELEX experiment which simply swaps the respective roles of the antibody and DNA target: A library of DNA targets, *e.g.* hairpins with random loop sequences, is selected against the immobilized antibody of interest which enriches DNA targets with high affinity for the antibody.

5.2.2 *In vitro* affinity maturation: from selection potentials to evolvability

For this project, we chose 3 antibody scaffolds that are the products of the *in vivo* immune response and were isolated from different HIV patients. However, the use of *in vivo* matured scaffolds comes with major caveats, whence the motivation for *in vitro* affinity maturation of antibodies: (i) Selective pressures in the in vivo affinity maturation are generally unknown: Selection for binding to pathogenic epitopes may be superposed with *e.g.* negative selection for autoreactivity (binding to self antigens) and may differ between maturation trajectories, especially between patients, and may vary over time, especially when the pathogen coevolves with the antibody. Evolvability can in principle be impacted by all these black box factors. (ii) Phylogenies of antibodies under *in vivo* maturation with sufficient temporal sampling (intervals at which blood is taken from a patient and sequenced), such as in [137], are rare.

To guarantee a meaningful definition of maturation degree, antibodies should therefore be matured *in vitro* under controlled selective pressures and mutational protocols; *in vitro* mutation and selection steps are repeatedly cycled to introduce fresh sequence diversity and enrich for sequences with improved performance, see figure 5.1(1). The selection steps would be identical to the ones performed here, or an improved version of it (see subsection 5.2.1). The mutation step could be simply performed by PCR replicating the antibody gene with a low-fidelity, error-prone DNA-polymerase which introduces (close-to) uniform, *i.e.* unbiased mutations along the gene sequence. The drawback of this *in vitro* mutagenesis method is the need for plasmid extraction and re-cloning the PCR-amplified gene back into the plasmid, which is a tedious task (see chapter A).

A recent technique developed with the group [148] uses *in vivo* mutagenesis which is faster,

but operates at low mutation rates and is far less controlled than the *in vitro* mutagenesis. Here, mutations can also occur outside the antibody gene, *e.g.* in the regulatory network of the M13 phage [166, 178], which is invisible in sequencing data and may induce unwanted amplification biases unrelated to the antibody phenotype. An example of an *in vitro* affinity maturation trajectory obtained with this method is shown in figure 5.2 where the third-best Germline V_H chain against the DNA1 target with CDR3 sequence `RTKH` was evolved for binding against DNA1 and resulted in the best Germline V_H chain with CDR3 sequence `RKKH`.

5.2.3 Selection potentials and evolvability *versus* other biophysical properties

In the literature, various biophysical and structural protein properties have been proposed to determine or correlate with evolvability, most notably thermal stability [11, 57], polarity of the fold [56], and modularity of functional organization within the fold [10, 179, 16]. Moreover, evolvability occasionally emerges as a by-product in theoretical models of protein evolution under fluctuating selection pressures for other properties [9, 10, 11, 12, 13]. However, none of these hypothesis have been put to the test on an experimental, real-life system as yet. We here designed a suitable model system in which we were able to define and infer selection potentials and which can be extended to measure evolvability (see subsection 5.2.2). Upon measuring both selection potentials/evolvability and biophysical properties of the same objects, precise dependencies between these properties can be revealed. Within the group, experimental techniques to quickly and efficiently measure thermal stability of proteins by limited proteolysis [180] and binding specificities by SELEX experiments [177] are being or have been set up. As a first step, these techniques could be used to measure thermal stabilities and recognition spectra of the antibodies that have been studied here and to confront these with the values of σ reported in this manuscript.

It should be noted that σ encodes for the binding specificity of the target molecule, not the one of the antibody. To measure the binding specificity of an antibody, the SELEX experiment reverses the respective roles of the antibody and the binding target in an otherwise identical experimental concept to the one we used here: The antibody of interest is immobilized on magnetic beads (or any other platform) and a library of potential binding targets is selected based on binding affinity towards the antibody; a schema of SELEX is shown in figure 5.1(2). In the literature, SELEX is used *e.g.* to measure the recognition spectra of transcription factors which are regulatory, DNA-binding proteins. Within our context, such an approach would be particularly meaningful to define and measure the specificities of our DNA-binding antibodies: DNA targets can themselves be defined on a sequence space and a library of DNA targets can be realized *e.g.* by randomizing the loop sequence of the DNA hairpin on each position to all four nucleotides. The DNA1 and DNA2 targets used here would be two distant sequences in such a target library.

It should also be noted that thermal stability possibly plays a particularly important role in

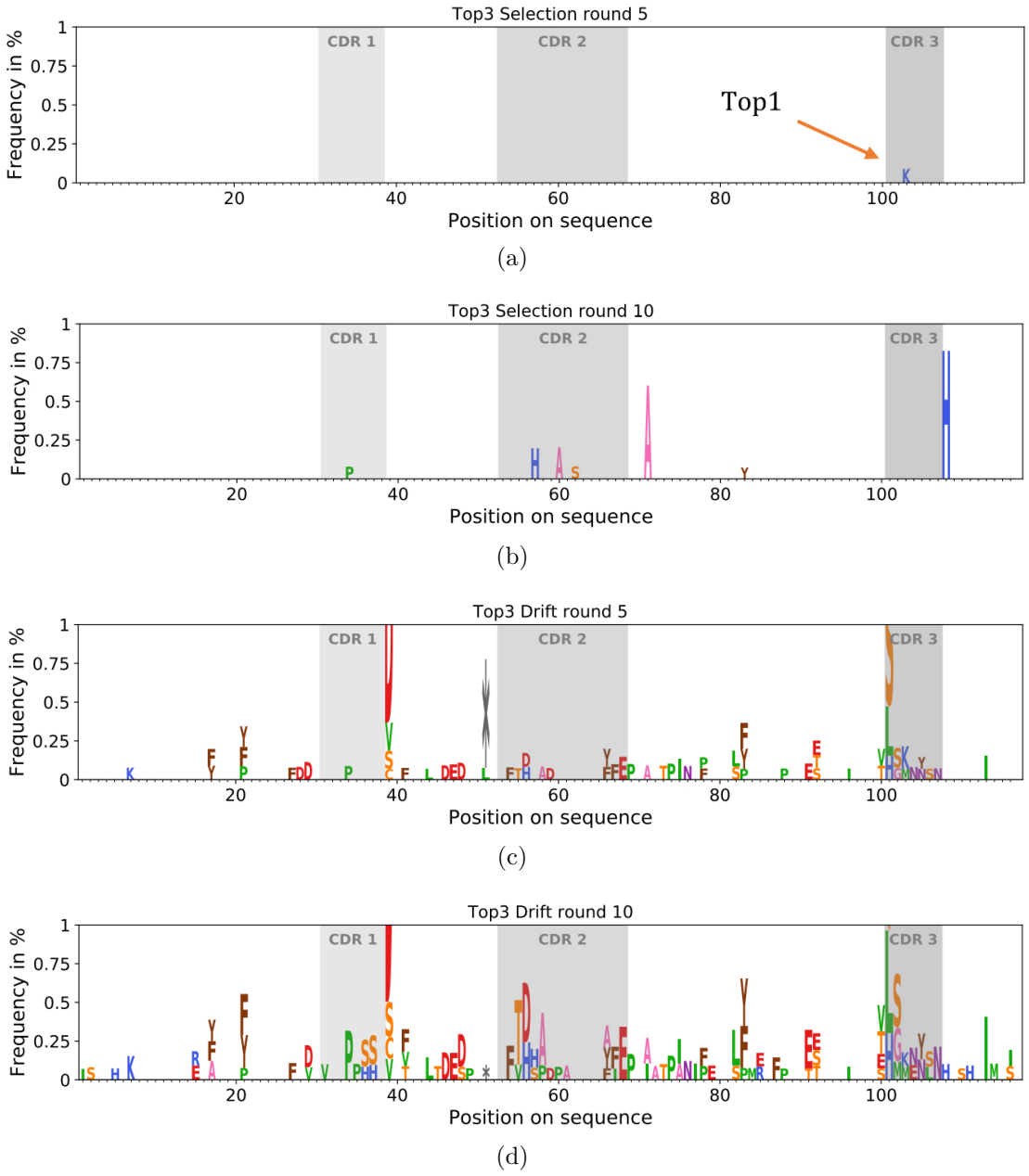


Fig. 5.2: Directed affinity maturation of an anti-DNA1 antibody. Sequence logos show the amino acids appearing in single mutants of the V_H chain after several rounds of random mutation and selection, starting from top-3 Germline antibody against DNA1 (CDR3 sequence `RTKH`). The letter heights represent their frequency in the evolved library. **(a)** After 5 rounds of mutation and selection for binding to DNA1: evolution of the top-3 sequence recovers the top-1 sequence (CDR3 sequence `RKKH`) which is only 1 nucleotide mutation away from top-3. **(b)** As (a), but after 10 rounds of maturation: Mutations in other CDRs and in framework regions (FWRs) are now selected. **(c)** After 5 rounds of drift, *i.e.* random mutation and no selection for binding, showing beneficial mutations unrelated to binding. **(d)** As (c), but after 10 rounds. The experiments were performed and the figures generated by Guillaume Villain [148].

our antibody model: Instead of using scFv fragments which consist of a V_H chain paired with a V_L chain, we are working with standalone V_H chains which puts them into a rather unnatural context. As noted earlier, scFv particles typically retain the properties of the full antibody, while V_H chains are known to be less stable. This loss in stability may be particularly pronounced in mature V_H chains, as somatic mutations are found to be not always affinity-enhancing and sometimes stability-enhancing [58], especially those far from the binding site. Such stability-enhancing mutations may rescue stability of the overall construct in the aftermath of affinity-enhancing but stability-impairing mutations by increasing the entanglement between the V_H and V_L chains (increased interface, increased number of inter-domain hydrogen bonds, etc.) [52, 53]. Taking away the V_L chains from mature V_H-V_L constructs may thus be more stability-affecting than in naïve V_H-V_L constructs. On the contrary, naïve V_H chains may be more self-sufficient as they require tolerance to a variety of V_L chains upon combinatorial primary repertoire formation (see section B.1).

5.2.4 Theoretical models of evolvability

Besides the experimental opportunities for future work, further study of theoretical models is required to understand the emergence and physical origins of evolvability, as well as its coevolution with other observables. Some other interesting questions are: How to reverse the affinity maturation process? That is, if we believe that affinity maturation (or any time-constant selective pressure) converts an evolvable antibody (or any object with evolutionary degrees of freedom) into a highly specialized, non-evolvable antibody, how to go the other way round? For this purpose, can evolvability be directed targeted by selection? These questions are somewhat related to the problem of inducing and maintaining "generalists" that are moderately fit across several, time-alternating environments, as opposed to "specialists" that are very fit in one but unviable in other environments. This problem has already been studied elsewhere [34, 30], but, in the most general formulation of the problem, the universe of possible environments (aka pathogenic challenges in the adaptive immune system) is virtually infinite and cannot be entirely sampled on relevant timescales. This turns the problem of inducing evolvable variants into a learning problem; the idea of the immune system predicting its future challenges has already been formulated [181]. The relevant framework to address all these questions is probably the one of spin glasses and/or neural and elastic networks, which have already been used in the past for the study of evolution in biological contexts [9, 10, 182]. Moreover, such abstract models should supposably allow to define and unify the concept of evolvability and selection potentials across various biological and non-biological contexts, such as material sciences [38, 39, 182].

Chapter A

Experimental protocols

In this chapter, we will provide the experimental protocols that are behind the results of this thesis. These methods are, up to customizations, part of the common experimental repertoire in molecular biology, including notably phage display [162], *in vitro* selection, bacterial genetics (regulation of cell growth using antibiotics and protein expression using the *lac* operon) and cloning [183], and (Illumina) sequencing. Standard tasks also performed here are cell cultures in liquid [184] and solid [185] cell growth media [186], as well as transformations [187]. A phage display and selection protocol customized to our antibody libraries was established and defined as part of a former PhD project within our group [1]. Here, the goal consists not only in providing a "manual" to reproduce or learn our experiments, but also to open up the black box that they may represent to theorists working on/with biological data.

The following sections cover all steps of a single selection round, starting from bacterial library cells stored at −80 °C and containing the randomized antibody genes on a plasmid (1 plasmid and thus 1 antibody sequence per cell). By the end of the selection round, we end up again with such cells, modulo the bias in sequence frequencies due to selection, *i.e.* more cells than before contain the beneficial sequences whereas less cells than before contain the deleterious sequences. These cells can then be used as input for another round of selection, be sequenced to measure the frequency of each antibody sequence, or be stored at −80 °C for later use. In short, the procedure is as follows: The initial cells express the antibodies, in such a way that they are displayed and released from the cell on phage particles containing the plasmid and thus the genetic information of the antibody on display. The target molecules will be placed on magnetic beads, so to be controllable by the experimenter. Antibodies and targets are brought into contact to let the binding reaction occur. Then, antibody-target complexes are held back by a magnet, whereas unbound antibodies are washed away. The complexes are destroyed using a suitable chemical and the selected phage are used to infect fresh cells (that have no plasmid yet), the plasmids being injected by the phage into the cells. A schematic of the experimental workflow is shown in figure 3.5(c). (See sections A.4, A.5, and A.6.)

We will start by some preparatory tasks required for the selection experiments, in particular the preparation of a number of reagents (see section A.1). In practice, these reagents may be produced in such quantities sufficient for lots of selection experiments, so they do not need to be repeated each time. Moreover, the cloning and mixing of isolated antibody sequences for mini library construction is described (see sections A.2 and A.3). Subsequently to the selections, the preparation steps for Illumina MiSeq sequencing involving several PCR reactions are described (see section A.7).

Following the protocols presented here, a single selection round (without preparatory tasks and sequencing preparation) is set to take 3 days, a lower bound being defined by the various incubation times in the protocol. This means that automatization would not reduce this length; it could at best allow for parallel realization of independent selection trajectories. There have been efforts within our group to accelerate the selections, *e.g.* by letting the so-far-consecutive steps of phage production and selection happen simultaneously, but these imply major changes in the biological constructs used here (cell strain, plasmid setup and combinations).

A.1 Reagents and materials

Plastics

Typical recipients for liquids used here are Eppendorf tubes (1.5 mL, 2.0 mL, 0.5 mL; Eppendorf, Hamburg, Germany), Falcon(R) tubes (50 mL, 15 mL; Corning Inc., Corning, NY, USA), PCR tubes (0.2 mL) and cold-resistent cryotubes (1.0 mL, 1.8 mL; Thermo Fisher Scientific, Waltham, MA, USA) for storage at −80 °C and −20 °C. Pipettes are used with standard tips or filtered tips (Sorenson and Starlab) when working with liquids containing phage. Falcon(R) pipettes (Corning Inc., Corning, NY, USA) are used for larger volumes. Eppendorf tubes intended to contain cells and those used during selecion are sterilized by autoclaving before use.

MilliQ water

We systematically use MilliQ water, *i.e.* distilled and deionized water, with resistivity of $\simeq$ 16 MΩ cm across all experimental steps. Notably for PCRs, digestions, and ligations, we use DNase-free MilliQ water (Invitrogen Life Technologies, Carlsbad, CA, USA).

Cell growth medium

A growth medium containing the required ingredients for the cells' metabolism is needed for cell growth and antibody production. Growth media are typically purchased in powder form. It is then dissolved in MilliQ water to obtain liquid growth media for liquid cell cultures. Alternatively, it is dissolved along with Agar powder in MilliQ water in order to obtain solid growth media that is used for solid cell cultures in Petri dishes. Here, we use 2xYT an LB growth medium (both from Sigma-Aldrich, Saint-Louis, MO, USA) for all our *E. coli* cell cultures. 2xYT is used for phage display and selection, while LB is used to prepare competent cells and transformation. For 2xYT, 15.5 g of powder, containing 5 g yeast extract (Y), 8 g tryptone (T), and 2.5 g NaCl, are dissolved in 1 L of MilliQ water. For LB, 25 g of powder, containing 5 g yeast extract (Y), 10 g tryptone (T), and 10 g NaCl, are dissolved in 1 L of MilliQ water. To obtain solid growth medium, 7.5 g of Agar powder are added. The solution is then autoclaved at 121 °C for 30 min in order to eliminate any contamination it may contain (otherwise background bacteria and fungi spores from the atmosphere may easily grow in there). If Agar is added, the liquid will solidify after autoclaving upon cooling down to room temperature. From now on, the growth media is systematically handled in sterile condition to prevent any contamination, *i.e.* under a biological hood.

Cell growth medium with glycerol

For the purpose of long-term storage of cells at −80 °C, the cells must be kept in around 25 % glycerol. These cell stocks are thus also called "glycerol stocks". The presence of glycerol notably prevents lethal cell membrane damage upon freezing and the increase in volume of water. We produced glycerol stock media by mixing (well!) 1 volume of unsterile 100 % glycerol (92.09 g.mol^{-1}; Sigma-Aldrich, Saint-Louis, MO, USA) with 1 volume of sterile 2xYT growth medium from A.1, thus 50 % glycerol final. Glycerol is a very viscuous liquid, thus much care is needed upon pipetting it and mixing it with the growth media (otherwise it separates into two phases of different viscosities with a visible interface). The glycerol stock medium then is then sterilized using a 0.22 µm vacuum-driven millipore Stericup(R) filter system (Sigma-Aldrich, Saint-Louis, MO, USA). Autoclaving is not recommended with glycerol. The glycerol stock medium is then stored at room temperature but handled only in sterile condition (biological hood). For the storage of cells from a liquid culture, 1 volume of glycerol stock medium is mixed (well!) with 1 volume of liquid culture (25 % glycerol final) in a cryotube (cold-resistent plastic tube), and finally stored in the −80 °C freezer.

Antibiotic stocks

We need the antibiotics ampicillin (Amp) and kanamycin (Km) to select for respectively the presence of the pIT2 plasmid carrying an Amp resistance gene and the helper plasmid featuring a Km resistance gene. Antibiotics are added into the cell growth medium from a stock. In liquid and solid cell cultures, Amp and Km are used at concentrations of $100\,\mu g.mL^{-1}$ and $50\,\mu g.mL^{-1}$, respectively. These antibiotics are also purchased as powders (ampicillin sodium salt, kanamycin sulfate; both Sigma-Aldrich, Saint-Louis, MO, USA) to be stored in the fridge at 4 °C. Stock solutions at $100\,mg.mL^{-1}$ for Amp or $50\,mg.mL^{-1}$ for Km are produced by weighing 2 g of Amp powder or 1 g of Km powder into a 50 mL Falcon tube and adding 20 mL of MilliQ water. The tube is then vortexed until complete dissolution of the powder and filtered through a 0.22 µm syringe-driven filter system (Biosigma) used with syringes from Terumo (Tokyo, Japan). The stocks are then stored in the −20 °C freezer. For use in a cell culture, they are diluted 1000 x, *i.e.* 1 volume of antibiotic stock solution for 1000 (more precisely: 999) volumes of growth medium.

Glucose stock solution

Glucose is used to regulate the *lac* operon in *E. coli* cells and thus the expression of the antibody or antibody-pIII fusion which is under the control of the *lac* promoter: Allowing cells to metabolize glucose ba adding it to the growth medium turns off the *lac* operon and thus represses expression of antibody(-pIII). We add glucose at all amplification (cell culture growth) steps where antibody(-pIII) expression is not needed and allows to prevent possible biases in antibody sequence frequencies due to antibody sequence-dependent effects of antibody(-pIII) expression on cell growth rates. Inversely, no glucose is added for antibody(-pIII) and displaying phage production steps which are however accompanied by unavoidable amplification biases. Glucose is used at a concentration of 1 % in liquid and solid cultures to prevent expression of antibody-pIII, *i.e.* 1 g/100 mL. Glucose stock solution is produced at 40 % concentration and then diluted 40 x in the liquid culture to obtain 1 % concentration final, that is 1 volume of glucose stock solution for 39 (more precisely: 40) volumes of growth medium. To make to stock solution, 80 g of D-(+)-glucose ($180.16\,g.mol^{-1}$; Sigma-Aldrich, Saint-Louis, MO, USA) powder is dissolved in MilliQ water to obtain 250 mL of glucose stock solution final. Note that this does not mean that the powder has to be dissolved in 250 mL of MilliQ water but less, due to the tare volume of the glucose powder. In practice, we dissolved the glucose powder first in 100 mL of MilliQ water and, after complete dissolution, added as much MilliQ water as needed to obtain 250 mL volume final. Dissolving that much glucose in such a small volume of MilliQ water takes several hours even under violent stirring. To accelerate the dissolution, the MilliQ water is slightly heated during dissolution (we worked at ≈ 40 °C which takes about 2 h), but not too much to prevent caramelization of the glucose. Also, the glucose powder needs to be poured continuously over an extended amount of time into the MilliQ water. Pouring the MilliQ water onto the glucose powder will yield in agglutination of the glucose powder and dissolution is set to fail. After dissolution, we filtered the

glucose stock solution through a 0.22 µm vacuum-driven Stericup filter system (Sigma-Aldrich, Saint-Louis, MO, USA) to filter out any contaminants. Autoclaving is not recommended due to the risk of caramelization. 40 % glucose is non-viable for bacteria, but contamination with fungi occurs very easily. After filtration, the glucose solution is systematically handled in sterile condition (biological hood) and kept at room temperature to quickly reveal any contamination.

Calcium chloride solution (with and without glycerol)

Calcium chloride ($CaCl_2$) is needed to make cells chemically competent (for DNA uptake). The $CaCl_2$ is purchased as powder (110.98 g.mol^{-1}; Sigma-Aldrich, Saint-Louis, MO, USA) and dissolved in sterilized MilliQ water. $CaCl_2$ at a concentration of 0.1 M is obtained by mixing 1 volume of $CaCl_2$ at 1 M with 9 volumes of sterilized MilliQ water. $CaCl_2$ at 0.1 M and 15 % glycerol is obtained by mixing 2 volumes of $CaCl_2$ at 1 M with 15 volumes of sterilized MilliQ water and 3 volumes of 100 % glycerol (92.09 g.mol^{-1}; Sigma-Aldrich, Saint-Louis, MO, USA). Finally, both the $CaCl_2$ at 0.1 M and the $CaCl_2$ at 0.1 M with glycerol are each one filtered through a 0.22 µm vacuum-driven millipore Stericup(R) filter system (Sigma-Aldrich, Saint-Louis, MO, USA).

IPTG

IPTG (Isopropyl β-D-1-thiogalactopyranoside) also regulates of the *lac* operon. Unlike the glucose, it induces expression of the protein construct which is cloned into an expression vector carrying an Amp resistance gene and which is under the control of the *lac* promoter (here, antibody(-pIII), eGFP and mCherry). It is purchased as powder (Sigma-Aldrich, Saint-Louis, MO, USA) and stored at −20 °C. IPTG stock solutions at a concentration of 300 mM are produced in MilliQ water and filtered similarly to the antibiotic stocks. It is diluted 1000 x into liquid cell cultures by adding 1 volume of IPTG stock for 1000 (more precisely: 999) volumes of cell culture, *i.e.* 300 µM. It is not stable and degrades upon frequent freezing and defrosting; therefore, it must not be defrosted for use in liquid cell cultures more than once or twice.

PBS

PBS (phosphate buffered saline; NaCl, KCl, Na_2HPO_4, KH_2PO_4 in MilliQ water; pH $\simeq$ 7.4) is a pH buffer solution used for the storage of proteins and phage. It is purchased either as 10 x concentrated liquid (Sigma-Aldrich, Saint-Louis, MO, USA; 1 volume of 10 x PBS concentrate is mixed with 9 volumes of MilliQ water to obtain 1 x PBS) or as pellets (Sigma-Aldrich, Saint-Louis, MO, USA) to be dissolved in MilliQ water (5 pellets in 1 L of MilliQ water for 1 L of 1 x PBS). The 1 x PBS is stored at room temperature.

Trizma

Trizma (2-Amino-2-(hydroxymethyl)-1,3-propanediol or Tris(hydroxy-methyl)aminomethane) is another pH buffer solution here used for the storage of phage. It is prepared at 1 M stock concentration and a pH of 7.4. 15.76 g Trizma hydrochloride (157.6 g.mol^{-1}; Sigma-Aldrich, Saint-Louis, MO, USA) powder is dissolved in MilliQ water until 100 mL of solution (Trizma acid) is obtained. In addition, 12.11 g Trizma base (121.14 g.mol^{-1}; Sigma-Aldrich, Saint-Louis, MO, USA) powder is dissolved in MilliQ water until 100 mL of solution (Trizma basic) is obtained. Then, Trizma acid and Trizma basic are mixed such that Trizma at pH = 7.4 is obtained. The pH is checked in real-time upon adding the Trizma basic to the Trizma acid (or *vice versa*) using a pH-meter. The Trizma is stored at room temperature.

Sodium hydroxyde

250 mL of sodium hydroxyde (NaOH) solution at 1 M final stock concentration is obtained by dissolving 10.0 g NaOH (40.00 g.mol^{-1}; Sigma-Aldrich, Saint-Louis, MO, USA) powder in MilliQ water until 250 mL of solution are obtained. The dissolution of the powder is difficult and is supported by stirring. The dissolution is performed under the chemical hood because vaporizing NaOH solution is very corrosive and damaging to the inhalatory organs of the body. The product is stored at room temperature.

EDTA

EDTA (disodium ethylenediamine tetraacetate, 2 H_2O) is a divalent cation chelator. 100 mL of EDTA solution at 0.5 M final stock concentration are obtained according to the protocol in [188]: 9.305 g EDTA (292.24 g.mol^{-1}; Sigma-Aldrich, Saint-Louis, MO, USA) powder are added to 40 mL of MilliQ water and stirred. The powder will not dissolve until the pH reaches 8.0. Using a pH-meter to scan the pH of the solution in real-time, NaOH solution at 1 M is added until pH = 8.0. Then, another 50 mL of MilliQ water are added and the solution is violently stirred until complete dissolution of the powder. The EDTA solution is stored at room temperature.

Bw1x washing buffer

Bw1x buffer washing liquid containing sodium chloride (NaCl) at 1 M final, Trizma at 5 mM final and pH = 7.4, as well as EDTA at 0.5 mM final is needed for the washing of streptavidin-coated magnetic beads. In order to produce 100 mL of Bw1x washing liquid, MilliQ water is poured onto 5.85 g NaCl (58.44g/mol; Sigma-Aldrich, Saint-Louis, MO, USA) powder until a volume of 50 mL

are obtained and stirred until all NaCl powder is dissolved. Then, 500 µL of Trizma at 1 M and pH = 7.4 and 100 µL EDTA at 0.5 M are added. Finally, MilliQ water is added until 100 mL of final volume are reached. The washing liquid is stored at room temperature.

Other reagents

Other notable chemicals and products needed later for phage display and selection comprise Dynabeads(R) M-280 Streptavidin and others (see figure 3.7; Invitrogen Life Technologies, Carlsbad, CA, USA), Na_2HPO_4 (sodium phosphate dibasic, 141.96 g.mol^{-1}; Sigma-Aldrich, Saint-Louis, MO, USA), NaH_2PO_4 (sodium phosphate monobasic, 119.98 g.mol^{-1}; Sigma-Aldrich, Saint-Louis, MO, USA), Tween20 surfactant (viscous liquid, 1.095 g.mL^{-1}; Sigma-Aldrich, Saint-Louis, MO, USA), as well as triethylamine (TEA) at >99% (101.19 g.mol^{-1}; Sigma-Aldrich, Saint-Louis, MO, USA), and Javel water for neutralization of phage. DNase I (Sigma-Aldrich, Saint-Louis, MO, USA) and lyzozyme (Sigma-Aldrich, Saint-Louis, MO, USA) are needed for the harvest of expressed fluorescent proteins from cells. For electrophoresis, Agar (Sigma-Aldrich, Saint-Louis, MO, USA), 6 x Gel Loading Dye (purple) without SDS (New England Biolabs, Ipswich, MA, USA), and 10,000 x SYBR(R) Safe DNA gel stain in DMSO (Invitrogen Life Technologies, Carlsbad, CA, USA) are needed. The restriction enzymes XhoI and BssHII, CutSmart 10 x buffer, as well as T4 DNA ligase and its buffer 10 x concentrated (all from New England Biolabs, Ipswich, MA, USA) are needed. All these chemicals and products are all used as purchased. The *E. coli* TG1 and C3019 cell strains relevant for cloning and phage display, respectively, we purchased (New England Biolabs, Ipswich, MA, USA).

Petri dishes for bacterial growth

Here, we describe how we make bacterial growth plates that are used for various purposes across the project: counting the number (or concentration) of cells in a liquid cell culture and obtaining isogenic cells. The bottled solid growth medium with Agar is heated and thus melted in a microwave. The heating must be slow and with regular shaking over a longer period of time in order to melt the bulk without boiling the boundaries. Upon cooling down, glucose and antibiotics, if needed, are added only when right above the solidification temperature in order to avoid caramelization of the glucose and degradation of the antibiotics. After shaking, the growth medium is poured or pipetted into the Petri dishes, approximately 25 mL per small Petri dish (round, 4.25 cm radius; Greiner Bio-One, Kremsmünster, Austria), 250 mL per large Petri dish (quadratic, 24.5 cm edge length; Corning Inc., Corning, NY, USA). Air bubbles are destroyed or at least moved to the boundary of the plate using pipette tips. After solidification, the Petri dishes are stored in the fridge at 4 °C until use and discarded when unused within roughly a month.

M13 helper phage stock

To start the production of displaying phage, library cells need to be infected with helper phage. A stock of several mL of helper phage at a concentration of approximately $10^{12}\,\mathrm{mL}^{-1}$ (meaning 10^{12} phage particle per mL) starting from a droplet of highly concentrated phage (5 µL phages at $10^{13}\,\mathrm{mL}^{-1}$) is produced according to the following protocol: First, the droplet is defrosted and diluted to 50 µL to a concentration of $10^{12}\,\mathrm{mL}^{-1}$ by adding 20 µL of 2xYT growth medium and 25 µL of 50 % glycerol in 2xYT.

Then, fresh exponential cells are infected with this helper phage solution in order to then produce even more helper phage: A fresh liquid culture of TG1 cells is started with 10 mL of 2xYT growth medium and 100 µL of an overnight TG1 liquid culture (100 x dilution) in a 50 mL Falcon tube. The culture is incubated at 37 °C temperature and a rotational speed of 180 rpm (rounds per minute) until a bacterial density of $\mathrm{OD}_{600} = 0.4$ is reached (exponential growth phase; OD_{600} measures the absorbance of monochromatic light of a wavelength of 600 nm shined through a liquid cell culture), which takes approximately 2 h. At this point, 10 µL of helper phage at $10^{12}\,\mathrm{mL}^{-1}$ are added to the cell culture. In order to let the infection happen, the culture is incubated for another 30 min at 37 °C, at rest. In the meantime, the next liquid culture for the phage production is prepared: 50 mL 2xYT growth medium and 50 µL kanamycin antibiotic are given to a 250 mL flask. By the end of 30 min, the infected cells are centrifuged at 3,200 g (a velocity leading to a centrifugal force equivalent to the force acting on a body in a gravitational field 3,200 x as strong as the one at the Earth's surface), 25 °C for 10 min. As a result, the heavy cells (heavy compared to all other things in the culture) accumulate at the bottom of the tube. The supernatant is then poured away, leaving behind only the cells in the tube. The supernatant is neutralized under the chemical hood using Javel water. The cells are now resuspended into the new culture (which was prepared during the 30 min wait). The culture is incubated overnight (meaning for $\geq$ 16 h) at 30 °C, 180 rpm. During this time, the infected cells will produce and release into the culture new helper phage particles. The presence of kanamycin antibiotic assures that only infected cells can survive and grow in the culture, as infected cells acquired a plasmid carrying a kanamycin resistance gene. In the absence of kanamycin, cells who lose the helper plasmid may have a fitness advantage, as they do not need to produce phage particles and are not penalized upon losing the resistance gene, and may therefore take over the cell culture.

On the following day, the cell culture is transferred to a new Falcon tube and centrifuged at 10'800 g, 25 °C for 10 min. This time, however, we are interested in the supernatant as it contains the produced phage particles. Thus, the supernatant is poured into yet another Falcon tube, while the one with the cell pellet is discarded.

In order to obtain an estimation of the helper phage concentration, we proceed with serial dilutions and infection of fresh exponential cells: 500 µL of $10^{2,4,6,8,9,10,11,12}$ x dilutions of helper phage solution is made in 1.5 mL Eppendorf safe-lock tubes by pipetting into each tube 495 µL

of 2xYT growth medium and 5 µL of previous dilution for a dilution factor of 100 x, or 450 µL of 2xYT growth medium and 50 µL of previous dilution for a dilution factor of 10 x. The dilutions are vortexed before going forward with the next dilution in order to have the phage well-mixed in the liquid. Then, 500 µL of a fresh exponential cell culture are added to each dilution and incubated for infection at 37 °C, at rest for 30 min. The helper phage are exposed to an excess of cells, meaning that each phage should give rise to one infected cell (no competition of phage for cells): The number of helper phage per tube for the 10^{-d} x dilution is $10^{-d} \cdot 10^{12}\,\mathrm{mL}^{-1} \cdot 5 \cdot 10^{-1}\,\mathrm{mL} = 5 \cdot 10^{11-d}$, whereas the number of cells at $\mathrm{OD}_{600} = 0.4$ is $10^{8}\,\mathrm{mL}^{-1} \cdot 5 \cdot 10^{-1}\,\mathrm{mL} = 5 \cdot 10^{7}$. Thus, the cells outnumber the helper phage for dilutions higher than 10^{5} x. As infection is very efficient, each phage present in the tube will give rise to an infection event. Moreover, each cell can be infected only once. As a result, the number of phage in the tube directly translates into the number of infected cells and cell colonies after cell growth on a plate. This allows to infer the number of helper phage indirectly by counting the number of cell colonies. In order to count the number of infected cells, 100 µL of each dilution is distributed over a Petri dish coated with 2xYT-kanamycin selective growth medium. Each infected cell should give rise to one colony upon cell growth, due to its kanamycin resistance, whereas uninfected cells should not grow. In the end, counting colonies is somewhat equivalent to counting the number of helper phage, see also figure 3.9(a). The plates are incubated overnight at 37 °C for cell growth. The supernatant is kept in the fridge at 4 °C.

On the following day, the helper phage concentration $[\mathrm{H}\phi]$ in the supernatant is estimated using

$$[\mathrm{H}\phi] = 2 \cdot 10^{d+1} \cdot N_{\mathrm{col.}}(d)\,\mathrm{mL}^{-1}, \tag{A.1}$$

for various d where $N_{\mathrm{col.}}(d)$ is the number of colonies observed for the 10^{-d} x dilution. If the estimated phage concentration is satisfactory, *i.e.* $\geq 10^{11}\,\mathrm{mL}^{-1}$, the helper phage at 4 °C is prepared for long-term storage at −80 °C for later use in phage display experiments: 1 volume of helper phage is mixed with 1 volume of glycerol stock medium from A.1, aliquoted into cryotubes, and moved to the −80 °C freezer.

A.2 Cloning

The goal of cloning is to insert a target gene into the expression vector plasmid pIT2 which in turn must be taken up by expression strain cells, here TG1 cells. For the purpose of this project, we need to clone $\mathrm{V_H}$ genes with specific CDR3 sequences into TG1 cells. To this goal, we streak cells from the relevant library cell glycerol stock on Petri dish with 2xYT-ampicillin-glucose growth medium and grow them overnight at 37 °C. The colonies seen on the following day contain the plasmid with the relevant library scaffold but random CDR3 sequences. Any colony can be used as a template for the cloning procedure in which the random CDR3 sequence is to be replaced by the target CDR3. An overnight liquid culture is grown according to and miniprepped on the

following day to extract the plasmid. The idea for what follows is to remove the random CDR3 sequence by cutting the plasmid DNA immediately up- and downstream of the CDR3 by digestion with restriction enzymes, removing it, and replacing it by the target CDR3 sequence by ligation, see also figure 3.3. We proceeded this way with all target CDR3 sequences presented in table 4.1. A similar strategy was followed by Boyer *et al.* [1] to clone the VH genes into the pIT2 plasmid using other restriction sites up- and downstream of the target gene region, where a default target gene in the plasmid was replaced by the VH genes.

As an example, consider here the cloning of the best CDR3 sequence of the Germline library against the DNA1 target which has the CDR3 nucleotide sequence `CGGAAGAAGCAT`. By picking randomly from the Germline library, a plasmid is obtained, that around the CDR3 has the sequence `(GCGCGC)XXXXXXXXXXXXTTCGACTACTGGGGTCAGGGTACCCTGGTTACCGT(CTCGAG)` with some random CDR3 denoted by `X`s, see also figure 3.4. Restriction enzymes can cut double-stranded DNA at specific sequences. For instance, the restriction enzyme XhoI recognizes the sequence `C'TCGA,G`, while BssHII recognizes `G'CGCG,C`. The cut of the DNA is performed at the location indicated by `'` on the $5' \to 3'$ strand; note that sequences recognized by restriction enzymes map onto themselves upon reverse-complementation and the $3' \to 5'$ strand is thus cut at a different position, indicated by `,`. These two sequences are contained in the V_H genes right upstream and a bit further downstream of the CDR3 sequence, indicated above by brackets `()`. Thus, the CDR3 sequence can be cut away from the template plasmid using these two restriction enzymes.

The target CDR3 sequences including the sequences recognized by XhoI and BssHII are purchased as dsDNA from IDT, Leuven, Belgium. In our example, the purchased sequence is thus `(GCGCGC)CGGAAGAAGCATTTCGACTACTGGGGTCAGGGTACCCTGGTTACCGT(CTCGAG)`. After cutting this sequence with the same restriction enzymes, an insert that fits into the template plasmid digested with the same restriction enzymes is obtained. This insert is glued into the template in a reaction called ligation.

Digestion, agarose gels

As a first step, template plasmids are extracted from an overnight grown liquid culture using a commercial DNA purification kit (Macherey-Nagel, Düren, Germany; see manual for a detailed protocol), and $\geq 1\,\mu g$ of template plasmids are digested with XhoI and BssHII. The concentration of template plasmids is measured by Nanodrop (BioPhotometer(R), Eppendorf, Hamburg, Germany), allowing to compute the volume required for $\geq 1\,\mu g$ of plasmid DNA, say $5.0\,\mu L$. XhoI and BssHII are active at two different temperatures, 37 °C and 50 °C, respectively. Hence, the digestions can not be done simultaneously, but need to be performed sequentially. On ice, the $5.0\,\mu L$ of plasmid DNA is pipetted into a PCR tube, along with $12.0\,\mu L$ of DNase-free MilliQ water, $2.0\,\mu L$ of 10 x CutSmart buffer (1 x CutSmart buffer final, as required), the buffer in which the enzyme is optimally active, and $1.0\,\mu L$ of BssHII enzyme ($20.0\,\mu L$ volume final). The enzyme

is added last. The tube is flipped, centrifuged briefly and then incubated at 50 °C in a PCR machine for 90 min. Subsequently, 3.5 µL of DNase-free MilliQ water, 0.5 µL of 10 x CutSmart buffer (again 1 x CutSmart buffer final), and 1.0 µL of XhoI enzyme (again lastly) are added to the tube (3.5 µL volume final), which is then flipped, centrifuged, and incubated at 37 °C in a PCR machine for another 90 min. A negative control is performed simultaneously where the above steps are followed in the exact same way except for the BssHII and XhoI which are respectively replaced by the same volume of additional DNase-free MilliQ water.

The product of the ligation is then submitted to electrophoresis, *i.e.* run on an agarose gel: In an electric field, the negatively charged DNA is dragged through the gel lattice, with smaller pieces of DNA going across more easily/rapidly than larger ones. Hence, the digestion products are separated according to their size and the CDR3 sequence is separated from the much larger remaining part of the plasmid. A 1 % agarose gel is prepared by dissolving 1 g of agarose powder in 100 mL of 1 x TAE running buffer upon heating in a microwave. Then, 10 µL of 10^4 x SYBR(R) Safe DNA gel stain are added (1 x SYBR(R) Safe final) and the liquid is poured into a gel mold, and let to cool down to room temperature to obtain the gel upon solidification. SYBR(R) Safe is a DNA intercalating dye which settles on the "ladders" of the double-stranded molecules and must, for this very reason and the associated suspected cancerogenicity, be handled very carefully by the experimenter. The gel is placed into the electrophoresis station, immersed in TAE running buffer, and connected at its extremities to electrodes.

1 volume of 6 x Purple Loading Dye are added to 5 volumes of digestion product so that the final product is 1 x Gel Loading Dye, *i.e.* 5 µL of Loading Dye for 25 µL of digestion product. After pipetting up and down until the volume is well-mixed, the sample is loaded into a well of the gel by careful pipetting. In addition, 5 µL of ladder, *i.e.* a sample containing DNA molecules of various known sizes, is loaded into a separate well. These are used later as a reference to estimate the size of the sample. An electric tension of 100 V is then applied to the gel for approximately 40 min, with the electric field being parallel to the direction of motion and the anode (positive electrode) on the opposite site of the gel, such that the DNA is forced to traverse the gel. The electrophoresis is finished when the visible violet spots from the Loading Dye have reached the lowest quarter of the gel. The gel is removed, wiped to remove any liquid, and shined with blue or UV light. Any DNA becomes visible as bands on the gel due to the SYBR(R) Safe which is fluorescent in blue and UV light. The location of the band(s) is compared with those of the reference bands in the lane with the ladder to estimate the size of the digestion product. The observed size is then compared with the expected size. Finally, the plasmid DNA is purified by cutting out the band from the gel and removing the agarose using a commercial gel extraction and purification kit (Macherey-Nagel, Düren, Germany; see manual for a detailed protocol). As a result, around 30 µL of purified plasmid DNA in MilliQ water (alternatively: in elution buffer) with DNA concentrations of typically around 20 ng μL^{-1} is obtained.

The similar procedure is followed for 500 ng of the purchased DNA with the new CDR3 sequence to obtain purified insert DNA. The purchased DNA is first resuspended in MilliQ water and

aliquoted so to have it at a concentration of $50\,\mathrm{ng}\,\mu\mathrm{L}^{-1}$. Thus, the first digestion is started with $10\,\mu\mathrm{L}$ of DNA, $2.5\,\mu\mathrm{L}$ of DNase-free MilliQ water, $1.5\,\mu\mathrm{L}$ of 10 x CutSmart buffer (1 x final), and $1.0\,\mu\mathrm{L}$ of BssHII enzyme ($15\,\mu\mathrm{L}$ volume final). For the second digestion, $3.5\,\mu\mathrm{L}$ of DNase-free MilliQ water, $0.5\,\mu\mathrm{L}$ of 10 x CutSmart buffer (1 x final), and $1.0\,\mu\mathrm{L}$ of XhoI enzyme ($20\,\mu\mathrm{L}$ volume final) are added. The enzymes are added last in each digestion. The digestion product is directly purified without running a gel, using a commercial PCR clean-up kit (Macherey-Nagel, Düren, Germany; see manual for a detailed protocol) which does not retain the tiny DNA fragments that were cut off at the 5′ and 3′ ends of the DNA sequences. The insert of a size of $\simeq 50\,\mathrm{bp}$ is too small to obtain a clear band on an agarose gel. The concentration of the purified insert is measured by Nanodrop, in this example it amounts to $10\,\mathrm{ng}\,\mu\mathrm{L}^{-1}$, but insert concentrations are way smaller and even indistinguishable in other cases. Still, this is not crucial for the following ligation and transformation as few copies of the insert DNA can in principle be sufficient. (This holds only for the cloning of a single sequence at a time; for the cloning of libraries, a sufficient oversampling is indeed required.)

Ligation

The two pieces of the "puzzle", *i.e.* the linearized template plasmid (backbone) and the CDR3 insert are now sticked together by a ligation reaction. The volume of insert DNA V_i required for the ligation is typically calculated by ligation calculators from

$$N_i = \frac{[i]V_i}{L_i} \stackrel{!}{=} r\frac{[b]V_b}{L_b} = rN_b m_{\mathrm{insert}} = \frac{\mathrm{length}_{\mathrm{insert}}}{\mathrm{length}_{\mathrm{backbone}}}\frac{1}{\text{insert ratio}} m_{\mathrm{backbone}}. \tag{A.2}$$

where, i and b refer to the insert and backbone, respectively, N, $[\cdot]$, V., and L. refer to number of copies, mass concentration, volume, and sequence length, respectively, and r is an insert (to backbone) ratio typically chosen to be $r = 3$. Here, we have typically $[b] \simeq [i] \simeq 10\,\mathrm{ng}\,\mu\mathrm{L}^{-1}$, $L_i \simeq 70\,\mathrm{bp}$, and $L_b \simeq 4500\,\mathrm{bp}$ and we typically choose $V_b = 3\,\mu\mathrm{L}$. The required volume of insert DNA V_i according to equation (A.2) is then much less than $1\,\mu\mathrm{L}$, so we simply provide insert DNA in excess by taking $3\,\mu\mathrm{L}$. Thus, $3.0\,\mu\mathrm{L}$ of plasmid backbone and $3.0\,\mu\mathrm{L}$ of insert DNA are pipetted to a PCR tube, along with $11.0\,\mu\mathrm{L}$ of MilliQ water, $2.0\,\mu\mathrm{L}$ of 10 x T4 DNA ligase buffer (thus 1 x final), and $1.0\,\mu\mathrm{L}$ of T4 DNA ligase ($20.0\,\mu\mathrm{L}$ volume final). The ligase is added last. The tube is incubated at 25 °C in a PCR machine for 15 min. A negative control is realized by performing an identical ligase reaction, but simply leaving out the insert. In this case, the $1.0\,\mu\mathrm{L}$ of insert DNA is replaced by $1.0\,\mu\mathrm{L}$ of additional DNase-free MilliQ water. This negative control should not give rise to an intact circular plasmid, unless at least one of the previous digestion steps was unsuccessful and the plasmid can recircularize by itself.

Competent cells

In order to transform the ligated plasmid into cells, *i.e.* incorporate the new plasmid into the cells, the cells must be "competent" for transformation. There are two major methods of transformation: 1) by heat shock using chemically competent cells where the cell membrane is fractured for a short period of time by a sudden increase in temperature so that the plasmid DNA can enter the cell, 2) by electroporation using electrocompetent cells, where the same is achieved by applying a voltage. Electroporation is more delicate in practice, but generally leads to a higher transformation efficiency, *i.e.* a larger amount of transformed cells. It is therefore the preferred choice for the transformation of libraries where many different sequences are transformed at the same time and a sufficient oversampling of sequence space is needed to avoid biases and elimination of sequences by chance. But here, we only need to clone ~ 20 single antibody sequences each at a time, and we therefore opt for chemical transformation. Competent can be either purchased (expensive!) or self-made, but commercial competent cells again generally have higher transformation efficiencies and are therefore again preferred for library transformation. We opt for self-made chemically competent cells as, in principle(!), a single transformant should be enough for successful transformation of a single sequence. We also decided to not directly transform into the expression strain for phage display *E. coli* TG1, but to first transform to the *E. coli* C3019 strain which is optimized for transformation efficiency and then transfer to TG1. The transfer of an intact plasmid by extraction from one cell strain and retransformation into another cell strain comes with higher efficiency and less pathologies (see A.2) than the transformation of ligation products.

In the following steps, the growth medium does not contain antibiotics; it is thus necessary to work under the biological hood to avoid growth of and contamination by bacteria/fungi from the environment. It is beneficial to use (close to) isogenetic cells for transformations. Therefore, we start a liquid cell culture with cells from a single cell colony rather than a glycerol stock: For both cell strains, C3019 and TG1, a few cells are streaked on a plate with LB growth medium and grown overnight at 37 °C. On the following day, a liquid cell culture in 5 mL of LB growth medium in a 50 mL Falcon tube is started using cells from a single colony on the LB plates and grown overnight at 30 °C, 180 rpm. On the following day, a new liquid culture is started in a larger volume of 100 mL of LB growth medium in a 500 mL flask with 1 mL of overnight culture (100 x dilution) and grown at 37 °C, 180 rpm until $OD_{600} = 0.4$ (not more!) which takes around $1.5 - 2$ h. In the meantime, the centrifuge is cooled to a temperature of 4 °C. As the cell culture reaches the required OD_{600}, it is partitioned into two equal volumes which are decanted into two 50 mL Falcon tubes. The culture is placed on brayed ice for 10 min to cool down. Henceforth, the cells need to be permanently kept in a cold environment. The cultures are centrifuged at 4000 rpm, 4 °C for 10 min. The supernatant is poured and the pellets are resuspended gently in 5 mL of cold 0.1 M $CaCl_2$ (which was stored in the fridge at 4 °C beforehand) each by slow pipetting up and down. The cells' membrane is mechanically fragile and prone to disruption. The cells are then further chilled on ice for 20 min. After another centrifugation at 4000 rpm, 4 °C for 10 min,

the supernatant is again poured away and the pellets resuspended gently in 5 mL of cold 0.1 M $CaCl_2$ with 15 % glycerol (which was also stored in the fridge at 4 °C). The cells are aliquoted into sterilized and pre-cooled (on ice) 1.5 mL Eppendorf tubes, 300 µL of cells per cells. Finally, the cell aliquots are shock-freezed at liquid nitrogen and stored at −80 °C. (The competence of cells decreases with time and the cells should no longer be used after 6 months.)

Transformation

The ligated plasmid with the replaced CDR3 is inserted into chemically competent C3019 cells by heat shock. Competent C3019 cells are taken from the −80 °C freezer and placed on ice for 10 min, so they can defrost. 50 µL of competent cells are mixed with 7.5 µL of ligation product in a tube and chilled on ice for another 30 min. Then, the cells are quickly moved to a water bath at 42 °C for 30 s, before being put back on ice for 1 min. Due to the heat shock, the cells are exhausted and must be fed and oxygenated: 400 µL of pre-heated LB growth medium is added to the cells, followed by incubation at 37 °C for 30 min under intense shaking, $\geq$ 300 rpm (1 h of incubation should be performed when transforming plasmids with resistances different from ampicillin). Finally, the transformed cells are centrifuged, resuspended in 150 µL of LB growth medium, and plated on selective LB-ampicillin plates with 1 % glucose for growth at 37 °C overnight. The ampicillin assures only transformed cells with the pIT2 plasmid and thus an ampicillin resistance gene would grow. The glucose inhibits expression of the antibody which is not needed at this step. Another transformation without ligation product, as well as a transformation with pUC19 plasmid are performed in parallel as negative and positive controls, respectively. The negative control should not give rise to colonies, while the positive control should if the transformation was successful. The next day, if the transformation was successful and colonies appear on the plate, liquid cultures are started in 6 mL of 2xYT growth medium, ampicillin, and 1 % glucose, and then grown overnight at 37 °C, 180 rpm using cells from one or several colonies (1 culture for each colony to be tested). On the following day, 1 mL of culture is used to make a glycerol stock of the transformed cells: mixing with 1 mL of 50 % glycerol in 2xYT and storage at −80 °C. The remaining 5 mL of culture are used to check if the plasmid has the expected sequence by Sanger sequencing: The plasmids are extracted from the cells using a commercial DNA extraction kit (Macherey-Nagel, Düren, Germany; see manual for a detailed protocol) and and sent for Sanger sequencing to GATC, Konstanz, Germany (now owned by Eurofins, Luxembourg, Luxembourg). If the sequencing yielded the expected sequence, the glycerol stock of transformed cells is used for transfer into TG1, otherwise it is discarded. While the testing of a single colony of transformants is in principle sufficient, if it carries the correct sequence right away, the cells do not necessarily carry the correct sequence and finding a colony with the correct sequence (if any) is more or less a matter of pure luck: In our case, we had to test many transformants per CDR3 sequence in some cases ($\simeq$ 200 colonies and Sanger sequencings in total for the $\simeq$ 20 sequences in table 4.1) to find the correct sequence and for some sequences the cloning was unsuccessful (see table 4.1). The cells tend to introduce many kinds of pathologies to the ligated plasmids upon transformation (ranging from

point mutations and indels of single or several base pairs, especially near the restriction sites, to the deletion of the entire antibody gene), presumably linked to the potential toxicity or at least uselessness of the expressed antibody for the cells. This is a general feature of the cloning of any non-self protein into cells and is not restricted to antibodies.

Transfer to the expression strain

If the correct antibody sequence was found in a tested cell colony in the previous step, a final transfer of the plasmid from the cloning strain C3019 to the expression strain TG1 for the phage display is performed. This is done by plasmid extraction using commercial DNA extraction kits (Macherey-Nagel, Düren, Germany) and transformation of the extracted plasmid to chemically competent TG1 cells. The transformation protocol is identical to the one of the first transformation, except that 1.5 µL of plasmid (instead of the 7.5 µL of ligation product) are added to the competent cells. Again, glycerol stocks of transformed TG1 cells are prepared after transformation and overnight liquid cultures, and the sequence of the antibody gene is checked by plasmid extraction and Sanger sequencing. Here, testing of a single colony to find the correct sequence was indeed sufficient in all cases.

A.3 Mini libraries

The mini libraries are obtained by cloning the CDR3 sequences in table 4.1 individually according to the protocol in A.2 and then mixing cells carrying plasmids with DNA1-specific, DNA2-specific, or randomly picked CDR3: One liquid culture per sequence is started from the corresponding TG1 cell glycerol stocks in 5 mL 2xYT, ampicillin, and 1 % glucose each, and grown overnight at 37 °C, 180 rpm. On the following day, 4 mL from each of the 10 DNA1-specific cultures are poured together (40 mL final), as well as 4 mL from each of the 9 top-DNA2 (36 mL final) and 4 mL from all 10 random clones (40 mL final), thus yielding the “top DNA1”, “top DNA2”, and “random” mini libraries. Each mini library is centrifuged at 3200 g, 25 °C for 10 min, the supernatants are discarded and the pellets resuspended in 2 mL of 2xYT growth medium and 2 mL of 50 % glycerol in 2xYT (thus 25 % glycerol final), aliquoted into cryotubes, and stored at −80 °C. The density of these glycerol stocks is measured by Nanodrop to be around $OD_{600} \approx 36$ (100 x dilutions had $OD_{600} \approx 0.36$).

A.4 Phage display

In this step, antibody displaying phage particles are produced by starting from library TG1 cells that are stored at −80 °C (the full Germ, Lmtd, and BnAb libraries, as well as the DNA1-specific, DNA2-specific, and random mini libraries are available in this format, see A.3). The steps to be followed here are somewhat similar to the ones in the helper phage production protocol in A.1: First, a liquid culture of library cells is started, which is then infected by helper phage. As a consequence, the infected library cells then produce and release antibody displaying phage. If several selection experiments on different libraries are planned in parallel, the following steps apply independently to all libraries to be screened.

A liquid culture with 5 mL 2xYT growth medium, 1 % glucose, and ampicillin is started from the library's glycerol stock and grown overnight at 30 °C, 180 rpm. On the following day, a new liquid culture of library cells is started from the overnight culture and grown until exponential growth phase is reached: 20 mL 2xYT growth medium with 1 % glucose and ampicillin are pipetted to a 250 mL erlenmeyer flask and 200 µL of library cells from the overnight culture are added (100 x dilution). The bacterial density OD_{600} is measured by Nanodrop and should be $OD_{600} < 0.1$. The culture is then incubated at 37 °C, 180 rpm until $OD_{600} = 0.4$ which takes approx. 2 h (bacterial density is regularly checked). Then, the culture is transferred to a 50 mL Falcon tube, 150 µL of helper phage stock from A.1 is defrosted and added to the culture, which is then further incubated for infection at 37 °C, at rest for $\geq$ 30 min. Here, it should be checked that helper phage are in excess, as we want all cells in the culture to be infected: Indeed, if the concentration of the helper phage stock is at least $10^{11}\,\mathrm{mL}^{-1}$, then the number of cells, $0.4 \cdot 8 \cdot 10^8\,\mathrm{mL}^{-1} \cdot 20\,\mathrm{mL} \simeq 6.4 \cdot 10^9$ ($OD_{600} = 1.0$ corresponds to $8 \cdot 10^8\,\mathrm{mL}^{-1}$), is by a factor of at least 2 less than the number of phage $10^{11}\,\mathrm{mL}^{-1} \cdot 1.5 \cdot 10^{-1}\,\mathrm{mL} = 1.5 \cdot 10^{10}$. Note that increasing the volume of helper phage stock added may be harmful to the culture: The helper phage stock contains kanamycin from the helper phage production step in A.1, but the library cells are not kanamycin-resistant before infection. In the meantime, the next liquid culture is prepared: 50 mL of 2xYT growth medium is pipetted along with both ampicillin and kanamycin to a 250 mL flask. After infection, the library cell culture is centrifuged at 25 °C, 3'200 g for 10 min, the supernatant is poured away and neutralized with Javel water (to kill remaining helper phage), and any remaining liquid is aspired with a filtered pipette in order to remove as much as helper phage as possible. The cell pellet is then resuspended into the new culture and incubated at 30 °C, 180 rpm for 7 h. The selective growth medium with ampicillin and kanamycin is viable only for infected library cells carrying both the pIT2 plasmid and the helper plasmid. These cells produce antibody-displaying phage particles during this incubation time. By the end of 7 h, the cultures are transferred to 50 mL Falcon tubes and centrifuged at 10'800 g, 25 °C for 10 min. The supernatant containing the displaying phage is poured into new Falcon tubes and stored in the fridge at 4 °C, the cell pellet is discarded. The supernatant can be kept and used for antibody screen for at most 24 h. We should expect a rather limited V_H stability and lifetime in this unnatural context and assume unfolding beyond 24 h from expression where they would be no longer functional. The selection step should thus be performed

within this time window of 24 h and a fresh phage production should be preferred otherwise. We do not precipitate the displaying phage before the selection step, as advised by Philippe Minard (Université Paris-Sud), in order to minimize phage-phage interactions which contribute to the noise level of the experiment.

A.5 Target production and immobilization

Target molecules are immobilized on streptavidin-coated magnetic beads (Dynabeads(R) M-280 Streptavidin). The immobilization is achieved by strong binding of the target molecules to the streptavidin molecules, either through an attached biotin in the case of the DNA targets or through a SBP tag in the case of the case of the protein targets. The targets bind to the beads (via streptavidin) sufficiently strongly to not dissociate upon any upcoming incubation and washing steps during immobilization and selection.

The biotinylated DNA hairpin targets (DNA1 and DNA2) are purchased from IDT (Leuven, Belgium); the fusion with a biotin is realized at their $5'$ ends. The affinity between biotin with streptavidin is of the order of $K \simeq 10^{-14}$ M and known to be one of the strongest naturally occuring non-covalent interactions. The DNA targets are shipped in purified, solid form. In order to have them in solution, DNase-free MilliQ water is added in such a quantity as needed for a stock concentration of 400 µM target DNA final (the volume to be added is typically indicated on a data sheet). The purified target DNA is scratched from the tube wall with a pipette tip and is then incubated in the water at room temperature for 1 h. Finally, the DNA target solution is aliquoted into DNA low-bind tubes (DNA LoBind tubes, Eppendorf, Hamburg, Germany) and stored in the −20 °C freezer.

The protein targets are produced by ourselves with a SBP tag sequence downstream of the protein sequence that binds to the streptavidin. The genes of protein targets (eGFP and mCherry, corresponding respectively to PDB IDs `2YOG` and `2H5Q`) in fusion with a SBP tag were kindly provided by Sandrine Moutel (Institut Curie, Paris, France). The genes are each one located on a plasmid with an ampicillin resistance cassette and under the control of the T7 promoter. The plasmids are transformed to T7 Express *E. coli* cells (similarly to the transformations in A.2) which have the T7 RNA polymerase inside the *lac* operon. Expression of the fluorescent proteins in T7 Express cultures can thus be induced by adding IPTG; this induces the *lac* operon and thus T7 RNA polymerase which in turn transcribes the fluorescent protein genes: First, a liquid culture of transformed T7 Express cells is started in 5 mL of 2xYT growth medium, 1 % glucose, and ampicillin and grown overnight at 37 °C, 180 rpm. On the next day, a 200 mL liquid culture is started in a 1 L flask with 2xYT growth medium, ampicillin, and 2 mL of cells from the overnight culture (100 x dilution). The culture is grown at 37 °C, 180 rpm until a density of $OD_{600} = 0.48$ is reached. At this point, 200 µL of IPTG at 300 mM are added to the culture (300 µM IPTG final). The culture is incubated overnight at 30 °C, 180 rpm. The IPTG induces the expression

of the fluorescent proteins. On the following day, the cultures visibly changed their color from brownish (the color of the 2xYT) into reddish and greenish respectively for the mCherry and GFP expressing cultures. The proteins need to be harvested, *i.e.* extracted and isolated from the cells. To this aim, the cultures are transferred to 50 mL Falcon tubes and centrifuged at 3220 g for 10 min and the supernatant is poured away. The cell pellets are resuspended in 5 mL of 1 x PBS. The cells are threefold flash-freezed in liquid nitrogen and quick thawed in a water bath at 42 °C in order to burst the cells. Then, the lysate is incubated at 30 °C, 180 rpm for 15 min with 5 µL of DNase I (2.5 $\mathrm{U.mL^{-1}}$ final) and 5 µL of lysozyme (50 $\mu\mathrm{g\,mL^{-1}}$ final). The DNase degrades all DNA in the lysate, the lysozyme degrades the cell wall. The lysate is centrifuged at very high speed (15000 g), 4 °C for 30 min in order to collect all cell debris at the bottom of the tube. The visibly red or green supernatants are poured into new 50 mL Falcon tubes and the pellets are discarded. The fluorescent proteins are then aliquoted into 1.5 mL Eppendorf tubes, flash-freezed in liquid nitrogen, and stored in the −80 °C freezer. We did not purify the proteins before immobilization on magnetic beads. However, everything with no affinity for streptavidin (including cell DNA, cell proteins, and cell membranes) is removed upon washing of the target-bead complexes in the next step.

The binding of target molecules to streptavidin-coated magnetic beads is 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. Compared to the usual Eppendorf tubes, these tubes minimize non-specific interactions between the tube walls and respectively the DNA and proteins. After vortexing the magnetic beads stock (where the magnetic beads are suspended in a buffer and tend to sediment) for 30 s, 50 µL of magnetic beads are pipetted to each low-bind tube which is then kept on ice during the remaining procedure to prevent streptavidin degradation. As a first step, the magnetic beads are washed: The beads are collected on the side of the tube using a magnet and the buffer liquid is removed and replaced by 500 µL of washing liquid. Bw1x is used as washing liquid for beads to be treated with DNA targets, 1 x PBS with 0.1 % of Tween20 (1 volume of Tween20 in 1000 volumes of PBS; careful shaking is required while dissolving the Tween20 in PBS due to the risk of generating foam) is used for beads to be treated with protein targets. This choice of washing solutions further helps screening respectively non-specific electrostatic interactions in the case of DNA targets and non-specific hydrophobic interactions in the case of protein targets. The tubes are vortexed for 5 s, then the washing liquid is removed after collecting the beads again on one side of the tube. Now, the beads are brought into contact with the targets: For DNA targets, 90 µL of washing liquid are added to the beads, as well as 10 µL of targets or 10 µL of MilliQ water for a null selection tube. Given that the DNA targets are at a stock concentration of 400 µM, we can check they are in excess compared to the number of available streptavidin binding sites in the tube which assures that all binding sites have high chance to be filled with target molecules: $6 \cdot 10^{23}\,\mathrm{DNA.mol^{-1}} \cdot 400.10^{-6}\,\mathrm{mol.L^{-1}} \cdot 10^{-5}\,\mathrm{L} = 2.4 \cdot 10^{15}\,\mathrm{DNA}$ target molecules compared to $6 \cdot 10^{23}\,\mathrm{DNA.mol^{-1}} \cdot 775\,\mathrm{pmol\ DNA.(mg\ beads)^{-1}} \cdot 10\,\mathrm{mg\ beads.mL^{-1}} \cdot 50 \cdot 10^{-3}\,\mathrm{mL} = 2 \cdot 10^{14}$ binding sites. Thus, we may expect the beads' surface to be saturated with target molecules. For protein targets, 50 µL of 1 x PBS is added to the beads, as well as 50 µL of protein targets or 50 µL

of MilliQ water for a null selection tube, thus 0.5 x PBS final. The beads are resuspended in the target solution by vortexing and incubated smoothly ($\simeq$ 30 rpm) shaking at room temperature for 15 min to let the target molecules bind to the beads' surface. Then, another washing of the beads is performed in order to remove all unbound targets from the tube: The beads are collected on one side of a tube with a magnet, the target solution is removed and 3 washing steps are performed. One washing step consists of adding 500 µL of washing liquid to the beads, vortexing for 5 s, brief centrifugation in order to collect all liquid at the bottom of the tube including those stuck in the lid after vortexing, and removal of all liquid while holding back the beads with a magnet. Finally, the beads are resuspended in 50 µL of washing liquid and stored in the fridge at 4 °C for use in selections on the following day. In the case of the protein targets, we confirm successful binding of fluorescent protein to the magnetic beads by fluorescence measurements of treated beads at green and red wavelengths *versus* naked beads using a fluorescence plate reader (Spark, Tecan, Männedorf, Switzerland) as described in the main text (see section 3.3.2).

A.6 Selection

The selection is performed on the day immediately following the one of the phage production and target immobilization steps. The workflow is as follows: The magnetic beads covered with target molecules from A.5 are suspended in the solution of antibody displaying phage particles from A.4. After some waiting to let the binding reaction happen, the unbound phage are removed, while the beads with bound and "stuck" phage particles are held back and washed. Finally, the antibody-target complexes are broken and the retrieved phage particles used to infect fresh exponential cells.

In a first place, a pH adjustment of the phage solution is performed by dissolving 236 mg of Na_2HPO_4 and 102 mg NaH_2PO_4 per 50 mL of library phage solution. The final pH should be around 7.0, as measured by a pH-meter. A null selection step is performed using bare magnetic beads that were treated with MilliQ water rather than targets: The magnetic beads are collected on one side of the tube with a magnet, and the washing liquid is removed and trashed using a pipette and replaced by 1 mL of phage solution. The beads are resuspended in the phage solution by vortexing and incubated at rest for 1 h, then smoothly shaking ($\simeq$ 30 rpm) for 30 min. Then, the beads are collected on the side both in the null selection tube and in the positive selection tube containing beads covered with target molecules. The washing liquid is removed from the latter one and the phage solution is transferred from the null selection tube to the positive selection tube. As before, the beads are resuspended in the phage solution by vortexing and incubated smoothly shaking ($\simeq$ 30 rpm) for 90 min (alternatively for 30 min). The shaking prevents sedimentation of the beads at the bottom of the tube. Meanwhile, a washing liquid is prepared by pipetting 100 µL of Tween20 surfactant into 100 mL of 1 x PBS (0.1 % Tween20 final) and shaked smoothly in order to prevent the formation of foam. Due to its high viscosity, the Tween20 is conveniently pipetted by cutting away the cusp of the pipette tip, thus increasing the size of the tip entry.

By the end of the binding step, the complete volume is transferred to a 15 mL Falcon tube, the beads are collected at the side using a magnet and the liquid is poured away. Then, a 10-fold washing step is performed by repeatedly adding 9 mL of 1 x PBS-Tween20 (0.1 %), quickly rotating the tube twice by an angle of π around its axis of symmetry while keeping the magnet fixed, and pouring away the PBS-Tween20; this way, the beads run across the liquid to the opposite site of the tube upon the quick rotation. We expect the washing to have an effect on two accounts: First, phage that are not bound but just "stuck" are removed from the beads and diluted into the liquid. Second, the tubes have a conical shape at the bottom, leading to some "dead" volume left behind upon pouring away the liquid due to surface tension effects. The phage contained in the dead volume are hence gradually diluted away as well during the repeated washing. Finally, elution of bound phage is achieved through resuspending the beads in 1 mL of triethylamine (TEA) dilution (14 µL of 100 % TEA in 1 mL of MilliQ water) and incubating smoothly shaking at ambient temperature for 10 min. The TEA solution has pH $\geq$ 11 which is harmless to the phage particles but breaks protein-protein (here: antibody-target and target-streptavidin) interactions, thus removing the phage particles from the beads. Then, the beads are collected at the side of the tube using a magnet and half of the volume (500 µL) containing eluted phage particles is transferred to another tube containing 500 µL of Trizma at 1 M, pH = 7.4. The beads are then resuspended in the remaining 500 µL of TEA by vortexing and incubated smoothly shaking for another 10 min. Once again, the beads are collected at the side, the remaining volume is transferred to the tube with Trizma as well, and the beads are resuspended in 200 µL of Trizma at 1 M, pH = 7.4. All steps involving the TEA are carried out under the chemical hood because of its high toxicity and penetrant odor.

The solution of eluted phage is now used to infect fresh exponential TG1 cells and thus inject them with the plasmids carrying the genes of the selected antibodies: 30 mL of liquid culture are started with 2xYT growth medium and 300 µL of an overnight liquid culture (100x dilution) of TG1 in a 250 mL flask and incubated at 37 °C, 180 rpm until OD_{600} = 0.4 (takes approx. 2 h). Then, 4 mL of culture are added to the beads, while another 10 mL of culture are infected in a 50 mL Falcon tube with half of the retrieved phage, *i.e.* with 750 µL of the 1.5 mL; the other half is stored in the fridge at 4 °C as backup in case another infection must be performed later. Cells are given onto the beads because some phage may not have been detached from the targets and infection of cells is done directly from bound phage. Infection is then let to occur during incubation at 37 °C, at rest for 30 min. Then, the 4 mL are added to the other 10 mL of culture and centrifuged at 3'200 g, 25 °C for 10 min. The supernatant is discarded and neutralized, and the pellet is resuspended in 1.2 mL of fresh 2xYT growth medium. All 1.2 mL are plated on a large Petri dish coated with selective 2xYT-ampicillin growth medium and 1 % glucose and incubated overnight at 37 °C. Growing the output library on a plate is preferred here to a liquid culture because it is expected to be less prone to additional amplification biases from competition between colonies for ressources (growth medium and space) which would lead to further unwanted shifts in frequencies unrelated to selection for binding.

We estimate the efficiency of the selection, *i.e.* the ratio of the number of phage particles in

the selection output (number of bound phage) and number of phage particles in the input, by an equivalent approach as for the helper phage stock production in A.1: To count the number of phages in the input (per mL), we infect TG1 cells with serial dilutions of the phage solution and count the number of colonies on 2xYT-ampicillin-1 % plates after overnight growth. To count the number of phages in the input (per mL), we perform serial dilutions of the infected cells (for the output) and plate and grow them as well on 2xYT-ampicillin-glucose (1 %) plates. For the input, dilutions of up to 10^{11} are useful, whereas dilutions up to 10^5 are sufficient for the output because the selection typically reduces the number of phage by a factor of 10^6.

On the following day, if the selection was successful, phage concentrations of around $10^{11}\,\mathrm{mL}^{-1}$ are observed in the input and around $10^5\,\mathrm{mL}^{-1}$ or more in the output. $10^5\,\mathrm{mL}^{-1}$ are typically counted in first selection rounds which are dominated by unspecific binders with binding probabilities of $\simeq 10^{-6}$. Moreover, the large Petri dish is then covered by a uniform carpet of library cell colonies that define the selected library. The library cells are brought into a storable, liquid form by pouring 5 mL of 25 % glycerol in 2xYT growth medium (made up of 1 volume of 2xYT growth medium and 1 volume of stock growth medium which is at 50 % glycerol). The cells are scraped into the liquid using a cell scraper. The liquid is collected at the corner of the plate and transferred into a Falcon tube using a pipette. This procedure is repeated once after rotating the plate by $\pi/2$ and scraping in the orthogonal direction. The outcome are cells in liquid form at a density of the order of $\mathrm{OD}_{600} \simeq 100$. After mixing and homogenization by vortexing, the cells are aliquoted into cryotubes and stored at −80 °C. At this point, we are technically back to the initial point of the experiment, *i.e.* glycerol stocks of library cells, modulo the bias in antibody sequence frequencies introduced by selection.

A.7 Illumina sequencing preparation

In order to sequence a glycerol stock of library cells using Illumina technology and to read out the antibody identities, "amplicons" must be produced starting from these glycerol stock of cells carrying various antibody sequences and whose frequencies are the matter of interest. Amplicons are pieces of double-stranded DNA carrying the sequence of interest in the center, as well as technically required barcode and adapter sequences at the extremities, see figure 3.11. These barcodes and adapters must be added by (two) PCR reactions. The PCR primers must be chosen as a function of the region of interest on the plasmid, as well as a few other technical and conceptual criteria: 1) The melting temperature T_m of the primer sequence (when it is understood as double-stranded, *i.e.* in fusion with its complementary-reversed sequence) must be around 55 °C, which leads to primer sequences of a typical length of $\simeq 20$ bp. 2) The primer sequence should optimally start and end with a `G` or a `C` since they imply 3 instead of 2 hydrogen bonds at the extremities of the primer sequence. Here, we are interested in the CDR3 sequence as well as a part of the library-specific scaffold sequence which allows to determine the scaffold identity (Germ, Lmtd, or BnAb). 3) In addition, since we opt for Illumina MiSeq 250 bp paired-end sequencing and

require fully overlapping forward and reverse reading of the sequence for higher confidence, the final amplicon must not be longer than 250 bp. This sets an upper bound to the effective readable sequence length for the region of interest of about 170 bp, the full amplicon length minus the length used up by adapters and primers. 4) The primer sequences must be common to all three libraries in play. Otherwise, either of the three libraries would be simply projected out in the PCR reactions and later in the sequencing results. The task thus consists in finding a common sequence within a scaffold that is globally library-specific. Figure 3.12 shows our choice of the region of interest and primer sequences according to these criteria: The forward primer is located downstream of CDR3, by the end of FWR4, and is directed against the antibody reading frame; the reverse primer is located right upstream of FWR3, by the end of CDR2.

The two primers are purchased only as single-stranded DNA from IDT, Leuven, Belgium. This is because directional symmetry is broken by the DNA polymerase during PCR reactions that copies the DNA only in $5' \to 3'$ direction. Thus, the primers in $3' \to 5'$ direction are not needed. Note that as the forward primer applies to the complemented strand from the one shown in figure 3.12, the reverse-complemented sequence of the one highlighted in this figure needs to be ordered for the forward primer. Thus, the ordered sequences are `ACAACCCGTCTCTTAAGTCTCGT` and `GCTCGAGACGGTAACCAGG`. The shipped primer DNA is resuspended in DNase-free MilliQ water in the same fashion as the DNA targets in A.5 to stock concentrations of 100 µM and/or 10 µM. The required volume of water needed is indicated on the data sheet.

All glycerol stocks of interest are defrosted and diluted to a volume of 5 mL with a bacterial density of $OD_{600} \approx 4.3$ using 2xYT growth medium, thus mimicking the outcome of an overnight culture which is skipped here to avoid introducing additional, unwanted biases in antibody sequence frequencies. The amount of glycerol stock needed is calculated as $4.3/OD_{600,\text{stock}} \cdot 5\,\text{mL}$. Thus, for a glycerol stock of $OD_{600,\text{stock}} = 150$, the diluted volume is 143 µL. Plasmids are extracted using commercial DNA purification kits (Macherey-Nagel, Düren, Germany; see manual for a detailed protocol) and gradually diluted to a DNA concentrations of $1\,\text{ng}\,\mu\text{L}^{-1}$. A calculation shows that 1 µL at $1\,\text{ng}\,\mu\text{L}^{-1}$, thus 1 ng of plasmids still amounts to 10^8 copies of the plasmids, thus conveniently higher than the upper limit from the sequencing depth of Illumina sequencing. The first PCR reaction (PCR1) is set up in 25 µL of final volume as follows: 1.0 µL of plasmid, 15.75 µL of DNase-free MilliQ water, 1.25 µL of forward primer at 10 µM, 1.25 µL of reverse primer at 10 µM, 5.0 µL of 5 x Q5 reaction buffer (1 x final), 0.5 µL of dNTPs at 10 mM, and 0.25 µL of Q5 HotStart HF (high-fidelity) DNA polymerase. Alternatively, it is set up with 1.0 µL of plasmid, 9.0 µL of DNase-free MilliQ water, 1.25 µL of forward primer at 10 µM, 1.25 µL of reverse primer at 10 µM, and 12.5 µL of 2 x Q5 HotStart Master Mix (1 x final). The Master Mix already contains the polymerase, its buffer, and dNTPs all mixed together. The parameters of the PCR reaction are chosen as follows: initial denaturation at 98 °C for 3 min, then 18-fold cycling of (denaturation at 98 °C for 10 s, annealing at 68 °C for 20 s, extension at 72 °C for 20 s), then final extension at 72 °C for 5 min and hold at 4 °C. The low number of PCR cycles (usually 30) is chosen to minimize potential biases from differential amplification of different sequences during PCR. The PCR product is run on an agarose gel with the same protocol as in A.2 to check for a band at the expected

size and the presence of unspecific other bands, see figure 3.13(a). The band at the expected size is excised from the gel with a sterile blade and purified using a commercial gel extraction and purification kit (Macherey-Nagel, Düren, Germany; see manual for a detailed protocol).

Then, a second PCR reaction (PCR2) is performed using the gel-purified products of PCR1 and primers containing sample-specific barcode called P5 and P7 indices, as well as adapter sequences down- and upstream of these indices linking to the product of PCR1 and the Illumina machinery, respectively (see main text in subsection 3.5.2 for details). The purified PCR1 product is diluted 10 x in DNase-free MilliQ water and the second PCR reaction is set up equivalently to the first one, except for using different primers and the purified PCR1 product instead of the plasmid. The parameters of the reaction are the same as well, except for the annealing temperature which is set to 65 °C for these primers. Equivalently to the first PCR, the products of the second PCR are run on an agarose gel, checked for correct size (see figure 3.13(b)), and purified, ending up with the amplicons required for Illumina sequencing. The amplicon concentrations are measured by Qubit, thus with higher precision than by Nanodrop. Knowing the concentrations of all samples, lengths of the amplicons, and given the intended number of sequencing reads for each of them, the samples are all mixed together in proportions that are conveniently calculated in an excel table, see figure E.2. The final product is again run on a TapeStation (Agilent Technologies, Santa Clara, CA, USA), a high-precision gel, to check for the correct size(s) of the amplicon mix (see figure 3.13(c), (d)), before being handed over to the sequencing platform at the Institut de biologie intégrative (I2BC) at Gif-sur-Yvette, France, where the amplicons are again checked for quality, sequenced, and demultiplexed. Demultiplexing means classifying reads according to their P5 and P7 indices defining together which sample a read belongs to.

All new selection experiments performed as part of this PhD were prepared as described here and distributed over 3 complete Illumina MiSeq 2×250 bp paired-end sequencing runs all being taken care of at Gif-sur-Yvette and yielding an overall dataset of about 40 million reads in size. Data that was obtained by Sébastien Boyer during his PhD used a different primer design (same forward primer, but different reverse primer such that the whole antibody could be sequenced, but without overlap of forward and reverse reads) and a commercial sequencing platform at Eurofins.

Chapter B

Antibody affinity maturation

Here, we convey the current picture of the process of antibody affinity maturation based on literature. In short, a primary repertoire of naïve antibody genes is assembled by recombination of the gene fragments with additional random sequence diversity at the fragment junctions (section B.1). Then, successful naïve variants with an initial affinity for epitopes located on the pathogen serve as starting point for the evolutionary process by which they are further improved upon somatic hyper-mutation and selection for beneficial, notably affinity-enhancing, mutations (section B.2). This way, naïve antibodies gradually fixate more and more beneficial somatic mutations over time and we define the "maturation degree" of an antibody as the time since initiation of the affinity maturation or, somehow equivalently, as the number of accumulated somatic mutations. It has to be noted that while many aspects of affinity maturation are known, some details are still debated, in particular the randomness of and mechanisms behind the somatic hyper-mutation. The typical timescale of the overall process is in the weeks and months [19] by which the variable domain of the antibody (V_H and V_L together) fixate up to 20 somatic mutations [189]. Some antibodies even evolve over years of chronic infection, leading to so-called broadly-neutralizing antibodies (bnAbs), which fixate between 40 and 100 somatic mutations [20] (section B.3).

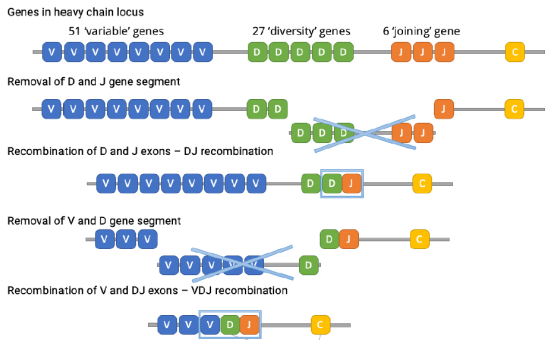
B.1 Primary repertoire formation upon VDJ recombination

Genetically, the variable V_H and V_L chains involved in antibody binding are not encoded into a single continuous gene in the germline. Rather, the V_H region is sectioned into, and assembled from 3 gene fragments, namely the variable (V), diversity (D), and joining (J) gene fragments. These fragments cover respectively FWR1 through FWR3 (V), the center part of CDR3 (D), and FWR4 (J). The genome does not contain a single, but several templates for each of these 3 gene

fragments, in human, 51 V [190], 27 D, and 6 J fragment templates that are sequentially grouped along the genome, see figure B.1(a). A full V_H gene is obtained from recombination of this pool of gene fragments (VDJ recombination), which is catalyzed by DNA recombinases and occurs upon initial repertoire formation: One template is randomly (but not uniformly) chosen for each of the segments in a way that is depicted in figure B.1(a); segments of DNA are removed in between the chosen V and D (D and J) gene fragments, including the rejected V and D (D and J) fragments. The blunt ends are ligated and additional random nucleotides may be inserted at the junctions, called the N regions (note however the strong constraint of continuance in reading frame). As a result, the three chosen fragments are recombined together to form the full V_H coding sequence (VDJ), including random sequences at the level of CDR3. Finally, there are two sources of diversity in the initial antibody repertoire, each unique variant of which can possibly serve as starting point for affinity maturation trajectories, see next subsection 1.1.3: (i) combinatorial diversity in the FWRs, as well as CDR1 and 2 with in total $51 \cdot 27 \cdot 6 = 8282$ possibilities for human V_H scaffolding, (ii) enormous diversity in the CDR3 stemming from the randomness of length and sequence of the insertions. Human V_L are recombined in a similar way as V_H, but using a reduced pool of only 40 V and 5 J gene fragments (no D fragment here; $40 \cdot 5 = 200$ possibilities for V_L scaffolding in the initial repertoire) to give rise to recombined VJ genes. Overall, the number of possible naïve antibodies is estimated to 10^{57}, 10^{45} alone for naïve V_H chains, while only 10^{14} of them can be sampled within a lifetime [191]. As a consequence, CDR3 stands out compared to CDR1 and 2 in terms of initial diversity and likelihood of epitope complementarity. The recombination is performed independently by each B cell; thus one repertoire sequence corresponds to one B cell.

B.2 Mechanistic details of affinity maturation

Upon encounter of a pathogen, the goal of the adaptive immune response is to produce antibodies with high binding affinity and specificity to epitopes located on the pathogenic particles. Specificity here, means in particular that the antibody should generally minimize binding affinities to self epitopes, while maximizing affinity to foreign epitopes; if this constraint is a strong one (*i.e.* when self and foreign epitopes are structurally similar), there may be trade-offs between affinity and specificity. Generally, the combinatorial and junctional diversity of the primary repertoire is sufficiently sampled by the naïve B cell population to have among them one or a few that display receptors with binding affinities that are high compared to most other B cells, but still weak in absolute terms. These are selected to serve as starting points of the affinity maturation process [192, 18], upon which their affinity is further improved. The fact that potential pathogens exceed realized naïve B cell receptor sequences in numbers, but (relatively) high-affinity naïve B cell receptors are still being identified, implies that one B cell receptor sequence accounts for many potential pathogens. This property is known as antibody multispecificity [193] and is mediated by conformational isomerism [194]. The affinity maturation is an evolutionary process that follows the rules of Darwinian evolution; a schematic of affinity maturation with biological details is shown in figure B.1(b): The naïve B cells iteratively undergo periods of (somatic hyper-)mutation of their



(a)

not included for copyright reasons

9_affmat.jpg

(b)

Fig. B.1: Primary repertoire formation through VDJ recombination and affinity maturation. **(a)** Genetic structure of the heavy chain of the antibody variable domain (V_H) and primary repertoire formation by VDJ recombination. Taken and adapted from [148]. **(b)** Details of the antibody affinity maturation process. Taken from [19].

receptor sequences and selection for (increased) binding affinity to the pathogen. Upon fixation of beneficial mutations, the naïve B cells (or the antibodies they encode for) "maturate", *i.e.* they are turned into mature B cells with high-affinity receptor sequences.

Naïve (coming from the bone marrow; after tolerance check for non-autoreactivity) and previously matured memory B cells (if any), displaying receptors with highest (in relative terms) binding affinity to relevant epitopes, obtain T cell help, which is limiting and thus, competitive. Antigen-stimulated (or -activated) naïve and memory B cells, as well as antigen-specific T helper cells migrate towards lymphoid tissue to form germinal centers [195] together with local antigen-displaying follicular dendritic cells (FDC). FDC non-specifically collect and display at their surface foreign antigens present in the body, *i.e.* they sample from all antigen types the body is currently coping with, generally provided by non-specific B cells. These 3 essential players take part in the germinal center reaction, which consists of clonal expansion of B cells accompanied by somatic hyper-mutation (SHM) and pre- and post-clonal selection for binding capacity and T cell help. Mutation and selection take place respectively in the dark and light zones of the germinal center. In the dark zone, B cells start to proliferate and the enzyme activation-induced cytidine deaminase (AID or AICDA) is turned on after some temporal delay, which results in somatic hyper-mutation in the immunoglobulin (Ig) region [196, 197, 198, 199]: Several mechanisms of somatic hyper-mutation are mediated by AID, which catalyzes the conversion of cytosines `C` in the DNA motif `WRC(Y)` (`W`=`A`/`T`, `R`=`A`/`G`, `Y`=`C`/`T`) into uracil `U` by deamination, while keeping the complemented guanine `G` in place. This creates a `U:G` mismatch in the double-stranded genomic DNA, which is resolved notably in 3 different ways: (i) The DNA is replicated as such by DNA polymerases with the mismatched `U:G` being replaced through two replication steps by `U:A` (and `C:G`; DNA polymerases consider `U` as equivalent to `T`), then `T:A` (and `U:A`). (ii) The `U:G` mismatch is excised by the enzyme uracil-DNA glycosylase (UNG) and randomly replaced upon repair by error-prone DNA polymerase η, also in two steps: from `-:G` to `-:N` (and `C:G`) to `N:N` (and `-:N`). (iii) The

mismatch is recognized by the DNA mismatch repair (MMR) system, which also involves repair by DNA polymerase η. As DNA polymerase η reinstates the DNA sequence in a neighborhood of the mismatch, it can insert wrong nucleotides not only by false negatives at the mismatched site, but also by false positives at close-by residues and lead to one or several point mutations along the sequence. Even if mechanisms (ii) and (iii) sample point mutations uniformly, there will be mutational hotspots along the sequence and a bias in favor of the transition `C:G` $\mapsto$ `T:A` due to mechanism (i). Overall, the somatic hyper-mutation operates at a mutation rate of $\simeq 10^{-3} - 10^{-2}$ mutations per bp and cell division, *i.e.* $\simeq 1$ mutation per gene per cell division [200], which is $10^3 - 10^4$ x higher than elsewhere in the genome under the usual DNA polymerases. This applies to somatic mutations upon introduction, *i.e.* before fixation of the mutations by selection; AID hotspots typically do not correlate with the location of fixed mutations [201]. Insertions and deletions are also possible. B cells may become nonfunctional upon frame shift, introduction of stop codons, or deleterious mutations corrupting *e.g.* B cell receptor stability. After hyper-mutation, B cells displaying folded receptors enter into the light zone, where they undergo positive selection for binding capacity and are required to collect two survival signals; failure to collect any of these signals results in apoptosis (B cell suicide). First, B cells obtain a survival signal upon binding to and internalizing antigen, displayed on the surface of antigen-displaying FDC. Success or failure to do so is binding affinity-dependent: Detachment and internalization of antigen requires applying a certain minimal force that the B cell can exert with little probability of breaking the receptor-antigen complex only if its binding affinity is sufficiently strong (in absolute terms, which implies that this survival signal is non-limiting, *i.e.* does not depend on the performance of other B cells). However, if availability of displayed antigen is limiting, B cells are in competition with each other for antigen, which may put selective pressure not on K directly, but on k_+ and k_- separately. Low-affinity B cells failing to internalize antigen are destined for apoptosis. Second, remaining B cells seek for T cell help, which provides the second survival signal: B cells display peptides from internalized and digested antigen in fusion with major histocompatibility complex (p and MHC, to give pMHC) at their surface, which is recognized and bound to by antigen-specific (but not peptide- or epitope-specific!) T helper cells. T cell help is a limiting ressource, with T cell help being most probably provided to those B cells displaying highest amounts of pMHC among all B cells, *i.e.* the ones most strongly binding to antigen in the first selection step (binding affinity in relative terms). Absence of T cell help again induces apoptosis. The majority of selected mature B cells returns to the dark zone and reiterates the mutation-selection procedure; the rest differentiates in equal proportions into (i) antibody-expressing and -secreting plasma cells, which are released into the blood and (ii) memory B cells for long-term storage of the genetic information of maturation outcome and reactivation into a germinal center upon later re-encounter of the same or a similar pathogen. Differentiation of mature B cells is accompanied by a change in antibody isotype from IgM to IgG, also mediated by AID.

To put affinity maturation to the numbers, the antibody binding affinity towards the cognate epitope is typically increased 10^1 x to 10^3 x from naïve to final mature. In absolute terms, the affinity given by the equilibrium constant $K = k_-/k_+$ of the antibody-antigen binding reaction is lowered from the µM to the nM range [202], with the limiting on- and off-rates being

$k_+ \simeq 10^4\,\mathrm{M.s}^{-1}$ and $k_- \simeq 10^{-4}\,\mathrm{s}^{-1}$ [100]. So achieve this, $\simeq 10-20$ (fixed) somatic mutations within the V_H or V_L chains [189], as well as a period of weeks to months of immune response are typically required to go from naïve to final mature. However, deeply matured broadly-neutralizing antibodies that may arise in the immune response against highly mutable pathogens after years of chronic infection (see next subsection 1.2.2), accumulate up to $\simeq 40-100$ (fixed) somatic mutations [20]. Altogether, combinatorial recombination and random junction during primary repertoire formation, as well as somatic hyper-mutation are able to yield such high-affinity antibodies to virtually any foreign target [203].

Antibody affinity maturation during the adaptive immune response, together with possible simultaneous, evasive co-evolution of the pathogen, stands out from all other known evolutionary processes that occur over unobservable (beyond human lifetime) time scales and with the precise meaning of "variation" and "fitness" being oftentimes unclear: In affinity maturation, the evolutionary process does not occur over many generations of the protected organism, but within a given organism's lifetime; the relevant notion of "generation" here is with respect to B cell populations. To understand this, it should be noted that adaptive immune protection is not inherited, that is, the genetic code of the product of affinity maturation (mature B cells), is not passed on from the parent organism to its progeny; only the pool of (naïve) V, D, and J gene fragments is. (ii) The variation step is realized by somatic point mutations in a sequence space and is thus well-defined, as is the selective pressure, which is *grosso modo* expected to act on binding (modulo the necessary conditions of protein stability, solubility, reading frame conservation, and possible affinity-specificity trade-offs as mentioned above). Note, that selection for binding does not necessarily imply selection on binding affinity $K = k_-/k_+$, but may also involve selection on both k_+ and k_-. The relevance of kinetic in addition to thermodynamic selection has been demonstrated [204].

B.3 Broadly-neutralizing antibodies

The adaptive immune responses against highly mutable pathogens, such as HIV [139], influenza [137], and hepatitis C, face particular challenges: These pathogens are able to evolve on similar time scales as B cell receptors and, thus, significantly diverge and evade from B cell specificities, leading to B cell-pathogen co-evolution [25]. Typical features of such pathogens are easily accessible, but highly structurally variable (conformationally within a pathogen variant and in mean between variants), low-density epitopes burying more hardly accessible, but conserved (within and between pathogen variants) epitopes that are required for pathogen stability and/or function [143, 6, 205]. Access of B cell receptor binding sites to these vulnerable epitopes is the key towards binding to various pathogen variants, but requires significant efforts to *e.g.* circumvent steric exclusion from variable epitopes. The solution to this task that is *sometimes* found by the adaptive immune response, are so-called broadly-neutralizing antibodies (bnAb) [142, 206, 147, 145, 146]. The name stems from their ability to bind to and neutralize a spectrum of different variants of a

given pathogen. Typical properties of bnAbs that distinguish them from usual mature antibodies against fixed pathogens, are high numbers of fixed somatic mutations both in FWR and CDR regions (typically $40-100$ across the entire variable region [20]), as well as significantly elongated CDR3 loops in V_H [140]. Despite the large number of fixed somatic mutations, bnAbs targeting the same epitope are found to be typically similar in sequence and structure [207]. Long CDRs are presumably required to bypass variable epitopes and access hidden conserved epitopes upon providing flexibility and/or contact [123]. However, mutations at contact positions are not sufficient: It was found that somatic mutations far from the actual binding site [123], which are most oftentimes in FWR regions, as well as both V_H and V_L together [208], are required for neutralization breadth of bnAbs. Interestingly, bnAb keep neutralization breadth upon reversal of most somatic mutations [144, 209], albeit somatic mutations being generally required [123]. Mature antibodies are usually specific to a cognate target epitope, but bnAbs display autoreactivity [210, 211], polyreactivity [210], and heteroligation [210] (meaning binding affinities for self epitopes, foreign epitopes unrelated to cognate, and different epitopes on cognate, respectively) to higher extents than in non-neutralizing antibodies (nnAb) [212]. These characteristics provide increased neutralization breadth [210], with a possible rationale being (selection for) structural mimicry of self antigens by the pathogen [212]. BnAbs have been a strong candidate for vaccine design in particular against HIV [141, 143, 5, 6, 139], but with no definitive success, as elicitation of bnAb turns out intricate [208]. As a matter of fact, bnAbs naturally appear in few HIV patients after years of chronic infection, a rationale being the subtle weighing between too similar and too dissimilar, contradictory (*i.e.* frustrated) selection pressures that direct B cell maturation towards epitopes that are specific to few ubiquitous pathogen variants, and extinction of the B cell lineage [5, 6], respectively. To understand mechanisms and possible controls of affinity maturation, mathematical models of affinity maturation have been considered in numerous studies [25, 79, 101, 82], interrogating *e.g.* fixation probabilities of bnAbs *versus* nnAbs in models of affinity maturation against variable pathogens [25].

Chapter C

Computations

C.1 Binding kinetics

We seek to solve the kinetics of the binding reaction $A + T \rightleftharpoons AT$ between a ligand A and its target T. The reaction equations are given by

$$\frac{\mathrm{d}}{\mathrm{d}t}[AT] = k_+[A][T] - k_-[AT], \tag{C.1}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}[A] = \frac{\mathrm{d}}{\mathrm{d}t}[T] = k_-[AT] - k_+[A][T], \tag{C.2}$$

where $[A]$, $[T]$, and $[AT]$ denote respectively the concentrations of the ligand, the target, and their complex formed upon binding. k_+ (k_-) is the association (dissociation) rate of the binding reaction. Equations (C.2) are consequences of equation (C.1) and the overall conservation of ligands and targets,

$$[A](t) + [AT](t) = [A]_{\mathrm{tot}}, \tag{C.3}$$

$$[A](t) + [AT](t) = [T]_{\mathrm{tot}}, \tag{C.4}$$

which holds for every $t \geq 0$. $[A]_{\mathrm{tot}}$ and $[T]_{\mathrm{tot}}$ are the initial total concentration of ligands and targets injected into the system. Two concentrations, say the concentrations of reactants $[A]$ and $[T]$, can thus be eliminated by inserting equations (C.3) and (C.4) into equation (C.1). This reduces the problem to solving a non-linear equation for $[AT]$,

$$\frac{\mathrm{d}}{\mathrm{d}t}[AT] = k_+\left([AT]^2 - \left([A]_{\mathrm{tot}} + [T]_{\mathrm{tot}}\right)[AT] + [A]_{\mathrm{tot}}[T]_{\mathrm{tot}}\right) - k_-[AT], \tag{C.5}$$

which is separable, i.e.

$$\int_{[AT]_0}^{[AT]} \frac{\mathrm{d}\xi}{\xi^2 - (K_{AT} + [A]_{\mathrm{tot}} + [T]_{\mathrm{tot}})\,\xi + [A]_{\mathrm{tot}}[T]_{\mathrm{tot}}} = k_+ \int_0^t \mathrm{d}\zeta = k_+ t, \tag{C.6}$$

where $K_{AT} = k_-/k_+$ denotes the dissociation constant. We can make use of the standard integral [213]

$$\int \frac{\mathrm{d}\xi}{a\xi^2 + b\xi + c} = \frac{1}{\sqrt{-\Delta}} \ln\left(\frac{2a\xi + b - \sqrt{-\Delta}}{2a\xi + b + \sqrt{-\Delta}}\right) = \frac{-2}{\sqrt{-\Delta}} \mathrm{artanh}\left(\frac{2ax + b}{\sqrt{-\Delta}}\right), \tag{C.7}$$

as the discriminant $\Delta = 4ac - b^2$ is strictly negative or zero in our problem,

$$\begin{aligned}
\Delta &= 4[A]_{\mathrm{tot}}[T]_{\mathrm{tot}} - (K_{AT} + [A]_{\mathrm{tot}} + [T]_{\mathrm{tot}})^2 \\
&= -K_{AT}^2 - ([A]_{\mathrm{tot}} - [T]_{\mathrm{tot}})^2 - 2K_{AT}\,([A]_{\mathrm{tot}} + [T]_{\mathrm{tot}}) \\
&= -\,(K_{AT} + [A]_{\mathrm{tot}} - [T]_{\mathrm{tot}})^2 - 4K_{AT}[T]_{\mathrm{tot}} \\
&\leq -4K_{AT}[T]_{\mathrm{tot}} \\
&\leq 0.
\end{aligned} \tag{C.8}$$

Note that Δ is symmetric in $[A]_{\mathrm{tot}}$ and $[T]_{\mathrm{tot}}$, reflecting a symmetry related to the arbitrariness of labeling the reagents "ligand" and "target". By solving for $[AT]$ using the functional identity $f(\xi + \zeta) = (f(\xi) + f(\zeta))\,/\,(1 + f(\xi)f(\zeta))$ of the hyperbolic tangent $f = \tanh$, this yields with the initial condition $[AT](t = 0) = 0$

$$\begin{aligned}
[AT](t) &= \frac{1}{2}\left(\gamma_1 - \sqrt{-\Delta}\,\frac{\tanh\left(\frac{k_+ t\sqrt{-\Delta}}{2}\right) + \frac{\gamma_1}{\sqrt{-\Delta}}}{1 + \frac{\gamma_1}{\sqrt{-\Delta}}\tanh\left(\frac{k_+ t\sqrt{-\Delta}}{2}\right)}\right), \\
&= \frac{2\gamma_0 \tanh\left(\frac{k_+ t\sqrt{-\Delta}}{2}\right)}{\sqrt{-\Delta} + \gamma_1 \tanh\left(\frac{k_+ t\sqrt{-\Delta}}{2}\right)},
\end{aligned} \tag{C.9}$$

where $\Delta = 4\gamma_0 - \gamma_1^2$, $\gamma_0 = K_{AT} + [A]_{\mathrm{tot}}[T]_{\mathrm{tot}}$ and $\gamma_1 = K_{AT} + [A]_{\mathrm{tot}} + [T]_{\mathrm{tot}}$. For generic initial condition $[AT](t = 0) = [AT]_0$, the solution reads

$$\begin{aligned}
[AT](t) &= \frac{1}{2}\left(\gamma_1 - \sqrt{-\Delta}\,\frac{\tanh\left(\frac{k_+ t\sqrt{-\Delta}}{2}\right) + \frac{\gamma_1 - 2[AT]_0}{\sqrt{-\Delta}}}{1 + \frac{\gamma_1 - 2[AT]_0}{\sqrt{-\Delta}}\tanh\left(\frac{k_+ t\sqrt{-\Delta}}{2}\right)}\right) \\
&= \frac{\sqrt{-\Delta}[AT]_0 + (2\gamma_0 - \gamma_1[AT]_0)\tanh\left(\frac{k_+ t\sqrt{-\Delta}}{2}\right)}{\sqrt{-\Delta} + (\gamma_1 - 2[AT]_0)\tanh\left(\frac{k_+ t\sqrt{-\Delta}}{2}\right)}.
\end{aligned} \tag{C.10}$$

At infinite time $t \to +\infty$, the concentration of the complex $[AT]$ converges to its equilibrium value

$$[AT]_\infty = \frac{2\gamma_0}{\sqrt{-\Delta} + \gamma_1} = \frac{\gamma_1 - \sqrt{-\Delta}}{2}$$

$$= \frac{1}{2}\left(K_{AT} + [A]_{\text{tot}} + [T]_{\text{tot}} - \sqrt{\left(K_{AT} + [A]_{\text{tot}} + [T]_{\text{tot}}\right)^2 - 4[A]_{\text{tot}}[T]_{\text{tot}}}\right)$$
$$= \frac{1}{2}\left(K_{AT} + [A]_{\text{tot}} + [T]_{\text{tot}} - \sqrt{\left(K_{AT} - [A]_{\text{tot}} + [T]_{\text{tot}}\right)^2 + 4K_{AT}[A]_{\text{tot}}}\right), \tag{C.11}$$

which is indeed a solution to $\frac{\mathrm{d}[AT]}{\mathrm{d}t} = 0$ in equation (C.5). The second equilibrium point at $(\gamma_1 + \sqrt{-\Delta})/2$ is not achievable because it is located above the maximum possible concentration of complex which is given by $\min([A]_{\text{tot}}, [T]_{\text{tot}})$ as can be seen by the following chain of inequalities,

$$\frac{\gamma_1 + \sqrt{-\Delta}}{2} \geq \frac{\gamma_1}{2} \geq \frac{[A]_{\text{tot}} + [T]_{\text{tot}}}{2} > \min([A]_{\text{tot}}, [T]_{\text{tot}})\,. \tag{C.12}$$

As expected, $[AT]_\infty$ depends on the rates $k_\pm$ only via their ratio which defines the equilibrium constant $K_{AT} = k_-/k_+$.

An expansion of the right-hand site of equation (C.5) around the equilibrium point $[AT]_\infty$ by setting $[AT] = [AT]_\infty + c$ and Taylor expanding to linear order in c yields

$$\dot{c} = (2C_\infty - \gamma_1)\, k_+ c = -\sqrt{-\Delta} k_+ c. \tag{C.13}$$

This equation is solved by $c \propto \exp(-t/\tau)$, where $\tau = \left(k_+\sqrt{-\Delta}\right)^{-1}$ defines the equilibration time scale of the binding reaction. Interestingly, τ diverges in the case of equal initial ligand and target concentrations and no dissociation, *i.e.* $[A]_{\text{tot}} = [T]_{\text{tot}} = C$ and $k_- = 0$. Here, the binding reaction continues until all ligands and all targets are engaged,

$$[AT]_\infty = \frac{1}{2}\left([A]_{\text{tot}} + [T]_{\text{tot}} - |[A]_{\text{tot}} - [T]_{\text{tot}}|\right) = \min([A]_{\text{tot}}, [T]_{\text{tot}}) = C, \tag{C.14}$$

where the first equality holds for $k_- = 0$ and the third equality if, in addition, $[A]_{\text{tot}} = [T]_{\text{tot}} = C$.

In another special case in which the quantity of target exceeds the quantity of ligands, *i.e.* $[A]_{\text{tot}} \ll [T]_{\text{tot}}$, we can expand equation (C.11) for small $\epsilon = [A]_{\text{tot}}/[T]_{\text{tot}} \ll 1$ to find

$$[AT]_\infty \simeq \frac{[T]_{\text{tot}}}{[T]_{\text{tot}} + K_{AT}}[A]_{\text{tot}} - \frac{K_{AT}[T]_{\text{tot}}}{\left(K_{AT} + [T]_{\text{tot}}\right)^3}[A]_{\text{tot}}^2 + \frac{K_{AT}\left(K_{AT} - [T]_{\text{tot}}\right)[T]_{\text{tot}}}{\left(K_{AT} + [T]_{\text{tot}}\right)^5}[A]_{\text{tot}}^3 + \mathcal{O}\left(\left(\frac{[A]_{\text{tot}}}{[T]_{\text{tot}}}\right)^4\right). \tag{C.15}$$

To first order, we thus obtain $[AT]_\infty \simeq \frac{1}{1+K_{AT}/[T]_{\text{tot}}}[A]_{\text{tot}}$. The quantity $\frac{[AT]_\infty}{[A]_{\text{tot}}} = \frac{1}{1+K_{AT}/[T]_{\text{tot}}}$ represents to first order the fraction among all ligands that is engaged in binding at equilibrium, or, equivalently, the equilibrium probability for a single copy of the ligand to be in bound state, and leads to the Fermi-Dirac distribution discussed in section 2.1.

The solution in equation (C.9) may also be reparametrized to a simpler form using the equilibrium complex concentration $[AT]_\infty$, the time scale τ, and the additional quantity $\alpha = \frac{\gamma_1}{\sqrt{-\Delta}}$,

instead of k_+, γ_1 and Δ or k_+, k_-, $[A]_{\text{tot}}$ and $[T]_{\text{tot}}$,

$$\frac{[AT](t)}{[AT]_\infty} = \frac{(\alpha+1)\tanh\left(\frac{t}{2\tau}\right)}{\alpha+\tanh\left(\frac{t}{2\tau}\right)}. \tag{C.16}$$

Chapter D

Supplementary tables

D.1 List of acronyms

acronym	explanation
Germ, Germline	VH library based on a naïve antibody scaffold
Lmtd, Limited	VH library based on an antibody matured against HIV
BnAb	VH library based on a profoundly matured, broadly neutralizing antibody against HIV
Mix3	uniform (at selection round $t = 0$) mix of the Germ, Lmtd, and BnAb libraries
DNA1	DNA hairpin with loop sequence `CCCATAGCG`
DNA2	DNA hairpin with loop sequence `TTGGTAATA`
prot1	eGFP green fluorescent protein (PDB accession `2Y0G`) in fusion with an SBP tag
prot2	mCherry red fluorescent protein (PDB accession `2H5G`) in fusion with an SBP tag
top DNA1	mini library of 10 clones from across Germ, Lmtd, BnAb selected against DNA1
top DNA2	mini library of 9 clones from across Germ, Lmtd, BnAb selected against DNA2
random	mini library of 10 randomly picked clones from across Germ, Lmtd, BnAb
PCR	polymerase chain reaction
V_H (V_L)	heavy (light) chain of the antibody variable region
FWR	framework regions in V_H and V_L chains, numbered from 1 to 4
CDR	complementary determining regions in V_H and V_L chains, numbered from 1 to 3
nt, bp, aa	nucleotide, base pair, amino acid (as sequence length units)
CDF	cumulative distribution function
PDF	probability distribution function
EVT	extreme-value theory
ML(E)	maximum-likelihood (estimation)
iid	independently and identically distributed
PP plot	probability-probability plot
QQ plot	quantile-quantile plot
PWM	position weight matrix

Tab. D.1: List of acronyms used throughout the manuscript.

D.2 List of variables

symbol	explanation
A, T, AT	ligand (antibody), binding target (epitope), and the complex of both formed upon binding
$[\cdot]$	concentration of species $\cdot$
β	(inverse) temperature, $\beta = (k_B T)^{-1}$
ΔG	free energy of binding
K, K_D	equilibrium constant of a binding reaction, $K = \exp(\beta \Delta G)$
L	length of a sequence given by the number of amino acid or nucleotide residues
q	number of Potts spin states per position, $q = 20$ for the alphabet of amino acids, $q = 4$ for the alphabet of nucleotides
x	sequence identity, given by the sequence of L letters, $x = \{x_1, x_2, \ldots, x_L\}$, where x_i takes on the alphabet of q letters for all $i \in \{1, \ldots, L\}$
ℓ	library/scaffold identity which takes on $\{\text{Germ}, \text{Lmtd}, \text{Bnab}\}$
s	enrichment/binding probability
$P(s)$	probability distribution function of enrichments
t	discrete time, $t = 0, 1, 2, \ldots$, where one unit of time is defined as one cycle of selection
$N_t(x)$	number of copies of sequence x in the population at selection round t
$n_t(x)$	number of occurrences of sequence x in a sequenced sample at selection round t
$f_t(x)$	frequency of sequence x in the population at selection round t: $f_t(x) = \frac{N_t(x)}{\sum_y N_t(y)} \simeq \frac{n_t(x)}{\sum_y n_t(y)}$
$f_{t,i}(a)$	frequency of letter a (among the alphabet of q letters) on residue i at time t (position weight matrix)
λ_t	amplification factor of a population after selection round t to recover its initial size
σ^2	parameter of the probability density function of a lognormal variable S; corresponds to the variance $\langle \ln(S)^2 \rangle - \langle \ln(S) \rangle^2$ of the Gaussian variable $\ln(S)$
μ	chemical potential OR parameter of the probability density function of a lognormal variable Z; corresponds to the mean $\langle \ln(S) \rangle$ of the Gaussian variable $\ln(S)$
κ, τ	shape and scale parameter, respectively, of a generalized Pareto distribution
s^*, y^*	threshold (log-)enrichment for the inference of generalized Pareto and lognormal models, respectively, from truncated data
p	order of a p-spin glass model (p-body interaction)
$h_i(a)$, $J_{ij}(a, b)$	local field functions and pairwise couplings in one- and two-spin glass models ($p = 1, 2$), respectively

Tab. D.2: Recap of the most prevalent variables and their definitions used throughout the manuscript.

D.3 List of P5 and P7 indices

library	target	round	P5 index (5′ → 3′)	P7 index (5′ → 3′)	run
Mix3		0	cctatcct	acgaattc	1
Mix3	DNA1	1	atagaggc	ttctgaat	3
Mix3	DNA1	2	aggcgaag	acgaattc	1
Mix3	DNA1	3	caggacgt	acgaattc	1
Mix3	DNA1	4	tatagcct	acgaattc	3
Mix3	DNA2	1	tatagcct	ttctgaat	3
Mix3	DNA2	2	ggctctga	acgaattc	1
Mix3	DNA2	3	taatctta	acgaattc	1
Mix3		0	cctatcct	ttctgaat	2
Mix3	prot1 replica 1	1	aggcgaag	ttctgaat	3
Mix3	prot1 replica 1	2	tatagcct	acgaattc	2
Mix3	prot1 replica 1	3	atagaggc	acgaattc	2
Mix3	prot1 replica 2	1	taatctta	ttctgaat	3
Mix3	prot1 replica 2	2	cctatcct	acgaattc	2
Mix3	prot1 replica 2	3	ggctctga	acgaattc	2
Mix3	prot1 replica 2	4	cctatcct	acgaattc	3
Mix3	prot2 replica 1	1	cctatcct	ttctgaat	3
Mix3	prot2 replica 1	2	ggctctga	ttctgaat	2
Mix3	prot2 replica 1	3	aggcgaag	ttctgaat	2
Mix3	prot2 replica 1	4	atagaggc	acgaattc	3
Mix3	prot2 replica 2	1	ggctctga	ttctgaat	3
Mix3	prot2 replica 2	2	taatctta	ttctgaat	2
Mix3	prot2 replica 2	3	caggacgt	ttctgaat	2
Mix3		2	ggctctga	acgaattc	3
Mix3	ampl replica 1	3	aggcgaag	acgaattc	3
Mix2	ampl replica 2	3	taatctta	acgaattc	3
topDNA1+rand		0	ggctctga	agcttcag	1
topDNA1+rand	DNA1 replica 1	1	taatctta	agcttcag	1
topDNA2+rand		0	cctatcct	agcttcag	1
topDNA2+rand	DNA2	1	aggcgaag	agcttcag	1
topDNA1+rand		0	ggctctga	agcttcag	3
topDNA1+rand	DNA1 replica 2	1	aggcgaag	agcttcag	3
topDNA1+rand	beads	1	taatctta	agcttcag	3
topDNA1+topDNA2		0	aggcgaag	acgaattc	2
topDNA1+topDNA2	DNA1	1	caggacgt	acgaattc	2
topDNA1+topDNA2	DNA2	1	taatctta	acgaattc	2
Germ RKKH			caggacgt	agcttcag	1
Germ RKKH			caggacgt	agcttcag	2
Germ RKKH			caggacgt	agcttcag	3

Tab. D.3: Combinations of P5 and P7 indices added during the second sequencing preparation PCR in order to identify corresponding library, target and selection round for all sequencing cluster. The primer sequences used are AATGATACGGCGACCACCGAGATCTACACxxxxxxxxACACTCTTTCCCTACACGAC (forward, 5′ → 3′) and CAAGCAGAAGACGGCATACGAGATxxxxxxxxGTGACTGGAGTTCAGACGTG (reverse, 5′ → 3′), where the xxxxxxxx invoke the listed barcodes and the sequences upstream and downstream the barcode are respectively the sequencing adapter and the primer annealing to the products of the first PCR. In total, three complete sequencing runs were performed to obtain the data presented here. Index combinations are unique within a given run.

D.4 List of model parameter

		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									1.00 ± 0.17	0.23 ± 0.21
Frog3	DNA1							0.99 ± 0.17	0.30 ± 0.18	N/A	N/A
Germ	DNA1									1.07 ± 0.10	0.27 ± 0.11
Germ	DNA2									0.80 ± 0.19	-0.14 ± 0.28
Frog3	PVP							0.45 ± 0.04	0.04 ± 0.05	N/A	N/A
NurseShark1	PVP									1.22 ± 0.45	0.52 ± 0.20

Tab. D.4: Parameters obtained from fits of the distribution of enrichments to generalized Pareto distributions (κ, τ) and lognormal distributions (σ, μ) for experiments performed by Sébastien Boyer [1] and within this project. N/A indicates that data was insufficient to make a meaningful fit. For enrichments against the protein targets between rounds $t = 2$ and $t + 1 = 3$, 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 Cramér-Rao bound.

Chapter E

Supplementary figures

E.1 Amplicon design and preparation for high-throughput sequencing

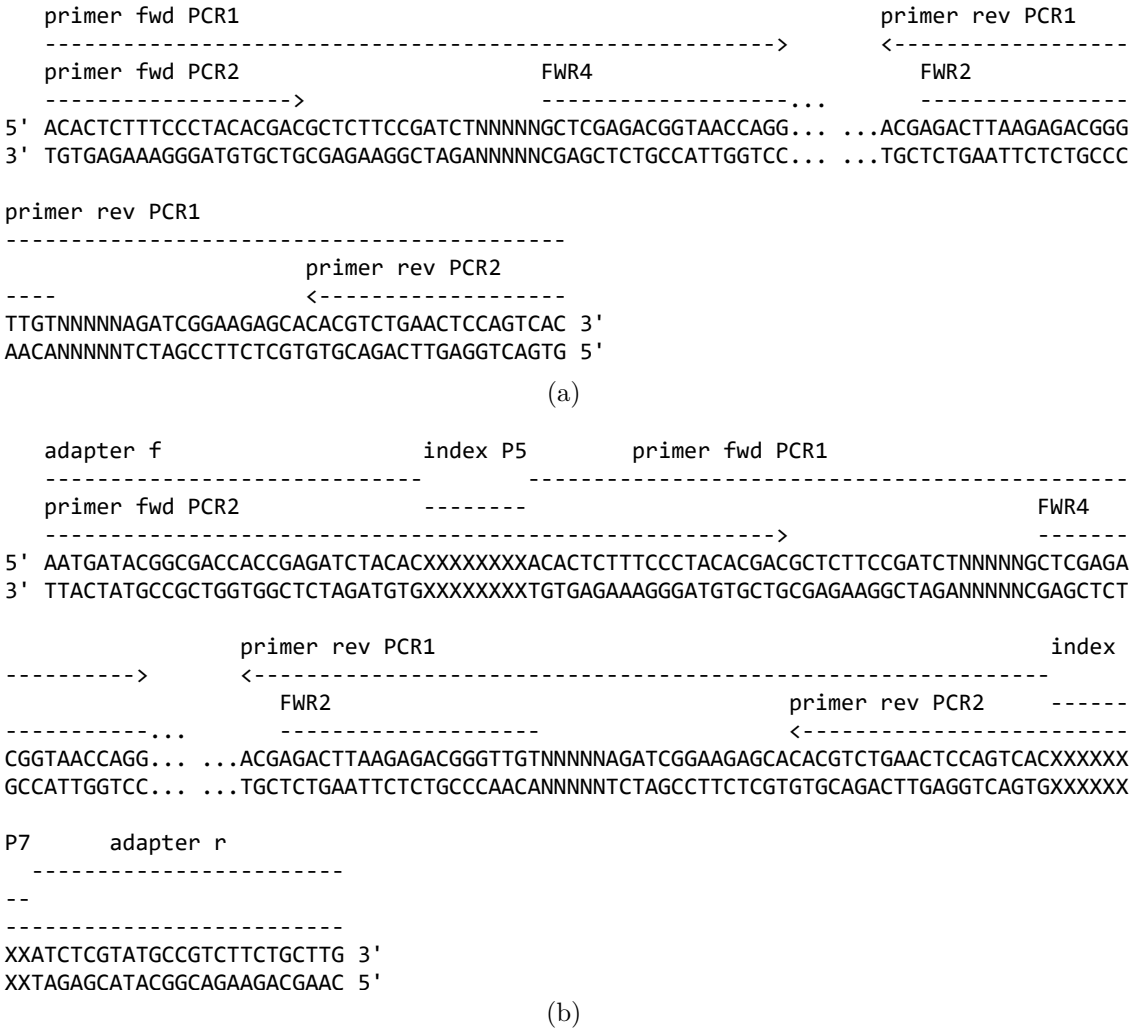


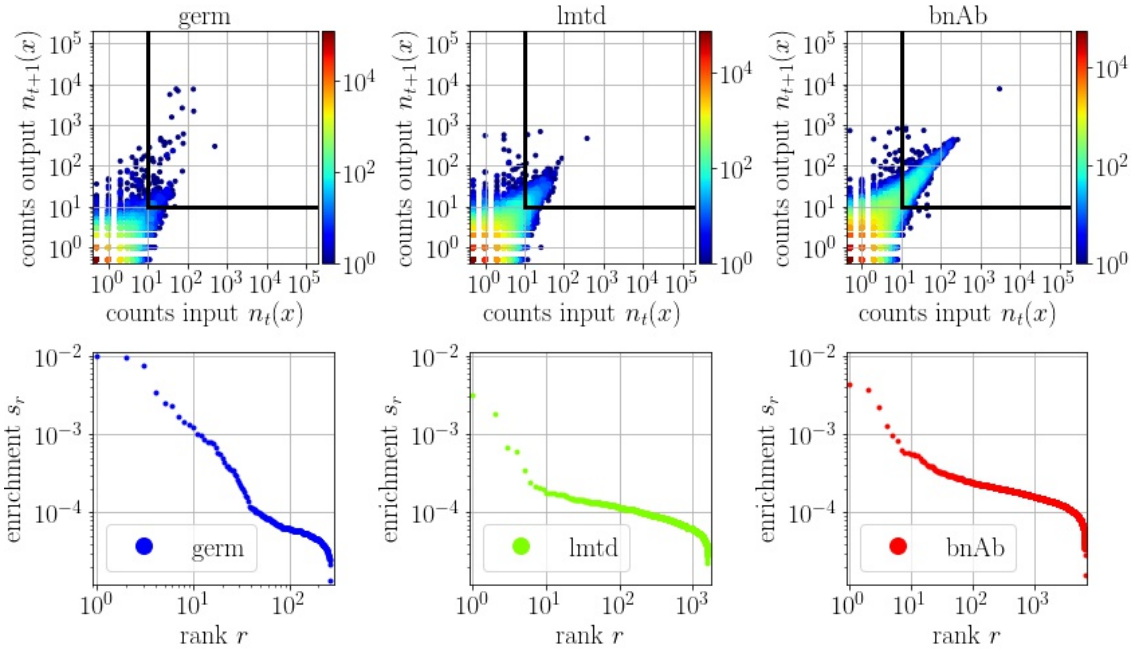
Fig. E.1: Design of the Illumina MiSeq sequencing amplicon. Only the flanking regions added up- and downstream of the region of interest on the antibody sequence and the primers linking to the antibody are shown; the region of interest in the center is replaced by dots. Both DNA strands are shown. **(a)** Amplicon after the first PCR: Two random 5 bp cluster barcodes indicated by `NNNNN` encompass the region of interest and allow for discrimination between clusters later during the sequencing. The sequences at the extremities of the amplicon are used as primers for the second PCR. **(b)** Amplicon after the second PCR: Two non-random, 8 bp sample-specific barcodes (indices) indicated by `XXXXXXXX`, as well as adapter sequences at the extremities defined by the Illumina platform now encompass the product of the first PCR. The sample barcodes allow to assign clusters back to samples they originate from. The sample barcodes used throughout the project are provided in table D.3.

		Total # of reads	Final molar concentration C_fin [nM]	Final concentration c_fin [ng/uL]	Total volume to submit V_tot [uL]	Final DNA mass m_fin [ng]
Input						
Output		1,00E+07	50,00	9,164	100,00	916,41

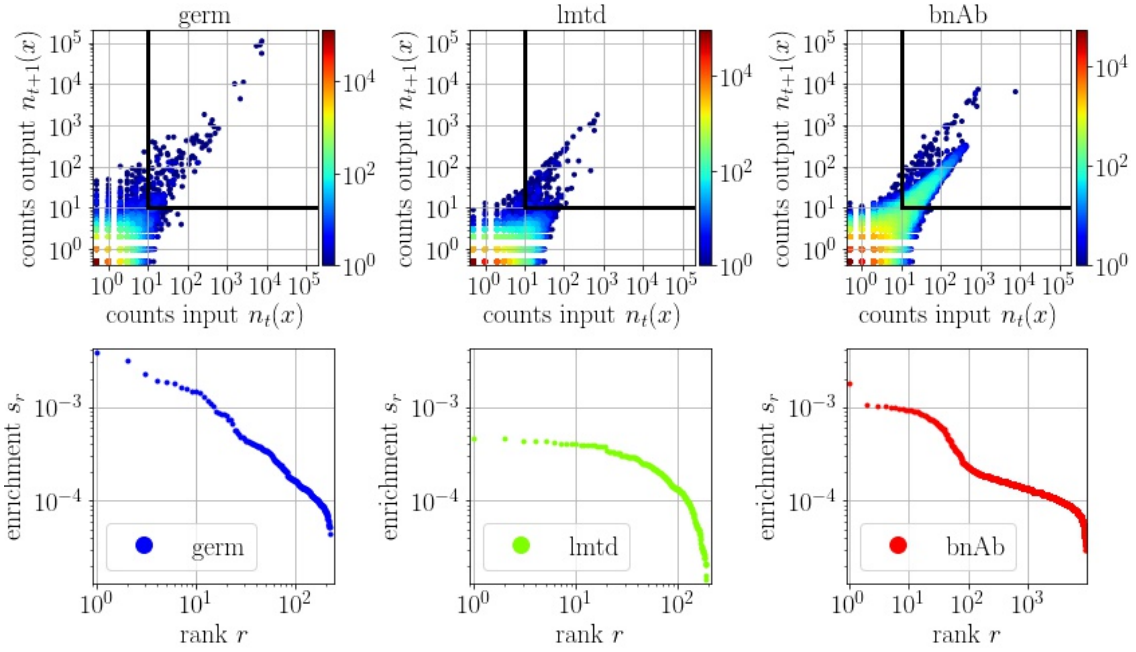
	Amplicon length L [bp]	Molar mass M [ng/nmol]	Sample concentration c [ng/uL]	Projected # of reads	Final molar concentration C_fin [nM]	Final concentration c_fin [ng/uL]	Sample volume to add V_tot [uL]	Final DNA mass m_fin [ng]
GLB 0 PCR2	316	208560	17,10	2,00E+05	1,00	0,209	1,22	20,86
GLB-B2 PCR2	316	208560	16,50	5,00E+05	2,50	0,521	3,16	52,14
GLB-N2 PCR2	316	208560	15,20	5,00E+05	2,50	0,521	3,43	52,14
GLB-B3 PCR2	316	208560	12,50	5,00E+05	2,50	0,521	4,17	52,14
GLB-N3 PCR2	316	208560	15,40	5,00E+05	2,50	0,521	3,39	52,14
TOP Bleue Input PCR2	316	208560	12,90	2,00E+05	1,00	0,209	1,62	20,86
TOP Noire Input PCR2	316	208560	13,70	2,00E+05	1,00	0,209	1,52	20,86
TOP Bleue Output PCR2	316	208560	17,00	2,15E+06	10,75	2,242	13,19	224,20
TOP Noire Output PCR2	316	208560	14,00	2,15E+06	10,75	2,242	16,01	224,20
Gm-N TOP1 PCR2	316	208560	16,20	1,00E+05	0,50	0,104	0,64	10,43
Megane 1	191	126060	18,70	2,00E+05	1,00	0,126	0,67	12,61
Megane 2	191	126060	10,90	2,00E+05	1,00	0,126	1,16	12,61
Megane 3	191	126060	15,50	2,00E+05	1,00	0,126	0,81	12,61
Megane 4	191	126060	9,46	2,00E+05	1,00	0,126	1,33	12,61
Megane 5	187	123420	7,74	1,00E+06	5,00	0,617	7,97	61,71
Megane 6	187	123420	10,80	1,00E+06	5,00	0,617	5,71	61,71
Megane 7	191	126060	13,30	1,00E+05	0,50	0,063	0,47	6,30
Megane 8	191	126060	12,50	1,00E+05	0,50	0,063	0,50	6,30
					Water to add: [uL]		33,01	

Fig. E.2: Example of an amplicon multiplexing for Illumina sequencing. Given the size and concentration of the final amplicon, as well as the desired number of counts for each sample, final volume, and final molar DNA concentration (all in yellow), the volume of each sample and of additional water to be mixed together is calculated (output quantities in green).

E.2 Sequence counts



(a)



(b)

Fig. E.3: Raw data from selection experiments. Mix3 library mix against the DNA1 target. **(a)** Selection round $t = 1$ versus round $t+1 = 2$, **(b)** $t = 2$, $t+1 = 3$. **Top** The number of counts after selection $n_{t+1}(x)$ is plotted against the number of counts before selection $n_t(x)$ for all libraries ℓ and CDR3 sequences x at several rounds of selection t. The color code encodes the number of sequences per dot. The solid black line defines the window, $n_{t+1}(\ell, x) \geq 10$, $n_t(\ell, x) \geq 10$, in which enrichments can be reliably computed from the ratios $n_{t+1}(\ell, x)/n_t(\ell, x)$. **Bottom** Enrichments $s(\ell, x) = an_{t+1}(\ell, x)/n_t(\ell, x)$ sorted in decreasing order plotted against their rank within the sample. Here, a is chosen such that $\sum_{\ell,r} s_r(\ell) = 1$. **Left** Germline library, **center** Limited library, **right** BnAb library. Continuation in figure E.4.

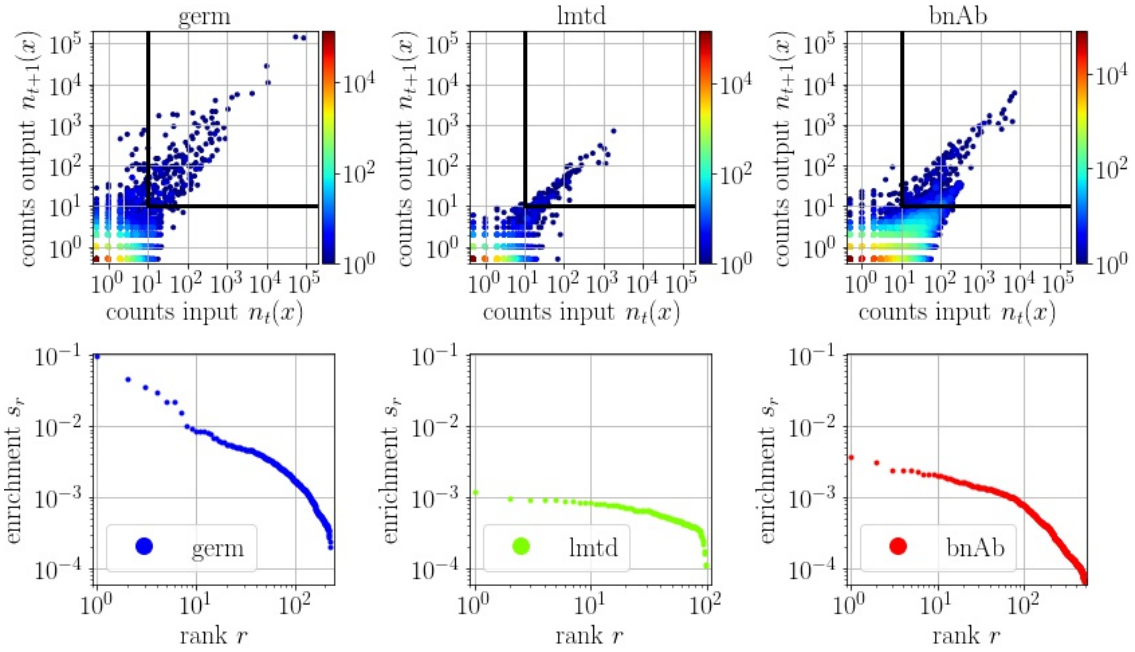


Fig. E.4: Continuation of figure E.3. Mix3 library mix against the DNA1 target, $t = 3$, $t + 1 = 4$.

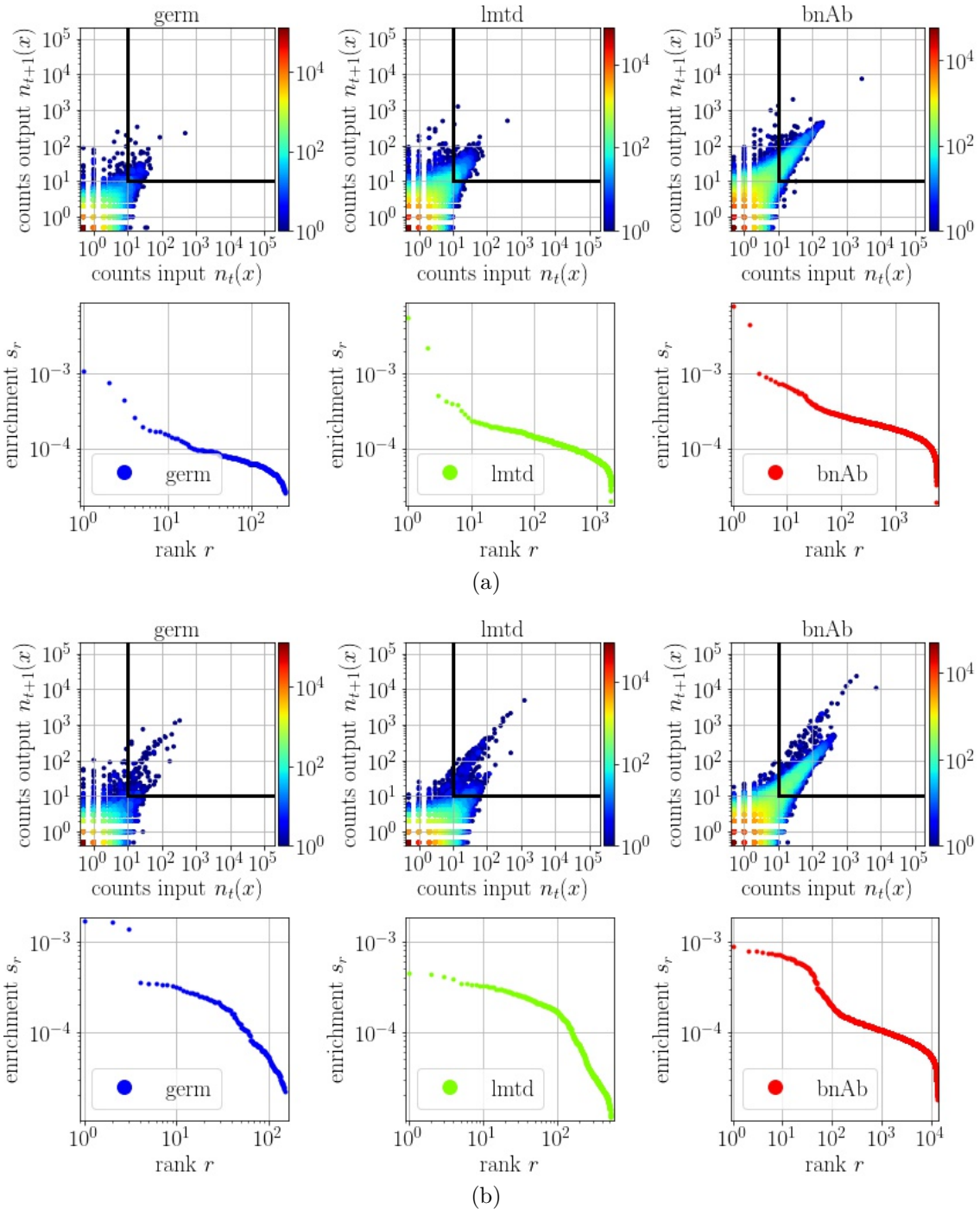


Fig. E.5: Raw data from selection experiments. Similar figure as figure E.3 for the Mix3 library mix against the DNA2 target. **(a)** Selection round $t = 1$ versus round $t + 1 = 2$, **(b)** $t = 2$, $t + 1 = 3$.

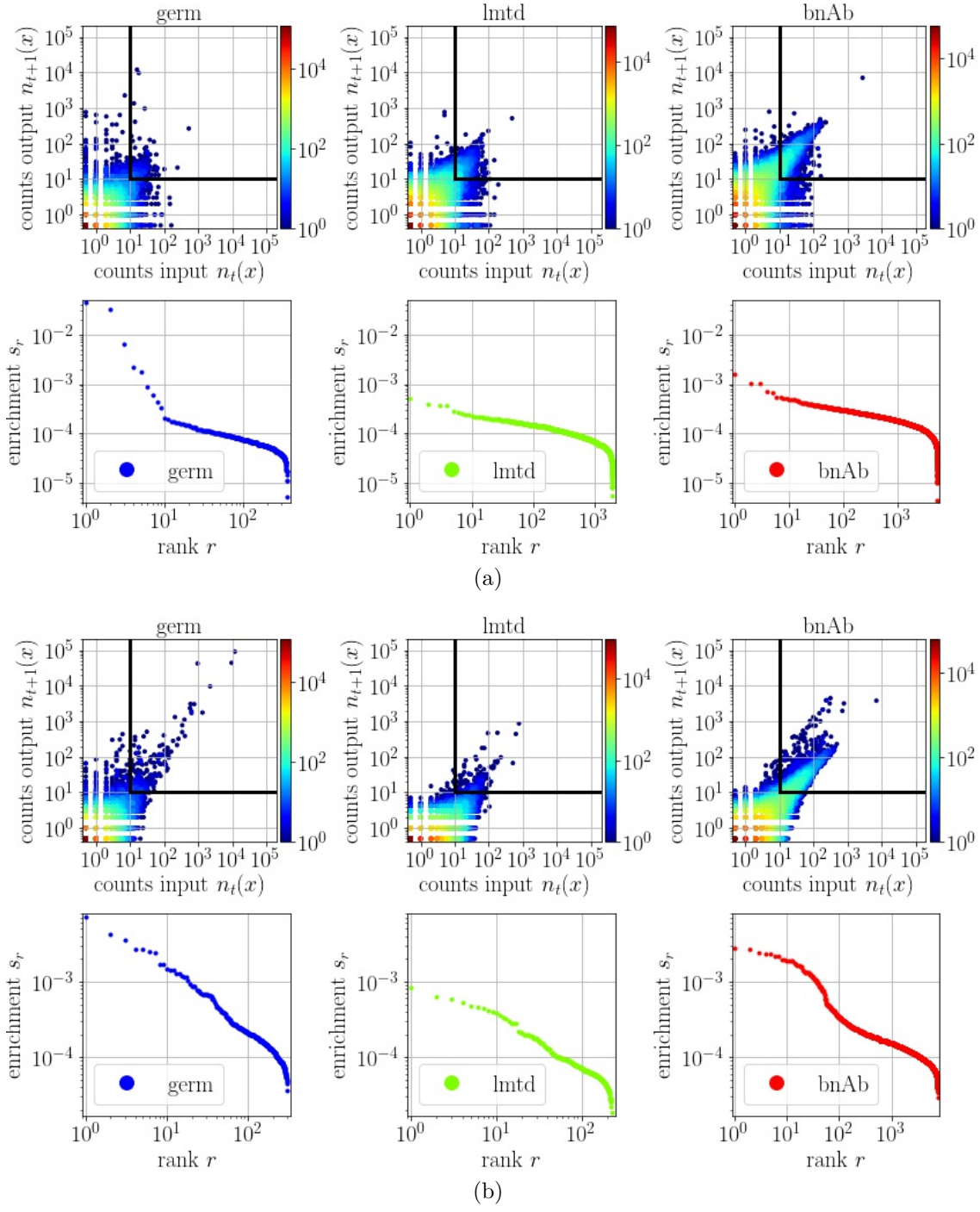


Fig. E.6: Raw data from selection experiments. Similar figure as figure E.3 for replica 1 of the Mix3 library mix against the prot1 (eGFP) target. **(a)** Selection round $t = 1$ versus round $t + 1 = 2$, **(b)** $t = 2$, $t + 1 = 3$.

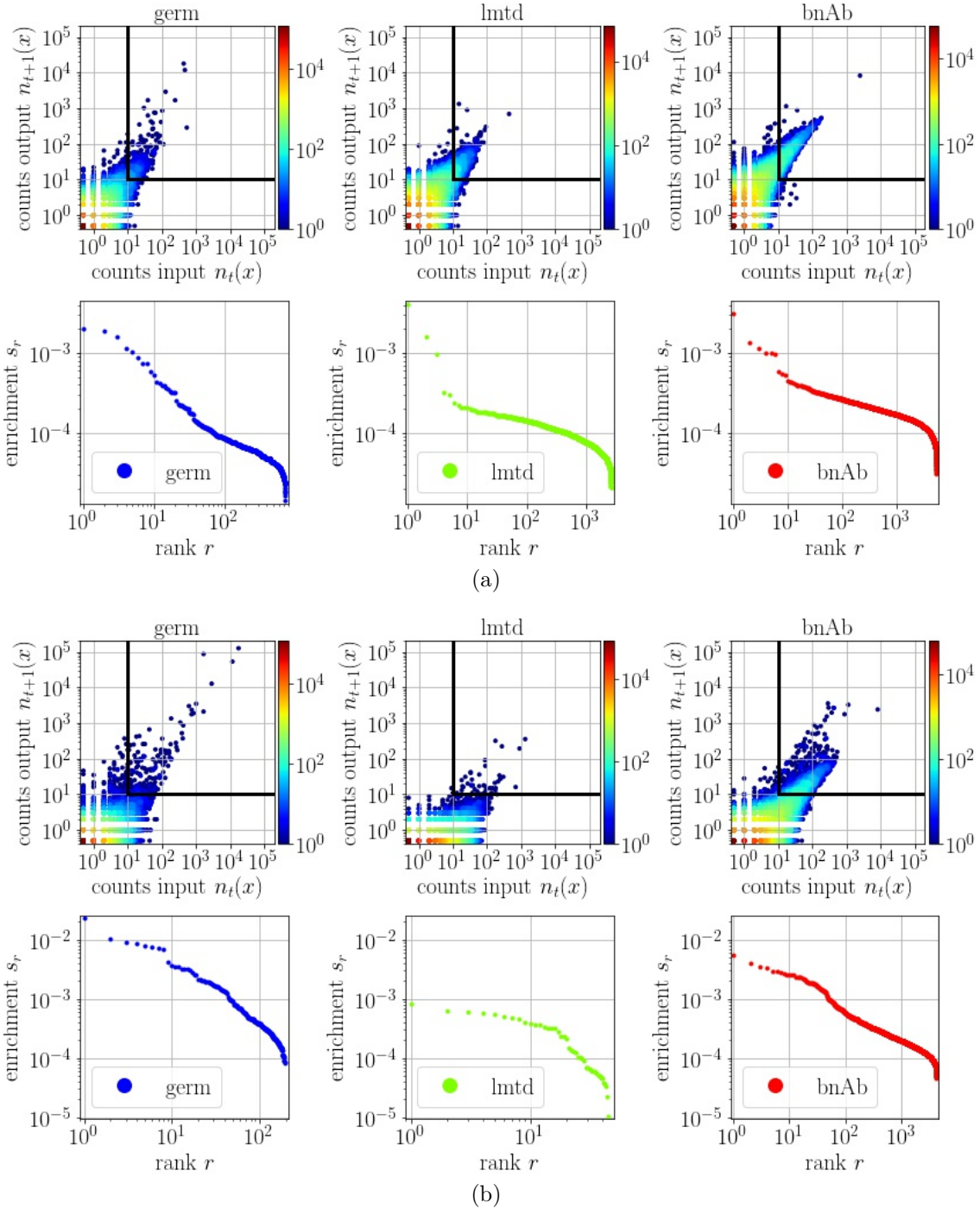


Fig. E.7: Raw data from selection experiments. Similar figure as figure E.3 for replica 2 of the Mix3 library mix against the prot1 (eGFP) target. **(a)** Selection round $t = 1$ versus round $t + 1 = 2$, **(b)** $t = 2$, $t + 1 = 3$.

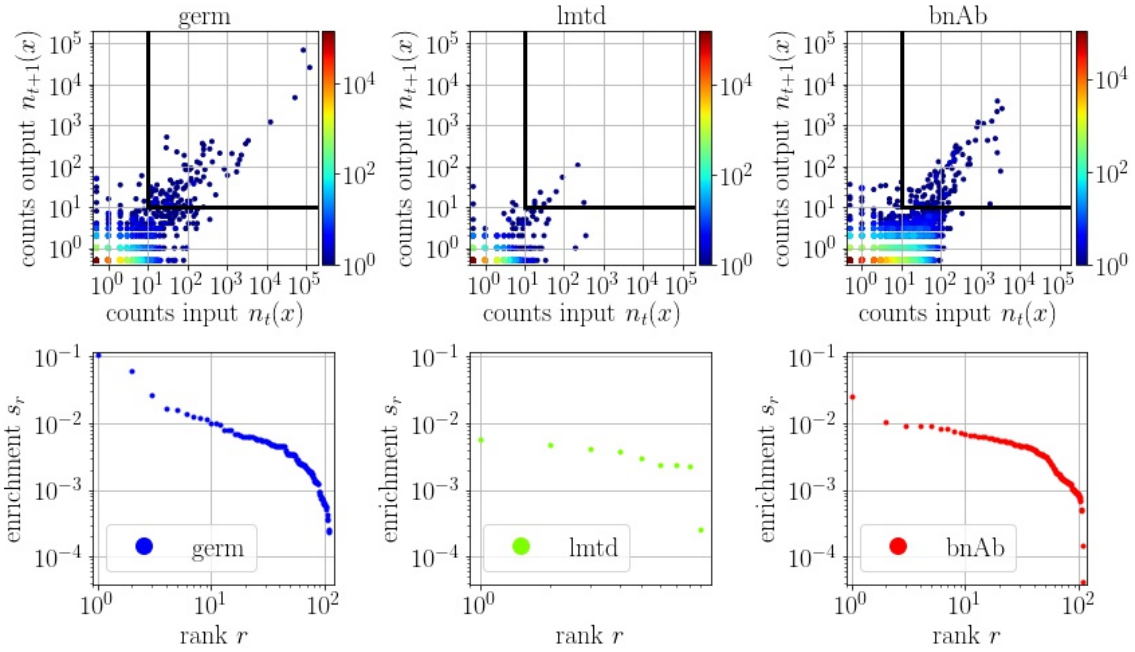


Fig. E.8: Continuation of figure E.7. Replica 2 of the Mix3 library mix against the prot1 (eGFP) target, $t = 3$, $t + 1 = 4$.

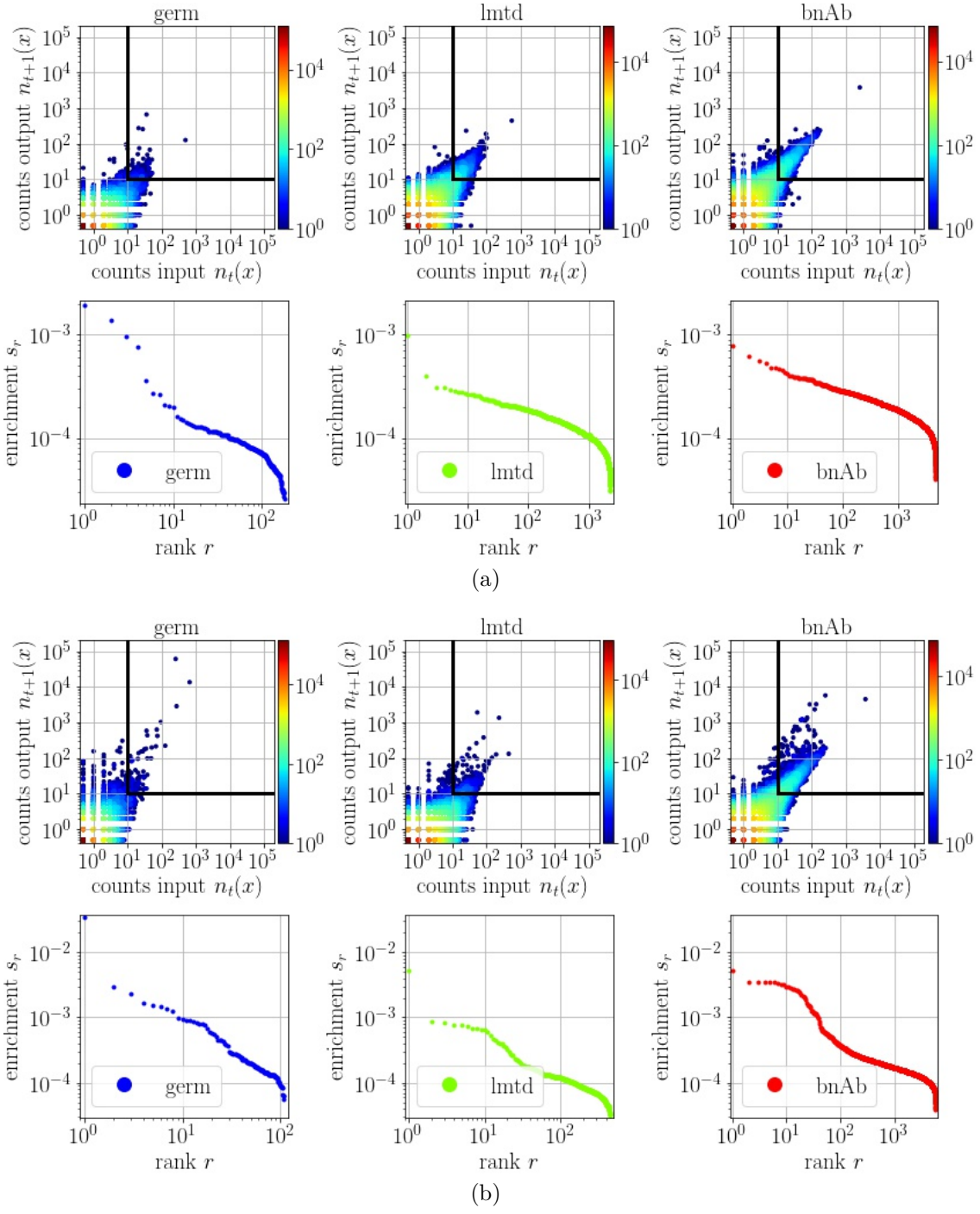


Fig. E.9: Raw data from selection experiments. Similar figure as figure E.3 for replica 1 of the Mix3 library mix against the prot2 (mCherry) target. **(a)** Selection round $t = 1$ versus round $t + 1 = 2$, **(b)** $t = 2$, $t + 1 = 3$.

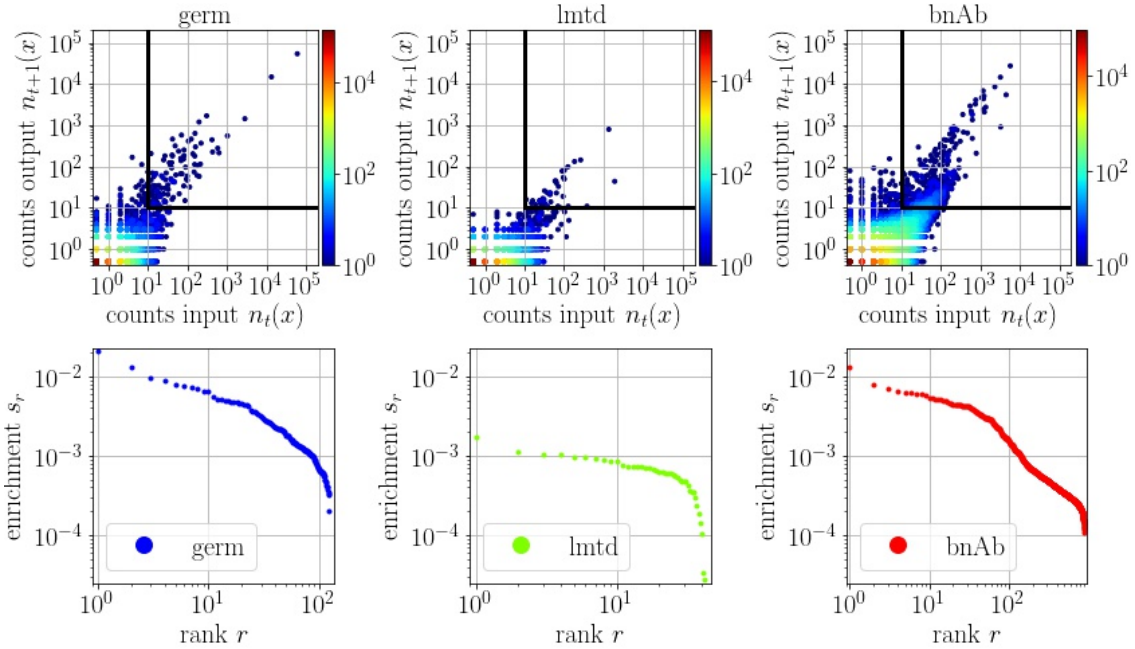


Fig. E.10: Continuation of figure E.9. Replica 1 of the Mix3 library mix against the prot2 (mCherry) target, $t = 3$, $t + 1 = 4$.

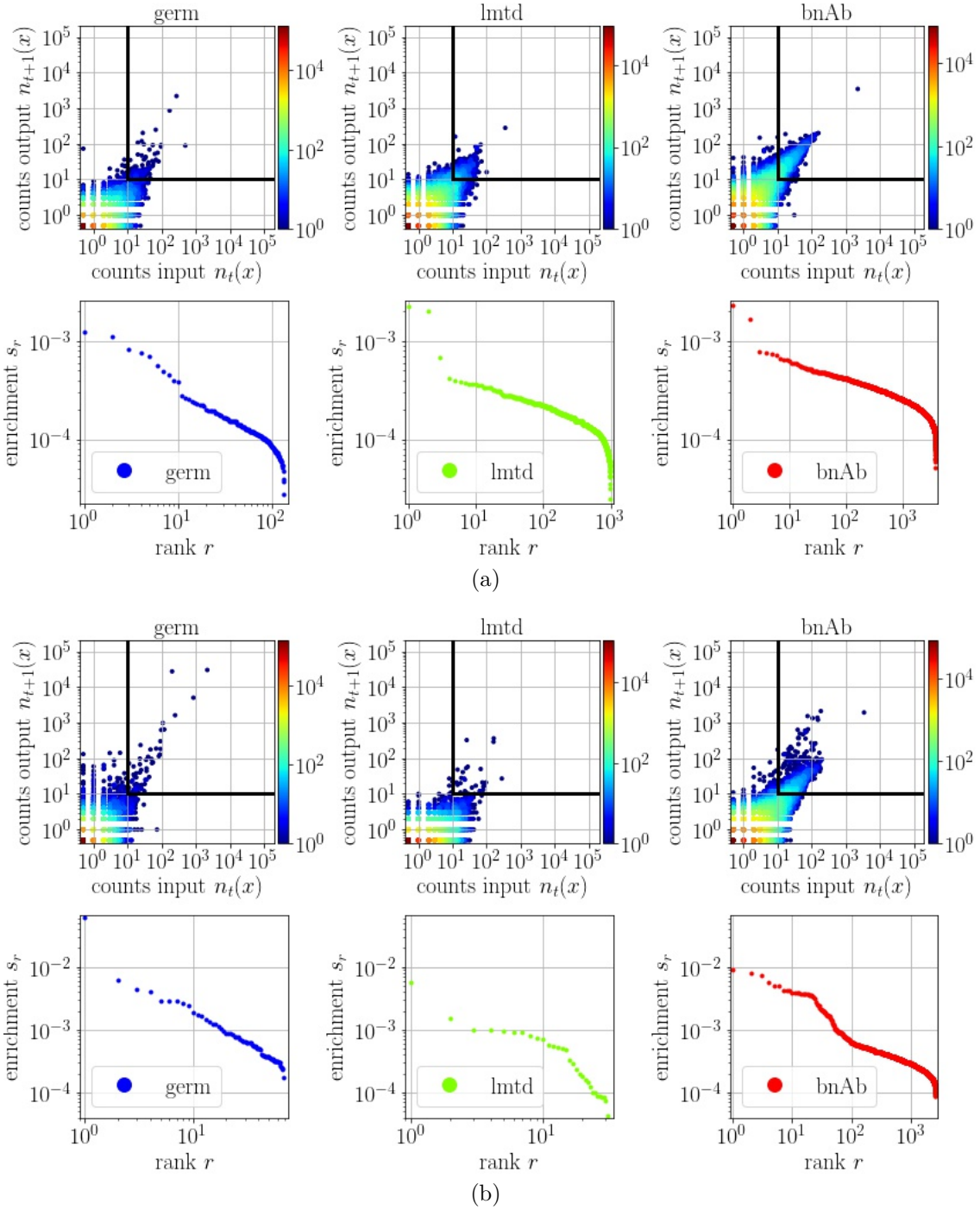


Fig. E.11: Raw data from selection experiments. Similar figure as figure E.3 for replica 2 of the Mix3 library mix against the prot2 (mCherry) target. **(a)** Selection round $t = 1$ versus round $t + 1 = 2$, **(b)** $t = 2$, $t + 1 = 3$.

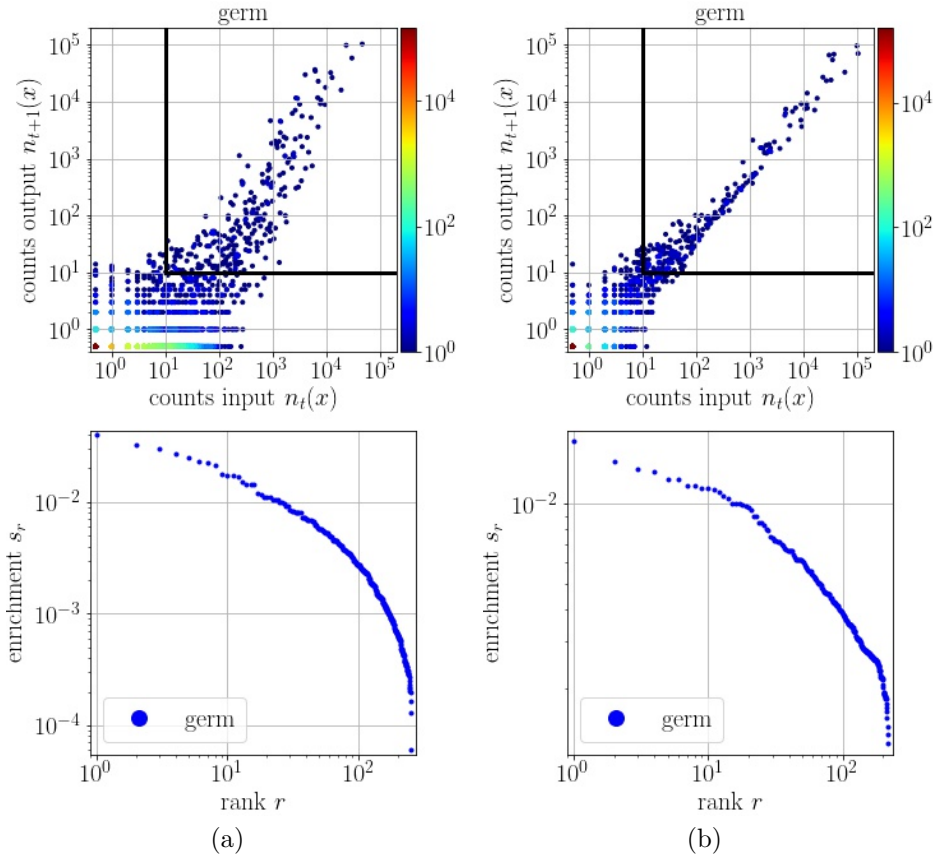


Fig. E.12: Raw data from selection experiments. Similar figure as figure E.3 for the Germline library (alone) against the DNA1 target. **(a)** Selection round $t = 1$ versus round $t + 1 = 2$, **(b)** $t = 2$, $t + 1 = 3$.

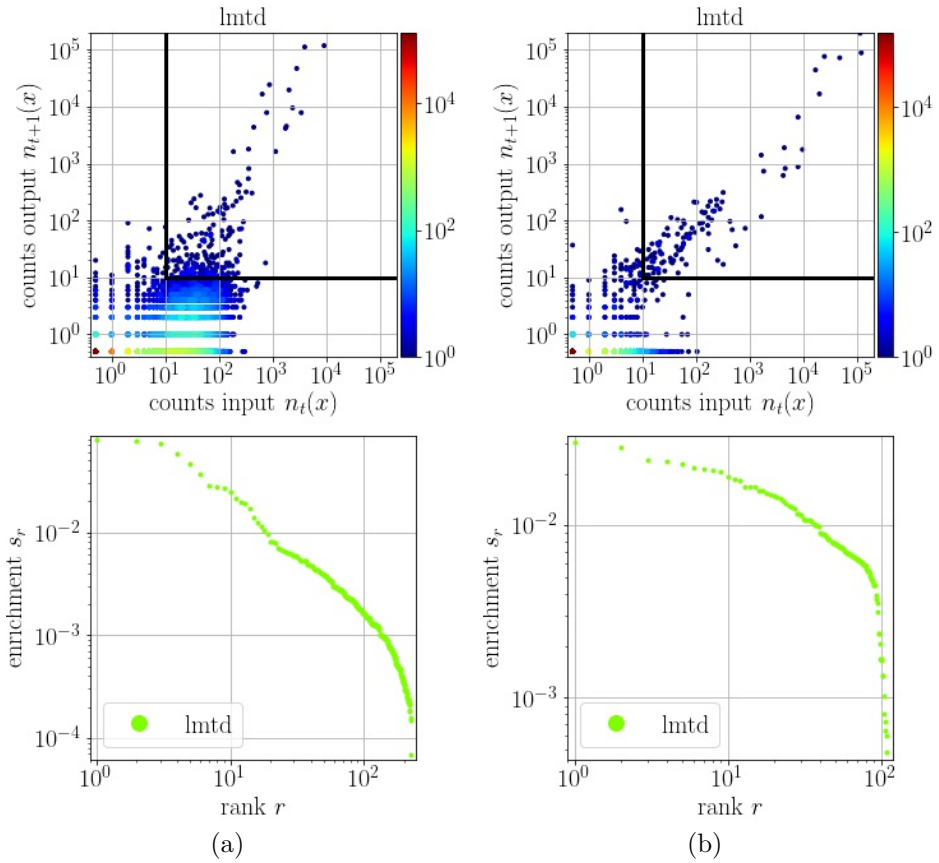


Fig. E.13: Raw data from selection experiments. Similar figure as figure E.3 for the Limited library (alone) against the DNA1 target. **(a)** Selection round $t = 1$ versus round $t + 1 = 2$, **(b)** $t = 2$, $t + 1 = 3$.

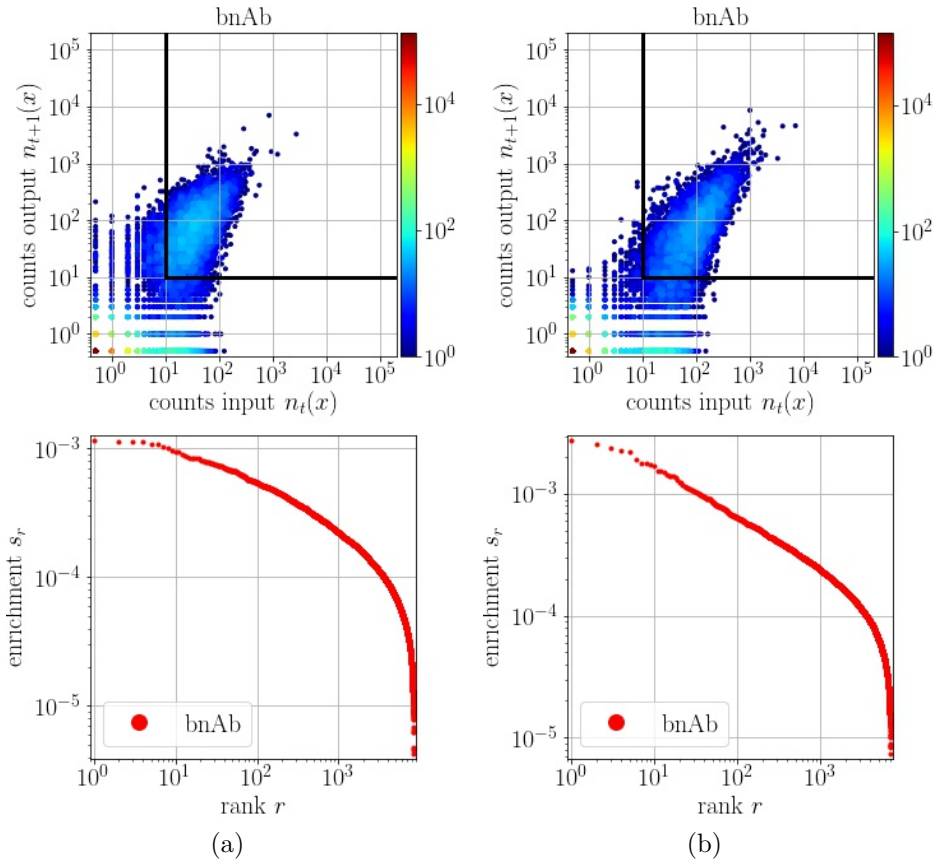


Fig. E.14: Raw data from selection experiments. Similar figure as figure E.3 for the BnAb library (alone) against the DNA1 target. **(a)** Selection round $t = 1$ versus round $t + 1 = 2$, **(b)** $t = 2$, $t + 1 = 3$.

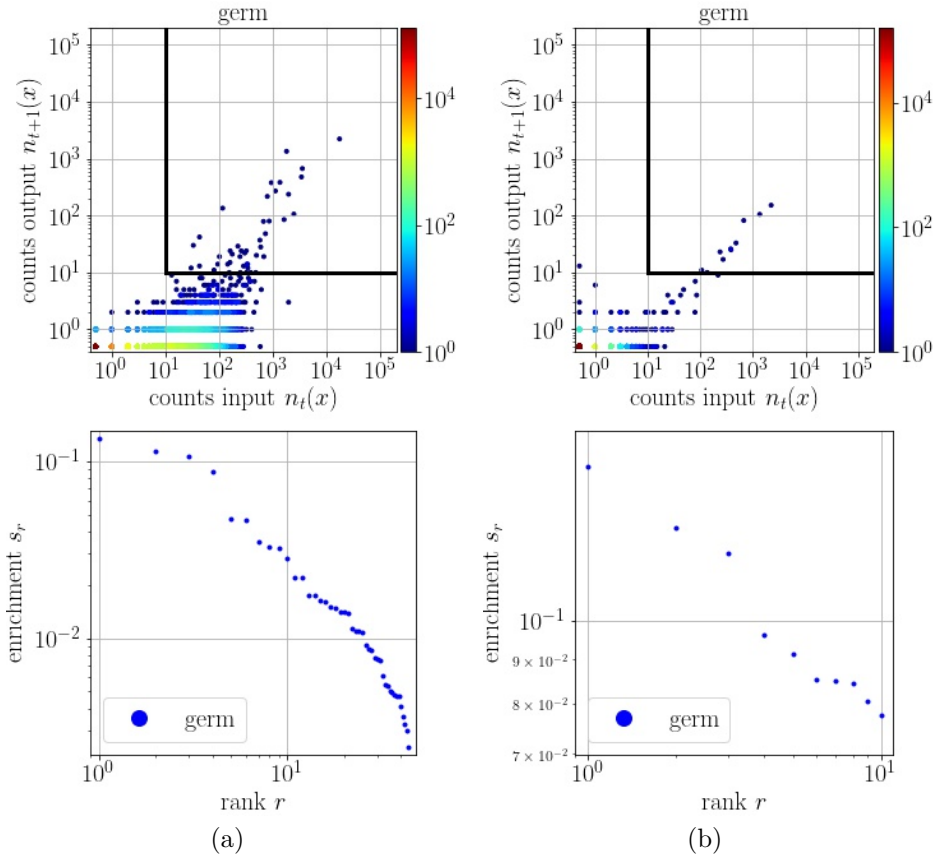


Fig. E.15: Raw data from selection experiments. Similar figure as figure E.3 for the Germline library (alone) against the DNA2 target. **(a)** Selection round $t = 1$ versus round $t + 1 = 2$, **(b)** $t = 2$, $t + 1 = 3$.

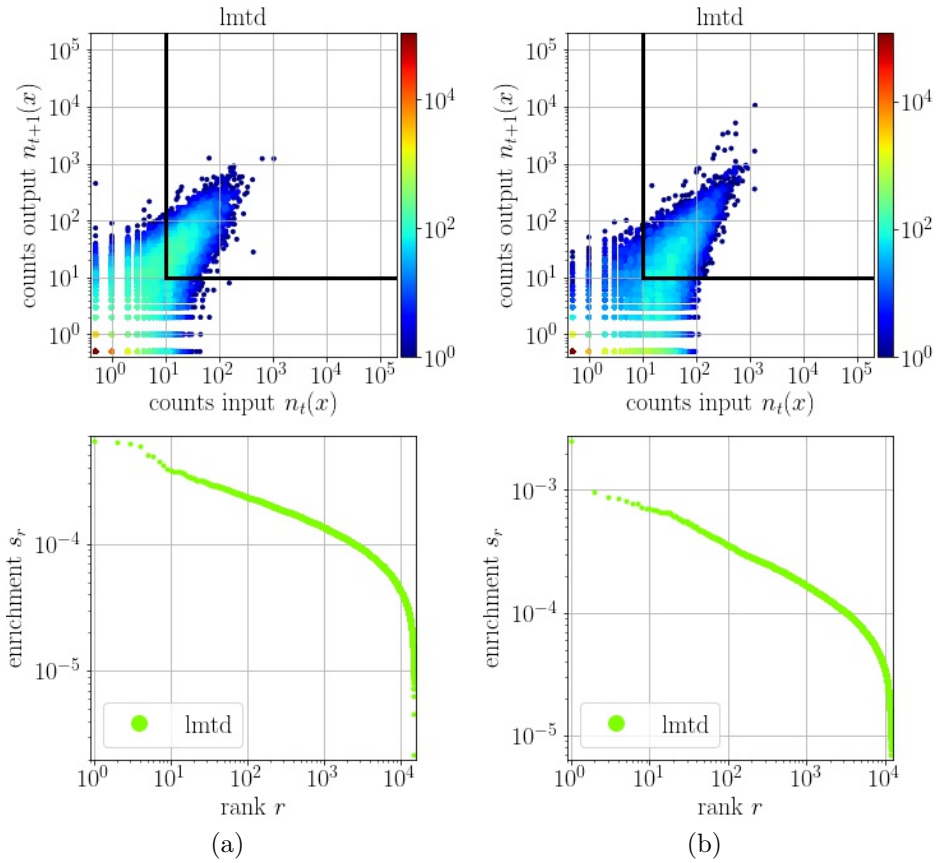


Fig. E.16: Raw data from selection experiments. Similar figure as figure E.3 for the Limited library (alone) against the DNA2 target. **(a)** Selection round $t = 1$ versus round $t + 1 = 2$, **(b)** $t = 2$, $t + 1 = 3$.

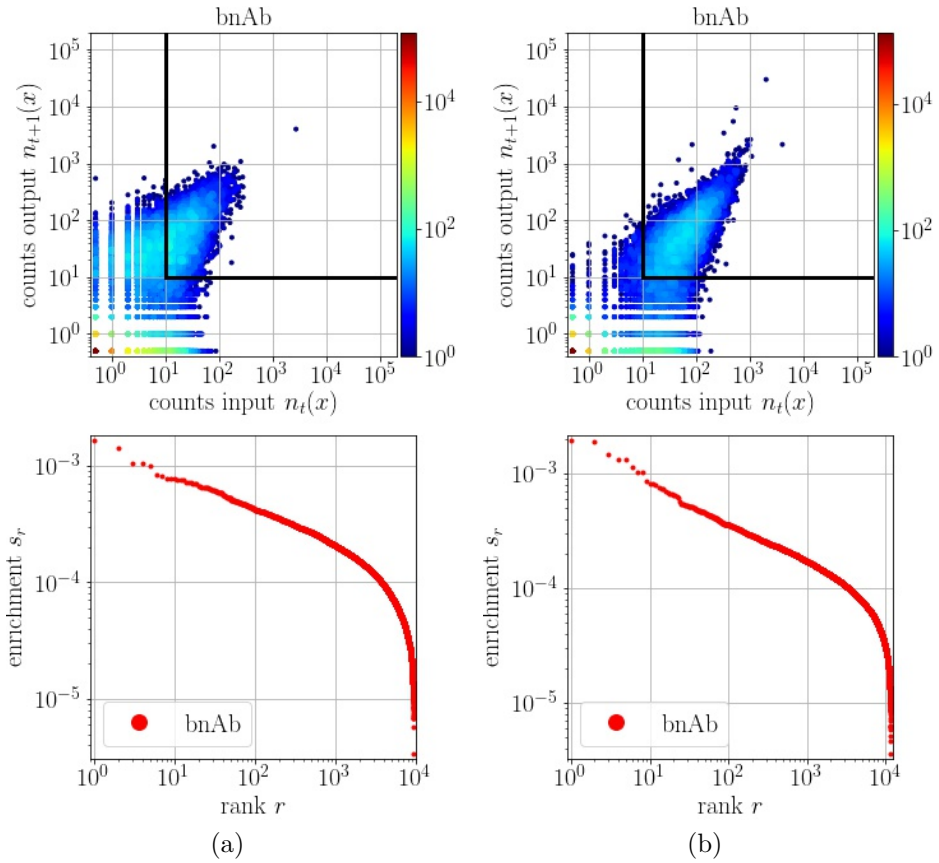


Fig. E.17: Raw data from selection experiments. Similar figure as figure E.3 for the BnAb library (alone) against the DNA2 target. **(a)** Selection round $t = 1$ versus round $t + 1 = 2$, **(b)** $t = 2$, $t + 1 = 3$.

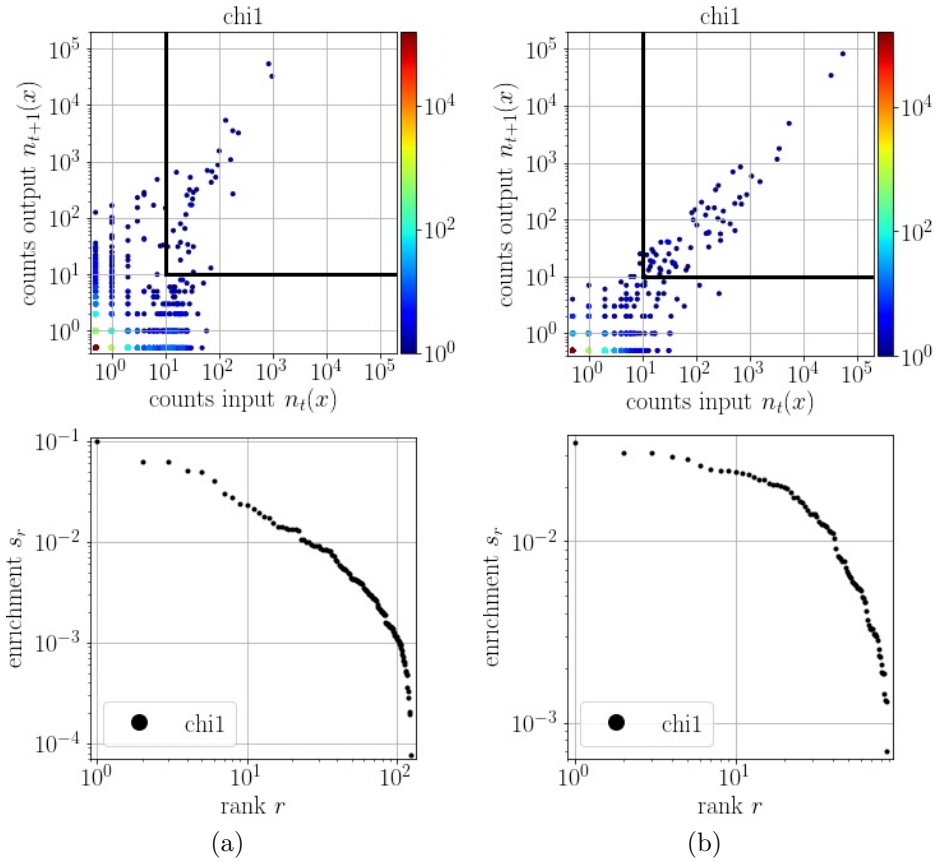


Fig. E.18: Raw data from selection experiments. Similar figure as figure E.3 for the Chicken library (in Mix21) against the DNA1 target. **(a)** Selection round $t = 1$ versus round $t + 1 = 2$, **(b)** $t = 2$, $t + 1 = 3$.

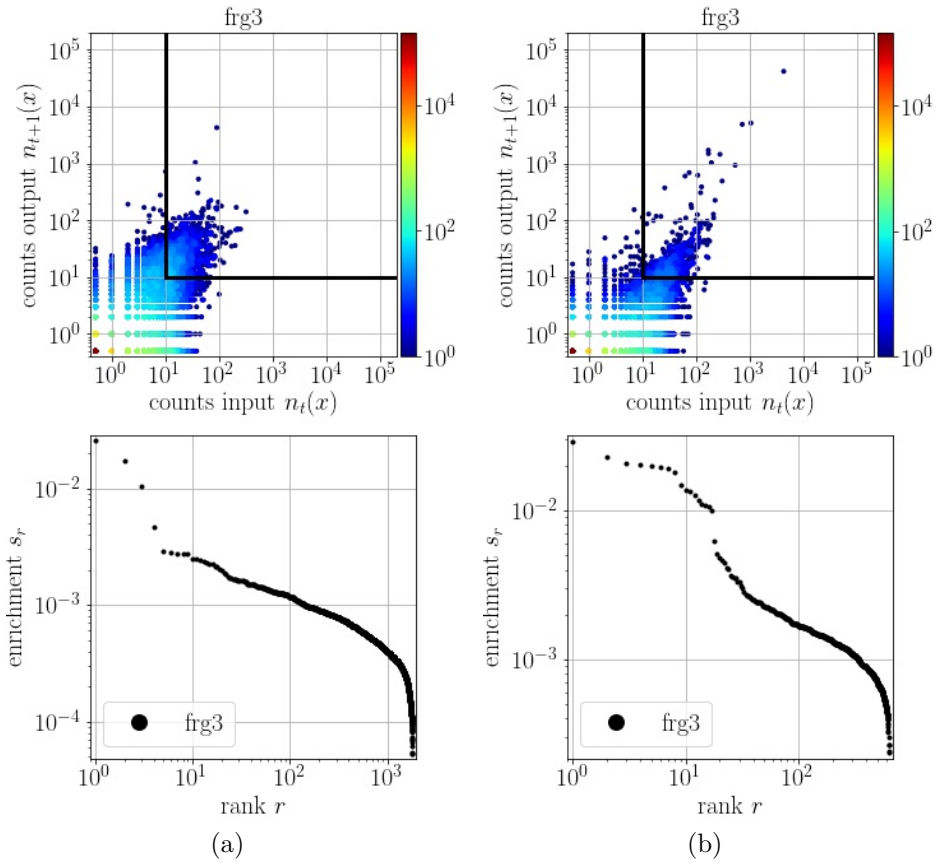


Fig. E.19: Raw data from selection experiments. Similar figure as figure E.3 for the Frog3 library (alone) against the DNA1 target. **(a)** Selection round $t = 1$ versus round $t + 1 = 2$, **(b)** $t = 2$, $t + 1 = 3$.

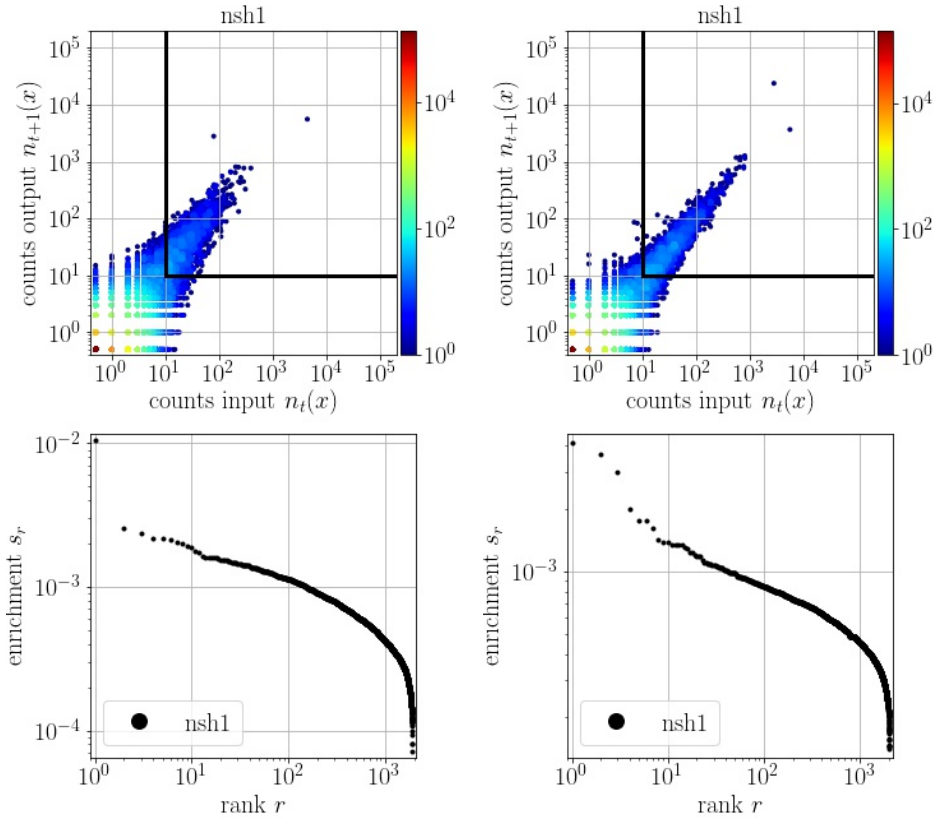


Fig. E.20: Raw data from selection experiments. Similar figure as figure E.3 for the NurseShark library (in Mix24) against the PVP target. **(a)** Selection round $t = 1$ versus round $t + 1 = 2$, **(b)** $t = 2$, $t + 1 = 3$.

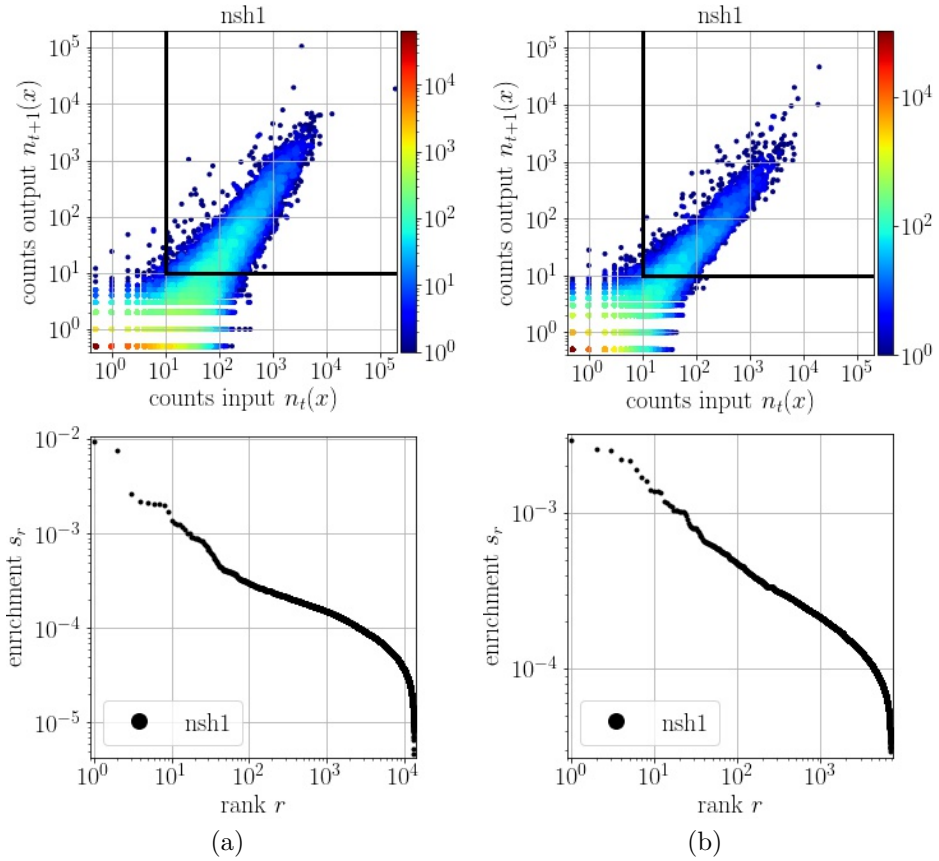


Fig. E.21: Raw data from selection experiments. Similar figure as figure E.3 for the NurseShark library (in Mix21) against the PVP target. **(a)** Selection round $t = 1$ versus round $t + 1 = 2$, **(b)** $t = 2$, $t + 1 = 3$.

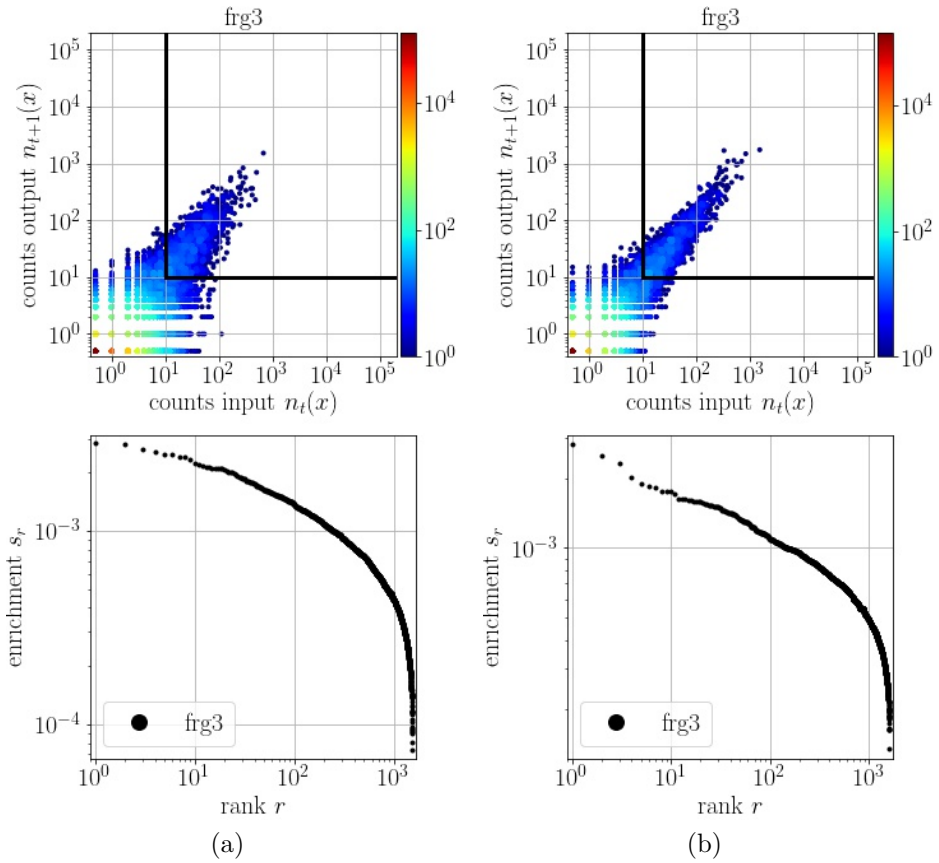


Fig. E.22: Raw data from selection experiments. Similar figure as figure E.3 for the Frog3 library (alone) against the PVP target. **(a)** Selection round $t = 1$ versus round $t + 1 = 2$, **(b)** $t = 2$, $t + 1 = 3$.

E.3 Amplification bias

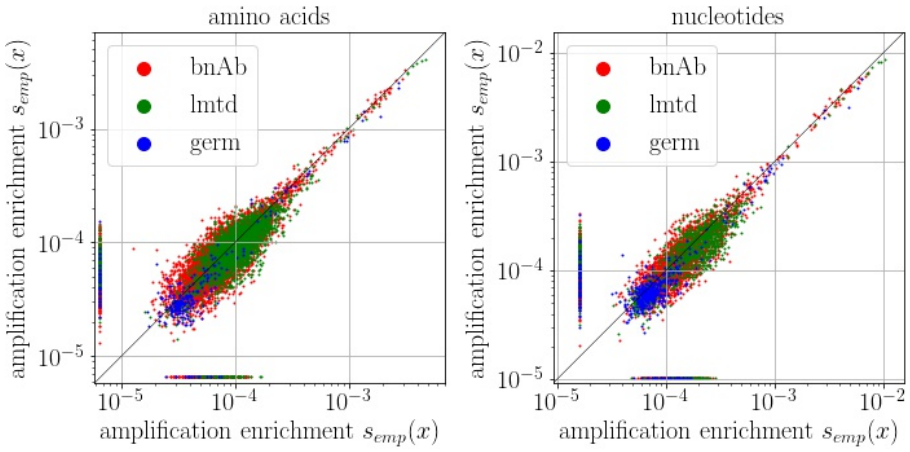
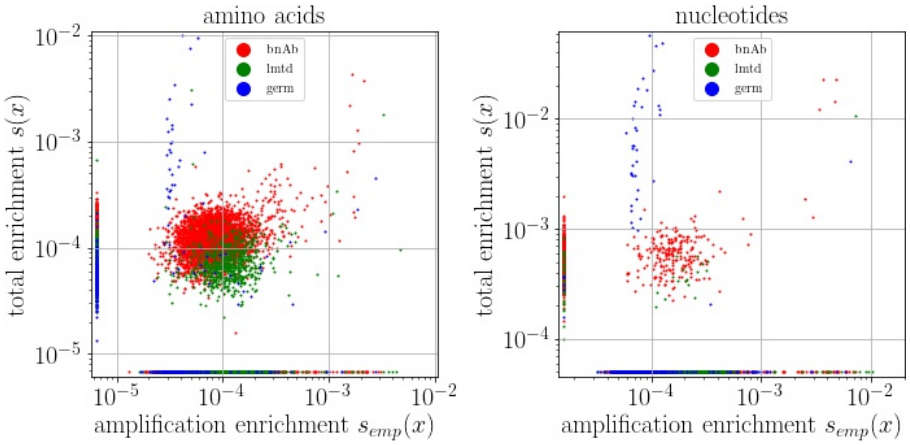
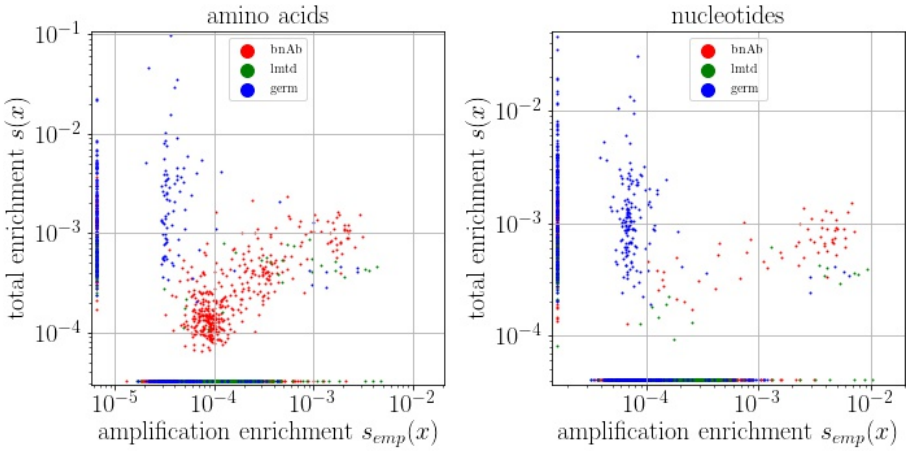


Fig. E.23: Reproducibility of amplification bias. Comparison of two replicates of the amplification step (complete experiment except for the selection for binding that is left out) showing reproducibility and thus sequence-dependent amplification. Enrichments due to amplification s_{ampl} are computed as count ratios from before/after amplification. **Left** Enrichments are computed for amino acid sequences. **Right** Enrichments are computed for nucleotide sequences.



(a)



(b)

Fig. E.24: Orthogonality of binding and amplification bias. Same as figure 4.5(b), but total enrichments at rounds **(a)** $t = 1$, $t + 1 = 2$ and **(b)** $t = 3$, $t + 1 = 4$ of the library mix (Mix3) against the DNA1 target are compared to amplification bias enrichments.

E.4 Choice of threshold enrichments s^*

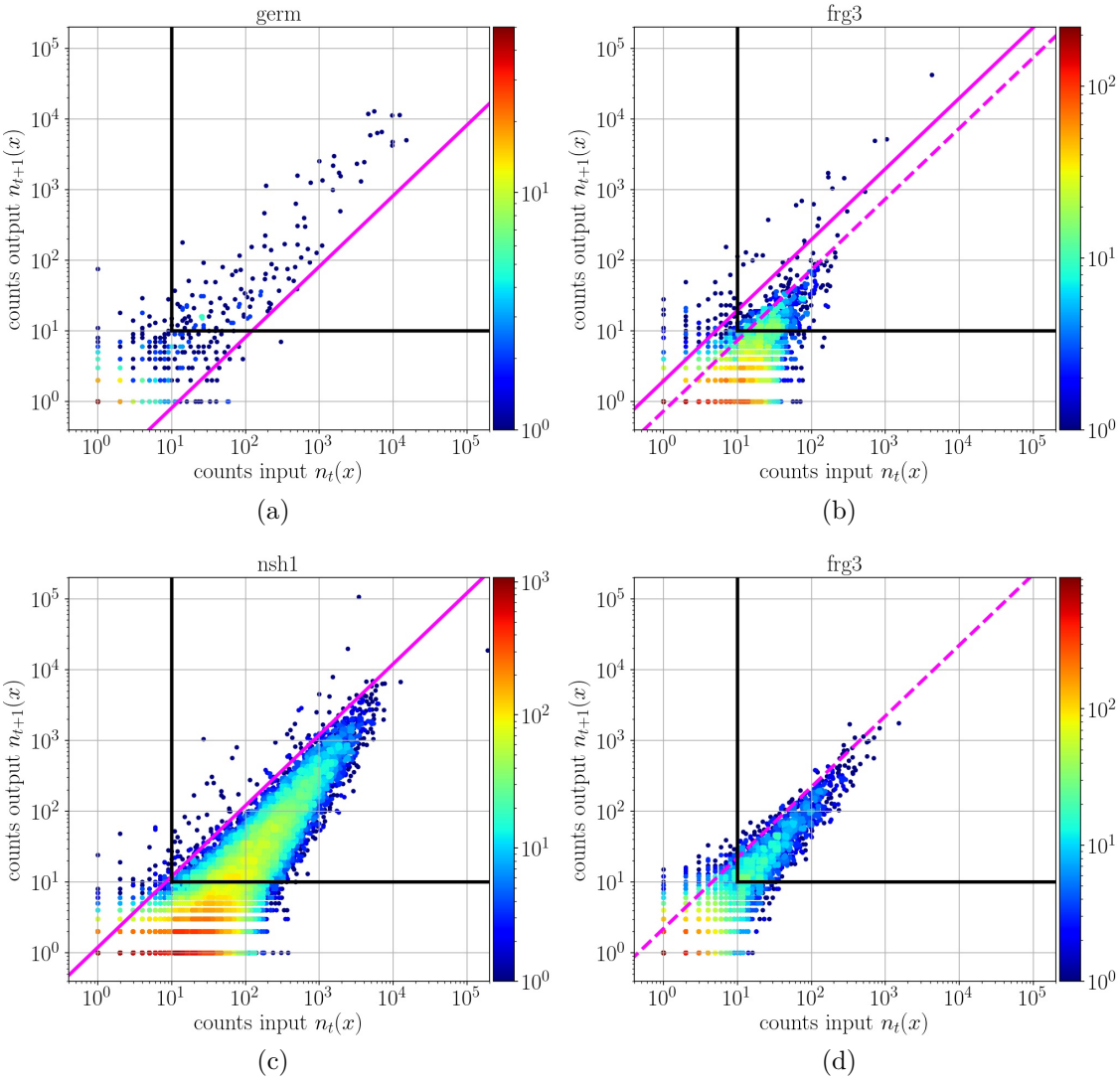


Fig. E.25: Choice of the threshold enrichment s^* for model inference for previous selection data published in [1]. The selection in- and output counts $n_t(x)$, $n_{t+1}(x)$ are plotted one against another along with the window defined by the upper triangle between the black and pink lines containing the sequences taken into account for model inference. The **black** and **pink** lines represent respectively the count threshold for reliable enrichment computation ($n_t(x) \geq 10$, $n_{t+1}(x) \geq 10$) and the choice of s^* (the line is parametrized by $n_{t+1}(x)/n_t(x) = s^*$). The **dashed** line represent values of s^* chosen in [1] but which are illegitimate because they cut through the unspecific binding mode. **Solid** lines represent valid or corrected values for s^*. **(a)** Germline library in Mix24 against the DNA1 target, $t = 2$, $t + 1 = 3$, **(b)** Frog3 library alone against the DNA1 target, $t = 2$, $t + 1 = 3$, **(c)** NurseShark library in Mix21 against the PVP target, $t = 1$, $t + 1 = 2$, **(d)** Frog3 library alone against the PVP target, $t = 2$, $t + 1 = 3$.

E.5 Threshold scans

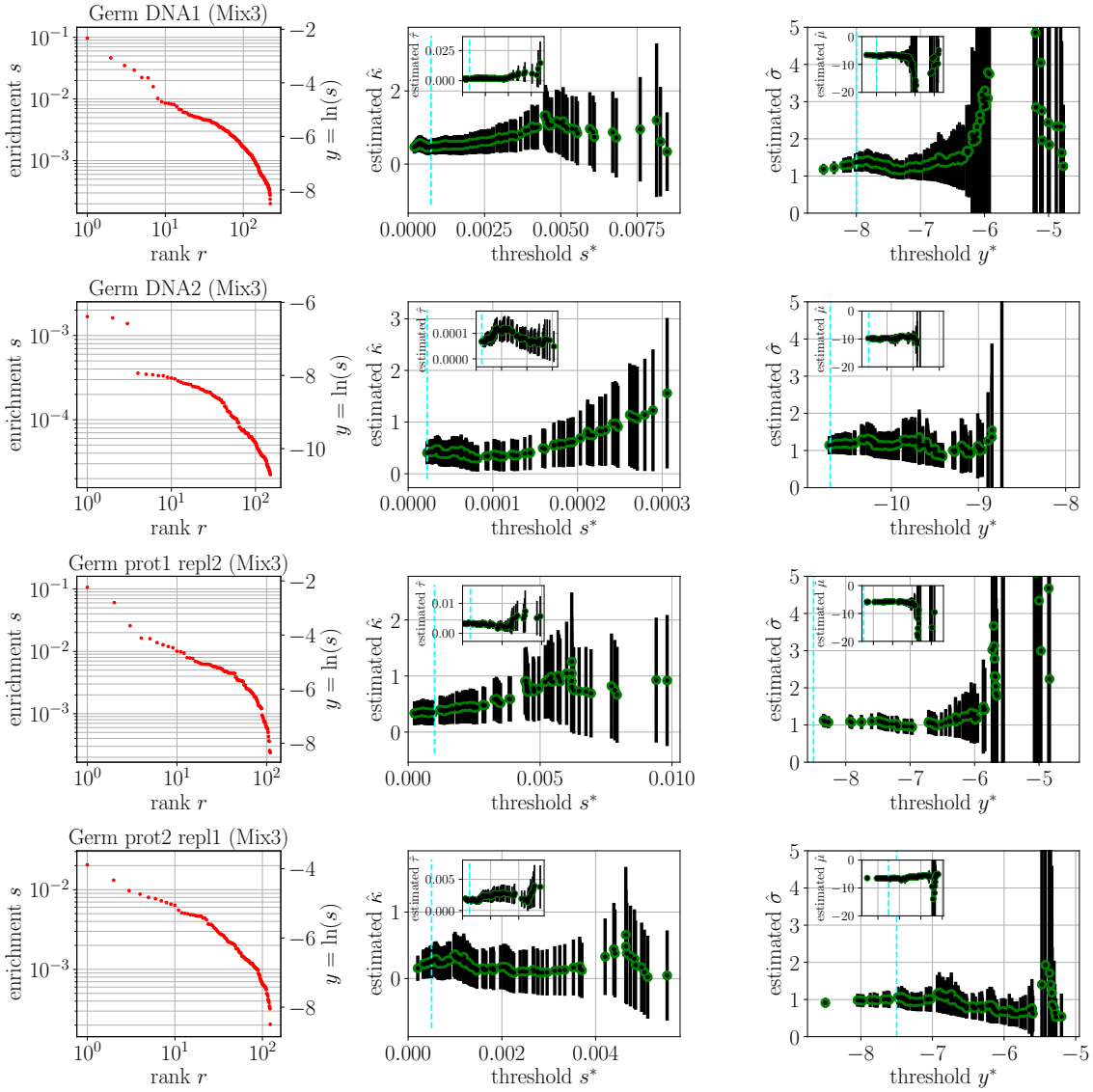


Fig. E.26: Threshold scan plots showing the values of model parameters as functions of truncation values for enrichment data from the Germline library in some of the Mix3 selections. **Left** Enrichments s and log-enrichments $y = \ln(s)$ sorted in decreasing order as a function of their rank r. **Center** ML EVT model parameter $\hat{\kappa}$ ($\hat{\tau}$ in inset) as a function of s^*. **Right** ML lognormal model parameter $\hat{\sigma}$ ($\hat{\mu}$ in inset) as a function of y^*. Error bars show 90 % confidence intervals estimated from the Fisher information matrix and the Cramer-Rao bound. The vertical dashed cyan lines indicate the chosen values of s^* and y^* used for the inference.

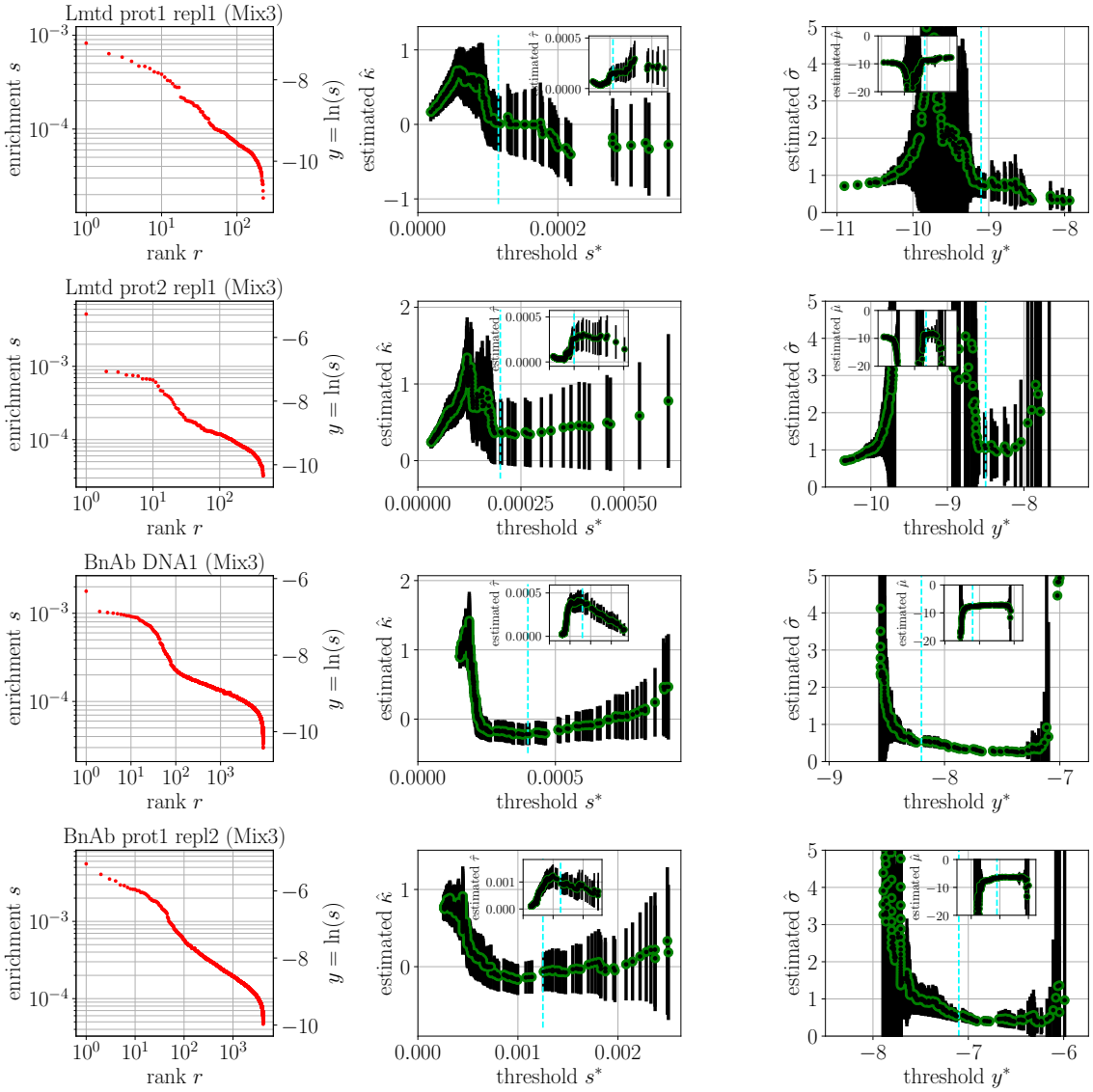


Fig. E.27: Threshold scan plots showing the values of model parameters as functions of truncation values for enrichment data from the Limited and BnAb library in some of the Mix3 selections. **Left** Enrichments s and log-enrichments $y = \ln(s)$ sorted in decreasing order as a function of their rank r. **Center** ML EVT model parameter $\hat{\kappa}$ ($\hat{\tau}$ in inset) as a function of s^*. **Right** ML lognormal model parameter $\hat{\sigma}$ ($\hat{\mu}$ in inset) as a function of y^*. Error bars show 90 % confidence intervals estimated from the Fisher information matrix and the Cramer-Rao bound. The vertical dashed cyan lines indicate the chosen values of s^* and y^* used for the inference.

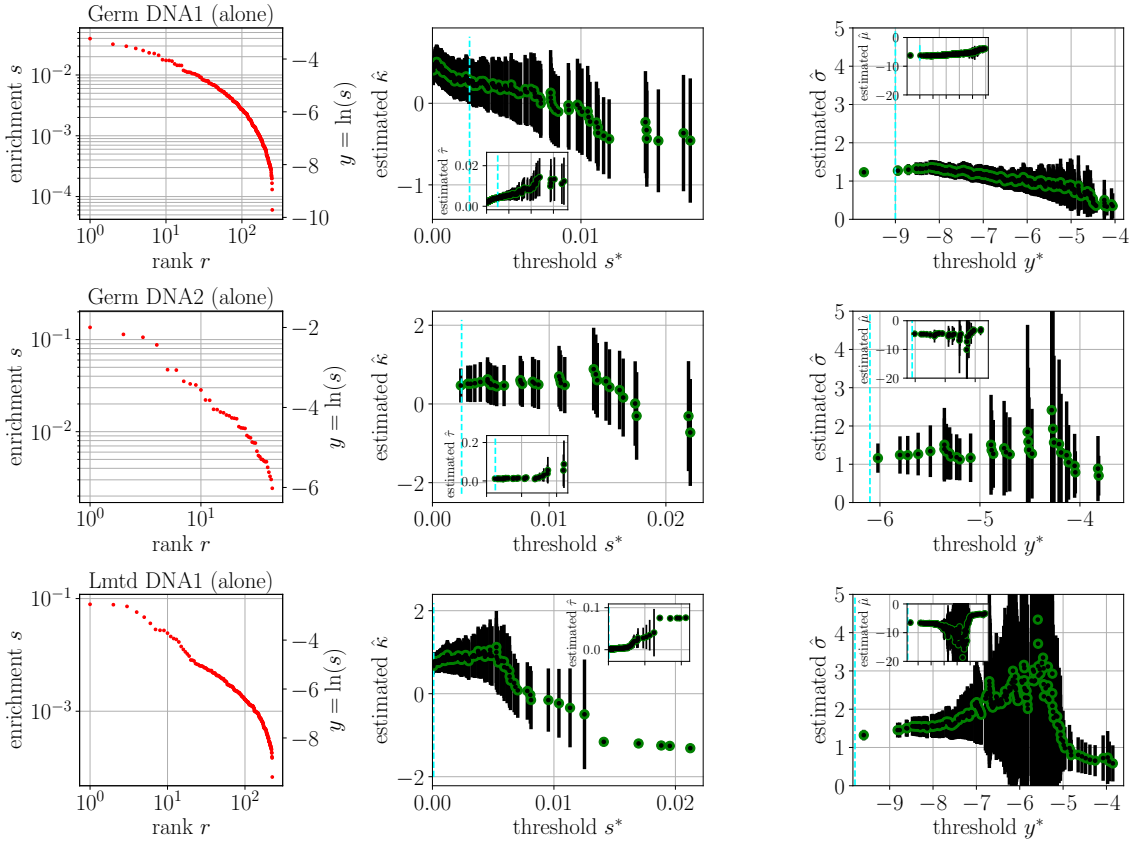


Fig. E.28: Threshold scan plots showing the values of model parameters as functions of truncation values for enrichment data from the Germline and Limited libraries selected alone. **Left** Enrichments s and log-enrichments $y = \ln(s)$ sorted in decreasing order as a function of their rank r. **Center** ML EVT model parameter $\hat{\kappa}$ ($\hat{\tau}$ in inset) as a function of s^*. **Right** ML lognormal model parameter $\hat{\sigma}$ ($\hat{\mu}$ in inset) as a function of y^*. Error bars show 90 % confidence intervals estimated from the Fisher information matrix and the Cramer-Rao bound. The vertical dashed cyan lines indicate the chosen values of s^* and y^* used for the inference.

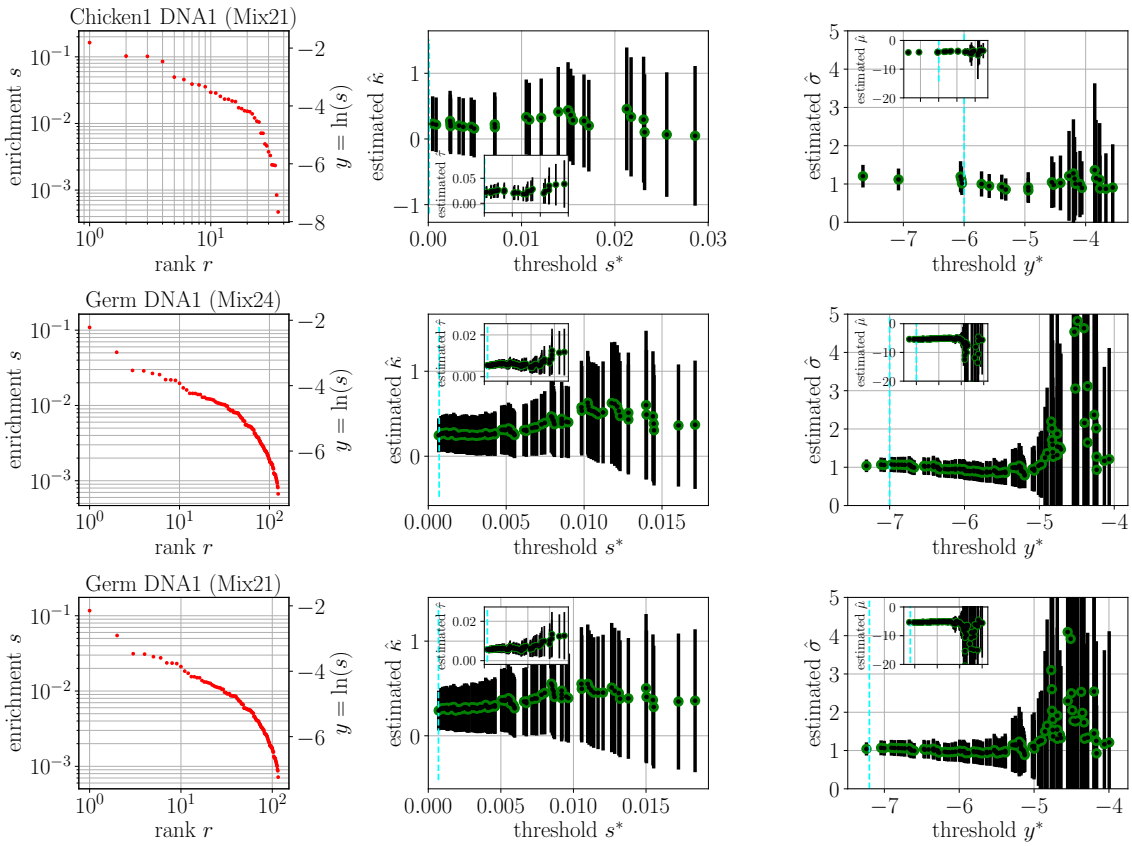


Fig. E.29: Threshold scan plots showing the values of model parameters as functions of truncation values for enrichment data from some of previously reported experiments [1]. **Left** Enrichments s and log-enrichments $y = \ln(s)$ sorted in decreasing order as a function of their rank r. **Center** ML EVT model parameter $\hat{\kappa}$ ($\hat{\tau}$ in inset) as a function of s^*. **Right** ML lognormal model parameter $\hat{\sigma}$ ($\hat{\mu}$ in inset) as a function of y^*. Error bars show 90 % confidence intervals estimated from the Fisher information matrix and the Cramer-Rao bound. The vertical dashed cyan lines indicate the chosen values of s^* and y^* used for the inference.

E.6 Enrichment histograms and model distributions $P(s)$

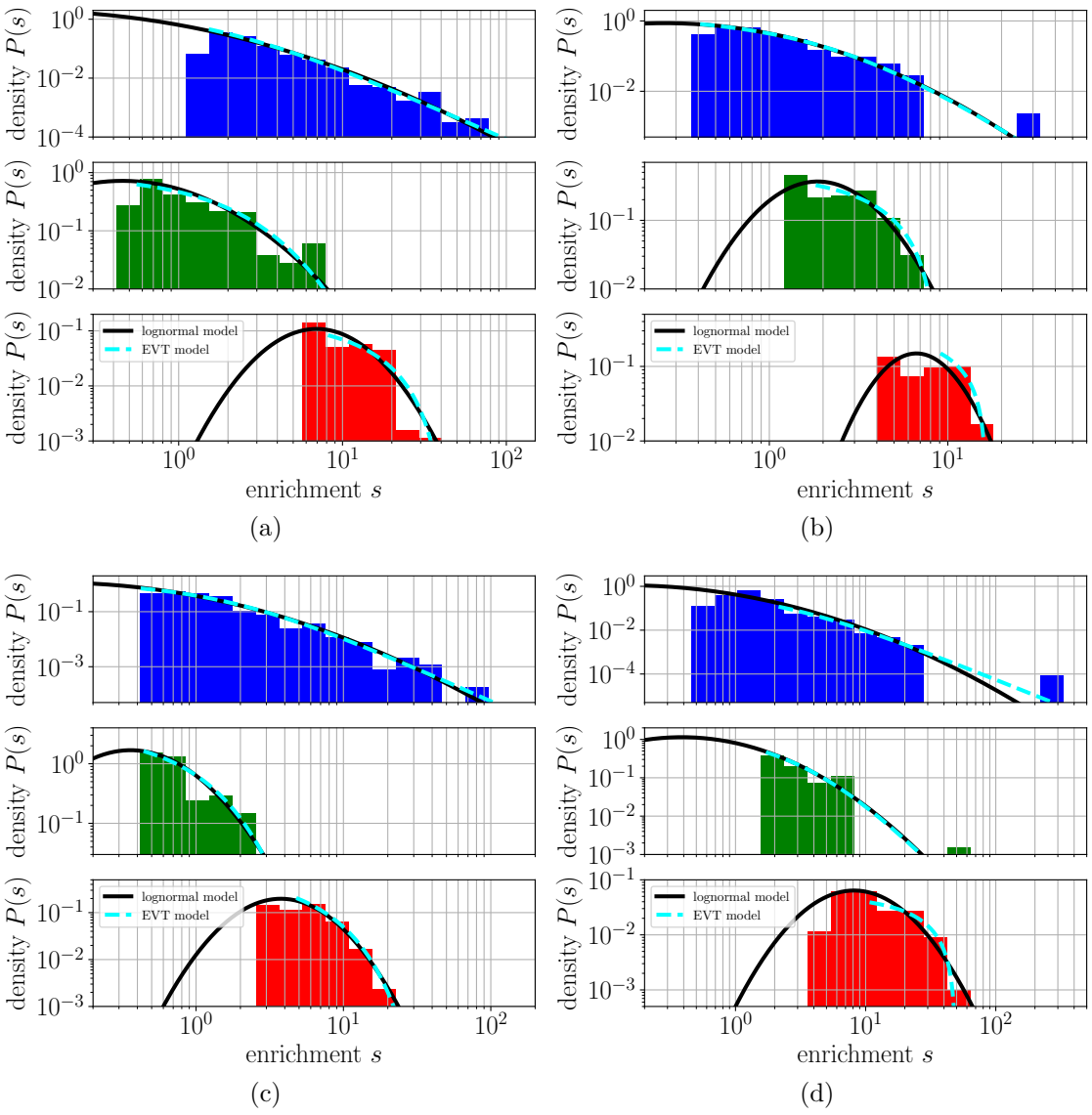


Fig. E.30: Enrichment histograms plotted with the fitted generalized Pareto and lognormal models. The histogram of enrichment values $s(x) \geq \max(s^*, \exp(y^*))$ is plotted for all three libraries of the Mix3 selections at selection round $t = 2$, $t + 1 = 3$ against the **(a)** DNA1, **(b)** DNA2, **(c)** prot1, **(d)** prot2 target. **Top, blue** Germline library, **center, green** Limited, **bottom, red** BnAb. The inferred models for $P(s)$ with the parameters from figure 4.10 are shown, **black solid** lognormal $P(s)$, **cyan dashed** generalized Pareto $P(s|s \geq s^*)$.

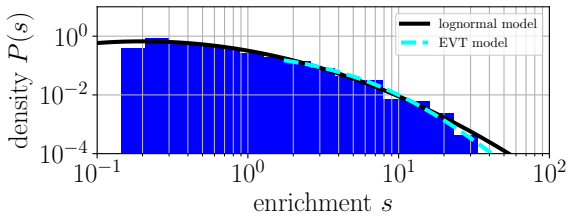


Fig. E.31: Enrichment histograms plotted with the fitted generalized Pareto and lognormal models. The histogram of enrichment values $s(x) \geq \max(s^*, \exp(y^*))$ is plotted for the Germline library selected alone against the DNA1 target at selection round $t = 1$, $t+1 = 2$. The inferred models for $P(s)$ with the parameters from figure 4.10 are shown, **black solid** lognormal $P(s)$, **cyan dashed** generalized Pareto $P(s|s \geq s^*)$.

E.7 Quality of fit: PP plots and QQ plots

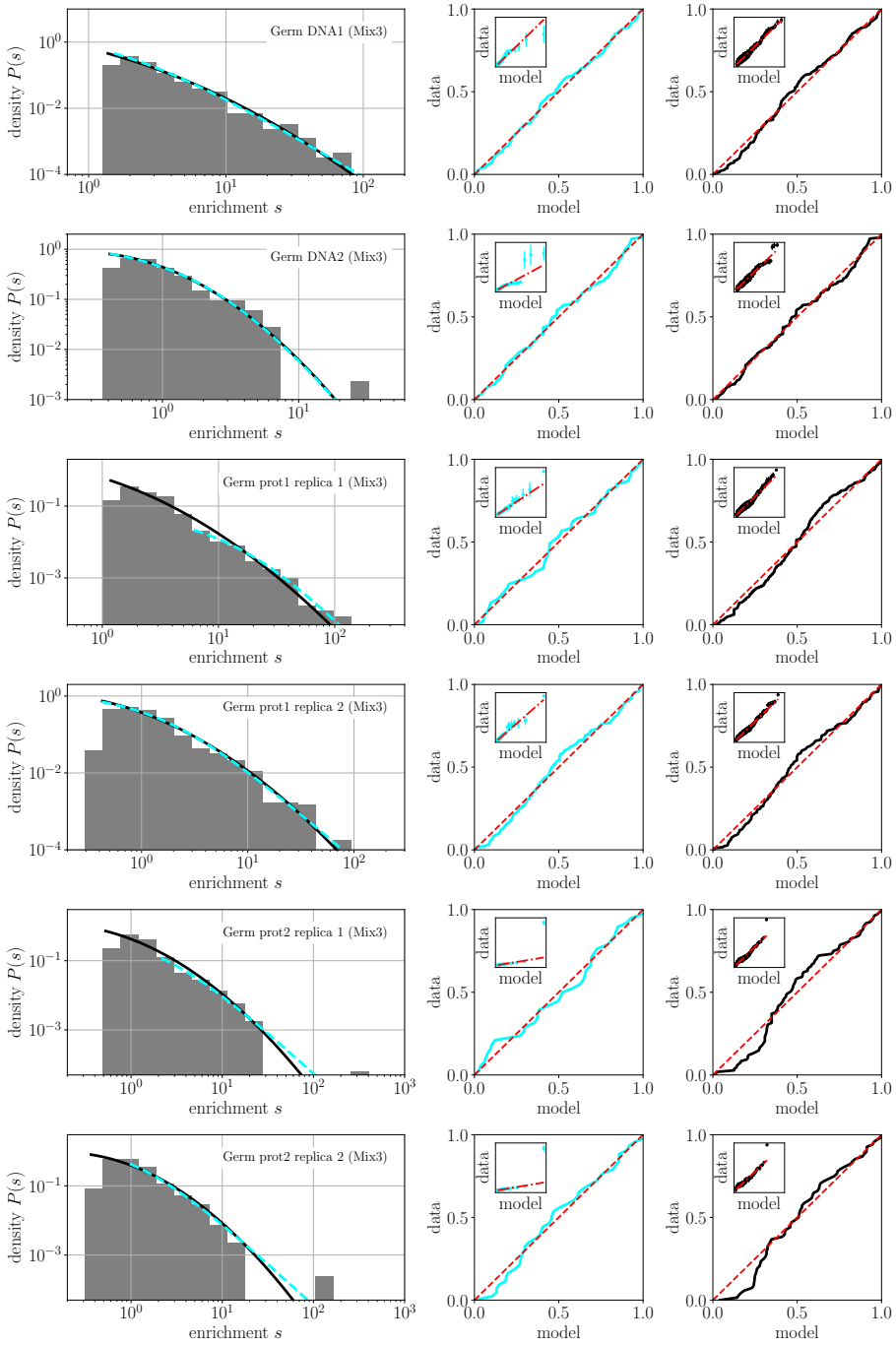


Fig. E.32: Quality of fit assessment for the generalized Pareto and lognormal distributions. Germline library selected in Mix3 against various targets (as indicated), at selection round $t = 2$, $t + 1 = 3$. **Left** Histograms of enrichment values $s(x) \geq \max(s^*, \exp(y^*))$ are plotted along with the inferred model probability densities, **black solid** lognormal $P(s)$, **cyan dashed** generalized Pareto $P(s|s \geq s^*)$. **Center** PP plot and QQ plot (inset) in cyan for the generalized Pareto distribution comparing respectively the model and empirical cumulative distribution functions, and the model and empirical enrichments. **Right** PP plot and QQ plot (inset) in black for the lognormal distribution. **Red dashed** and **red dash-dotted** lines represent the expected plots in case of perfect agreement between model and data.

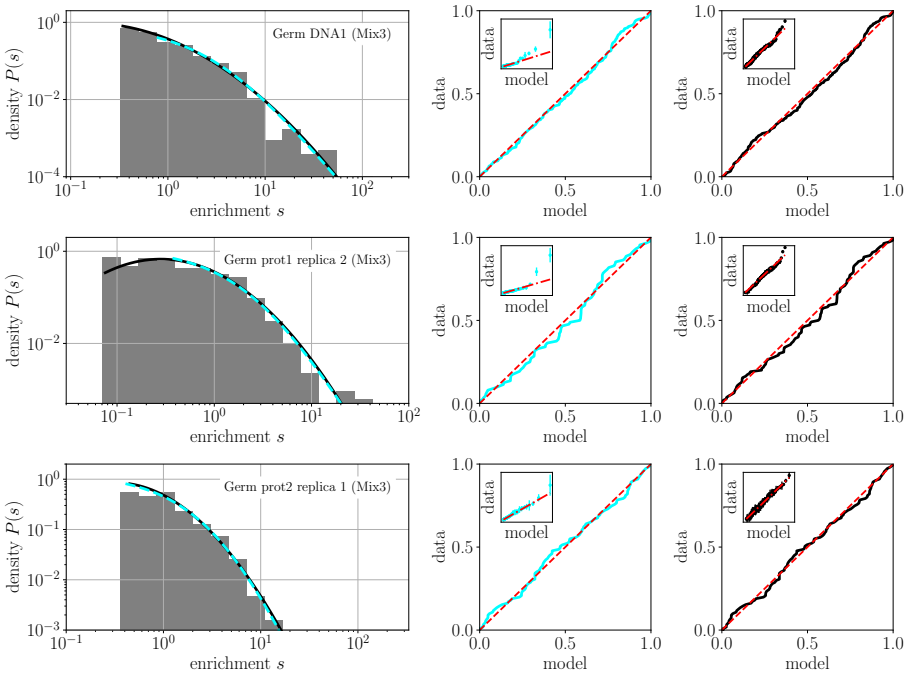


Fig. E.33: Quality of fit assessment for the generalized Pareto and lognormal distributions. Similar plots as in figure E.32 for the Germline library selected in Mix3 against various targets (as indicated), at selection round $t = 3$, $t + 1 = 4$.

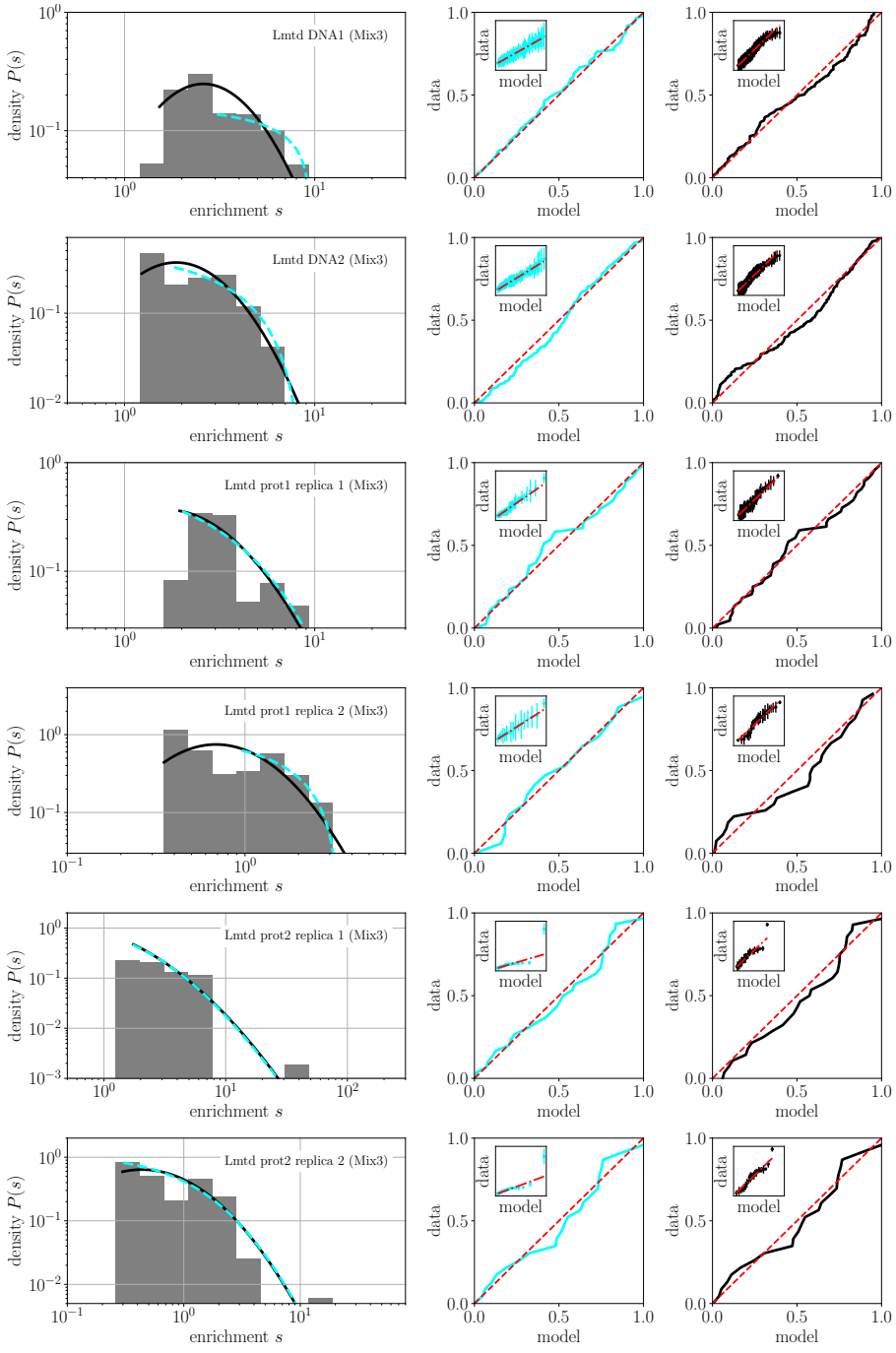


Fig. E.34: Quality of fit assessment for the generalized Pareto and lognormal distributions. Similar plots as in figure E.32 for the Limited library selected in Mix3 against various targets (as indicated), at selection round $t = 2$, $t + 1 = 3$.

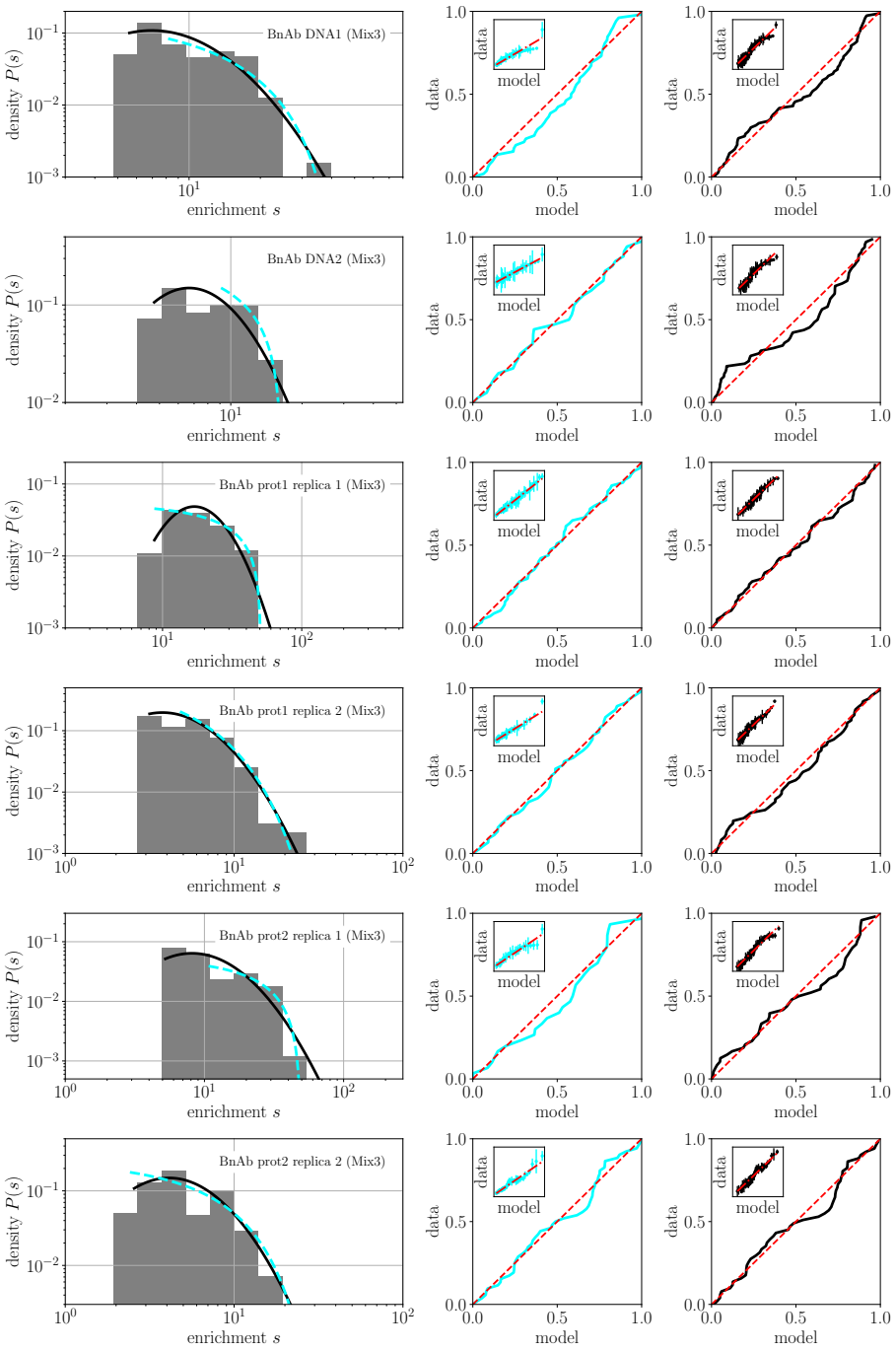


Fig. E.35: Quality of fit assessment for the generalized Pareto and lognormal distributions. Similar plots as in figure E.32 for the BnAb library selected in Mix3 against various targets (as indicated), at selection round $t = 2$, $t + 1 = 3$.

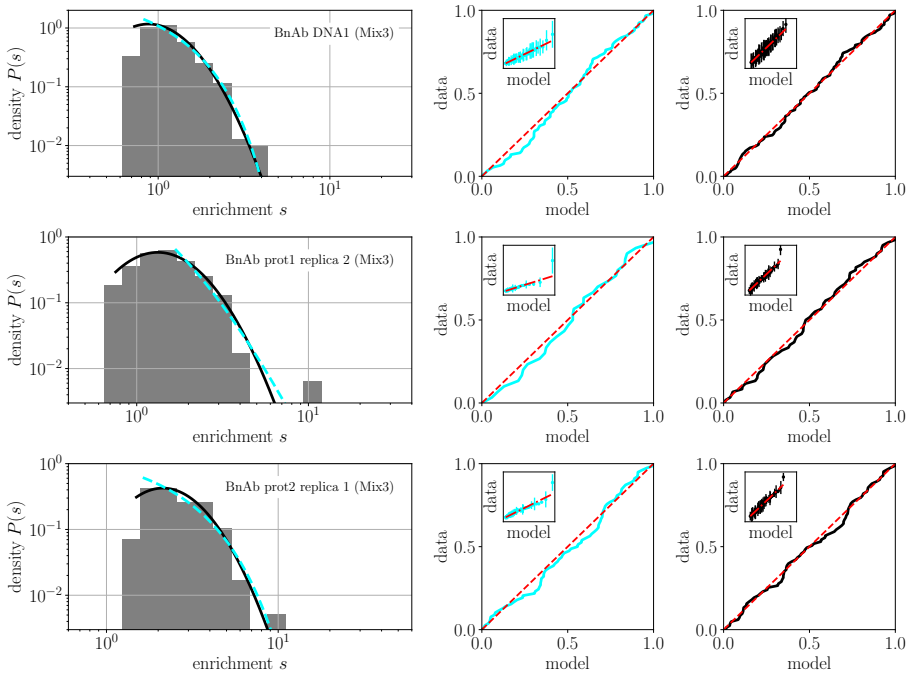


Fig. E.36: Quality of fit assessment for the generalized Pareto and lognormal distributions. Similar plots as in figure E.32 for the BnAb library selected in Mix3 against various targets (as indicated), at selection round $t = 3$, $t + 1 = 4$.

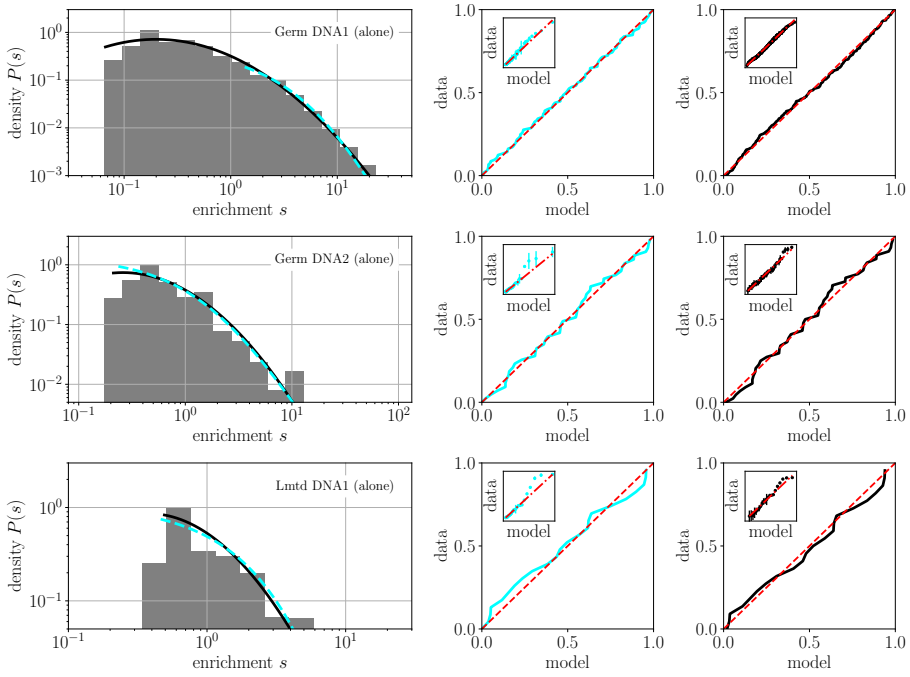


Fig. E.37: Quality of fit assessment for the generalized Pareto and lognormal distributions. Similar plots as in figure E.32 for libraries selected alone against DNA targets (as indicated), at selection round $t = 1$, $t + 1 = 2$.

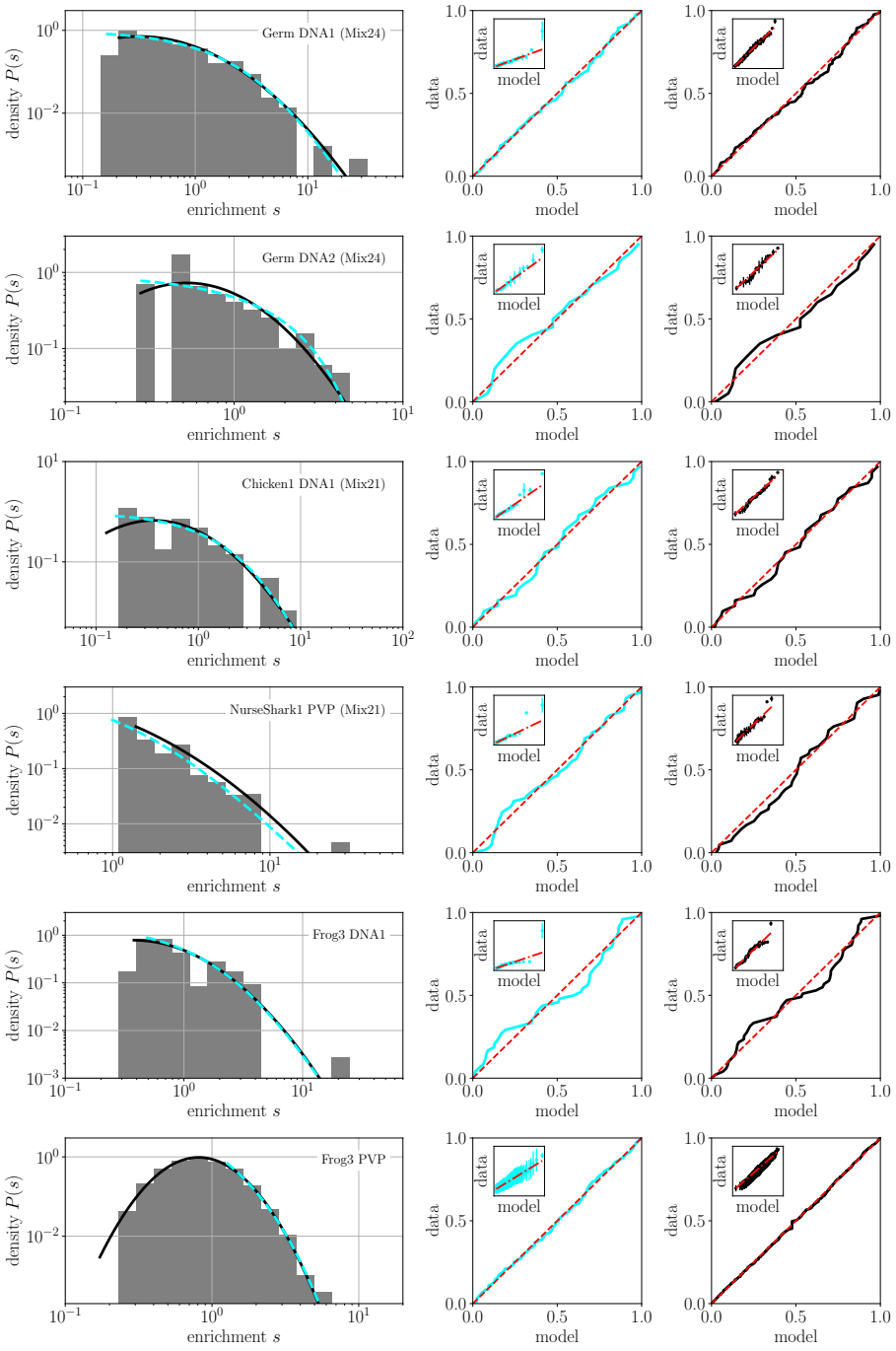


Fig. E.38: Quality of fit assessment for the generalized Pareto and lognormal distributions. Similar plots as in figure E.32 for previously reported experiments (as indicated) [1].

E.8 κ versus σ

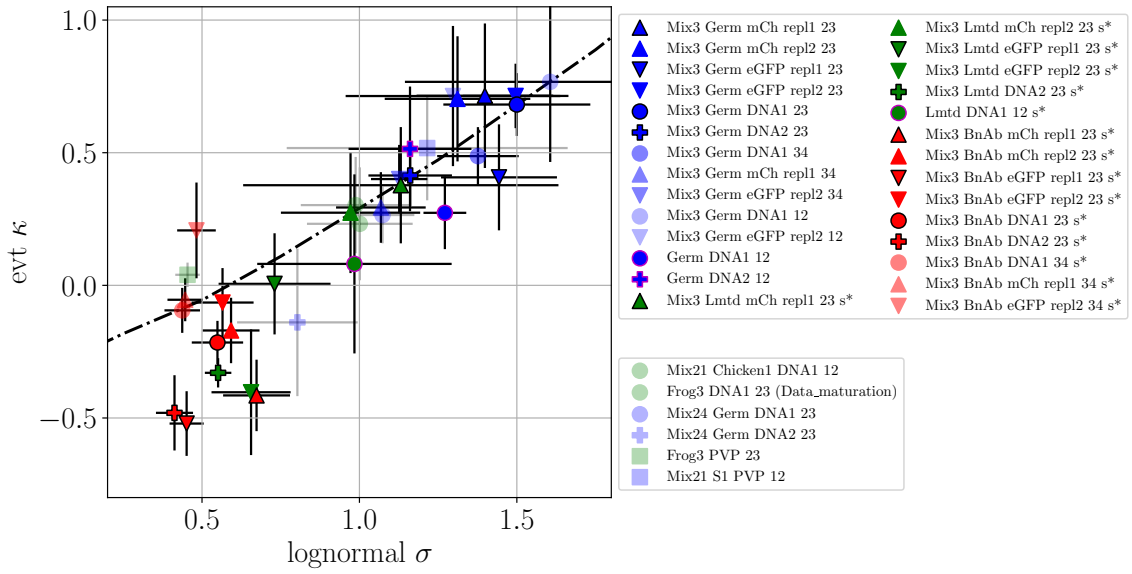


Fig. E.39: The EVT parameter $\hat{\kappa}$ is plotted against the lognormal parameter $\hat{\sigma}$ for various selection experiments reported here and elsewhere [1]. Different colors encode different libraries as indicated in the legend. Different symbols encode different targets, **circle** DNA1, **cross** DNA2, **triangle down** prot1, **triangle up** prot2. **Black encircled** and **white encircled** points are from mixed selections (different replica), **pink encircled** points are from separate selections. The precise experiments the points originate from are listed in the **legend**. The behaviour is compared to the apparent $\hat{\kappa}$ as a function of σ as found from a numerical experiment in which truncated iid lognormal numbers with given σ were fitted to a generalized Pareto distribution.

E.9 Mini library selections

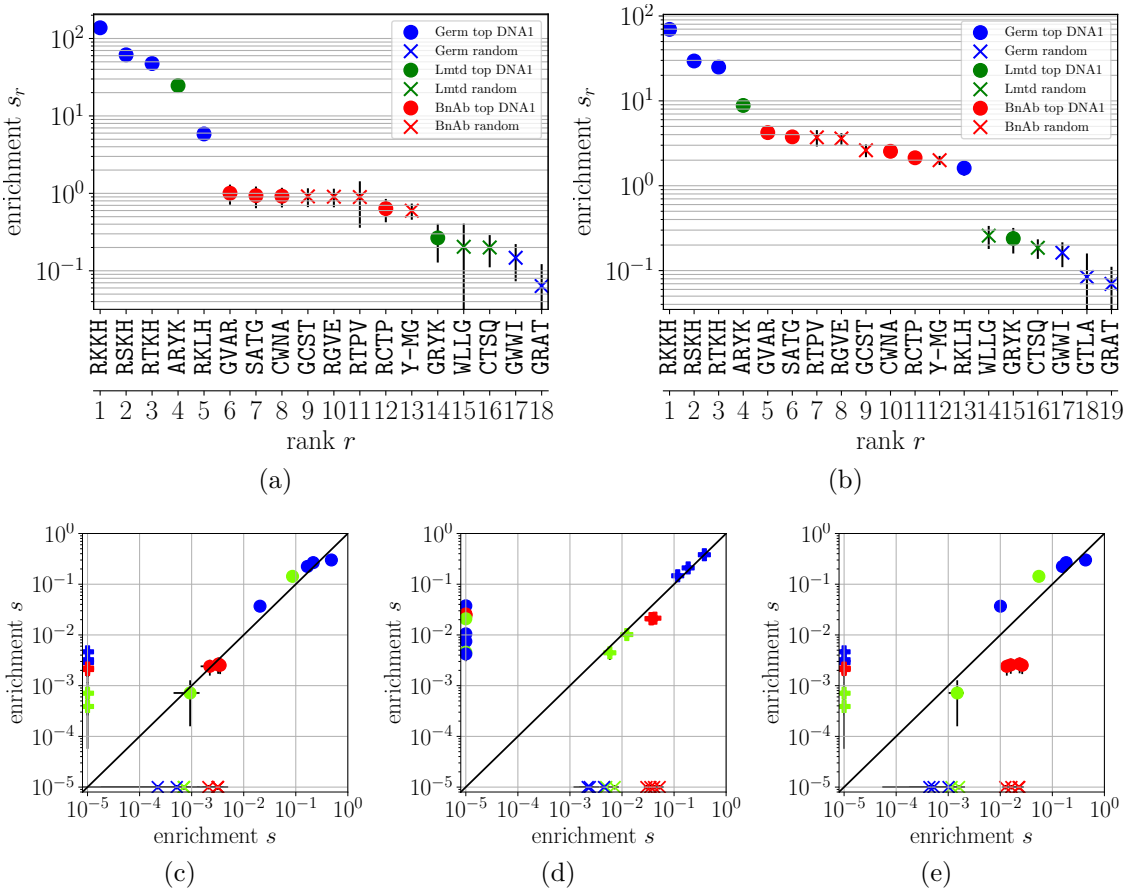


Fig. E.40: Reproducibility and correlation of mini library selections against DNA targets. Continuation of figure 4.12. High-precision enrichments from libraries with around 20 different sequences are plotted in decreasing order and the CDR3 sequences are indicated. **(a)**, **(b)** Two replica of DNA1-specific and random clones from all three libraries selected against DNA1. Correlating enrichments between experiments shown in figure 4.12: **(c)** Enrichments from 4.12(a) versus 4.12(c), **(d)** 4.12(b) versus 4.12(d), **(e)** (a) versus (b). Error bars are 20 x enlarged.

E.10 Selection dynamics

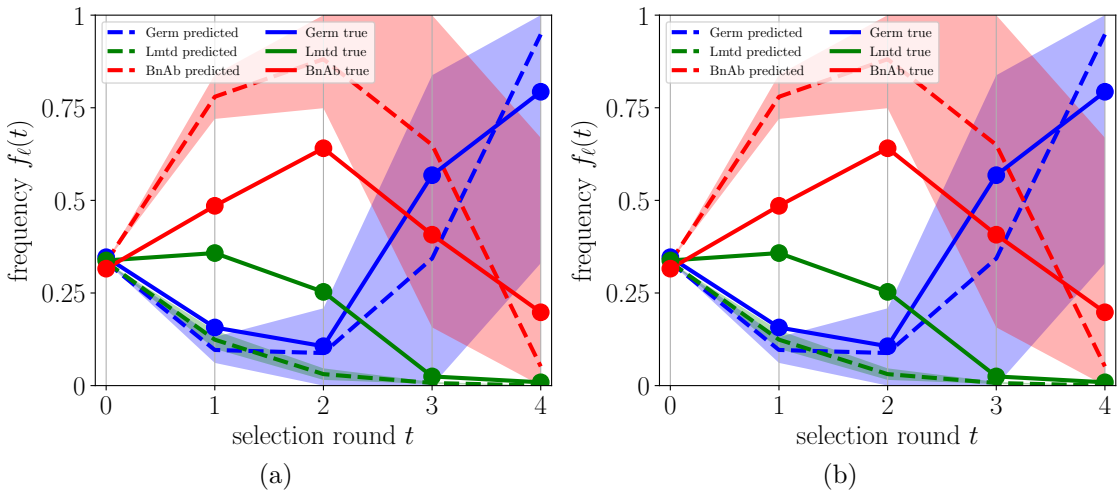


Fig. E.41: Observed *versus* predicted selection dynamics. For the Mix3 selection against the **(a)** prot1 and **(b)** prot2 targets, the frequencies for all three libraries (see legend) within the mix is shown as a function of the selection round t. The observation (**solid**) is compared to the prediction of the lognormal model (**dashed**, shaded area corresponding to 68 % confidence interval in the parameters μ and σ) under the assumption of initially (at $t = 0$) uniform distribution of sequences within the libraries.

E.11 Frequency sequence logos

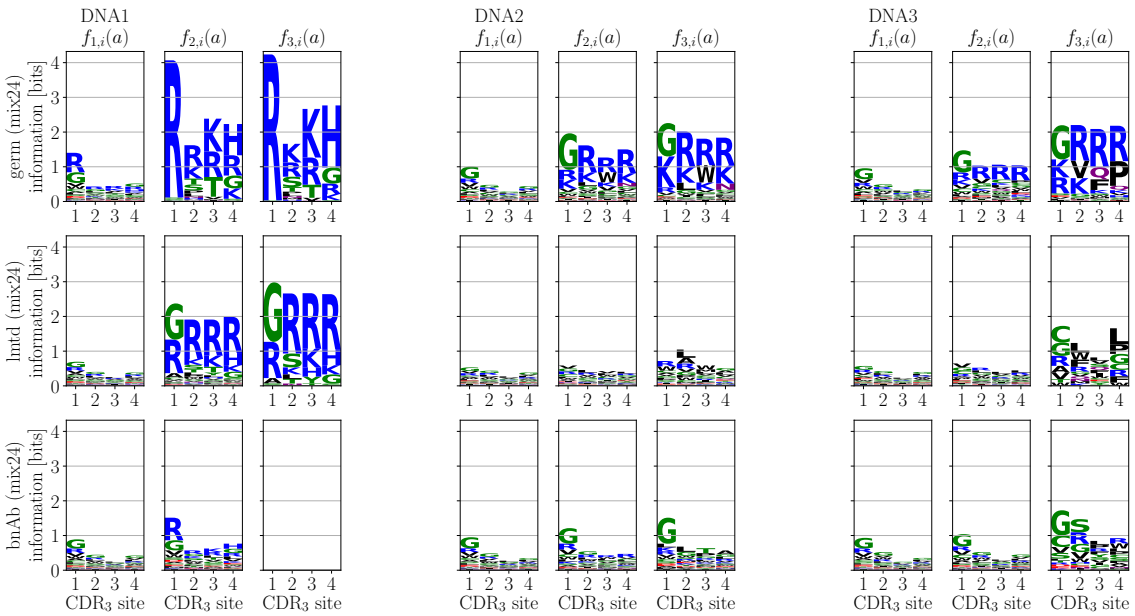


Fig. E.42: Sequence logos based on amino acid frequencies $f_{t,i}(a)$. Similar to figure 4.14. Data from all rounds $t \geq 1$ of previously reported library mix selections (Mix24) against the DNA targets [1] is shown.

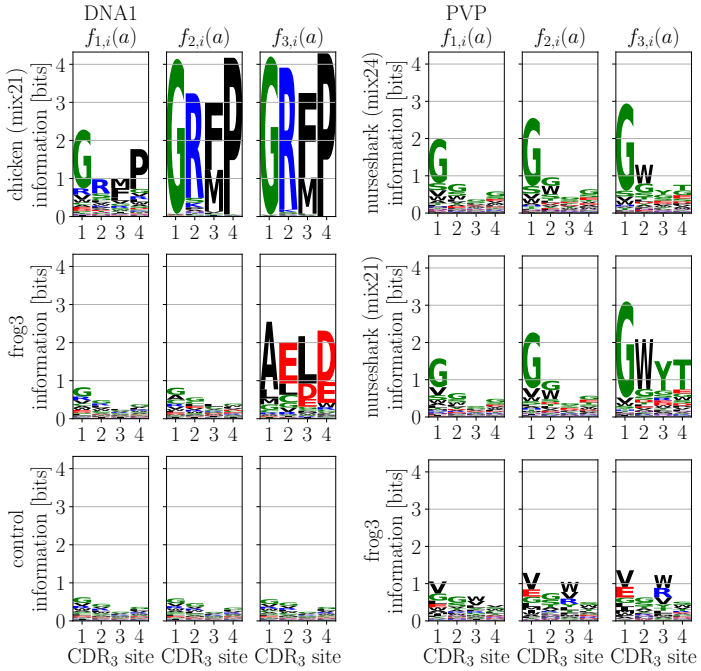


Fig. E.43: Sequence logos based on amino acid frequencies $f_{t,i}(a)$. Similar to figure 4.14. Data from all rounds $t \geq 1$ of previously reported selections against **(a)** the DNA target, **(b)** the PVP target [1] is shown.

E.12 Enrichment sequence logos (with truncation)

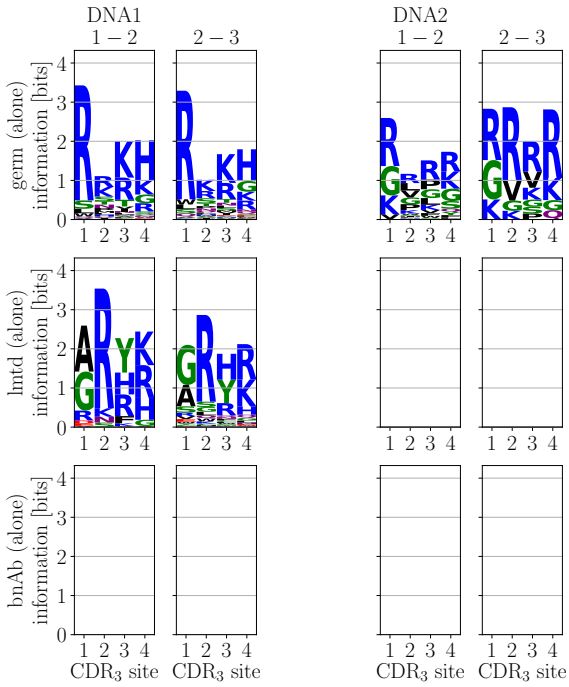


Fig. E.44: Sequence logos based on enrichments $s(x)$. Similar to figure 4.17. Data from all rounds t of separate selections against the DNA targets is shown. Logo is empty if there is no specific signal.

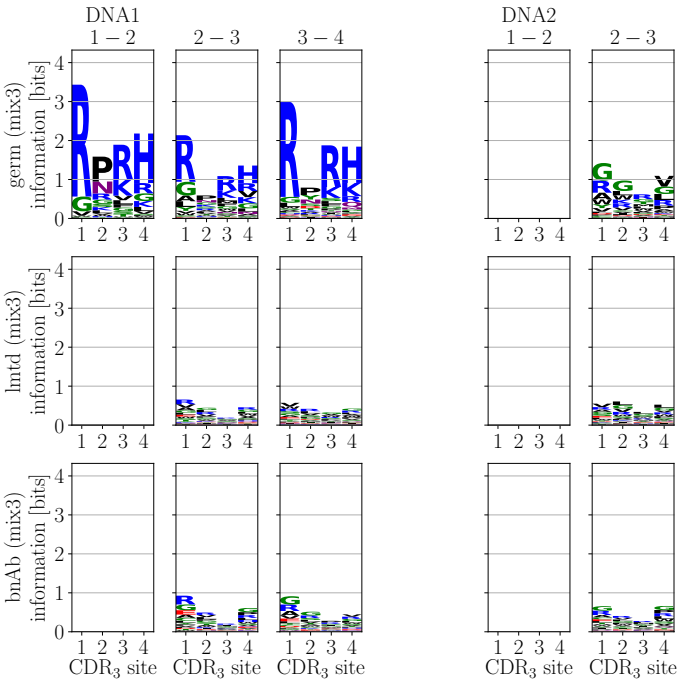


Fig. E.45: Sequence logos based on enrichments $s(x)$. Similar to figure 4.17. Data from all rounds t of library mix (Mix3) selections against the DNA targets is shown. Logo is empty if there is no specific signal.

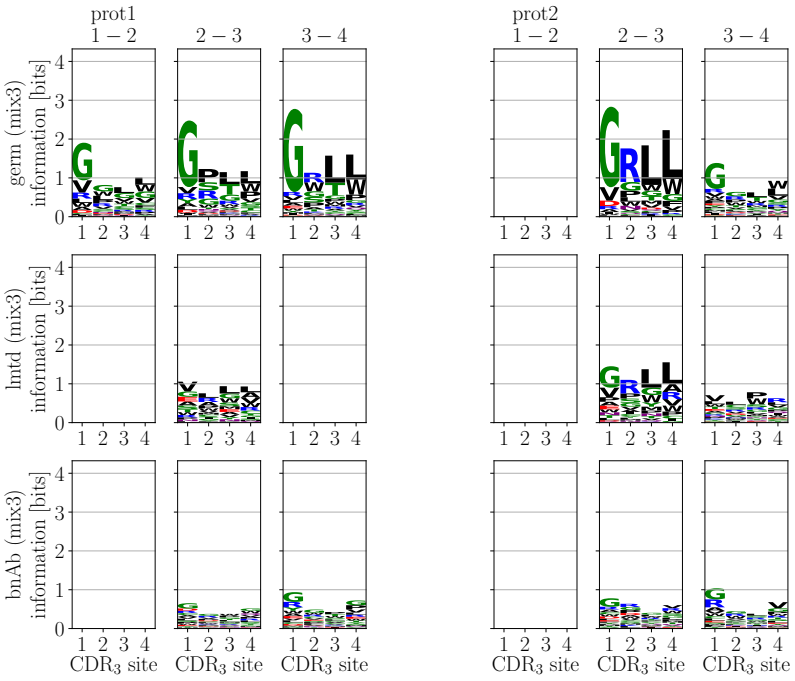


Fig. E.46: Sequence logos based on enrichments $s(x)$. Similar to figure 4.17. Data from all rounds t of library mix (Mix3) selections against the protein targets is shown. Logo is empty if there is no specific signal.

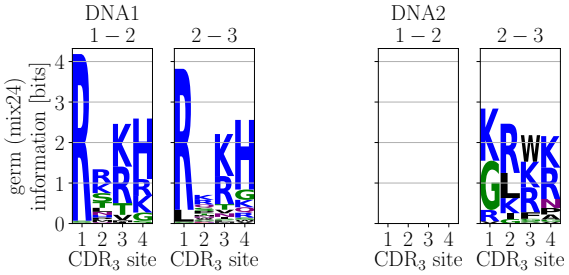


Fig. E.47: Sequence logos based on enrichments $s(x)$. Similar to figure 4.17. Data from all rounds $t \geq 1$ of previously reported library mix selections (Mix24) against the DNA targets [1] is shown. Logo is empty if there is no specific signal.

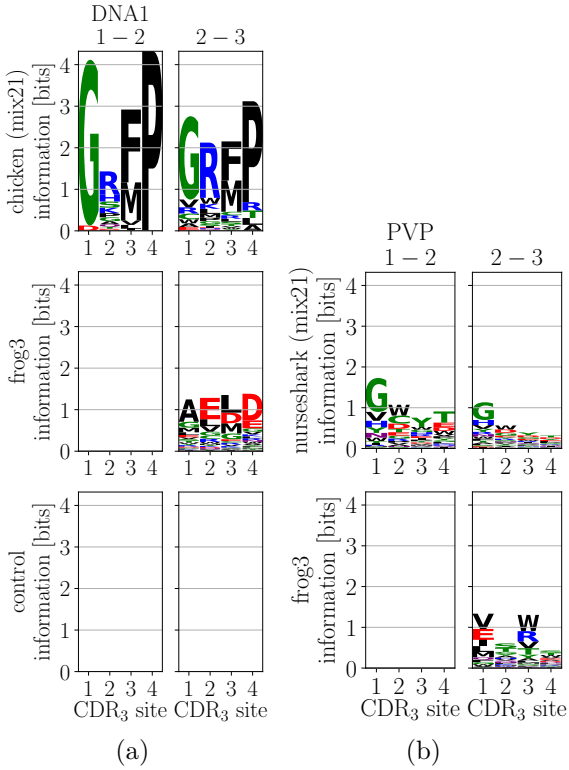


Fig. E.48: Sequence logos based on enrichments $s(x)$. Similar to figure 4.17. Data from all rounds $t \geq 1$ of previously reported selections against **(a)** the DNA target, **(b)** the PVP target [1] is shown. Logo is empty if there is no specific signal.

E.13 Enrichment sequence logos (without truncation)

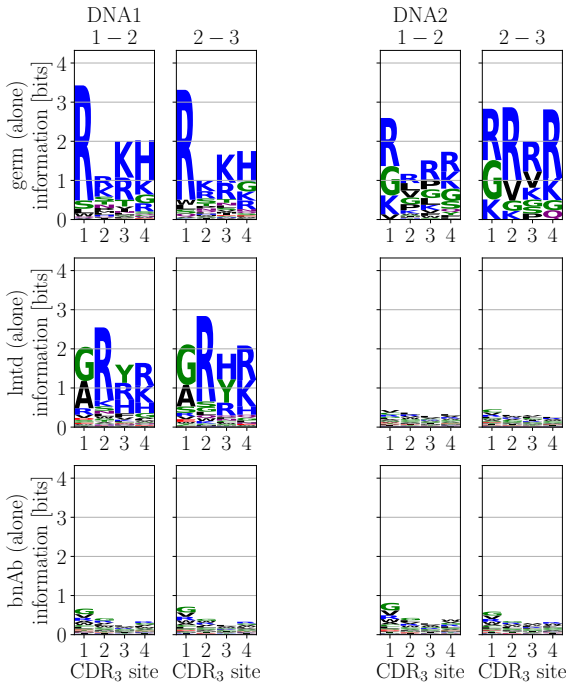


Fig. E.49: Sequence logos based on enrichments $s(x)$. Same as figure E.44, but using all available enrichment values (*i.e.* including those with $s(x) < s^*$). Data from all rounds t of separate selections against the DNA targets is shown.

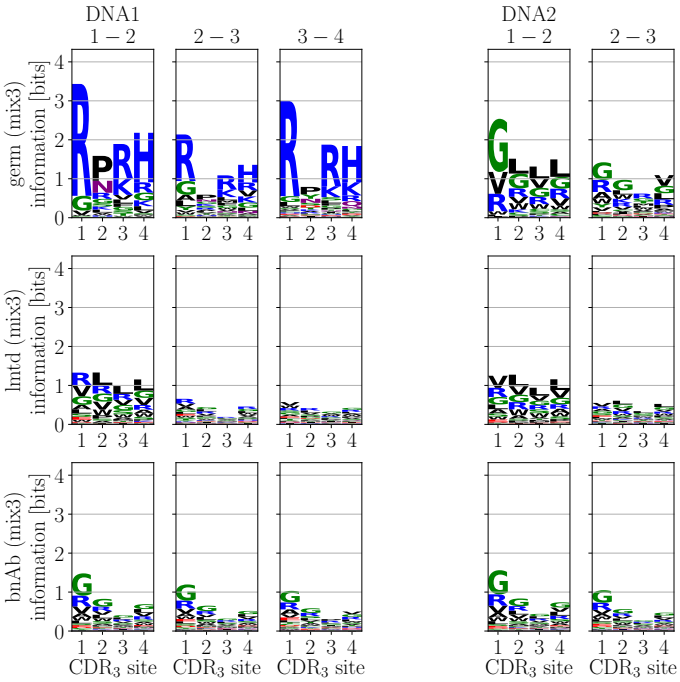


Fig. E.50: Sequence logos based on enrichments $s(x)$. Same as figure E.45, but using all available enrichment values (*i.e.* including those with $s(x) < s^*$). Data from all rounds t of library mix (Mix3) selections against the DNA targets is shown.

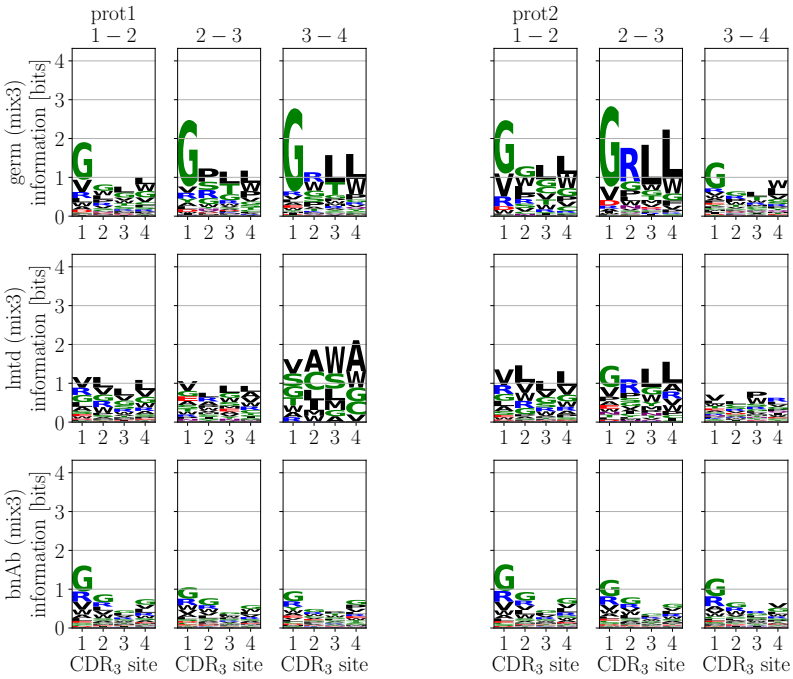


Fig. E.51: Sequence logos based on enrichments $s(x)$. Same as figure E.46, but using all available enrichment values (*i.e.* including those with $s(x) < s^*$). Data from all rounds t of library mix (Mix3) selections against the protein targets is shown.

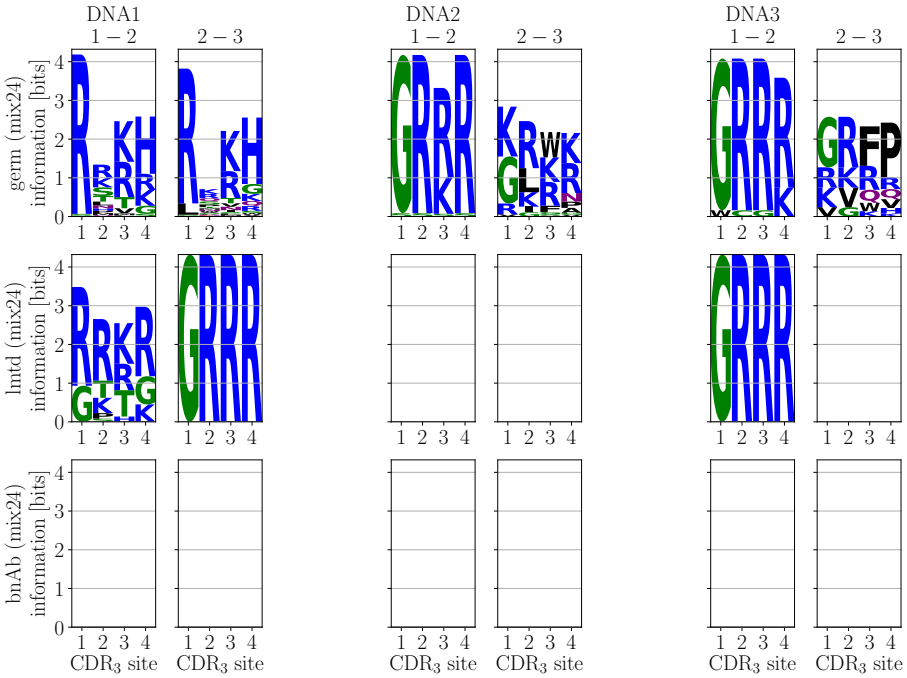


Fig. E.52: Sequence logos based on enrichments $s(x)$. Same as figure E.47, but using all available enrichment values (*i.e.* including those with $s(x) < s^*$). Data from all rounds $t \geq 1$ of previously reported library mix selections (Mix24) against the DNA targets [1] is shown.

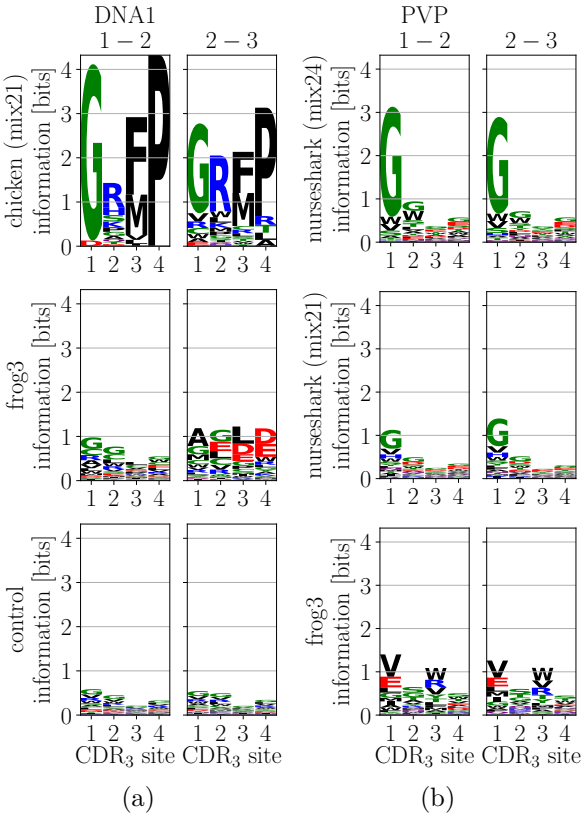


Fig. E.53: Sequence logos based on enrichments $s(x)$. Same as figure E.48, but using all available enrichment values (*i.e.* including those with $s(x) < s^*$). Data from all rounds $t \geq 1$ of previously reported selections against **(a)** the DNA target, **(b)** the PVP target [1] is shown.

Chapter F

Code

F.1 Sequencing data preprocessing

Code F.1: The class `Illumina` is used to preprocess the demultiplexed Illumina MiSeq 2 x 250 bp paired-end sequencing data. It takes as input the pairs of `.fastq` files listing respectively the measured forward and reverse sequences and quality reads of all sequencing cluster of a sample. It tests for sufficient reading quality, extracts the region of interest, and defines consensus sequences from the forward and reverse reads. For each sample, an output file containing the preprocessed sequencing reads and their respective quality reads is written. See section 3.5.3 for more details.

```
import numpy as np
from Bio.Seq import Seq

def hamdist(s1, s2):
    # Hamming distance between two sequences of same length
    assert len(s1) == len(s2), 'strings of different lengths: %g, %g' % (len(s1), len(s2))
    dist = 0
    for l1, l2 in zip(s1, s2):
        if l1 != l2: dist += 1
    return dist

class Illumina:
    # a class to represent Illumina sequencing raw data files
    def __init__(self, filename = '', nr_reads = np.inf, rev_compl = False):
        # load the fwd (and rev if any) raw sequencing data
        self.ids, self.seqs, self.quals, self.notes, tmp = [], {}, {}, {}, 0
        if filename:
            print(filename[1+filename.rfind('/'):filename.rfind('.')])
            for line in open(filename, 'r'):
                if tmp%4 == 0: self.ids.append(line[:-10])#-25])
                elif tmp%4 == 1: self.seqs[self.ids[-1]] = line[:-1]
                elif tmp%4 == 3: self.quals[self.ids[-1]] = line[:-1]
                tmp += 1
                if tmp == 4*nr_reads: break
            # reverse-complement the sequencing data
            if rev_compl:
                for ID in self.ids:
                    self.seqs[ID] = Seq(self.seqs[ID]).reverse_complement()._data
                    self.quals[ID] = self.quals[ID][::-1]
            print('number of reads:\t%g' % len(self.ids))

    def mk(self, ID, data = '', note = ''):
        # add a read to a sequencing dataset
        #assert ID not in self.ids, 'a sequence with this id already exists'
        '''if ID in self.ids:
            self.seqs[ID], self.quals[ID] = [self.seqs[ID]], [self.quals[ID]]
```

```
            if type(data) is list:
                if type(data[0]) is str:
                    self.seqs[ID].append(data[0])
                    self.quals[ID].append(data[1])
                    for dat in data:
                        self.seqs[ID].append(dat.seqs[ID])
                        self.quals[ID].append(dat.quals[ID])
            elif type(data) is Illumina: self.seqs[ID], self.quals[ID] = data.seqs[ID], data.quals[ID]
        else:'''
        self.ids.append(ID)
        if type(data) is list:
            if type(data[0]) is str: self.seqs[ID], self.quals[ID] = data[0], data[1]
            else: self.seqs[ID], self.quals[ID] = [dat.seqs[ID] for dat in data], [dat.quals[ID] for dat in data]
        else: self.seqs[ID], self.quals[ID] = data.seqs[ID], data.quals[ID]
        if note: self.notes[ID] = note

    def rm(self, ID):
        # remove a read from a sequencing dataset
        if ID in self.ids:
            self.ids.remove(ID)
            del self.seqs[ID], self.quals[ID]
            if ID in self.notes: del self.notes[ID]

    def crop(self, p1, p2, tol = 0, out = 10000):
        # extract region of interest between two primer sequences (if found; including the primers)
        cnt = 1
        for ID in self.ids:
            if cnt%out == 0: print('cropping: %g/%g' % (cnt, len(self.ids)))
            seq = self.seqs[ID]
            xcut1, xcut2 = seq.find(p1), seq.find(p2)
            if xcut1 == -1:
                xscan = [hamdist(p1, seq[x:x+len(p1)]) for x in range(len(seq)-len(p1))]
                if min(xscan) in np.unique(xscan) and min(xscan) <= tol: xcut1 = np.argmin(xscan)
            if xcut2 == -1:
                xscan = [hamdist(p2, seq[x:x+len(p2)]) for x in range(len(seq)-len(p2))]
                if min(xscan) in np.unique(xscan) and min(xscan) <= tol: xcut2 = np.argmin(xscan)
            if xcut1 >= 0 and xcut2 > xcut1+len(p1):
                self.seqs[ID] = seq[xcut1:xcut2+len(p2)]#seq[xcut1+len(p1):xcut2]
                self.quals[ID] = self.quals[ID][xcut1:xcut2+len(p2)]#self.quals[ID][xcut1+len(p1):xcut2]
            cnt += 1

def cleaning(L, fout, fwd, rev = Illumina(), out = 10000):
    # cleaning the raw data from sequences with wrong length, bad quality read and absence of restriction sites
    clean, trash = Illumina(), Illumina()
    # check for correct sequence length
    cnt, trash1 = 1, 0
    if rev:
        for ID in fwd.ids:
            if cnt%out == 0: print('cleaning step 1: %g/%g' % (cnt, len(fwd.ids)))
            l_fwd, l_rev = len(fwd.seqs[ID]), len(rev.seqs[ID])
            if l_fwd == L and l_rev == L:
                if fwd.seqs[ID] == rev.seqs[ID]:
                    clean.mk(ID, fwd)
                else:
                    cseq, cqual = [], []
                    for x in range(L):
                        if ord(fwd.quals[ID][x]) >= ord(rev.quals[ID][x]):
                            cseq.append(fwd.seqs[ID][x])
                            cqual.append(fwd.quals[ID][x])
                        else:
                            cseq.append(rev.seqs[ID][x])
                            cqual.append(rev.quals[ID][x])
                    clean.mk(ID, [''.join(cseq), ''.join(cqual)])
            elif l_fwd == L:
                clean.mk(ID, fwd)
                trash.mk(ID, rev, 'incorrect length: %g (only rev)' % l_rev)
            elif l_rev == L:
                trash.mk(ID, fwd, 'incorrect length: %g (only fwd)' % l_fwd)
                clean.mk(ID, rev)
            else:
                trash.mk(ID, [fwd, rev], 'incorrect lengths: %g, %g (fwd and rev)' % (l_fwd, l_rev))
                trash1 += 1
            cnt += 1
    else:
        for ID in fwd.ids:
            if cnt%out == 0: print('cleaning step 1: %g/%g' % (cnt, len(fwd.ids)))
            l = len(fwd.seqs[ID])
            if l == L: clean.mk(ID, fwd)
            else:
                trash.mk(ID, fwd, 'incorrect length: %g' % l)
                trash1 += 1
            cnt += 1
    cnt, cnttot, trash2, trash3, trash4 = 1, len(clean.ids), 0, 0, 0
```

```
5183        for ID in clean.ids:
5184            if cnt%out == 0: print('cleaning step 2: %g/%g' % (cnt, cnttot))
5185            # check for sufficient mean quality read
5186            avg_qual = np.mean([ord(clean.quals[ID][x]) for x in range(L)])
5187            tag_qual = (avg_qual > (33+25))
5188            # check for presence of restriction sites
5189            tag_restr = (hamdist('TGTGCGCGC', clean.seqs[ID][107:116]) <= 4 and hamdist('TTCGACTAC', clean.seqs[ID][128:137]) <= 4)
5190            if not tag_qual and not tag_restr:
5191                trash.mk(ID, clean, 'bad average quality read & unrecognizable restriction sites')
5192                clean.rm(ID)
5193                trash2 += 1
5194            elif not tag_qual:
5195                trash.mk(ID, clean, 'bad average quality read: %.2f' % avg_qual)
5196                clean.rm(ID)
5197                trash3 += 1
5198            elif not tag_restr:
5199                trash.mk(ID, clean, 'unrecognizable restriction sites')
5200                clean.rm(ID)
5201                trash4 += 1
5202            cnt += 1
5203        print('-----------------------------------')
5204        print('total number of reads:\t%g' % len(fwd.ids))
5205        print('number of clean reads:\t%g' % len(clean.ids))
5206        print('number of trash reads:\t%g' % sum([trash1, trash2, trash3, trash4]))
5207        print('-----------------------------------')
5208        print('incorrect length: %g' % trash1)
5209        print('bad average quality read: %g' % trash3)
5210        print('unrecognizable restriction sites: %g' % trash4)
5211        print('bad average quality read & unrecognizable restriction sites: %g' % trash2)
5212        print('-----------------------------------')
5213        fout.write('-----------------------------------\n')
5214        fout.write('total number of reads:\t%g\n' % len(fwd.ids))
5215        fout.write('number of clean reads:\t%g\n' % len(clean.ids))
5216        fout.write('number of trash reads:\t%g\n' % sum([trash1, trash2, trash3, trash4]))
5217        fout.write('-----------------------------------\n')
5218        fout.write('incorrect length: %g\n' % trash1)
5219        fout.write('bad average quality read: %g\n' % trash3)
5220        fout.write('unrecognizable restriction sites: %g\n' % trash4)
5221        fout.write('bad average quality read & unrecognizable restriction sites: %g\n' % trash2)
5222        fout.write('-----------------------------------\n')
5223        return [clean, trash]
5224
5225    def writefile(res, filename):
5226        fout = open(filename + '.txt', 'w')
5227        if type(res) is list:
5228            tag1 = False
5229            for j in range(len(res)):
5230                if tag1: fout.write('\n')
5231                tag2 = False
5232                for k in range(len(res[0])):
5233                    if tag2: fout.write('\t')
5234                    fout.write(str(res[j][k]))
5235                    tag2 = True
5236                tag1 = True
5237        elif type(res) is Illumina:
5238            tag1 = False
5239            for ID in res.ids:
5240                if tag1: fout.write('\n')
5241                if type(res.seqs[ID]) is list:
5242                    tag2 = False
5243                    for j in range(len(res.seqs[ID])):
5244                        if tag2: fout.write('\n')
5245                        fout.write(res.seqs[ID][j] + '\n' + res.quals[ID][j])
5246                        tag2 = True
5247                else:
5248                    fout.write(res.seqs[ID] + '\n' + res.quals[ID])
5249                if ID in res.notes: fout.write('\n' + res.notes[ID])
5250                fout.write('\n&')
5251                tag1 = True
5252        fout.close()
```

Code F.2: For a given Illumina run, the following code loops over all samples and makes use of the `Illumina` class to preprocess the raw data. It then counts the number of occurrences in a sample of each unique sequence defined by the scaffold identity and the CDR3 sequence. One output file is written per sample containing the identity and number of counts of all unique sequences that were observed at least once organized in three columns and sorted in decreasing order: scaffold

identity, CDR3 nucleotide sequence, number of counts. See section 3.5.3 for more details.

```
import time
import numpy as np
import os
from Bio.Seq import Seq
import illumina_TOOLS as illu

# primers to extract region of interest from sequencing reads
miseq_human_fwd = Seq('GCTCGAGACGGTAACCAGG').reverse_complement()._data # 5' -> 3'
miseq_human_rev = 'ACAACCCGTCTCTTAAGTCTCGT' # 5' -> 3'

# framework sequence references
fwk_refs = \
{'germ': 'ACAACCCGTCTCTTAAGTCTCGTGTTACCATCTCTGTTGACACCTCTAAAAACCAGTTCTCTCTGAAACTGTCTTCTGTTACTGCGGCGGACACTGCGGTTTACTACTGTGCGCGC',
 'lmtd': 'ACAACCCGTCTCTTAAGTCTCGTGTTACCATCTCTATCGACACCTCTAAAAACCACTTCTCTCTGCGTCTGATCTCTGTTACTGCGGCGGACACTGCGGTTTACCACTGTGCGCGC',
 'bnAb': 'ACAACCCGTCTCTTAAGTCTCGTCTGACCCTGGCGCTGGACACCCCGAAAAACCTGGTTTTCCTGAAACTGAACTCTGTTACTGCGGCGGACACCGCGACCTACTACTGTGCGCGC'}

# location of raw sequencing data and output files
loc_in, loc_out = 'illumina_2018_12/fastq/', 'illumina__clean/clean/'

# list of all raw sequencing datafiles
files_fwd, files_rev = [], []
for file in sorted(os.listdir(loc_in)):
    if file.endswith('_L001_R1_001.fastq'): files_fwd.append(file)
    if file.endswith('_L001_R2_001.fastq'): files_rev.append(file)
#files_fwd, files_rev = ['CTRL_S10_L001_R1_001.fastq'], ['CTRL_S10_L001_R2_001.fastq']

# number of reads to take into account from each file; screen output produced during raw data processing
nr_reads, out = np.inf, 100000

# log file
lout = open('1_cleaning_counting_log.txt', 'w')

# measure processing time
T0 = time.time()

# data processing
for f_fwd, f_rev in zip(files_fwd, files_rev):
    lout.write(f_fwd + ' & ' + f_rev + '\n')

    # reading in the raw sequencing data from files (forward reads)
    data_fwd = illu.Illumina(loc_in + f_fwd, nr_reads, True)
    data_fwd.crop(miseq_human_rev, miseq_human_fwd, 4, out)
    # reading in the raw sequencing data from files (reverse reads)
    data_rev = illu.Illumina(loc_in + f_rev, nr_reads, False)
    data_rev.crop(miseq_human_rev, miseq_human_fwd, 4, out)

    # cleaning the raw sequencing data
    clean, trash = illu.cleaning(170, lout, data_fwd, data_rev, out)

    # identifying framework and CDR3
    cntnt, cntaa, cnt = {}, {}, 1
    for ID in clean.ids:
        if cnt%out == 0: print('counting: %g/%g' % (cnt, len(clean.ids)))
        cnt += 1
        fwk_read = clean.seqs[ID][:116]
        cdr3_read = clean.seqs[ID][116:128]

        # framework
        fwk_ds = {lib: illu.hamdist(fwk_refs[lib], fwk_read) for lib in fwk_refs}
        ref1, ref1_d = min(fwk_ds, key=fwk_ds.get), min(fwk_ds.values())
        fwk_ds[ref1] = np.inf
        ref2, ref2_d = min(fwk_ds, key=fwk_ds.get), min(fwk_ds.values())
        if ref1_d <= 7 and ref2_d - ref1_d >= 3: fwk = ref1
        else: fwk = '????'

        # CDR3
        cdr3 = cdr3_read

        # add read to result dicts
        IDnt, IDaa = fwk + '\t' + cdr3, fwk + '\t' + Seq(cdr3).translate()._data
        if IDnt in cntnt: cntnt[IDnt] += 1
        else: cntnt[IDnt] = 1
        if IDaa in cntaa: cntaa[IDaa] += 1
        else: cntaa[IDaa] = 1

    # sort result dicts
    print('sorting')
    cntnt = sorted(cntnt.items(), key=lambda x: x[1], reverse = True)
    cntaa = sorted(cntaa.items(), key=lambda x: x[1], reverse = True)

    # output files
    foutname = f_fwd[1+f_fwd.rfind('/'):f_fwd.rfind('.')-12]
```

```
    illu.writefile(clean, loc_out + foutname + '_clean')
    illu.writefile(trash, loc_out + foutname + '_trash')
    illu.writefile(cntnt, loc_out + foutname + '_counted_nt')
    illu.writefile(cntaa, loc_out + foutname + '_counted_aa')
    print('time elapsed: %.2f s' % (time.time() - T0))
    lout.write('time elapsed: %.2f s\n\n' % (time.time() - T0))

lout.close()
```

F.2 Lognormal and generalized Pareto model parameter inference

Code F.3: These functions were used to fit truncated lognormal and generalized Pareto distributions to a histogram of enrichments. The input is a list of enrichments that was computed from any two consecutive rounds of selections. The function `do_threshold_scans` plots the fit parameter as a function of the lower enrichment threshold s^*. The function `do_fits` performs the parameter inference given a value for s^* and assesses the quality of fit by QQ- and PP-plots. The code for the generalized Pareto case was written by Sébastien Boyer and adapted by me. The idea was extended to the lognormal case and the code was written by me. See section 4.2 for more details.

```
import numpy as np
np.random.seed(1)
import scipy
import scipy.stats as ss
from scipy.special import erf, erfc, erfinv
import numdifftools as nd

#%matplotlib inline
import matplotlib.pyplot as plt

import warnings
warnings.filterwarnings('ignore')

''' Functions for the lognormal model '''

def log_likelihood_fct_logn(para, data_sorted):
    ''' Log-likelihood function for the lognormal model '''
    mu, sigma, estar = para[0], para[1], min(data_sorted)
    result = -len(data_sorted[:-1]) * ( np.log(sigma) + np.log( erfc((estar-mu)/(np.sqrt(2)*sigma)) ) )
    result -= sum([(x-mu)**2/(2*sigma**2) for x in data_sorted[:-1]])
    return -result

def info_mat_logn(para, data_sorted):
    ''' Fisher information matrix for the lognormal model '''
    return np.linalg.inv(nd.Hessian(log_likelihood_fct_logn)(para, data_sorted)).diagonal()

def threshold_scan_logn(en_sorted, min_points, max_points, para=[-7.,1.]):
    ''' Fit to the lognormal model for different values of threshold
    min_points sets the minimum number of points that are kept '''
    para_list, err_list = list(), list()
    for i in range(min(max_points, len(en_sorted)), min_points, -1):
        para_hat = scipy.optimize.fmin(log_likelihood_fct_logn, para,\
                                       args=(en_sorted[:i],), disp=False, maxiter=1000)
        para_list.append(para_hat)
        err = info_mat_logn(para_hat, en_sorted[:i])
        err_list.append([1.96*np.sqrt(err[0]), 1.96*np.sqrt(err[1])])
    return para_list, err_list

def cumF_logn(e, mu, sigma, estar):
    ''' Cumulative distribution function for the lognormal model '''
    e = (e-mu)/(np.sqrt(2.)*sigma)
    estar = (estar-mu)/(np.sqrt(2.)*sigma)
    return ( erf(e) - erf(estar) ) / ( 1. - erf(estar) )

def invcumF_logn(y, mu, sigma, estar):
    ''' Inverse cumulative distribution function for the lognormal model '''
    estar2 = (estar-mu)/(np.sqrt(2.)*sigma)
    return mu + np.sqrt(2.)*sigma*erfinv( (1.-erf(estar2))*y + erf(estar2) ) - estar
```

```
5393
5394  def log_likelihood_fct_logn2(para, data_sorted, mode):
5395      ''' Log-likelihood function for the lognormal model (when the mode is fixed) '''
5396      mu, sigma, estar = mode+para[0]**2, para[0], min(data_sorted)
5397      result = -len(data_sorted[:-1]) * ( np.log(sigma) + np.log( erfc((estar-mu)/(np.sqrt(2)*sigma)) ) )
5398      result -= sum([(x-mu)**2/(2*sigma**2) for x in data_sorted[:-1]])
5399      return -result
5400
5401  def info_mat_logn2(para, data_sorted):
5402      ''' Fisher information matrix for the lognormal model (when the mode is fixed) '''
5403      return np.linalg.inv(nd.Hessian(log_likelihood_fct_logn2)(para, data_sorted)).diagonal()
5404
5405
5406  ''' Functions for the EVT model '''
5407
5408  def log_likelihood_fct_exp(para, data_sorted):
5409      ''' Log-likelihood function for the exponential model '''
5410      tau, mu = para[0], min(data_sorted)
5411      return -sum([np.log(np.exp(-(x-mu)/tau)/tau) for x in data_sorted[:-1]])
5412
5413  def log_likelihood_fct_evt(para, data_sorted):
5414      ''' Log-likelihood function for the general model'''
5415      kappa, tau, mu = float(para[0]), para[1], min(data_sorted)
5416      return -sum([np.log((1+(x-mu)*(kappa/tau))**(-((kappa+1))/kappa)/tau)\
5417                   for x in data_sorted[:-1]])
5418
5419  def info_mat_exp(tau, data_sorted):
5420      ''' Fisher information matrix for the exponential model '''
5421      mu = min(data_sorted)
5422      data = [x-mu for x in data_sorted[:-1]]
5423      return -1/sum([(tau-2*x)/tau**3 for x in data])
5424
5425  def info_mat_evt(para, data_sorted):
5426      ''' Fisher information matrix for the general model '''
5427      matrix = np.zeros((2,2))
5428      kappa, tau, mu = para[0], para[1], min(data_sorted)
5429      data = [x-mu for x in data_sorted[:-1]]
5430      matrix[0][0] = -sum([(-kappa*x**2+tau**2-2*tau*x)/(tau*(kappa*x+tau))**2 for x in data])
5431      matrix[0][1] = -sum([x*(tau-x)/(tau*(kappa*x+tau)**2) for x in data])
5432      matrix[1][0] = -sum([x*(tau-x)/(tau*(kappa*x+tau)**2) for x in data])
5433      matrix[1][1] = -sum([(kappa*x*(kappa*(kappa+3)*x +2*tau)\
5434          -2*(kappa*x+tau)**2*np.log(1+kappa*x/tau))/(kappa**3*(kappa*x+tau)**2) for x in data])
5435      return np.linalg.inv(matrix)
5436
5437  def threshold_scan_evt(sel_sorted, min_points, max_points, para=[1,0.001]):
5438      ''' Fit to the general model for different values of threshold
5439      min_points sets the minimum number of points that are kept '''
5440      para_list, err_list = list(), list()
5441      for i in range(min(max_points, len(sel_sorted)), min_points, -1):
5442          para = scipy.optimize.fmin(log_likelihood_fct_evt, para,\
5443                                     args=(sel_sorted[:i],), disp=False, maxiter=1000)
5444          para_list.append(para)
5445          err = info_mat_evt(para, sel_sorted[:i])
5446          err_list.append([1.96*np.sqrt(err[1][1]), 1.96*np.sqrt(err[0][0])])
5447      return para_list, err_list
5448
5449
5450  ''' Repeated code '''
5451
5452  def do_threshold_scans(dataset, min_points, max_points_evt, max_points_logn):
5453      # reading selectivities and errors into dictionaries:
5454      sel_dict, err_dict = dict(), dict()
5455      for line in open(dataset, 'r'):
5456          seq, sel, err = line.split('\t')
5457          sel_dict[seq], err_dict[seq] = float(sel), float(err)
5458
5459      # sorting the data by decreasing values of selectivies:
5460      seq_sorted = sorted(sel_dict, key=lambda s: -sel_dict[s])
5461      sel_sorted = [sel_dict[s] for s in seq_sorted]
5462      #err_sorted = [err_dict[s] for s in seq_sorted]
5463      en_sorted = [np.log(sel) for sel in sel_sorted]
5464
5465      # selectivity-rank and energy-rank plots
5466      plt.rcParams['figure.figsize'] = 11, 5; plt.rc('font', size=16)
5467      plt.subplot(121)
5468      plt.loglog(range(1,1+len(sel_sorted)), sel_sorted,'or', lw = 2);
5469      plt.xlabel('rank', fontsize=20); plt.ylabel('selectivity', fontsize=20)
5470      plt.subplot(122)
5471      plt.scatter(range(1,1+len(en_sorted)), en_sorted);
5472      plt.xlabel('rank', fontsize=20); plt.ylabel('- energy', fontsize=20)
5473      plt.xscale('log')
5474      plt.show()
5475
```

```
    # threshold scan (EVT model)
    para_list, err_list = threshold_scan_evt(sel_sorted, min_points, max_points_evt)

    plt.rcParams['figure.figsize'] = 12, 5; plt.rc('font', size=12)
    plt.subplot(121)
    plt.errorbar(sel_sorted[min_points:min(max_points_evt,len(en_sorted))][::-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:min(max_points_evt,len(en_sorted))][::-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:min(max_points_evt,len(en_sorted))][::-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:min(max_points_evt,len(en_sorted))][::-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();
    plt.show()

    # threshold scan (lognormal model)
    para_list, err_list = threshold_scan_logn(en_sorted, min_points, max_points_logn)

    plt.rcParams['figure.figsize'] = 12, 5; plt.rc('font', size=12)
    plt.subplot(121)
    plt.errorbar(en_sorted[min_points:min(max_points_logn,len(en_sorted))][::-1],[p[1] for p in para_list],\
                 [e[1] for e in err_list],fmt='k.',linewidth=3)
    plt.plot(en_sorted[min_points:min(max_points_logn,len(en_sorted))][::-1],[p[1] for p in para_list],'go',markersize=8)
    plt.ylim(0., 5.)
    plt.xlabel(r'threshold $e^*$',fontsize=20); plt.ylabel(r'estimated $\sigma$',fontsize=20)
    plt.subplot(122)
    plt.errorbar(en_sorted[min_points:min(max_points_logn,len(en_sorted))][::-1],[p[0] for p in para_list],\
                 [e[0] for e in err_list],fmt='k.',linewidth=3)
    plt.plot(en_sorted[min_points:min(max_points_logn,len(en_sorted))][::-1],[p[0] for p in para_list],'go',markersize=8)
    plt.ylim(-20., 0.)
    plt.xlabel(r'threshold $e^*$',fontsize=20); plt.ylabel(r'estimated $\mu$',fontsize=20)
    plt.tight_layout();
    plt.show()

def do_fits(dataset, sel_star, en_star, plotflag = False, nr_bin = 50., logscalex = False, logscaley = False):
    # reading selectivities and errors into dictionaries:
    sel_dict, err_dict = dict(), dict()
    for line in open(dataset, 'r'):
        seq, sel, err = line.split('\t')
        sel_dict[seq], err_dict[seq] = float(sel), float(err)

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

    # truncation of the data given the selectivity threshold:
    sel_trunc = [s for s in sel_sorted if s > sel_star]
    mu = min(sel_trunc)
    N_samples = len(sel_trunc)

    # fit to the EVT model (2 parameters):
    print('General EVT fit:')
    para = scipy.optimize.fmin(log_likelihood_fct_evt, [.5,.01], args=(sel_trunc,), maxiter=10000)
    kappa, tau = para[0], para[1]
    para_u = info_mat_evt(para, sel_trunc)
    kappa_u, tau_u = 1.96*np.sqrt(para_u[1,1]), 1.96*np.sqrt(para_u[0,0])
    print('kappa = %.3f +- %.3f, tau = %.5f +- %.5f' % (kappa, kappa_u, tau, tau_u))

    if plotflag:
        # plotting the results for the EVT model:
        x_range = np.linspace(0, 1-1./N_samples, num=N_samples)
        sel_model = [((1-x)**(-kappa)-1)/kappa*tau for x in x_range]
        sel_pp = [1-(1+kappa*(s-mu)/tau)**(-1/kappa) for s in sel_trunc[::-1]]
        fig, ax = plt.subplots(figsize=(14,7))
        plt.rc('font', size=14)
        # Q-Q plot
        plt.subplot(121); plt.title('Q-Q plot (general model)', fontsize=18);
        plt.plot([0,max(sel_model)],[mu,mu+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)
        # P-P plot
        plt.subplot(122); plt.title('P-P plot (general model)', fontsize=18);
        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);
        plt.show()

    # sorting the data by decreasing values of energies:
    en_sorted, err_sorted = [np.log(sel) for sel in sel_sorted], [err_sorted[j]/sel_sorted[j] for j in range(len(sel_sorted))]
```

```
    # truncation of the data given the selectivity threshold:
    en_trunc = [en for en in en_sorted if en > en_star]
    #mu = min(en_trunc)
    N_samples = len(en_trunc)

    # fit to the lognormal model (2 parameters):
    print('Lognormal fit:')
    para = scipy.optimize.fmin(log_likelihood_fct_logn, [-7.,1.], args=(en_trunc,), maxiter=10000)
    mu, sigma = para[0], para[1]
    para_u = info_mat_logn(para, en_trunc)
    mu_u, sigma_u = 1.96*np.sqrt(para_u[0]), 1.96*np.sqrt(para_u[1])
    print('mu = %.3f +- %.3f, sigma = %.3f +- %.3f' % (-mu, mu_u, sigma, sigma_u))

    if plotflag:
        # plotting the results for the lognormal model:
        x_range = np.linspace(0, 1-1./N_samples, num=N_samples)
        fig, ax = plt.subplots(figsize=(14,7))
        plt.rc('font', size=14)
        # Q-Q plot
        plt.subplot(121); plt.title('Q-Q plot', fontsize=18); # Q-Q plot
        qq_data = [invcumF_logn(x, mu, sigma, en_star) for x in x_range]
        en_trunc_shifted = [en - en_star for en in en_trunc]
        plt.errorbar(qq_data, en_trunc_shifted[::-1],\
                     yerr=err_sorted[:N_samples][::-1], fmt='b.', markersize=15)
        plt.plot([min(qq_data), max(qq_data)], [min(en_trunc_shifted), max(qq_data)], 'k-.', linewidth=2)
        plt.axis([min(qq_data)-.1, max(qq_data)+.1, min(en_trunc)-.1, max(en_trunc)+.5])
        unten, oben = min([min(en_trunc_shifted), min(qq_data)])-.25, max([max(en_trunc_shifted), max(qq_data)])+.25
        plt.axis([unten, oben, unten, oben])
        plt.xlabel('model')
        plt.ylabel('data')
        plt.gca().set_aspect('equal', adjustable='box')
        # P-P plot
        plt.subplot(122); plt.title('P-P plot', fontsize=18); # P-P plot
        plt.plot([cumF_logn(e, mu, sigma, en_star) for e in en_trunc[::-1]], x_range, 'r-', lw=3)
        plt.plot([0.,1.], [0.,1.], 'k--', lw=2)
        plt.axis([0., 1., 0., 1.])
        plt.xlabel('model')
        plt.ylabel('data')
        plt.gca().set_aspect('equal', adjustable='box')
        plt.show()

    if plotflag:
        # density plot
        fig, ax = plt.subplots(figsize=(10,7))
        # histogram of data
        if logscalex: bins = np.logspace(np.log10(min(sel_sorted)), np.log10(1.2*max(sel_sorted)), nr_bin)
        else: bins = np.arange(min(sel_sorted),max(sel_sorted)+1e-3,(max(sel_sorted)-min(sel_sorted))/nr_bin)
        n, b, patches = plt.hist([sel for sel in sel_sorted if sel > min(sel_star, np.exp(en_star))], normed=True, bins=bins)
        # generalized Pareto pdf
        sel_range_evt = np.arange(sel_star,1.2*max(sel_sorted),(1.2*max(sel_sorted)-sel_star)/1000.)
        normalization = sum([1. if s>sel_star else 0. for s in sel_sorted]) / len(sel_sorted)
        sel_pdf_evt = normalization*ss.genpareto.pdf(sel_range_evt, kappa, sel_star, tau)
        # lognormal pdf + renormalization
        sel_range_lognorm = np.arange(np.exp(en_star),1.2*max(sel_sorted),(1.2*max(sel_sorted)-np.exp(en_star))/1000.)
        normalization = sum([1. if s>en_star else 0. for s in en_sorted]) / len(en_sorted) / (.5*erfc((en_star-mu)/(np.sqrt(2)*sigma)))
        sel_pdf_lognorm = normalization*ss.lognorm.pdf(sel_range_lognorm, sigma, 0., np.exp(mu))

        plt.plot(sel_range_evt, sel_pdf_evt, color='chartreuse', lw=3)
        plt.plot(sel_range_lognorm, sel_pdf_lognorm, '--', color='magenta', lw=3)
        plt.axis([min(sel_sorted), 1.2*max(sel_sorted), 1e-1, 1.25*max([max(sel_pdf_evt), max(sel_pdf_lognorm), max(n)])])#sel_star+10.*tau
        plt.title('density plot')
        plt.xlabel('selectivity')
        plt.ylabel('probability density')
        plt.legend(['EVT model', 'truncated lognormal model'], loc='upper right', fontsize=15)
        if logscalex: plt.xscale('log')
        if logscaley: plt.yscale('log')
        plt.show()
        #print(max(sel_sorted))
    return kappa, kappa_u, tau, tau_u, mu, mu_u, sigma, sigma_u
```

Chapter G

Preprint

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

The sequences of antibodies from a given repertoire are highly diverse at few sites located on the surface of a genome-encoded larger scaffold. The scaffold is often considered to play a lesser role than highly diverse, non-genome-encoded sites in controlling binding affinity and specificity. To gauge the impact of the scaffold, we carried out quantitative phage display experiments where we compare the response to selection for binding to four different targets of three different antibody libraries based on distinct scaffolds but harboring the same diversity at randomized sites. We first show that the response to selection of an antibody library may be captured by two measurable parameters. Second, we provide evidence that one of these parameters is determined by the degree of affinity maturation of the scaffold, affinity maturation being the process by which antibodies accumulate somatic mutations to evolve towards higher affinities during the natural immune response. In all cases, we find that libraries of antibodies built around maturated scaffolds have a lower response to selection to other arbitrary targets than libraries built around germline-based scaffolds. We thus propose that germline-encoded scaffolds have a higher selective potential than maturated ones as a consequence of a selection for this potential over the long-term evolution of germline antibody genes. Our results are a first step towards quantifying the evolutionary potential of biomolecules.

Significance statement

Antibodies in the immune system consist of a genetically encoded scaffold that exposes a few highly diverse, non-genetically encoded sites. This focused diversity is sufficient to produce antibodies that bind to any target molecule. To understand the control of the scaffold, which acquires hypermutations during the immune response, over the selective response, we analyze quantitative in vitro experiments where large antibody populations based on different scaffolds are selected against different targets. We show that selective responses are described statistically by two parameters, one of which depends on prior evolution of the scaffold as part of a previous response. Our work provides methods to assay whether naïve antibody scaffolds are endowed with a distinctively high selective potential.

*Corresponding authors.

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, 10] and the modularity and polarity of their native fold [11]. 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 introduce a quantitative approach to address this issue and present experimental results that point towards an evolutionary determinant of evolvability in the case of antibodies. Antibodies are particularly well suited to devise and test new approaches to measure and control evolvability, as diverse libraries of billions of different antibodies can be manipulated in vitro by well-established screening techniques [12]. The natural diversity of antibodies is remarkable. Their variable regions span a large phenotypic diversity, specific binding to virtually any molecular target. At the sequence level, this diversity has different origins. First, the variable regions of naïve antibody genes are formed by combining two or three out of tens of genomic segments, with additional randomization at the junction between segments. Second, variable regions of antibodies undergo random somatic mutations along their sequence and selection for higher affinity through the fast evolutionary process of affinity maturation [13]. At the structural level, antibody variable regions consist of a framework displaying variable surface loops called complementary determining regions (CDRs), the most variable one, CDR3, being partially encoded by the randomized sites at junctions between segments [14]. The surface loops, which contain most but not all of the substitutions found in maturated antibodies, and especially the CDR3 loop, are thought to be the primary determinants of binding affinity and specificity [14]. However, the framework has been shown to play an essential role in several cases. In particular the large fraction of framework somatic mutations found in many broadly neutralizing antibodies to HIV have been reported to be required to confer neutralization towards a broad range of viral strains [15].

Antibody variable regions are thus subject to evolution by natural selection on two distinct time scales: their genome-encoded segments evolve on the time scale of many generations of their host, as all other genes, while naïve antibodies assembled from those genome-encoded segments additionally evolve on a much shorter time scale as part of the immune response in the process of affinity maturation. Importantly, affinity maturation-associated mutations are somatic and the sequences of maturated antibodies are not transmitted to subsequent generations. Germline antibody genomic segments, 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 scaffolds [16]. We define scaffold as the genome-encoded sites of an antibody sequence. In a naïve antibody, the scaffold amino acids are identical to germline amino acids, in affinity maturated antibodies some scaffold sites are somatically mutated.

We took for these scaffolds the heavy chains (V_H) of natural antibodies, including their framework regions and CDR1 and CDR2 loops, and built libraries by introducing all combinations of amino acids at four consecutive sites in their CDR3 loop. Using phage display [17], we selected sequences from these libraries for their ability to bind different molecular targets and analyzed the relative enrichment of different antibody sequences through successive cycles of selection and re-amplification by high-throughput sequencing [18]. 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 enrichments 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 enrichments between and within libraries? (ii) How to rationalize the differences between scaffolds that are all homologous?

Here, we address 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. One of these parameters, σ, reports the phenotypic variability within a library and thus quantifies the potential of a library to respond to selection. Second, we present new experimental results and re-analyze previous results to provide evidence that the degree of maturation of an antibody scaffold is a control parameter for its selective potential. Our approach thus provides a general and quantitative framework to study experimentally the selective potential of biomolecules. Our results are also, to our knowledge, the first to indicate that long-term evolution may have endowed germline antibodies 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. Our initial populations are libraries made of sequences with a common part, which we call a scaffold, and 4 positions $x = (x_1, x_2, x_3, x_4)$ that are randomized to all $N = 20^4$ combinations, where 20 is the number of natural amino acids. We subject these populations to successive cycles of selection for binding against a target T and amplification. The critical property of a sequence x present in the initial population is its enrichment $s(x)$, the factor by which it is enriched or depleted from one cycle to the next (see Box). The mapping $x \mapsto s_{L,T}(x)$ from 4-position sequences x to enrichments generally depends both on the scaffold that defines the library L and on the target T that defines the selective pressure.

Experiments are designed for $s(x)$ to reflect the binding affinity of an antibody with CDR3 sequence x to the chosen target T (SI 1.1). In effect, however, selection does not depend exclusively on the CDR3 sequence x and the target T as phage-displayed antibodies may also be selected because they bind to something else than the target (the recipient or another phage) or because they bind to the target through their antibody scaffold. Such non-specific binding is generally negligible for the CDR3 sequences x of antibodies with top binding affinities to the target, but it dominates the selection of the majority of antibodies, which typically show no or weak CDR3 sequence-specific binding to the target. Following common practice in the field, we therefore perform three cycles of selection, which effectively enriches the population in strong binders,

and interpret only the top enrichments $s(x) > s^*$, computed at the last cycle, as resulting from specific binding to the target (SI 3.2 and 3.3). We are interested in properties of the scaffold that favor these large enrichment values, 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 [16]. These fragments originate from the germline or the B cells of organisms of various species. Scaffolds from the germline have not been subject to any affinity maturation, while scaffolds from B cells are taken from 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 [16]. 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-based scaffolds, the final population of one experiment was dominated by antibodies built on one of the two germline-based scaffolds, and the second by the other one. This suggests that germline scaffolds may have an intrinsically higher selective potential.

To investigate this hypothesis, we performed the selection against 4 different targets of 3 libraries built on scaffolds with varying degrees of maturation. The 3 single-domain V_H libraries are based on V genes from the heavy chain of 3 human antibodies that have evolved to different degrees as part of the immune response to HIV (Fig. S1). They bear identically randomized CDR3 at 4 sites (upstream of a common human framework FWR4 region JH4 and no light chain). The Lim and Bnab scaffolds are derived from antibodies isolated from patients (6-187 and PGT128) [19, 20] and have respectively limited and broad spectrum of neutralization of HIV strains [21, 15]. Previous studies [15] concluded that the heavy chain V genes of these antibodies result from distinct affinity maturation trajectoires originating from a common germline origin (IGHV4-39) on which our Germ scaffold is based. Our Germ scaffold has thus not undergone any maturation. The Lim scaffold differs from Germ, from which it originates, by 14 % of its amino acids. The Bnab scaffold 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 CDR2 of the Bnab scaffold also includes an insertion of 6 amino acids. The 3 single-domain V_H libraries, which are built around these V_H 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 [16]. 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 structurally related 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 enrichments. These parameters derive from a statistical approach that considers only the distribution $P(s)$ of values that enrichments take across the different sequences of a library [23, 24, 25]. They correspond to

the assumption that this distribution is log-normal,

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

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 parametrization of the distributions of enrichments by log-normal distributions has several motivations. 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 [26] and protein-DNA interactions [27], including studies that had access to the complete distribution $P(s)$ [27]. Second, log-normal distributions are stable upon iteration of the selective 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, then s also follows a log-normal distribution; more generally, log-normal distributions are attractors of evolutionary dynamics [28]. 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 $\beta\Delta G(x) = \sum_{i=1}^{\ell} h_i(x_i)$ with contributions $h_i(x_i)$ from each position i and amino acid x_i, and thus its enrichment $s(x) \simeq \mathrm{e}^{-\beta\Delta G(x)}$, where T is the temperature and k_B the Boltzmann constant (SI 1.1). At thermal equilibrium and for sufficiently large ℓ, a log-normal distribution of the affinities is then expected with $\mu \sim -\ell\langle h\rangle$ and $\sigma \sim \ell^{1/2}(\langle h^2\rangle - \langle h\rangle^2)^{1/2}$, where $\langle\epsilon\rangle$ and $\langle h^2\rangle - \langle h\rangle^2$ are respectively the mean and variance of the values of binding energies per position $h_i(x_i)$. This additive model, which ignores epistasis between the sites i is not expected to be exact but can provide a first approximation of the data (SI 3.3). The limit central theorem, on which the above argument is based, in fact remains valid in presence of weak epistasis. We also note that the model does not exclude epistasis between the sites i and the scaffold, which will be shown to be essential. The parameter σ, which quantifies the diversity of enrichment values within a library, also corresponds to a natural measure of diversity from the standpoint of information theory (SI 1.3). These multiple empirical and theoretical justifications motivate a description of the distributions of enrichments from selections of antibody libraries by log-normal distributions. We show below that our data does not exclude descriptions by other distributions, from which the same main conclusions can be drawn.

2.3. Inference of parameters

The enrichment $s(x)$ of a sequence x is obtained from comparing the frequency of x in the population before and after a round of selection. As only the largest enrichments are expected to reflect specific binding to the target (SI 3.2 and 3.3), we obtain the parameters σ and μ by fitting the values with truncated log-normal distributions, when $s(x)$ exceeds a threshold s^* (Fig. 1A and SI 3.4). The exact value of this threshold is not critical, provided it is large enough (Figs. S26-27), but it must be determined independently for each selection of each library as non-specific binding may depend on the scaffold (Fig. S24). An additional complication is that enrichments are defined 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 enrichment sequences (Fig. 1B), which allows for a very precise estimation of the relative

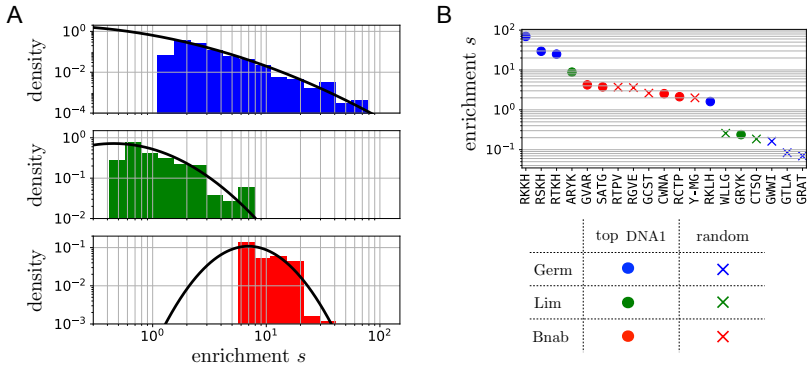


Figure 1: Fitting empirical distributions of enrichments with log-normal distributions. **A.** The selection of a library L against a target T provides the enrichments of the sequences in L that are best selected against T (the other ones are eliminated). Here, the histograms show the enrichments obtained 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. The fit is made only to the upper part of the distribution as experiments provide only the top enrichments. The quality of the fits is validated by probability-probability and quantile-quantile plots (Figs. S16-S22). **B.** To locate precisely the mode of the distributions (maximum of the black curves in A), we performed experiments where the initial population consists in a mixture of very few top (dots) and random (crosses) sequences. Because these experiments involve very few sequences, they provide very precise estimations of the relative enrichments (Fig. S25). Top sequences are identified from A based on the largest enrichments against the target. Random sequences, on the other hand, are picked at random in the libraries and are expected to have typical enrichments 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).

enrichments beyond the top sequences (Fig. S25): as the random sequences typically reflect the mode of the distributions, (the most likely enrichment value), these experiments provide an independent estimation of μ that we can profitably use (see details in SI 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 enrichments are measured by comparing frequencies between the 2nd and 3rd cycles, or between the 3rd and 4th cycles (Figs. S20-S21).

3. Results

3.1. Intra-library hierarchy

The hierarchy of enrichments 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 $p = (3!)^{-4} \simeq 7.10^{-4}$.

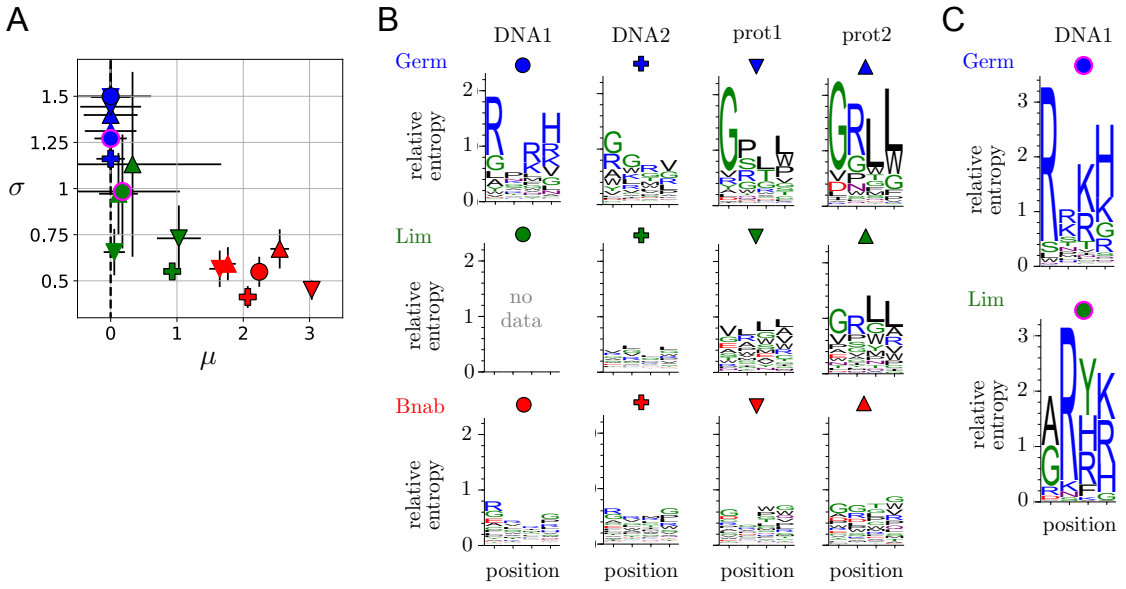


Figure 2: Comparing selections of libraries built on scaffolds with different degrees of maturation – **A.** Parameters (μ, σ) of the distributions of enrichments for our 3 libraries selected against 4 targets. The color of the symbols indicates the library (Germ, Lim or Bnab) and its shape the target (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 (SI 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 enrichments (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 the same motif with $x_1 = R$, $x_3 = R$ or K and $x_4 = H$ dominates 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).

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. This is illustrated through sequence logos (Fig. 2B-C). These sequence logos do not fully capture the specificity against each target, as they ignore any epistasis between the sites, but observing that they are different is sufficient to conclude that selection is target-specific. 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 enrichments (see also Fig. S6). Target-specific selection for binding, which is dominating the top enrichments in the Germ library (Fig. S11), is

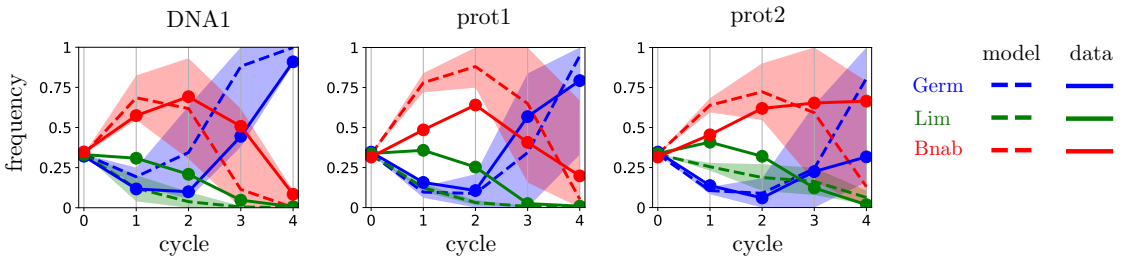


Figure 3: Dynamics of library frequencies – A mixture of the three libraries, Germ (blue), Lim (green) and Bnab (red) was subject to four successive cycles of selection and amplification against different targets. The full lines report the evolution of the relative frequencies of the three scaffolds. 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 (SI 1.5). The shaded area correspond to one standard deviation in the estimation of the parameters σ, μ. The fit is only qualitative as we assume here that sequences are uniformly represented in each initial library, which is not the case in experiments. The trends, which are controlled by the two parameters σ and μ, are nevertheless well reproduced.

thus of the same order of magnitude as amplification biases for the top enrichments in the Bnab library.

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 enrichments between libraries is quantified by the parameter μ. This parameter also shows a pattern that is independent of the target: $\mu_{\mathrm{Germ},T} \simeq \mu_{\mathrm{Lim},T} < \mu_{\mathrm{Bnab},T}$ and even $\max_T(\mu_{\mathrm{Germ},T}, \mu_{\mathrm{Lim},T}) < \min_T(\mu_{\mathrm{Bnab},T})$ (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. 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 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 are globally

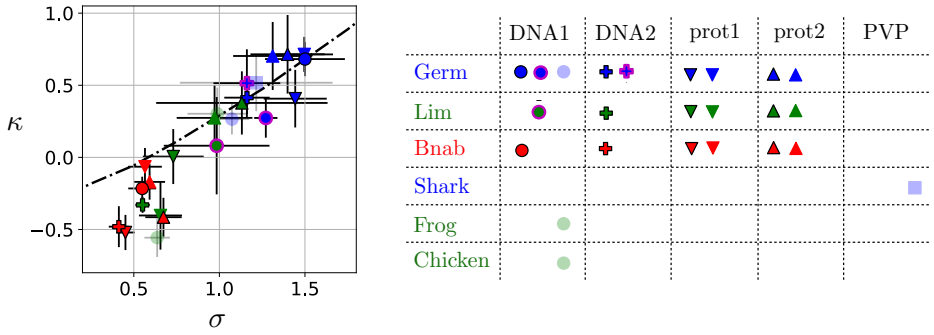


Figure 4: Shape parameter κ from fits of the enrichments to generalized Pareto distributions versus σ from fits to log-normal distributions – Results from different libraries selected against different targets are represented here with the same convention as in Figure 2: blue, green and red plain colors 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 [16] 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 SI Table 1).

predicted by a calculation of library frequencies in the mix based on the parameters (μ_L, σ_L) inferred for each library L independently (SI 1.5). We verify that the short-term dynamics are dominated by the library with largest μ while the long-term dynamics are dominated with the library with largest σ: which of the two parameters is most important thus depends on the considered time scale. The predictions reported in Fig. 3 are based on two assumptions: (i) the distributions of enrichments in different libraries L are log-normal; (ii) the sequences in the initial population have equal frequencies. This second hypothesis is only an approximation for our experiments, which limits the validity of the predictions. Nevertheless, the results illustrate how parametrizing the response to selection of a library by the two parameters (μ,σ) is not only useful to characterize its intrinsic response but also to rationalize the evolutionary dynamics of mixtures of libraries.

3.4. Additional data

Beyond the 3 libraries analyzed so far, our conclusions are supported by re-analyzing our previous results [16]. These previous results involved a library based on another germline scaffold, 19 libraries built on other maturated scaffolds, and a completely different target, in addition to some of the same frameworks and targets presented in this work. Inferring σ from these data, we observe again that libraries built around germline scaffolds have larger σ than libraries built around maturated scaffolds (Fig. 4 and SI Table 1). These supplementary results corroborate the hypothesis that our measure of selective potential σ decreases in the course of affinity maturation.

3.5. Extreme value statistics

In our previous work [16], we fitted the tail of the distribution of enrichments with generalized Pareto distributions, a family of distributions with two parameters, a shape parameter κ and a scaling parameter τ. This was motivated by extreme value theory, which establishes that these parameters are sufficient to describe the tail of any distribution (SI 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 [29]. 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 [30]. 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 [30]. 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 [16]: 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 S1).

4. Conclusion

In summary, we propose the hypothesis that naïve antibodies which are constructed from germline genes are endowed with a special evolutionary ability to generate selectable diversity, which they lose when undergoing affinity maturation. To study this hypothesis, we introduced an experimental and statistical approach that quantifies the selective potential of antibody scaffolds. In this approach, the response to selection of an antibody library against a given target is summarized by two parameters, σ and μ, which have different interpretations and implications. The parameter σ describes the variability of the responses between sequences in the library, while μ describes their common response. These two parameters may be viewed as quantifying the selective potential of a library over different time scales: when competing two libraries, the library with largest μ is initially more enriched but in the long-run sequences from the library with largest σ eventually dominate.

Applying this approach to data from our high-throughput selection experiments, we find results in favor of the hypothesis that germline-based antibody scaffolds have a higher potential to generate selectable diversity, corresponding to a higher σ. In particular, we analyzed new data centered onto 3 libraries, one built on a germline-based scaffold and two built on scaffolds derived from this germline-based scaffold with different degrees of maturation, which we selected against 4 different targets, all unrelated to the target against which the scaffold was originally maturated. We find that σ decreases with the degree of maturation. Our hypothesis is also corroborated by a re-analysis of our previous results, which involved a library built on

another germline-based scaffold, 19 libraries built on other maturated scaffolds, and a completely different target [16]. Further experiments with additional scaffolds and targets are needed to assess the generality of these results and the limitations of our statistical description by means of only two parameters. The present work provides the motivation and the methodology to generate and analyze such data and study alternative scenarios. We also stress that our analysis is generally applicable to antibody library screening beyond testing our hypothesis, in particular to compare quantitatively in a single plot, as in Figure 2A, the outcome of many selection experiments involving several libraries and/or several targets.

Quantifying the selective potential of an antibody scaffold is a first step towards designing libraries with optimized selectable diversity. Once the property of a biomolecule is measurable, one can indeed resort to directed evolution to attempt to optimize it. Here, the starting point would be a population comprising different libraries with different scaffolds but identical random variations. We previously competed for binding to a target 24 such libraries [16], a number that could be increased. By alternating such selections with the introduction of new mutations in the scaffolds, one may be able to evolve scaffolds with increased μ and/or σ.

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 [32, 33, 34, 35, 36, 37, 38, 39]. 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 [40]. 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. Another biophysical property is also known to correlate with evolvability, thermal stability [9, 10]. The loss of selective potential that we observe may thus derive from a loss of thermal stability [41, 42]. Destabilization during affinity maturation might for instance arise from the interaction between the heavy and light chains of antibodies: germline heavy chains, which have to be robust to various light chain pairings, may be more stable than maturated heavy chain whose stability may depend on their associated light chain. Our results may thus be tied to the fact that we are studying heavy chains in isolation. Additional studies are needed to test this and other hypotheses and to identify the mechanisms behind the differences of selective potential that we measure.

Irrespective of mechanisms, our hypothesis and methodology may find applications beyond antibodies, to understand more generally what controls the selective potential of 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?

Supporting information legends

The supporting information file provides a supporting text, a supporting table and 27 supporting figures.

Data availability

Raw sequencing data was deposited to the NCBI Sequence Read Archive (SRA accession: PRJNA592656).

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] S. Bershtein, M. Segal, R. Bekerman, N. Tokuriki, D. S. Tawfik, Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein, Nature 444 (7121) (2006) 929.

[11] 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.

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

[15] 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.

[16] 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.

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

[18] 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.

[19] J. F. Scheid, H. Mouquet, N. Feldhahn, M. S. Seaman, K. Velinzon, J. Pietzsch, R. G. Ott, R. M. Anthony, H. Zebroski, A. Hurley, A. Phogat, B. Chakrabarti, Y. Li, M. Connors, F. Pereyra, B. D. Walker, H. Wardemann, D. Ho, R. T. Wyatt, J. R. Mascola, J. V. Ravetch, M. C. Nussenzweig, Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals., Nature 458 (7238) (2009) 636–640.

[20] L. M. Walker, M. Huber, K. J. Doores, E. Falkowska, R. Pejchal, J.-P. Julien, S.-K. Wang, A. Ramos, P.-Y. Chan-Hui, M. Moyle, J. L. Mitcham, P. W. Hammond, O. A. Olsen, P. Phung, S. Fling, C.-H. Wong, S. Phogat, T. Wrin, M. D. Simek, Protocol G Principal Investigators, W. C. Koff, I. A. Wilson, D. R. Burton, P. Poignard, Broad neutralization coverage of HIV by multiple highly potent antibodies., Nature 477 (7365) (2011) 466–470.

[21] 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.

[22] N. S. Longo, M. S. Sutton, A. R. Shiakolas, J. Guenaga, M. C. Jarosinski, I. S. Georgiev, K. McKee, R. T. Bailer, M. K. Louder, S. O'Dell, M. Connors, R. T. Wyatt, J. R. Mascola, N. A. Doria-Rose, Multiple Antibody Lineages in One Donor Target the Glycan-V3 Supersite of the HIV-1 Envelope Glycoprotein and Display a Preference for Quaternary Binding., Journal of virology 90 (23) (2016) 10574–10586.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

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

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

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] R. Henderson, B. E. Watts, H. N. Ergin, K. Anasti, R. Parks, S.-M. Xia, A. Trama, H.-X. Liao, K. O. Saunders, M. Bonsignori, et al., Selection of immunoglobulin elbow region mutations impacts interdomain conformational flexibility in hiv-1 broadly neutralizing antibodies, Nature communications 10 (1) (2019) 654.

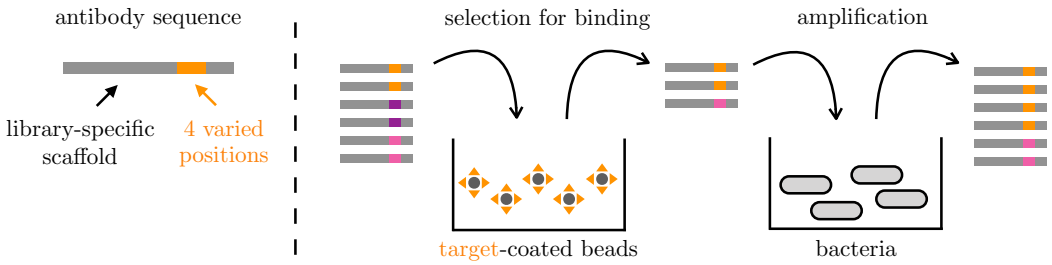
[42] L. Shehata, D. P. Maurer, A. Z.Wec, et al, Affinity maturation enhances antibody specificity but compromises conformational stability. Cell reports, 28(13), (2019) 3300-3308.

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

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 [17]. 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 SI 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 enrichment $s(x)$ 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, enrichments are independent of the cycle c (see SI 1.1). In the limit of infinite population sizes, $s(x)$ is proportional to the ratio $f^c(x)/f^{c-1}(x)$ of the frequencies $f^c(x)$ after any two successive cycles $c-1$ and c. To estimate these enrichments, 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 enrichment of x as

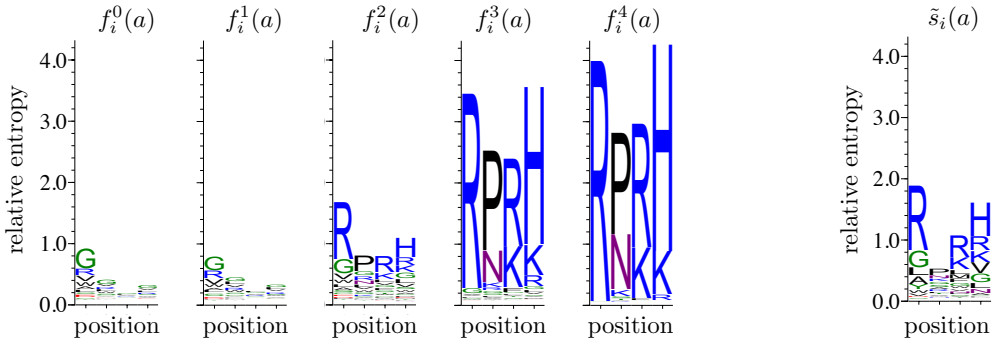
$$s(x) = \alpha_c^{-1} \frac{n^c(x)}{n^{c-1}(x)} \tag{2}$$

where α_c is an arbitrary multiplicative factor.

In practice, two types of noise must be taken into account when applying Eq. (2): 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 enrichments. We therefore compute $s(x)$ between the second and third cycles as $s(x) = \alpha_3^{-1} n^3(x)/n^2(x)$ by restricting to sequences x that satisfy $n^2(x) \geq 10$ and $n^3(x) \geq 10$. Additionally, as the smallest enrichments are due to experimental noise, we retain only the sequences with $s(x) > s^*$ where s^* is determined self-consistently (SI 3.2 and Fig. S3). Enrichments $s(x)$ 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 enrichments, we use sequence logos [43]. 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
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*

Contents

1. Theoretical methods

1.1. Physics of selection

1.1.1. Enrichments and binding energies

When assuming that selection is controlled by equilibrium binding to the target, the distribution of enrichments 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

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

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

$$s_{AT} = \frac{1}{1 + {\rm e}^{\beta(\Delta G_{AT} - \mu)}} \tag{2}$$

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 [1]. This Fermi-Dirac statistics is approximated by Boltzmann statistics

$$s_{AT} \simeq {\rm e}^{-\beta(\Delta G_{AT} - \mu)}. \tag{3}$$

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 enrichments 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

$$\sum_A [A]_{\rm tot} \ll [T]_{\rm tot} \ll \min_A K_{AT}. \tag{4}$$

The first constraint $\sum_A [A]_{\rm tot} \ll [T]_{\rm tot}$ guarantees an absence of competition between antibodies so that the enrichments 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. (1). 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,

$\sum_A [A]_{\text{tot}} \simeq 10^{11}\,\text{mL}^{-1}$ and $[T]_{\text{tot}} \simeq 10^{14}\,\text{mL}^{-1}$, which satisfies the first constraint. The concentration $\sum_A [AT]_{\text{eq}}$ of selected antibodies before amplification is estimated between $10^5\,\text{mL}^{-1}$ at the first round of selection and $10^7 - 10^8\,\text{mL}^{-1}$ at the fourth. Considering this last number to reflect properties of the best binders, we estimate that $\min_A K_{AT}/[T]_{\text{tot}} \simeq \sum_A [A]_{\text{tot}}/\sum_A [AT]_{\text{eq}} \simeq 10^3$, 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 $\Delta G(x) = \sum_{i=1}^{\ell} h_i(x_i)$ with the $h_i(x_i)$ taking random values, the central limit theorem indicates that for sufficiently large ℓ the energies $\Delta G(x)$ are distributed normally with a mean $\mu \simeq -\ell\langle h\rangle$ and a variance $\sigma^2 \simeq \ell(\langle h^2\rangle - \langle h\rangle^2)$, where $\langle h\rangle$ and $\langle h^2\rangle - \langle h\rangle^2$ are respectively the mean and variance of the values of binding energies per position $h_i(x_i)$. Given Eq. (3), this leads to a log-normal distribution for the enrichments $s(x) \propto e^{-\beta\Delta G(x)}$.

The assumptions involved in this derivation may not be justified, starting from the assumption that enrichment 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 [2]. Deviations from a log-normal distribution of enrichments can therefore, at least in principle, be systematically analyzed and understood.

1.2. Alternative statistical model from extreme value theory

In our previous work [3], we fitted the tail of the distribution of enrichments with generalized Pareto distributions from extreme value theory. For different libraries L and different targets T, we found that generalized Pareto distributions provide a good fit of the upper tail of the distribution of enrichments, 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. Here, we show that κ captures essentially the same information as σ, one of the two parameters of the model based on the log-normal distribution.

1.2.1. Extreme value statistics

Extreme value theory states that for any random variable 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 $s^\star \to \infty$ [4], where

$$f_\kappa(x) = \begin{cases} (1+\kappa x)^{-\left(1+\frac{1}{\kappa}\right)} & \text{if} \quad \kappa \neq 0, \\ \mathrm{e}^{-x} & \text{if} \quad \kappa = 0. \end{cases} \tag{5}$$

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

$$\mathbb{E}[\ln s_r] \simeq \mu + \left((2 \ln N)^{1/2} - \frac{\ln \ln N + \ln 4\pi}{2\sqrt{2 \ln N}} \right) \sigma - \frac{\sigma}{(2 \ln N)^{1/2}} \ln r, \tag{6}$$

which corresponds to an apparent shape parameter $\kappa_N = \sigma (2 \ln N)^{-1/2}$ [5]. 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. (6) 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.

1.2.3. κ versus σ in the data

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). These results are consistent with theoretical expectations. The (effective) shape parameter κ from extreme value theory and the parameter σ from log-normal distributions thus report essentially the same information (Fig. 4). The data thus support the hypothesis that maturation lead to a loss of selective potential irrespective of the statistical model used to quantify selective potentials, whether it is a generalized Pareto distribution motivated by extreme-value theory or a log-normal distribution.

1.3. Information theory of selection

Several relationships are known between evolutionary dynamics and information theory. In particular, the change of Malthusian fitness $\ln s(x)$ satisfies [8]

$$\Delta \ln s = D(\tilde{s} \| N^{-1}) + D(N^{-1} \| \tilde{s}). \tag{7}$$

Here we present a different relationship, which also involves the relative entropy $D(\tilde{s} \| N^{-1})$. This new relationship extends the work of Ref. [9].

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 [7], which is defined by

$$D(P^1\|P^0) = \sum_x P^1(x) \ln \frac{P^1(x)}{P^0(x)}. \tag{8}$$

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 $e^{-ND(P^1\|P^0)}$ [7].

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 [7]. 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

$$P^1(A,T) = \frac{[AT]_{\rm eq}}{[A]_{\rm eq} + \sum_{T'}[AT']_{\rm eq}} \simeq \frac{[AT]_{\rm eq}}{[A]_{\rm eq}} = K_{AT}^{-1}[T]_{\rm eq} \simeq K_{AT}^{-1}[T]_{\rm tot} \tag{9}$$

where K_{AT} is the dissociation constant between A and T and where the approximations are justified in Section 1.1. A random association is defined here by considering equal dissociation constants,

$$P^0(A,T) = \frac{[A]_{\rm tot}[T]_{\rm tot}}{\sum_{A',T'}[A']_{\rm tot}[T']_{\rm tot}}. \tag{10}$$

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

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

$$D(P^1\|P^0) = \left\langle \ln\left(\frac{P^1}{P^0}\right)\right\rangle_1 = \sum_{A,T} P^1(A,T) \ln \frac{P^1(A,T)}{P^0(A,T)} = \sum_{A,T} P^0(A,T) s_{AT} \ln s_{AT} = \langle s \ln s\rangle_0 \tag{11}$$

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,

$$D(P^1\|P^0) = \left\langle \frac{s}{\langle s\rangle_0} \ln \frac{s}{\langle s\rangle_0} \right\rangle_0 \tag{12}$$

When a single target T is considered with $P^0(A,T) = 1/N$ and $P^1(A,T) = s(x)$ where x represents the sequence of A, this becomes

$$D(\tilde{s}\|N^{-1}) = \left\langle \frac{s}{\langle s\rangle} \ln \frac{s}{\langle s\rangle} \right\rangle \tag{13}$$

Eq. (12) is valid for any initial distribution $f^0(x)$ as long as $f^1(x) \propto s(x)f^0(x)$ while Eq. (13), where averages $\langle\cdot\rangle$ are taken with a distribution $P(s)$ of the enrichments over the different sequences x, is 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

$$D(s\|N^{-1}) = \frac{\sigma^2}{2} \tag{14}$$

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 [9]. It is justified, however, only within the special model considered in [9] 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 $P^0(A,T) = 1/M^2 \simeq P^1(A)P^1(T)$.

1.3.4. Sequence motifs

Assuming that the different sites i along the sequence contribute independently to the enrichment, $\tilde{s}(x) = \prod_i \tilde{s}_i(x_i)$, 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 enrichments 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. (14) 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.

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

$$n^c(x) = \alpha_c s(x) n^{c-1}(x) \tag{15}$$

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

$$f^c(x) = \frac{s(x)f^{c-1}(x)}{\sum_{x'} s(x')f^{c-1}(x')} = \frac{(s(x))^c f^0(x)}{\sum_{x'}(s(x'))^c f^0(x')}. \tag{16}$$

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

$$f^c(L) = \frac{\langle s^c\rangle_L}{\sum_{L'}\langle s^c\rangle_{L'}} \tag{17}$$

with

$$\langle s^c\rangle_L = \sum_{x\in L}(s(x))^c f^0(x) = \int_0^\infty \mathrm{d}s\, P_L(s) s^c = \exp\left(c\mu_L + \frac{c^2\sigma_L^2}{2}\right). \tag{18}$$

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 enrichments s within library L are described by a distribution of enrichments $P_S(s)$ 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

$$f^c(L) = \left(\sum_{L'} \mathrm{e}^{c(\mu_{L'}-\mu_L)+c^2\left(\sigma_{L'}^2-\sigma_L^2\right)/2}\right)^{-1}. \tag{19}$$

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 assumptions, 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 [3], 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) and 50 μg/mL kanamycin (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 for protein targets and 0.9x PBS for DNA targets 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 [3]. 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 CDR3 (`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 ≤ 3 nt and the difference in Hamming distance between nearest and next-nearest was ≥ 1 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`

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

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 with a threshold

Enrichments are computed from sequencing counts as indicated in Eq. (2) in the Box. To account for sampling noise, only sequences whose count is ≥ 10 both at round c and $c+1$ are considered. Moreover, we ignore enrichments $s(x)$ below a threshold $s^\star$, which arise from unspecific binding. Unspecific binding modifies the expression for the enrichment of sequence x to include a sequence-independent unspecific binding energy ΔG_{us},

$$s(x) = \frac{\mathrm{e}^{-\beta\Delta G(x)} + \mathrm{e}^{-\beta\Delta G_{\text{us}}}}{1 + \mathrm{e}^{-\beta\Delta G(x)} + \mathrm{e}^{-\beta\Delta G_{\text{us}}}}. \tag{20}$$

It sets a lower bound for the enrichment given by

$$s_{\text{us}} = \frac{\mathrm{e}^{-\beta\Delta G_{\text{us}}}}{1 + \mathrm{e}^{-\beta\Delta G_{\text{us}}}} = \frac{1}{1 + \mathrm{e}^{\beta\Delta G_{\text{us}}}}. \tag{21}$$

The argument for log-normality of enrichment distributions applies only when the specific binding contribution $\Delta G(x)$ dominates the enrichment. We therefore eliminate the enrichments 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 enrichments 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 $s = s_{\text{us}}$ 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 enrichments extend beyond the tail of the distribution. In our previous work [3], 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 enrichments being dominated by unspecific binding (Fig. S15).

3.3. Noise cleaning with a stochastic model

Another approach was previously proposed to clean the noise when analyzing data comparable to ours, which introduces a stochastic model for the sequencing bias and for the mapping $x \mapsto \Delta G(x)$ [10, 11]. We illustrate here how it gives results consistent with those of our simpler approach.

In this alternative approach, the sampling noise from sequencing noise is described by a Poissonian distribution (a description that may be elaborated to take into account the non-Poissonian effects of PCR amplification [12]). Given reads between two successive rounds $\{n^{c-1}(x)\}$, $\{n^c(x)\}$, the log-likelihood of the enrichments $s(x)$ has the form [10]

$$\mathcal{L}(s(x)|\{n^{c-1}(x)\},\{n^c(x)\}) = \sum_x n^c(x)\ln s(x) - (n^{c-1}(x)+n^c(x))\ln(1+s(x)) \tag{22}$$

where the sum is over all sequences x. Optimizing over $s(x)$ for each x independently gives $\hat{s}(x) = n^c(x)/n^{c-1}(x)$, consistent with Eq. (4) in the Box.

To identify specific binding, the relation between $s(x)$ and $\Delta G(x)$ must be specified. In the approximation of weak binding (see Sec. 1.1), Eq. (20) takes the form

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

To differentiate specific from unspecific binding, we further assume that $\Delta G(x)$ takes the form [10, 11]

$$\beta\Delta G(x) = \sum_{i=1}^{\ell} h_i(x_i), \tag{24}$$

which corresponds to ignoring epistasis between the sites i. With $\ell = 4$ denoting the length of the variable sequence x and $q = 20$ denoting the number of possible amino acids, this model has one parameter $\Delta G_{\rm us}$ for unspecific binding and Lq parameters $h_i(x_i)$ for specific binding, of which only $\ell(q-1)+1 = 77$ are independent due of the invariance under the transformation $h_i(a) \leftarrow h_i(a) + g_i$ for any g_i satisfying $\sum_{i=1}^{\ell} g_i = 0$.

These parameters are obtained from the data by optimizing the log-likelihood given in Eq. (22). In practice, we find as in Ref. [11] that introducing a small ℓ_2 regularization on the fields $h_i(x_i)$ is necessary to prevent them from taking excessively large values. Two sets of parameters are of interest: the parameters

maximizing the log-likelihood when $e^{-\beta\Delta G_{\rm us}} = 0$, corresponding to a model $s_0(x)$ without unspecific binding, and the parameters maximizing the log-likelihood with $\Delta G_{\rm us}$ treated as a variable, corresponding to a model $s_1(x)$ integrating unspecific binding. Which solution is most relevant depends on whether unspecific binding is negligible or not. If unspecific binding is negligible, $s_1(x)$ tends to under-fit the data while if unspecific binding is not negligible, $s_0(x)$ tends to over-fit it.

This is demonstrated in Fig. S23 with the example of the Germ library selected against the DNA1 target, where unspecific binding is significant between rounds 1 and 2 but becomes negligible between rounds 2 and 3. In any case, the results are consistent with the choice of a cut-off s^*.

The results indicate that an additive model can provide a valid approximation of $\Delta G(x)$. Fig. S23G also illustrates how the data is consistent between rounds: a refined analysis may infer a model that fits the data over all available rounds.

3.4. 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) in the Box, we focus on the best available enrichments $s_i > s^*$, the log-normal distribution is under-sampled. In practice, it is more convenient to work with the log of the enrichments, $y_i = \ln(s_i)$, and to fit them with a normal distribution. If restricting to values y_i larger than a given threshold y^*, the probability density $P(Y = y|Y \geq y^*)$ of observing y_i given that $y_i \geq y^*$ is

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

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

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

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 previous section). Finally, we obtain a lower bound on the uncertainty of the parameter values using the Fisher information matrix and the Cramér-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 $y = F^{-1}(z|y^*)$.

3.5. Normalization of μ across libraries

The selection of a library L against a target T yields only the values of the highest enrichments $s(x)$ 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 enrichments 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 enrichments $s_{L,T}(x)$ as possible (see Box and Fig. 1A). We then identify 2 to 4 sequences with largest enrichment 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 enrichments 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 enrichment 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 unnecessary to estimate the enrichments of typical sequences against all targets once we understood that these values are not controlled by the target.

References

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

[2] 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.

[3] 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.

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

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

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

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

[8] S. A. Frank, Natural selection. V. How to read the fundamental equations of evolutionary change in terms of information theory., Journal of evolutionary biology 25 (12) (2012) 2377–2396.

[9] 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.

[10] J. Otwinowski, Biophysical inference of epistasis and the effects of mutations on protein stability and function. Molecular biology and evolution., 35 (10) (2018) 2345–2354.

[11] G. Rastogi et al. Accurate and sensitive quantification of protein-DNA binding affinity.. Proceedings of the National Academy of Sciences 115 (16) (2018) E3692–E3701.

[12] S. F. Levy, J. R. Blundell, S. Venkataram, D. A. Petrov, D.S. Fisher, , G. Sherlock, Quantitative evolutionary dynamics using high-resolution lineage tracking. Nature 519(7542) (2015) 181–186.

		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 enrichments to generalized Pareto distributions (κ, τ) and log-normal distributions (σ, μ) for experiments presented here and in our previous work [3]. N/A indicates that the data was insufficient to make a meaningful fit. For enrichments 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 Cramér-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 [3] for reasons explained in Section 3.2 and Figure S15.

SUPPLEMENTARY FIGURES

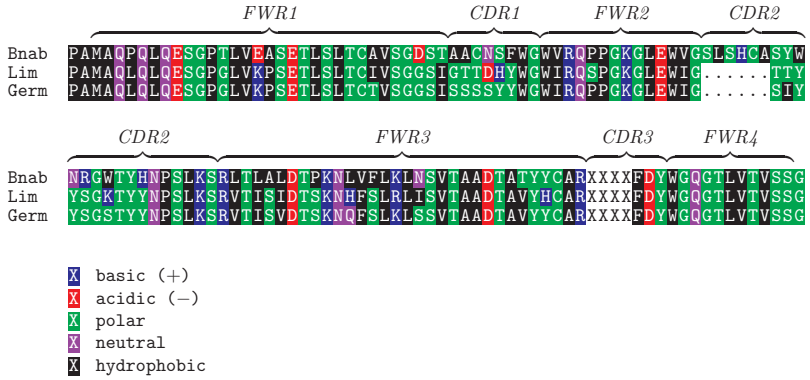


Figure S1: Alignment of the sequences of the three scaffolds, Bnab, Lim and Germ. The 4 randomized positions correspond to the part of the CDR3 indicated by `XXXX`.

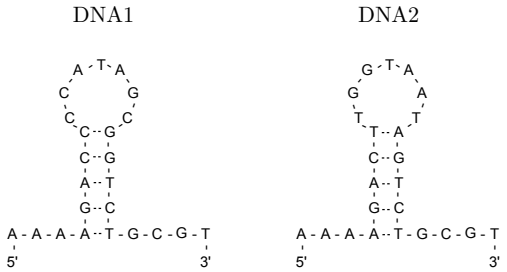


Figure S2: DNA1 and DNA2 binding targets. The targets display a hairpin structure at room temperature. They share a common stem sequence but the sequence of their loop differ. A biotin is placed at the 5' ends to allow for immobilization on streptavidin-coated magnetic beads.

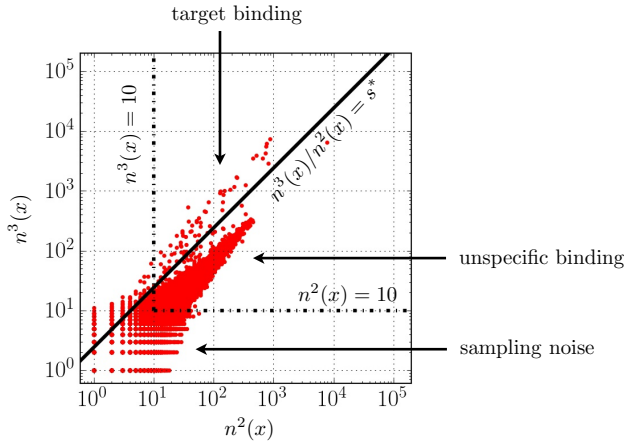


Figure S3: Illustration of the choice of the cutoff s^* below which measured enrichments are attributed to unspecific binding. The number $n^3(x)$ of counts in the sequencing data at round $c = 3$ is plotted against the number $n^2(x)$ of counts at round $c - 1 = 2$ for a selection of the Bnab library mixed with the two other libraries against the DNA1 target. An accumulation of sequences with similar enrichments 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 enrichment 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 enrichments are not dominated by sampling noise.

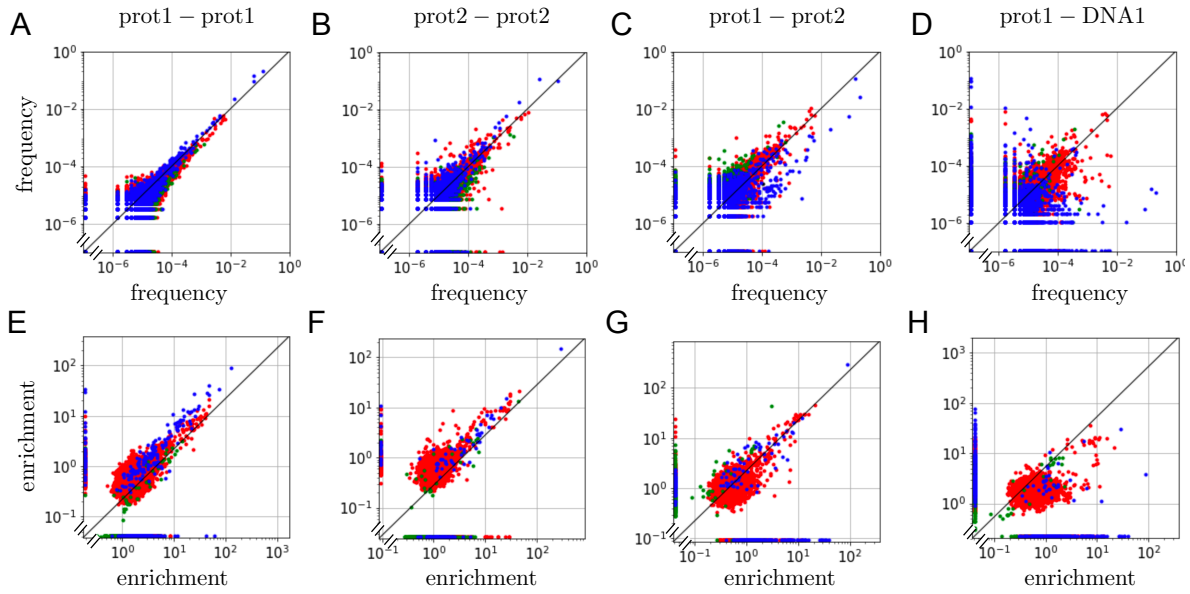


Figure S4: Comparisons between results of replicate and non-replicate experiments. **A.** Comparison of the frequencies $f^3(x) = n^3(x)/\sum_{x'} n^3(x')$ computed after the third cycle ($c = 3$) between two independent replicate experiments where a mixture of the Germ (in blue), Lim (in green) and 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 enrichments $s(x)$ 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 enrichment 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 enrichments 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.

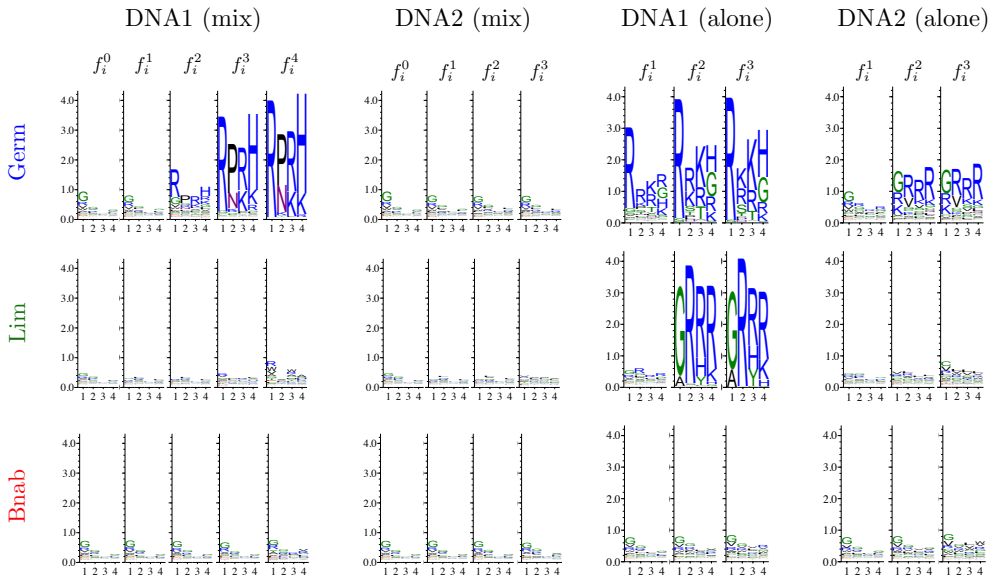


Figure S5: Extension of the figure in the Box to the 3 libraries Germ, Lim, Bnab selected either in a mixture (mix) or on their own (alone) against the DNA1 and DNA2 targets. The sequences logos represent the frequencies $f_i^c(a)$ of amino acids at each successive cycle $c = 0, 1, 2, 3, 4$.

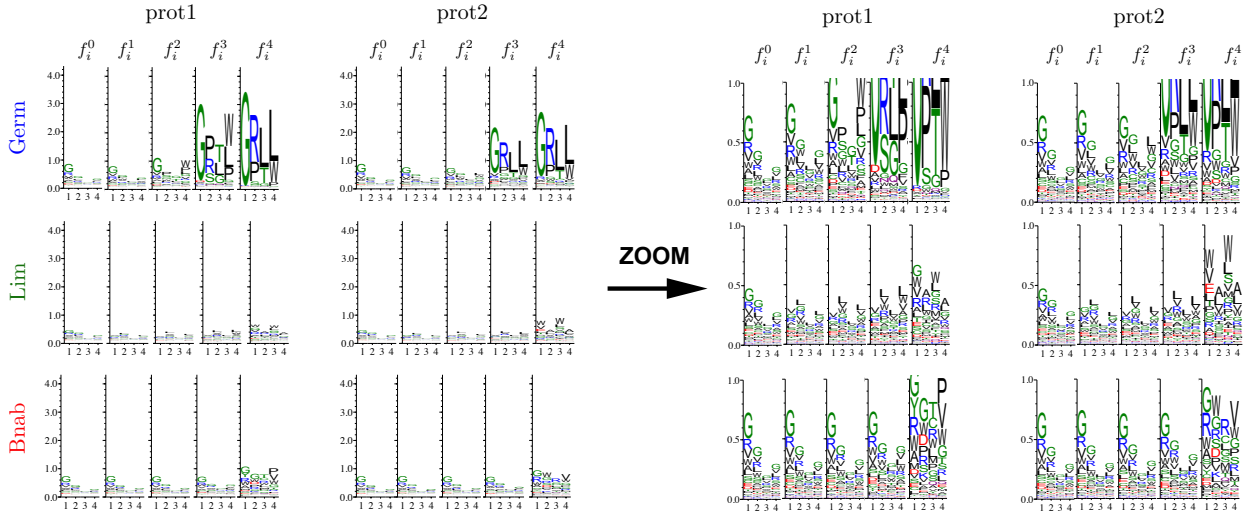


Figure S6: Extension of the figure in the Box to the 3 libraries Germ, Lim, Bnab selected in mixture against the prot1 and prot2 targets. The sequences logos represent the frequencies $f_i^c(a)$ of amino acids at eamotifs ch successive cycle $c = 0, 1, 2, 3, 4$. The data is presented at two different scales for better readability.

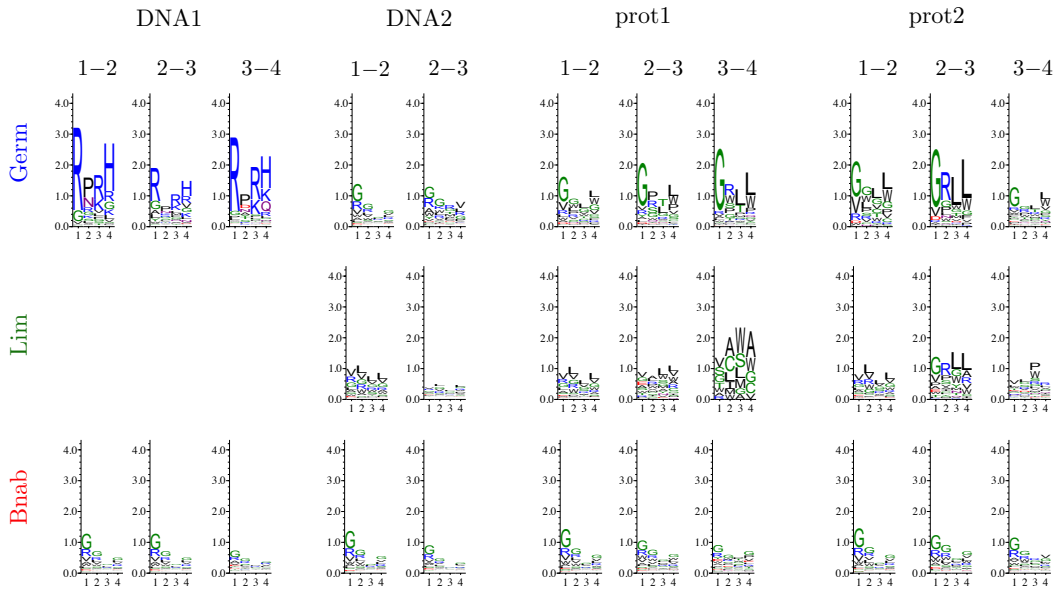


Figure S7: Sequence logos for the enrichments $\tilde{s}(x)$ computed between two successive rounds (1-2, 2-3 or 3-4). The differences between rounds reflect sampling fluctuations.

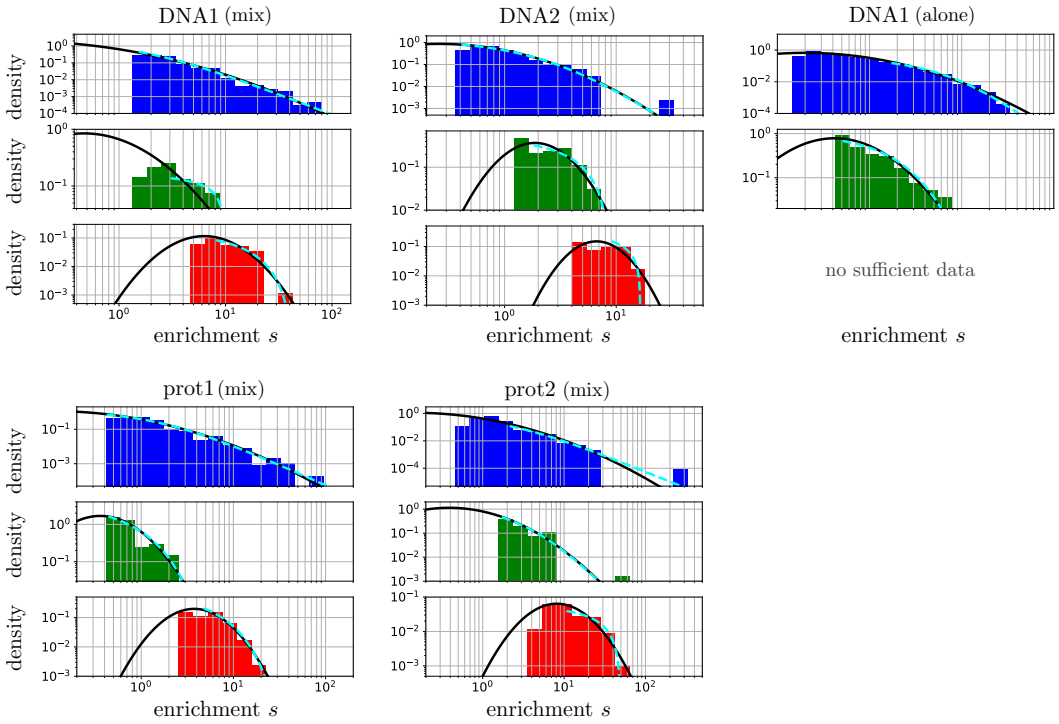


Figure S8: Distributions of enrichments of the three libraries (Germ in blue, Lim in green, Bnab in red) when selected either in a mixture (mix) or on their own (alone) against the different targets. This figure extends Fig. 1A that reports the selection against the DNA1 target of the Germ and Bnab libraries 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.

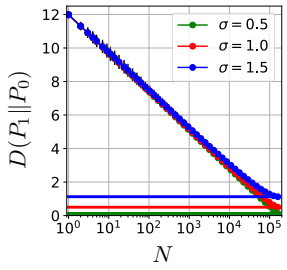


Figure S9: How the estimation of the entropy is biased by finite sampling. 10^5 values were drawn from a log-normal distribution with parameters $\mu = 0$ and $\sigma = 0.5$ (green), 1 (red) and 1.5 (blue). The relative entropy $D(P_1 \| P_0)$ was then estimated using a random subsample of size N. For any $N < 10^5$, this leads to an overestimation of $D(P_1 \| P_0)$ whose actual value $\sigma^2/2$ (see Eq. (14)) is represented by the horizontal lines at the bottom.

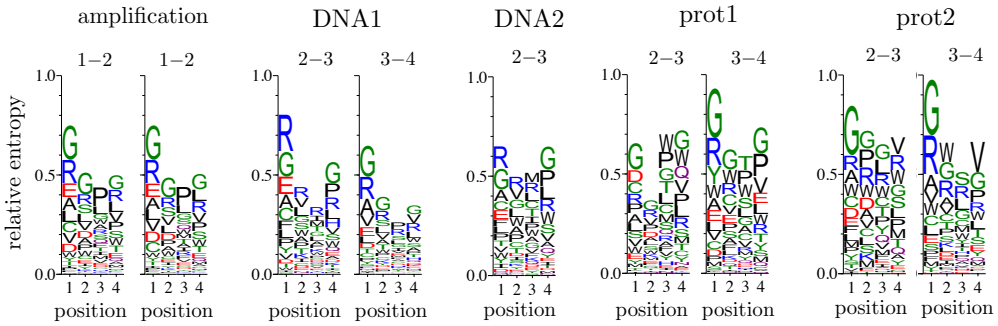


Figure S10: Sequence logos for the enrichments $\tilde{s}_i(a)$ of the Bnab library subject to either amplification only or to amplification and selection for binding against the DNA1, DNA2, prot1 or prot2 targets. The enrichments are computed between the first and second cycles (1-2) or between the third and fourth cycles (3-4); for amplification only, the results of two replicate experiments are shown. The sequence logos of enrichments 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.

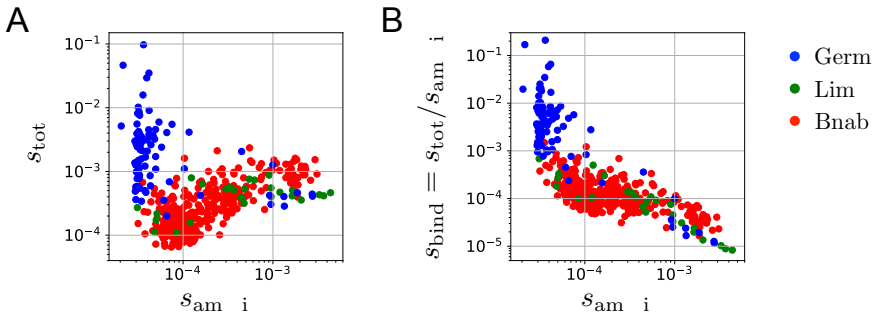


Figure S11: Contribution of amplification biases to the enrichments in selection against the DNA1 target. A separate experiment without any selection for binding was performed to estimate the difference of enrichments arising from the amplification step alone. **A.** The resulting s_{amplif} is here compared to the enrichments s_{tot} from an experiment including a selection for binding. The sequences with top s_{tot}, which all belong to the Germ library (in blue), are among the sequences with lowest s_{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_{tot} from the Lim and Bnab libraries (respectively in green and red), have also top s_{amplif}, which indicate a significant contribution from amplification biases. **B.** The ratio $s_{\mathrm{tot}}/s_{\mathrm{amplif}}$ represents the contribution to enrichment of binding alone. The two selective pressures, binding and amplification, appear here to be orthogonal.

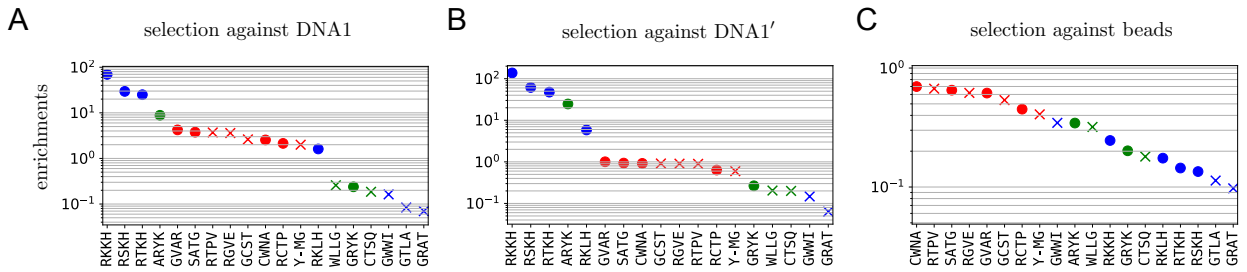


Figure S12: Supplementary experiments with minimal libraries. **A.** Enrichments of top and random sequences from the three libraries, Germ (in blue), Lim (in green) and Bnab (in red), against DNA1. This graph is identical to Fig. 1B. **B.** Results from a replicate experiment using a different stock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

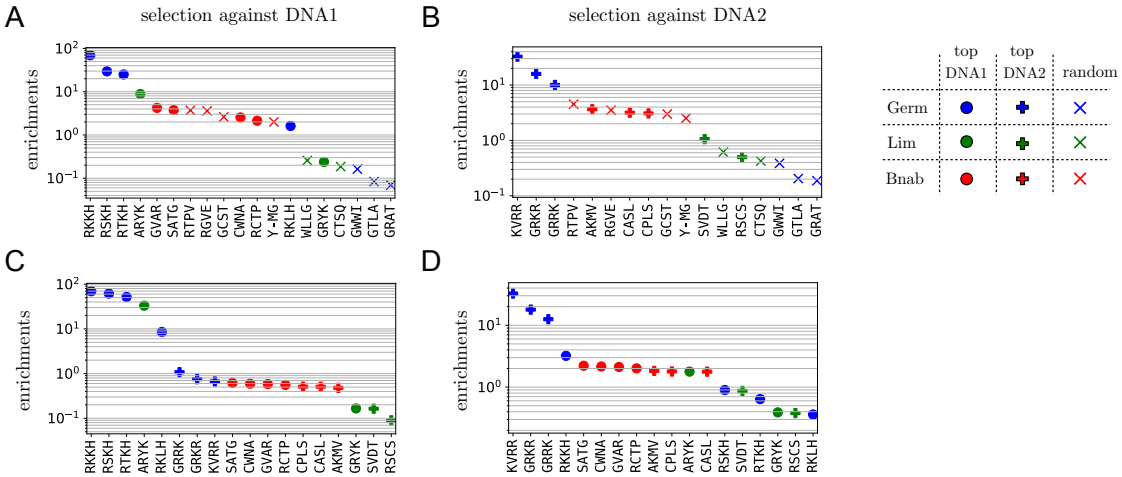


Figure S13: Cross selections with minimal libraries consisting of mixtures of top sequences against the DNA1 target (full circles) and top sequences against the DNA2 target (full crosses). **A,C.** Selection against the DNA1 target (same as Fig. 1B). **B,D.** Selection against the DNA2 target. The results 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).

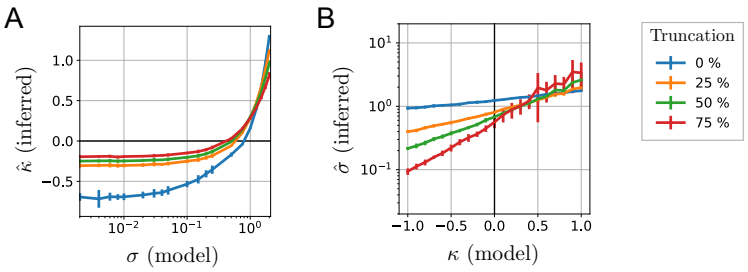


Figure S14: Relation between the parameter σ from log-normal fits and the parameter κ_N from generalized Pareto fits from numerical simulations. **A.** $N = 10^4$ values were drawn from a log-normal distribution with parameters $\mu = 0$ and varying σ (x-axis). The largest 25, 50, 75, 100 % of these 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).

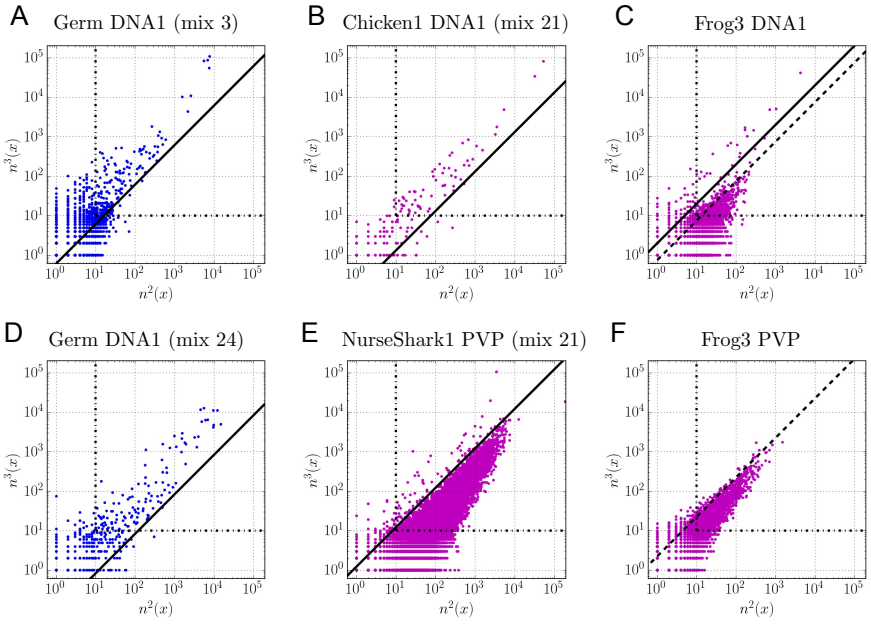


Figure S15: Definition of the threshold s^* above which enrichments s are considered for the experimental results reported here (A) and in Ref. [3] (B-F). As in Figure S3, the definition is based on a comparison between counts at the 2nd and 3rd cycles. The horizontal and vertical lines correspond to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

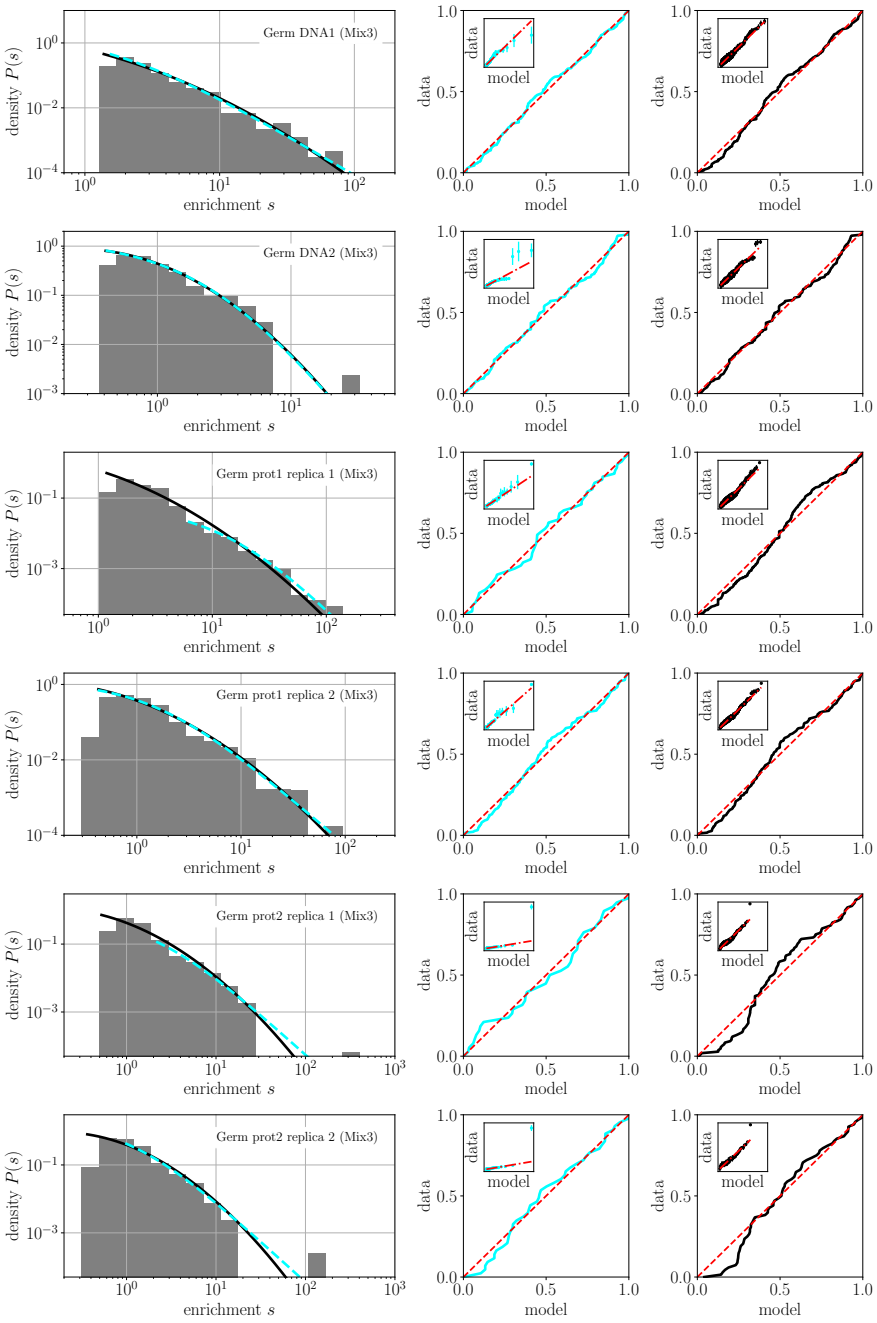


Figure S16: Assessments of the qualities of the fits of the enrichments to generalized Pareto distributions (cyan) and to log-normal distributions (black) for selections of the Germ library. The different graphs correspond to selections against different targets. For the protein targets prot1 and prot2, results from two replicate experiments are presented. All enrichments 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.

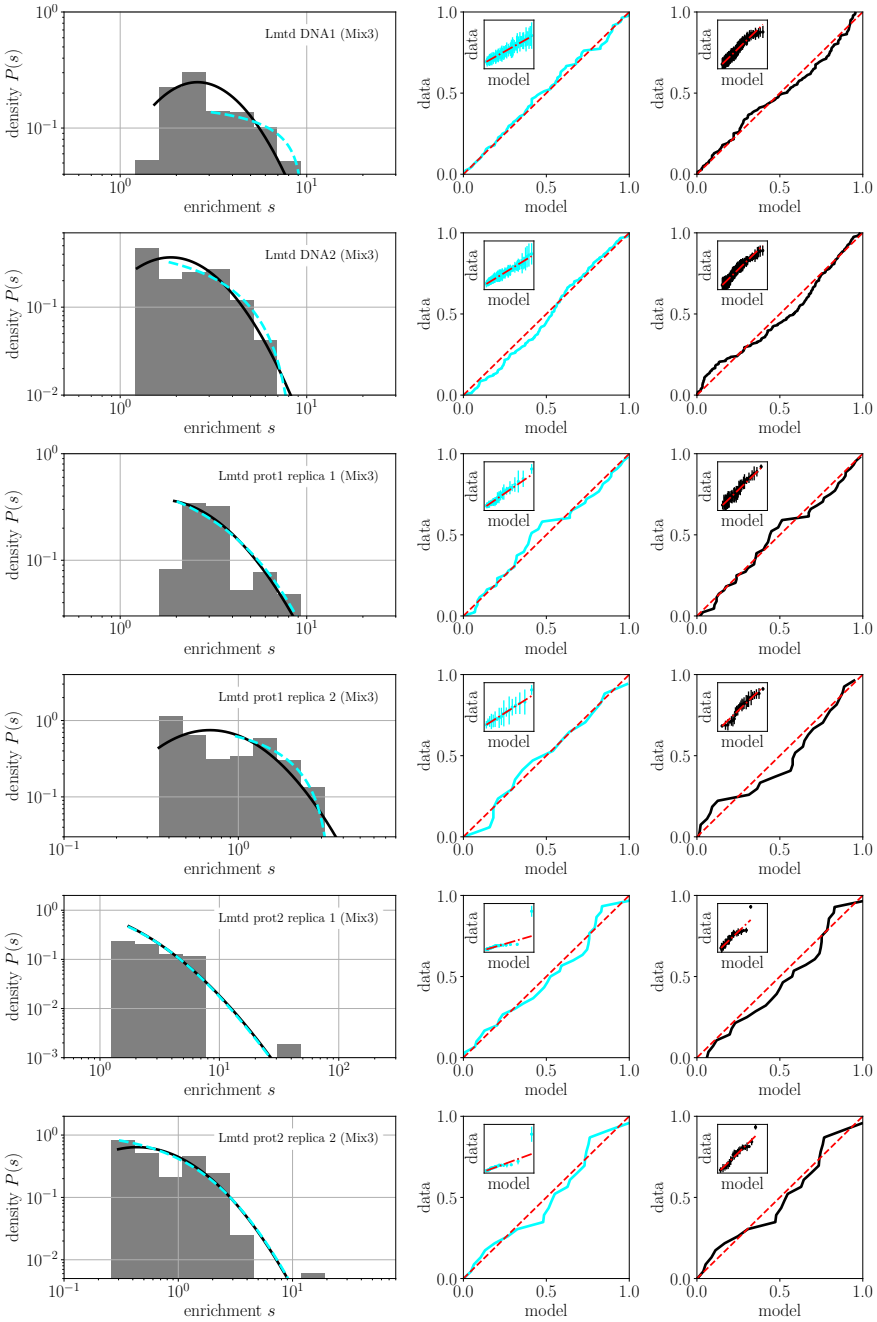


Figure S17: Same as Fig. S16 but for the Lim library instead of the Germ library.

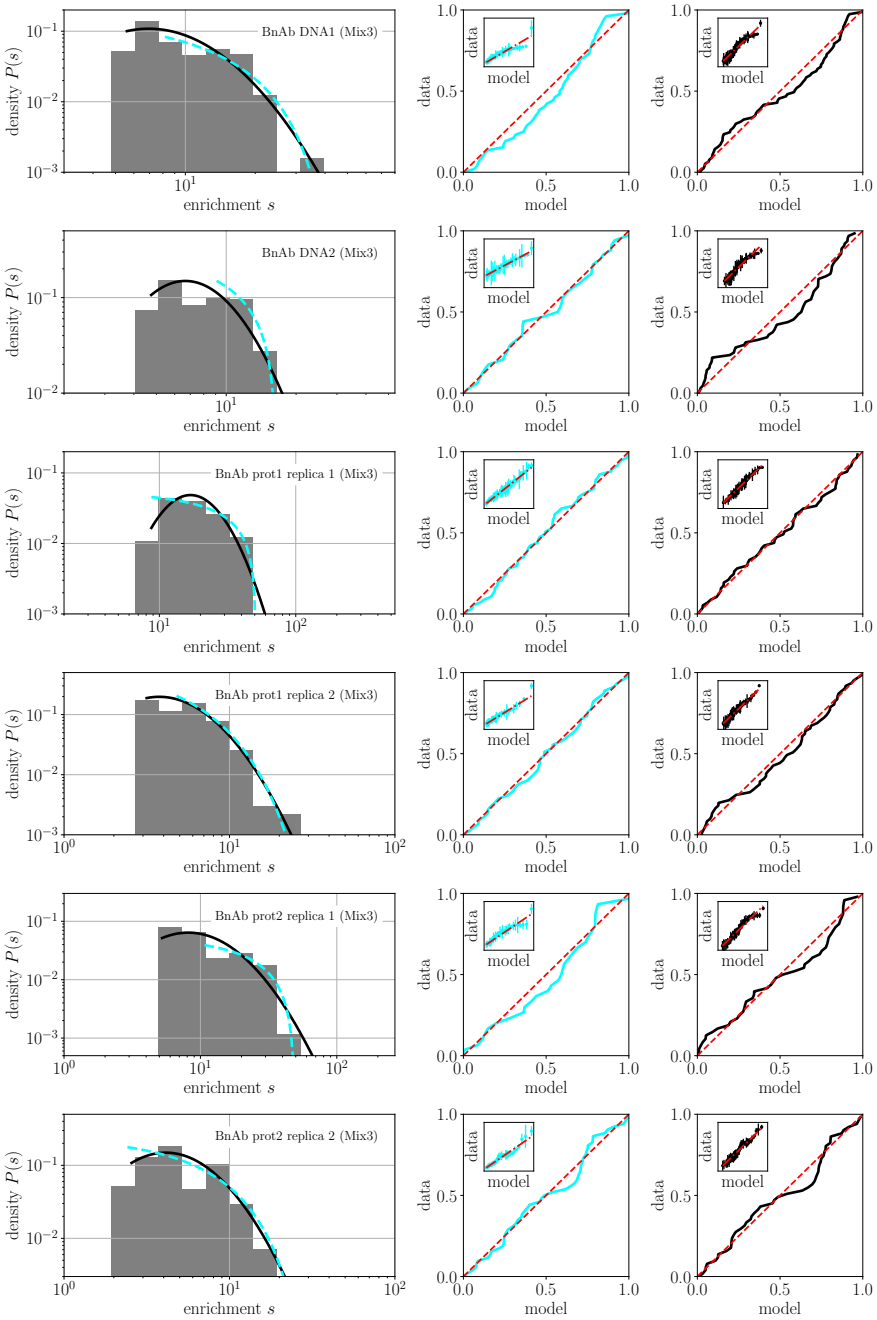


Figure S18: Same as Fig. S16 but for the Bnab library instead of the Germ library.

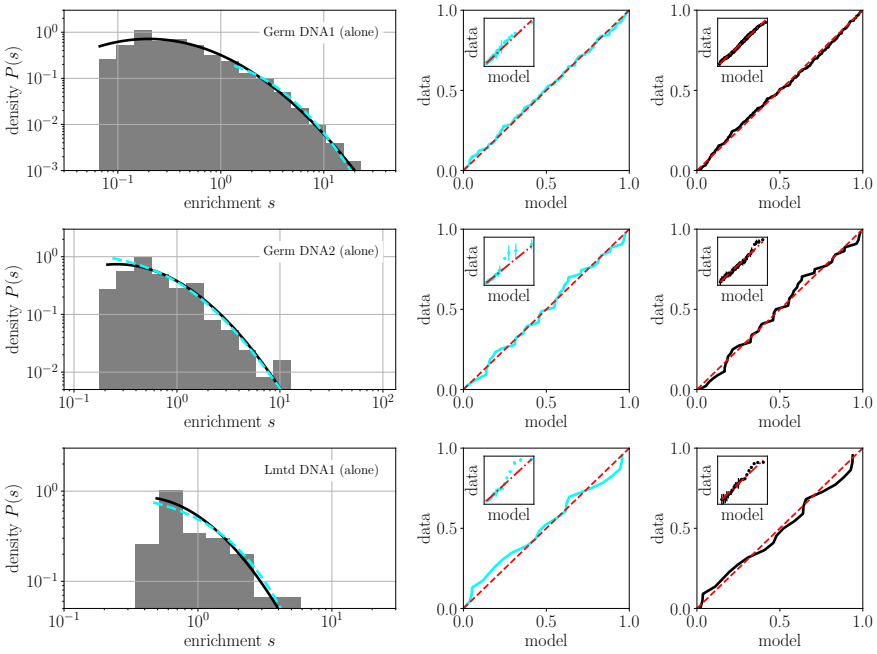


Figure S19: Same as Fig. S16 for the Germ library selected in isolation rather in a mixture with the two other libraries.

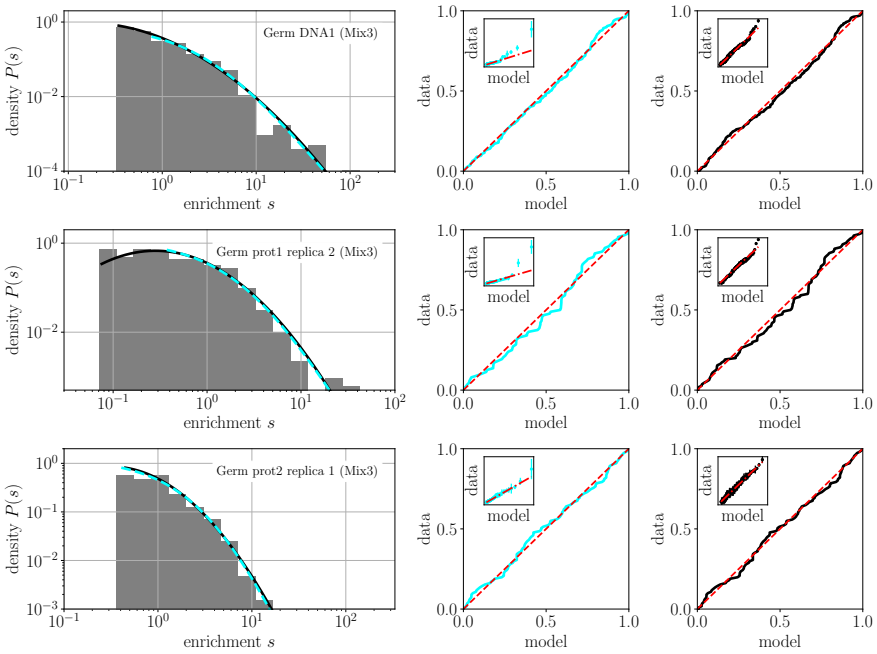


Figure S20: Same as Fig. S16 but for enrichments computed from a comparison between the 3rd and 4th cycle instead of the 2nd and 3rd cycle.

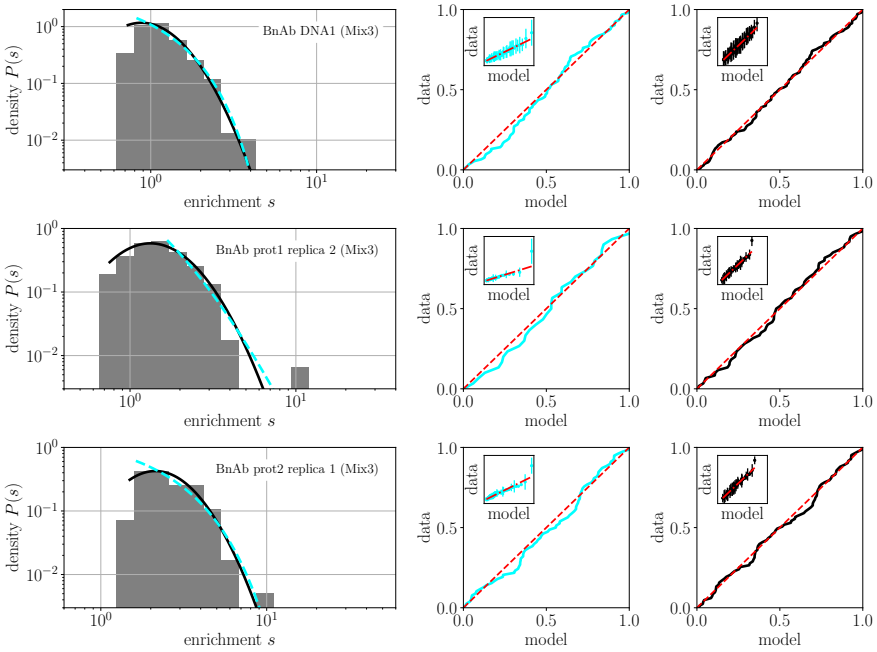


Figure S21: Same as Fig. S20 (enrichments computed from a comparison between the 3rd and 4th cycle) but for the Bnab library instead of the Germ library.

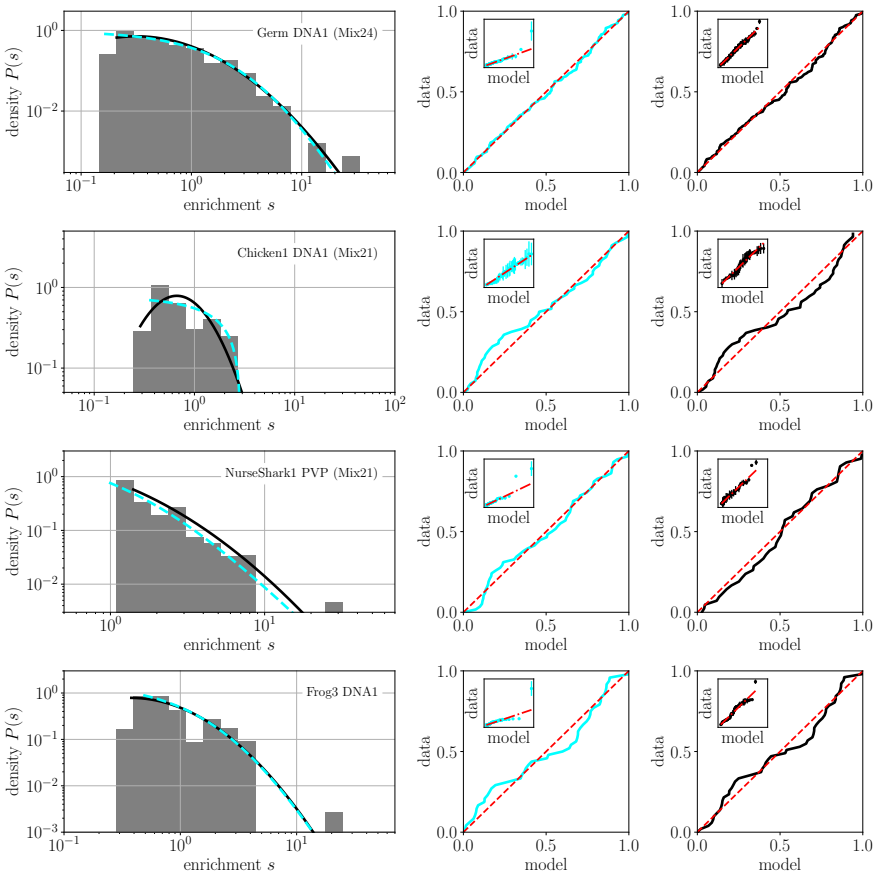


Figure S22: Same as Fig. S20 but for the experimental results reported in Ref. [3].

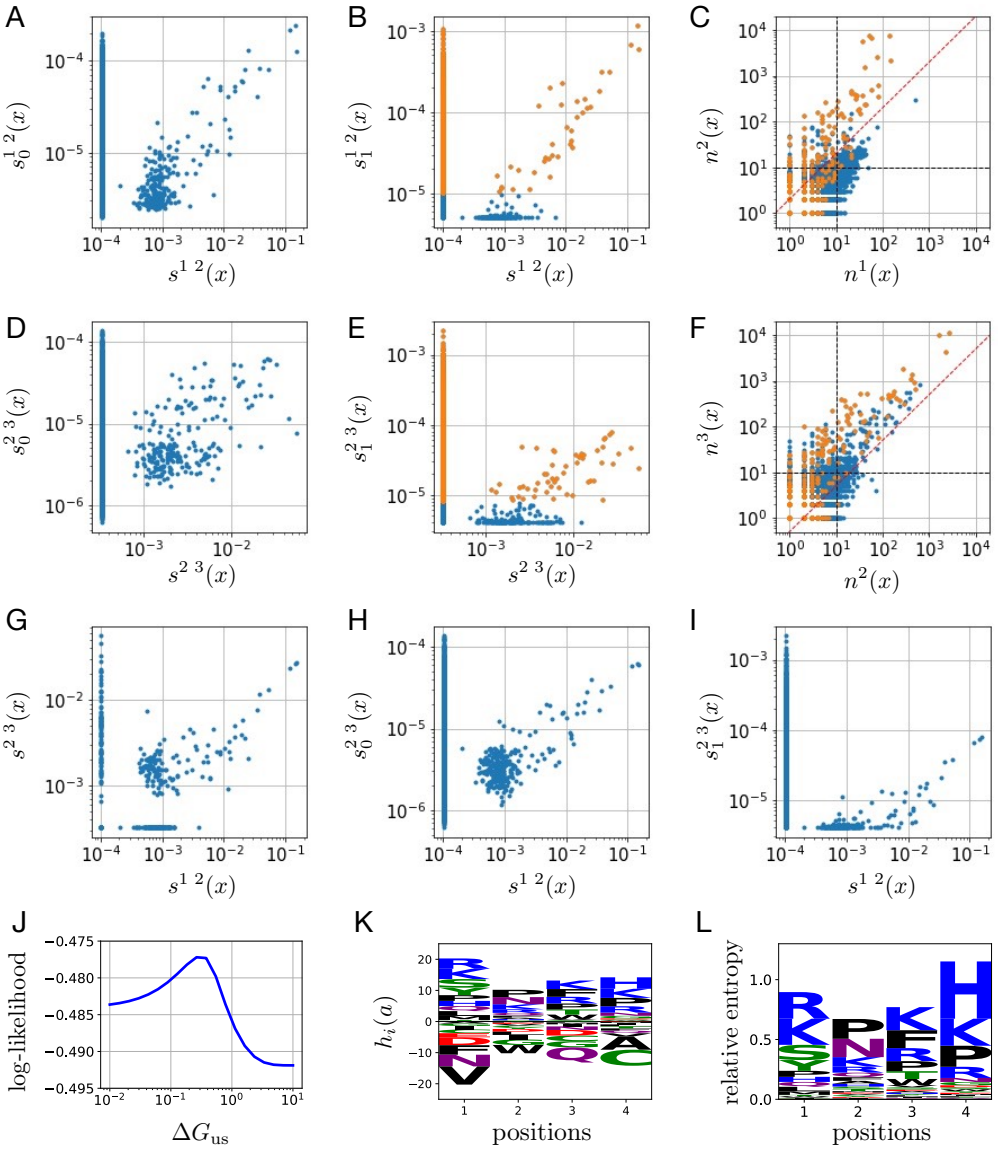


Figure S23: Analysis of data from the Germ library selected against the DNA1 target (in Mix) with the stochastic model presented in Sec. 3.3. The data consists in the counts $n^1(x)$, $n^2(x)$, $n^3(x)$ at the different rounds (panels C and F), from which enrichments are inferred in different ways that we compare. As in the main text, we define $s^{1\text{-}2}(x) \propto n^2(x)/n^1(x)$ when $n^1(x) \geq 10$ and $n^2(x) \geq 10$, and $s^{2\text{-}3}(x) \propto n^3(x)/n^2(x)$ when $n^2(x) \geq 10$ and $n^3(x) \geq 10$: they are shown in panel G to give consistent results (undefined values are represented as small values). Alternatively, we can infer enrichments by maximum likelihood using the model of Sec. 3.3. For each successive rounds c-$(c+1)$ with $c = 1$ or 2, two solutions are considered: $s_0^{c\text{-}(c+1)}(x)$ where unspecific binding is neglected ($\Delta G_{\rm us} = \infty$) and $s_1^{c\text{-}(c+1)}(x)$ where it is not ($\Delta G_{\rm us}$ treated as variable in addition to the $h_i(a)$). They are compared to $s^{c\text{-}(c+1)}$ in panels A, B, D, E. In B and E, where unspecific binding is present, the sequences that are predicted to be selected through specific binding ($e^{-\beta G(x)} > e^{-\beta G_{us}}$ in Eq. (23)) are represented in orange. When considering data between rounds 1-2, a good agreement is found between $s^{1\text{-}2}(x)$ and $s_1^{1\text{-}2}(x)$ (panel B) and the sequences identified as binding specifically (in orange) correspond indeed to those above a threshold, $s^{1\text{-}2}(x) > s^*$(panel C). This is not the case when considering the data between rounds 2-3 where the model predicts many sequences with high enrichments $s_1^{2\text{-}3}(x)$ that are not reported in $s^{2\text{-}3}(x)$ (panel E). In this case, the solution without non-specific binding $s_0^{2\text{-}3}(x)$ appears to be more relevant. This is confirmed in panels H and I where $s^{1\text{-}2}(x)$ is seen to correlate better with $s_0^{2\text{-}3}(x)$ than with $s_1^{2\text{-}3}(x)$. Panel J represents the maximum value of the log-likelihood for fixed values of $\Delta G_{\rm us}$, showing the presence of a non-trivial optimum (data from rounds 1-2). The fields $h_i(a)$ of this model are shown in panel K in the zero-sum gauge where $\sum_{a=1}^{q} h_i(a) = 0$ for all i. The same information can also be represented in the form a sequence logo (panel L), to be compared to the sequence logo obtained from $s(x)$ (Fig. 2B, Germ-DNA1).

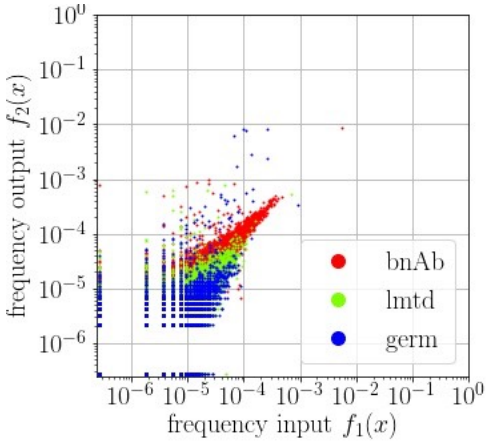


Figure S24: Relative frequencies at round 1 (x-axis) and round 2 (y-axis) of sequences from the 3 libraries, Germ (blue), Lim (green) and Bnab (red) when selected in mixture against the DNA1 target. This figure shows that each library has a different background noise.

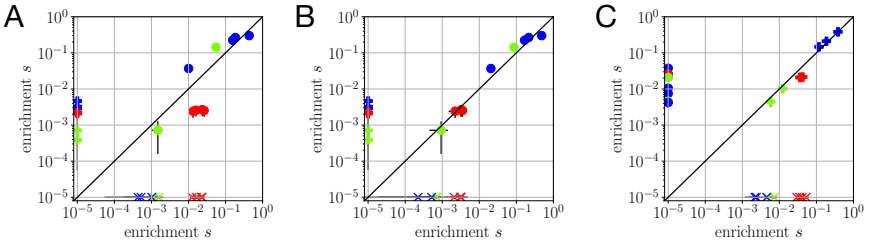


Figure S25: Reproducibility of enrichments inferred from experiments with mini-libraries. **A.** Enrichments from Fig. S13A versus Fig. S13C: the results from the two experiments are highly reproducible except for the bnAb sequences in red. This difference is due to the different batches of beads used in these two experiments. **B.** Enrichments from Fig. S13B versus Fig. S13D. Here the two experiments use the same batch of beads and the inferred enrichments are all very reproducible. **C.** Enrichments from Fig. S12B versus Fig. S12A, showing again high reproducibility. Error bars are enlarged 20 times to make them visible.

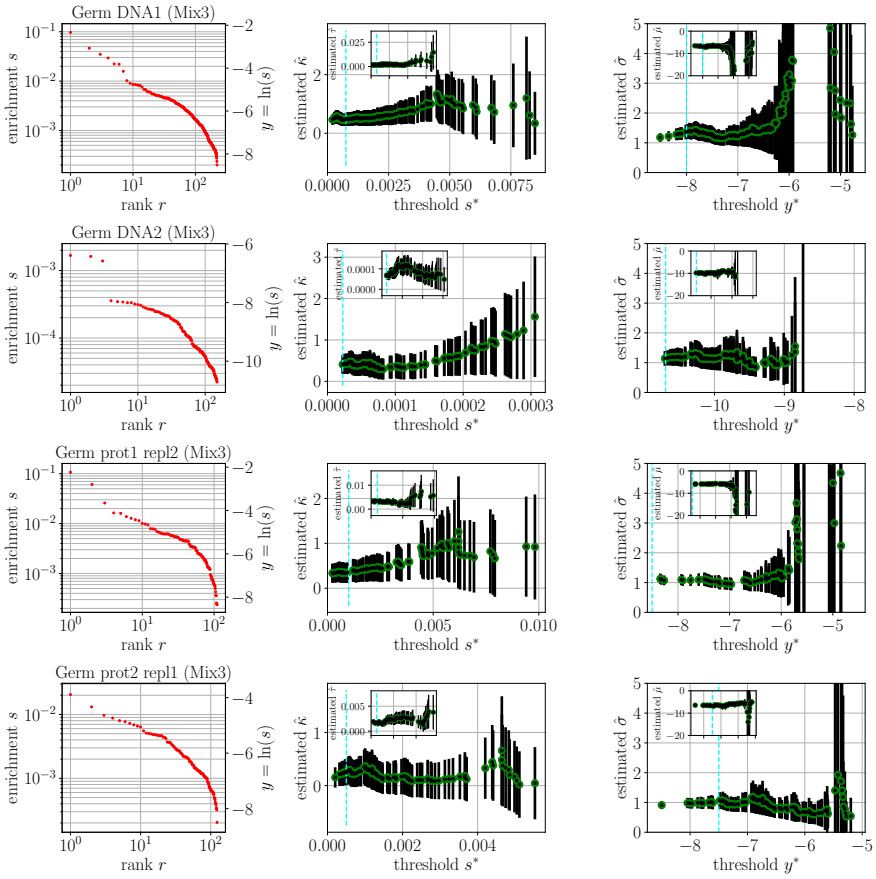


Figure S26: Dependence of the inferred values of $\hat{\kappa}$, when fitting the tail of the distribution of enrichments to a generalized Pareto distribution, and $\hat{\sigma}$, when fitting them to a truncated log-normal distribution, on the choice of the threshold s^* or $y^* = \ln(s^*)$ that defines the tail. Here for the Germ library selected against different targets. When the threshold is too large, very few data points are left and the error bars, obtained from the Fisher information matrix via the Cramér-Rao bound, are large. In any case, however, the estimation of $\hat{\kappa}$ and $\hat{\sigma}$ is consistent across a range of values of the thresholds.

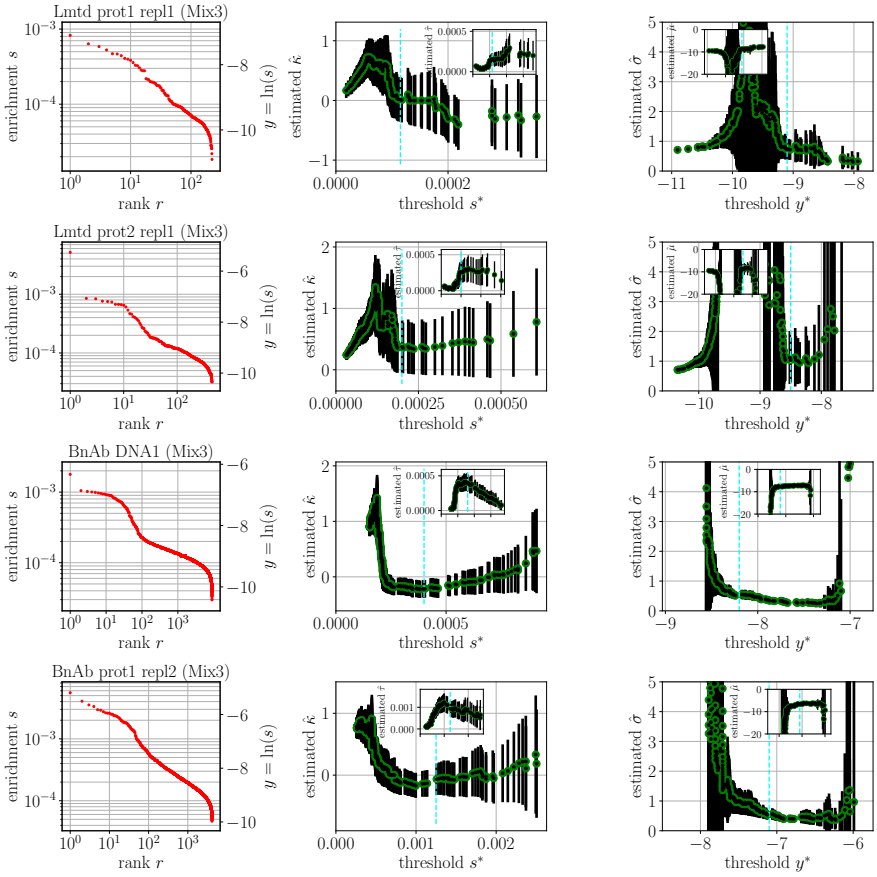


Figure S27: Similar to Fig. S26 but for the Lim and Bnab libraries.

Bibliography

[1] S. Boyer, D. Biswas, A. Kumar Soshee, N. Scaramozzino, C. Nizak, O. Rivoire, A. K. Soshee, N. Scaramozzino, C. Nizak, O. Rivoire, B. I. Shraiman, 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. arXiv:1509.02450, doi:10.1073/pnas.1517813113.

[2] 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. doi:10.1021/bi800374q.

[3] J. Zimmermann, E. L. Oakman, I. F. Thorpe, X. Shi, P. Abbyad, C. L. Brooks, S. G. Boxer, F. E. Romesberg, Antibody evolution constrains conformational heterogeneity by tailoring protein dynamics, Proceedings of the National Academy of Sciences of the United States of America 103 (37) (2006) 13722–13727. doi:10.1073/pnas.0603282103.

[4] R. Jimenez, G. Salazar, J. Yin, T. Joo, F. E. Romesberg, Protein dynamics and the immunological evolution of molecular recognition, Proceedings of the National Academy of Sciences of the United States of America 101 (11) (2004) 3803–3808. doi:10.1073/pnas.0305745101.

[5] J. S. Shaffer, P. L. Moore, M. Kardar, A. K. Chakraborty, Optimal immunization cocktails can promote induction of broadly neutralizing Abs against highly mutable pathogens, Proceedings of the National Academy of Sciences of the United States of America 113 (45) (2016) E7039–E7048. doi:10.1073/pnas.1614940113.

[6] S. Wang, J. Mata-Fink, B. Kriegsman, M. Hanson, D. J. Irvine, H. N. Eisen, D. R. Burton, K. D. Wittrup, M. Kardar, A. K. Chakraborty, Manipulating the selection forces during affinity maturation to generate cross-reactive HIV antibodies, Cell 160 (4) (2015) 785–797. arXiv:15334406, doi:10.1016/j.cell.2015.01.027.

[7] C. R. Darwin, The Origin of Species, 6th Edition, John Murray, 1872.

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

[9] O. Rivoire, Parsimonious evolutionary scenario for the origin of allostery and coevolution patterns in proteins, Physical Review E 100 (3) (2019) 1–20. arXiv:1812.01524, doi:10.1103/PhysRevE.100.032411.

[10] 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) 1–10. arXiv:1408.3240, doi:10.1103/PhysRevE.91.042704.

[11] J. D. Bloom, S. T. Labthavikul, C. R. Otey, F. H. Arnold, Protein stability promotes evolvability, Proceedings of the National Academy of Sciences of the United States of America 103 (15) (2006) 5869–5874. doi:10.1073/pnas.0510098103.

[12] N. Kashtan, U. Alon, Spontaneous evolution of modularity and network motifs, Proceedings of the National Academy of Sciences of the United States of America 102 (39) (2005) 13773–13778. doi:10.1073/pnas.0503610102.

[13] D. J. Earl, M. W. Deem, Evolvability is a selectable trait, Proceedings of the National Academy of Sciences of the United States of America 101 (32) (2004) 11531–11536. doi:10.1073/pnas.0404656101.

[14] A. Wagner, Robustness and evolvability in living systems, Vol. 24, Princeton University Press, 2013.

[15] M. Pigliucci, Is evolvability evolvable?, Nature Reviews Genetics 9 (1) (2008) 75–82. doi:10.1038/nrg2278.

[16] G. P. Wagner, L. Altenberg, Complex adaptations and the evolution of evolvability, Evolution 50 (3) (1996) 967–976. doi:10.1111/j.1558-5646.1996.tb02339.x.

[17] J. Tubiana, S. Cocco, R. Monasson, Learning protein constitutive motifs from sequence data, eLife 8 (2019) 1–19. arXiv:1803.08718, doi:10.7554/eLife.39397.

[18] K. Rajewsky, Clonal selection and learning in the antibody system (1996). arXiv:381(6585):751-8., doi:10.1038/381751a0.

[19] G. Altan-Bonnet, T. Mora, A. M. Walczak, Quantitative immunology for physicists, Physics Reports 849 (2020) 1–83. arXiv:1907.03891, doi:10.1016/j.physrep.2020.01.001.

[20] I. Mikell, D. N. Sather, S. A. Kalams, M. Altfeld, G. Alter, L. Stamatatos, Characteristics of the earliest cross-neutralizing antibody response to HIV-1, PLoS Pathogens 7 (1) (2011). doi:10.1371/journal.ppat.1001251.

[21] J. L. Xu, M. M. Davis, Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities, Immunity 13 (1) (2000) 37–45. doi:10.1016/S1074-7613(00)00006-6.

[22] T. B. Kepler, Codon bias and plasticity in immunoglobulins, Molecular Biology and Evolution 14 (6) (1997) 637–643. doi:10.1093/oxfordjournals.molbev.a025803.

[23] A. Mayer, T. Mora, O. Rivoire, A. M. Walczak, Diversity of immune strategies explained by adaptation to pathogen statistics, Proceedings of the National Academy of Sciences of the United States of America 113 (31) (2016) 8630–8635. arXiv:1511.08836, doi:10.1073/pnas.1600663113.

[24] G. Georgiou, G. C. Ippolito, J. Beausang, C. E. Busse, H. Wardemann, S. R. Quake, The promise and challenge of high-throughput sequencing of the antibody repertoire, Nature Biotechnology 32 (2) (2014) 158–168. doi:10.1038/nbt.2782.

[25] A. Nourmohammad, J. Otwinowski, J. B. Plotkin, Host-Pathogen Coevolution and the Emergence of Broadly Neutralizing Antibodies in Chronic Infections, PLoS Genetics 12 (7) (2016) 1–23. arXiv:1512.06296, doi:10.1371/journal.pgen.1006171.

[26] E. A. Padlan, Anatomy of the antibody molecule, Molecular Immunology 31 (3) (1994) 169–217. doi:10.1016/0161-5890(94)90001-9.

[27] W. Nolting, Grundkurs Theoretische Physik 2: Analytische Mechanik, Springer-Verlag, 2014.

[28] P. A. Romero, F. H. Arnold, Exploring protein fitness landscapes by directed evolution, Nature Reviews Molecular Cell Biology 10 (12) (2009) 866–876. doi:10.1038/nrm2805.

[29] D. A. Drummond, J. D. Bloom, C. Adami, C. O. Wilke, F. H. Arnold, Why highly expressed proteins evolve slowly, Proceedings of the National Academy of Sciences of the United States of America 102 (40) (2005) 14338–14343.

[30] V. Sachdeva, K. Husain, J. Sheng, S. Wang, A. Murugan, Tuning environmental timescales to evolve and maintain generalists, Proceedings of the National Academy of Sciences of the United States of America 117 (23) (2020) 12693–12699. arXiv:1906.11924, doi:10.1073/pnas.1914586117.

[31] A. Crombach, P. Hogeweg, Evolution of evolvability in gene regulatory networks, PLoS Computational Biology 4 (7) (2008) 1–13. doi:10.1371/journal.pcbi.1000112.

[32] 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). doi:10.1371/journal.pcbi.1000206.

[33] L. Ancel Meyers, F. D. Ancel, M. Lachmann, Evolution of Genetic Potential, PLoS Computational Biology 1 (3) (2005) 236–243. doi:10.1371/journal.pcbi.0010032.

[34] S. Wang, L. Dai, Evolving generalists in switching rugged landscapes, PLoS Computational Biology 15 (10) (2019) 1–21. doi:10.1371/journal.pcbi.1007320.

[35] A. Aharoni, L. Gaidukov, O. Khersonsky, S. M. Q. Gould, C. Roodveldt, D. S. Tawfik, The 'evolvability' of promiscuous protein functions, Nature Genetics 37 (1) (2005) 73–76. doi:10.1038/ng1482.

[36] O. Rivoire, K. A. Reynolds, R. Ranganathan, Evolution-Based Functional Decomposition of Proteins, PLoS Computational Biology 12 (6) (2016) 1–27. doi:10.1371/journal.pcbi.1004817.

[37] N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan, Protein Sectors: Evolutionary Units of Three-Dimensional Structure, Cell 138 (4) (2009) 774–786. doi:10.1016/j.cell.2009.07.038.

[38] M. Stern, C. Arinze, L. Perez, S. E. Palmer, A. Murugan, Supervised learning through physical changes in a mechanical system, Proceedings of the National Academy of Sciences of the United States of America 117 (26) (2020) 202000807. doi:10.1073/pnas.2000807117.

[39] M. Stern, M. B. Pinson, A. Murugan, Learned multi-stability in mechanical networks (2019). arXiv:1902.08317.

[40] D. Kern, E. R. Zuiderweg, The role of dynamics in allosteric regulation, Current Opinion in Structural Biology 13 (6) (2003) 748–757. doi:10.1016/j.sbi.2003.10.008.

[41] M. S. Celej, G. G. Montich, G. D. Fidelio, Protein stability induced by ligand binding correlates with changes in protein flexibility, Protein Science 12 (7) (2003) 1496–1506. doi:10.1110/ps.0240003.

[42] I. F. Thorpe, C. L. Brooks, Molecular evolution of affinity and flexibility in the immune system, Proceedings of the National Academy of Sciences of the United States of America 104 (21) (2007) 8821–8826. doi:10.1073/pnas.0610064104.

[43] J. Yin, A. E. Beuscher, 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. doi:10.1016/S0022-2836(03)00631-4.

[44] 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. doi:10.1016/S1074-7613(00)00061-3.

[45] L. T. Chong, Y. Duan, L. Wang, I. Massova, P. A. Kollman, Molecular dynamics and free-energy calculations applied to affinity maturation in antibody 48G7, Proceedings of the National Academy of Sciences of the United States of America 96 (25) (1999) 14330–14335. doi:10.1073/pnas.96.25.14330.

[46] 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. doi:10.1126/science.276.5319.1665.

[47] T. Sagawa, M. Oda, M. Ishimura, K. Furukawa, T. Azuma, Thermodynamic and kinetic aspects of antibody evolution during the immune response to hapten, Molecular Immunology 39 (13) (2003) 801–808. doi:10.1016/S0161-5890(02)00282-1.

[48] A. L. Notkins, Polyreactivity of antibody molecules, Trends in Immunology 25 (4) (2004) 174–179. doi:10.1016/j.it.2004.02.004.

[49] J. R. Jeliazkov, A. Sljoka, D. Kuroda, N. Tsuchimura, N. Katoh, K. Tsumoto, J. J. Gray, Repertoire analysis of antibody CDR-H3 loops suggests affinity maturation does not typically result in rigidification, Frontiers in Immunology 9 (2018) 413. doi:10.3389/fimmu.2018.00413.

[50] 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) 1–23. doi:10.1371/journal.pcbi.1004327.

[51] T. M. Davenport, J. Gorman, M. G. Joyce, T. Zhou, C. Soto, M. Guttman, S. Moquin, Y. Yang, B. Zhang, N. A. Doria-Rose, S. L. Hu, J. R. Mascola, P. D. Kwong, K. K. Lee, Somatic Hypermutation-Induced Changes in the Structure and Dynamics of HIV-1 Broadly Neutralizing Antibodies, Structure 24 (8) (2016) 1346–1357. doi:10.1016/j.str.2016.06.012.

[52] R. Henderson, B. E. Watts, H. N. Ergin, K. Anasti, R. Parks, S. M. Xia, A. Trama, H. X. Liao, K. O. Saunders, M. Bonsignori, K. Wiehe, B. F. Haynes, S. M. Alam, Selection of immunoglobulin elbow region mutations impacts interdomain conformational flexibility in HIV-1 broadly neutralizing antibodies, Nature Communications 10 (1) (2019). doi:10.1038/s41467-019-08415-7.

[53] P. Koenig, C. V. Lee, B. T. Walters, V. Janakiraman, J. Stinson, T. W. Patapoff, G. Fuh, Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding, Proceedings of the National Academy of Sciences of the United States of America 114 (4) (2017) E486–E495. doi:10.1073/pnas.1613231114.

[54] V. Ovchinnikov, J. E. Louveau, J. P. Barton, M. Karplus, A. K. Chakraborty, Role of framework mutations and antibody flexibility in the evolution of broadly neutralizing antibodies, eLife 7 (2018) 1–24. doi:10.7554/eLife.33038.

[55] Á. Tóth-Petróczy, D. S. Tawfik, The robustness and innovability of protein folds, Current Opinion in Structural Biology 26 (1) (2014) 131–138. doi:10.1016/j.sbi.2014.06.007.

[56] E. Dellus-Gur, A. Toth-Petroczy, 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. doi:10.1016/j.jmb.2013.03.033.

[57] S. Bershtein, M. Segal, R. Bekerman, N. Tokuriki, D. S. Tawfik, Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein, Nature 444 (7121) (2006) 929–932. doi:10.1038/nature05385.

[58] M. C. Julian, L. Li, S. Garde, R. Wilen, P. M. Tessier, Efficient affinity maturation of antibody variable domains requires co-selection of compensatory mutations to maintain thermodynamic stability, Scientific Reports 7 (2017) 1–13. doi:10.1038/srep45259.

[59] M. Heyne, N. Papo, J. M. Shifman, Generating quantitative binding landscapes through fractional binding selections combined with deep sequencing and data normalization, Nature Communications 11 (1) (2020) 2–8. doi:10.1038/s41467-019-13895-8.

[60] R. M. Adams, J. B. Kinney, A. M. Walczak, T. Mora, Epistasis in a Fitness Landscape Defined by Antibody-Antigen Binding Free Energy, Cell Systems 8 (1) (2019) 86–93. doi:10.1016/j.cels.2018.12.004.

[61] J. B. Kinney, G. Tkačik, C. G. Callan, Precise physical models of protein-DNA interaction from high-throughput data, Proceedings of the National Academy of Sciences of the United States of America 104 (2) (2007) 501–506. doi:10.1073/pnas.0609908104.

[62] 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 of the United States of America 90 (8) (1993) 3715–3719. doi:10.1073/pnas.90.8.3715.

[63] 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. doi:10.1073/pnas.1714376115.

[64] Y. Zhao, D. Granas, G. D. Stormo, Inferring binding energies from selected binding sites, PLoS Computational Biology 5 (12) (2009). doi:10.1371/journal.pcbi.1000590.

[65] M. Djordjevic, A. M. Sengupta, B. I. Shraiman, A biophysical approach to transcription factor binding site discovery, Genome Research 13 (11) (2003) 2381–2390. doi:10.1101/gr.1271603.

[66] U. Gerland, J. D. Moroz, T. Hwa, Physical constraints and functional characteristics of transcription factor-DNA interaction, Proceedings of the National Academy of Sciences of the United States of America 99 (19) (2002) 12015–12020. arXiv:0112083, doi:10.1073/pnas.192693599.

[67] H. X. Zhou, Rate theories for biologists, Quarterly Reviews of Biophysics 43 (2) (2010) 219–293. doi:10.1017/S0033583510000120.

[68] J. Otwinowski, Biophysical inference of epistasis and the effects of mutations on protein stability and function, Molecular Biology and Evolution 35 (10) (2018) 2345–2354. arXiv:1802.08744, doi:10.1093/molbev/msy141.

[69] W. Greiner, L. Neise, H. Stöcker, Thermodynamics and statistical mechanics, Springer Science & Business Media, 2012.

[70] R. P. Bhattacharyya, A. Reményi, B. J. Yeh, W. A. Lim, Domains, Motifs, and Scaffolds: The Role of Modular Interactions in the Evolution and Wiring of Cell Signaling Circuits, Annual Review of Biochemistry 75 (1) (2006) 655–680. doi:10.1146/annurev.biochem.75.103004.142710.

[71] M. C. Good, J. G. Zalatan, W. A. Lim, Scaffold proteins: Hubs for controlling the flow of cellular information, Science 332 (6030) (2011) 680–686. doi:10.1126/science.1198701.

[72] R. H. Garrett, C. M. Grisham, Biochemistry, Fourth Edition, Cengage Learning, 2010. doi:10.1016/s0968-0004(00)89112-4.

[73] C. E. Tinberg, S. D. Khare, J. Dou, L. Doyle, J. W. Nelson, A. Schena, W. Jankowski, C. G. Kalodimos, K. Johnsson, B. L. Stoddard, D. Baker, Computational Design of Ligand Binding Proteins with High Affinity and Selectivity, Nature 501 (7466) (2013) 212–216.

[74] L. Giger, S. Caner, R. Obexer, P. Kast, D. Baker, N. Ban, D. Hilvert, Evolution of a designed retro-aldolase leads to complete active site remodeling, Nature Chemical Biology 9 (8) (2013) 494–498. doi:10.1038/nchembio.1276.

[75] L. Jiang, E. A. Althoff, F. R. Clemente, L. Doyle, D. Röthlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, C. F. Barbas III, D. Hilvert, K. N. Houk, B. L. Stoddard, D. Baker, De Novo Computational Design of Retro-Aldol Enzymes, Science 319 (5868) (2008) 1387–1391.

[76] M. L. Mehta, F. J. Dyson, Statistical theory of the energy levels of complex systems, Journal of Mathematical Physics 4 (5) (1963).

[77] T. Guhr, A. Müller-Groeling, H. A. Weidenmüller, Random-matrix theories in quantum physics: common concepts, Physics Reports 299 (4-6) (1998) 189–425.

[78] A. Sakata, K. Hukushima, K. Kaneko, Funnel landscape and mutational robustness as a result of evolution under thermal noise, Physical Review Letters 102 (14) (2009) 1–5. arXiv:0807.1216, doi:10.1103/PhysRevLett.102.148101.

[79] L. M. Childs, E. B. Baskerville, S. Cobey, Trade-offs in antibody repertoires to complex antigens, Philosophical Transactions of the Royal Society B: Biological Sciences 370 (1676) (2015) 1–10. doi:10.1098/rstb.2014.0245.

[80] A. Sakata, K. Hukushima, K. Kaneko, Replica symmetry breaking in an adiabatic spin-glass model of adaptive evolution, Europhysics Letters 99 (6) (2012). arXiv:1111.5770, doi:10.1209/0295-5075/99/68004.

[81] J. Sun, D. J. Earl, M. W. Deem, Glassy dynamics in the adaptive immune response prevents autoimmune disease, Physical Review Letters 95 (14) (2005) 1–4. arXiv:0508020, doi:10.1103/PhysRevLett.95.148104.

[82] S. A. Kauffman, E. D. Weinberger, The NK model of rugged fitness landscapes and its application to maturation of the immune response, Journal of Theoretical Biology 141 (2) (1989) 211–245. doi:10.1016/S0022-5193(89)80019-0.

[83] H. C. Nguyen, R. Zecchina, J. Berg, Inverse statistical problems: from the inverse Ising problem to data science, Advances in Physics 66 (3) (2017) 197–261. arXiv:1702.01522, doi:10.1080/00018732.2017.1341604.

[84] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, M. Weigt, Inverse statistical physics of protein sequences: A key issues review, Reports on Progress in Physics 81 (3) (2018) 1–18. arXiv:arXiv:1703.01222v1, doi:10.1088/1361-6633/aa9965.

[85] R. G. Smock, O. Rivoire, W. P. Russ, J. F. Swain, S. Leibler, R. Ranganathan, L. M. Gierasch, An interdomain sector mediating allostery in Hsp70 molecular chaperones, Molecular Systems Biology 6 (414) (2010). doi:10.1038/msb.2010.65.

[86] E. De Leonardis, B. Lutz, S. Ratz, S. Cocco, R. Monasson, A. Schug, M. Weigt, Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction, Nucleic Acids Research 43 (21) (2015) 10444–10455. doi:10.1093/nar/gkv932.

[87] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, C. Sander, Protein 3D structure computed from evolutionary sequence variation, PLoS ONE 6 (12) (2011). doi:10.1371/journal.pone.0028766.

[88] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, Proceedings of the National Academy of Sciences of the United States of America 108 (49) (2011). doi:10.1073/pnas.1111471108.

[89] L. Asti, G. Uguzzoni, P. Marcatili, A. Pagnani, Maximum-Entropy Models of Sequenced Immune Repertoires Predict Antigen-Antibody Affinity, PLoS Computational Biology 12 (4) (2016) 1–20. doi:10.1371/journal.pcbi.1004870.

[90] H. L. Zeng, E. Aurell, Inferring genetic fitness from genomic data, Physical Review E 101 (5) (2020) 1–9. arXiv:2001.02173, doi:10.1103/PhysRevE.101.052409.

[91] B. Levitan, Models and Search Strategies for Applied Molecular Evolution, Annual Reports in Combinatorial Chemistry and Molecular Diversity (1997) 95–152.

[92] F. Morcos, N. P. Schafer, R. R. Cheng, J. N. Onuchic, P. G. Wolynes, Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection, Proceedings of the National Academy of Sciences of the United States of America 111 (34) (2014) 12408–12413. doi:10.1073/pnas.1413575111.

[93] J. Otwinowski, I. Nemenman, Genotype to Phenotype Mapping and the Fitness Landscape of the E. coli lac Promoter, PLoS ONE 8 (5) (2013). arXiv:1206.4209, doi:10.1371/journal.pone.0061570.

[94] B. Derrida, Random-energy model: An exactly solvable model of disordered systems, Physical Review B 24 (1981) 2613–2626. doi:10.1103/PhysRevB.24.2613.

[95] B. Derrida, Random-energy model: Limit of a family of disordered models, Physical Review Letters 45 (2) (1980) 79–82. doi:10.1103/PhysRevLett.45.79.

[96] M. Kardar, Statistical physics of fields, Cambridge University Press, 2007.

[97] J. Zinn-Justin, Phase transitions and renormalization group, Oxford University Press, 2007.

[98] M. Smerlak, A. Youssef, Universal statistics of selected values, Europhysics Letters 117 (5) (2017). arXiv:1612.00843, doi:10.1209/0295-5075/117/50003.

[99] D. S. Fisher, Asexual evolution waves: Fluctuations and universality, Journal of Statistical Mechanics: Theory and Experiment 2013 (1) (2013). arXiv:1210.6295, doi:10.1088/1742-5468/2013/01/P01011.

[100] T. R. Poulsen, A. Jensen, J. S. Haurum, P. S. Andersen, Limits for Antibody Affinity Maturation and Repertoire Diversification in Hypervaccinated Humans, The Journal of Immunology 187 (8) (2011) 4229–4235. doi:10.4049/jimmunol.1000928.

[101] M. Heo, K. B. Zeldovich, E. I. Shakhnovich, Diversity Against Adversity: How Adaptive Immune System Evolves Potent Antibodies, Journal of Statistical Physics 144 (2) (2011) 241–267. doi:10.1007/s10955-011-0241-8.

[102] M. M. Tanaka, S. A. Sisson, G. C. King, High affinity extremes in combinatorial libraries and repertoires, Journal of Theoretical Biology 261 (2) (2009) 260–265. doi:10.1016/j.jtbi.2009.07.041.

[103] B. Vant-Hull, L. Gold, D. A. Zichi, Theoretical Principles of In Vitro Selection Using Combinatorial Nucleic Acid Libraries, Current Protocols in Nucleic Acid Chemistry 00 (1) (2000) 9.1.1–9.1.16. doi:10.1002/0471142700.nc0901s00.

[104] B. Goldstein, Theory of hapten binding to IgM: The question of repulsive interactions between binding sites, Biophysical Chemistry 3 (4) (1975) 363–367.

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

[106] J. Otwinowski, D. M. McCandlish, J. B. Plotkin, Inferring the shape of global epistasis, Proceedings of the National Academy of Sciences of the United States of America 115 (32) (2018) E7550–E7558. doi:10.1073/pnas.1804015115.

[107] A. Clauset, C. R. Shalizi, M. E. Newman, Power-law distributions in empirical data, SIAM Review 51 (4) (2009) 661–703. arXiv:0706.1062, doi:10.1137/070710111.

[108] T. Mora, A. M. Walczak, W. Bialek, C. G. Callan, Maximum entropy models for antibody diversity, Proceedings of the National Academy of Sciences of the United States of America 107 (12) (2010) 5405–5410. arXiv:0912.5175, doi:10.1073/pnas.1001705107.

[109] T. Mora, A. M. Walczak, L. Del Castello, F. Ginelli, S. Melillo, L. Parisi, M. Viale, A. Cavagna, I. Giardina, Local equilibrium in bird flocks, Nature Physics 12 (12) (2016) 1153–1157. arXiv:1511.01958, doi:10.1038/nphys3846.

[110] I. Mastromatteo, M. Marsili, On the criticality of inferred models, Journal of Statistical Mechanics: Theory and Experiment 2011 (10) (2011). arXiv:1102.1624, doi:10.1088/1742-5468/2011/10/P10012.

[111] R. Perline, Strong, weak and false inverse power laws, Statistical Science 20 (1) (2005) 68–88. doi:10.1214/088342304000000215.

[112] S. Coles, An Introduction to Statistical Modeling of Extreme Values, Springer, 2001. doi:10.1007/978-1-4471-3675-0.

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

[114] B. C. Arnold, N. Balakrishnan, H. N. Nagaraja, A first course in order statistics, Vol. 54, SIAM, 1992.

[115] P. Embrechts, C. Klüppelberg, T. Mikosch, Modelling extremal events: for insurance and finance, Vol. 33, Springer Science & Business Media, 2013.

[116] G. D. Stormo, Modeling the specificity of protein-DNA interactions, Quantitative Biology 1 (2) (2013) 115–130. doi:10.1007/s40484-013-0012-4.

[117] T. D. Schneider, R. M. Stephens, Sequence logos: A new way to display consensus sequences, Nucleic Acids Research 18 (20) (1990) 6097–6100. doi:10.1093/nar/18.20.6097.

[118] 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. arXiv:1602.05649, doi:10.1073/pnas.1520969113.

[119] T. M. Cover, J. A. Thomas, Elements of Information Theory, John Wiley & Sons, 2006. doi:10.1002/047174882X.

[120] S. A. Frank, Natural selection. V. How to read the fundamental equations of evolutionary change in terms of information theory, Journal of Evolutionary Biology 25 (12) (2012) 2377–2396. doi:10.1111/jeb.12010.

[121] R. A. Fisher, The genetical theory of natural selection, Ripol Klassik, 1958.

[122] G. Winter, A. D. Griffiths, R. E. Hawkins, H. R. Hoogenboom, Making Antibodies by Phage Display Technology, Annual Review of Immunology 12 (1) (1994) 433–455. doi:10.1146/annurev.iy.12.040194.002245.

[123] 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. 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. arXiv:NIHMS150003, doi:10.1016/j.cell.2013.03.018.

[124] Z. A. Ahmad, S. K. Yeap, A. M. Ali, W. Y. Ho, N. B. M. Alitheen, M. Hamid, ScFv antibody: Principles and clinical application, Clinical and Developmental Immunology 2012 (2012). doi:10.1155/2012/980250.

[125] A. R. Bradbury, S. Sidhu, S. Dübel, J. McCafferty, Beyond natural antibodies: The power of in vitro display technologies, Nature Biotechnology 29 (3) (2011) 245–254. doi:10.1038/nbt.1791.

[126] H. R. Hoogenboom, Selecting and screening recombinant antibody libraries, Nature Biotechnology 23 (9) (2005) 1105–1116. doi:10.1038/nbt1126.

[127] X. Cai, A. Garen, Comparison of fusion phage libraries displaying V(H) or single-chain Fv antibody fragments derived from the antibody repertoire of a vaccinated melanoma patient as a source of melanoma-specific targeting molecules, Proceedings of the National Academy of Sciences of the United States of America 94 (17) (1997) 9261–9266. doi:10.1073/pnas.94.17.9261.

[128] D. Gussow, E. S. Ward, A. D. Griffiths, P. T. Jones, G. Winter, Generating binding activities from Escherichia coli by expression of a repertoire of immunoglobulin variable domains, Cold Spring Harbor Symposia on Quantitative Biology 54 (1) (1989) 265–272. doi:10.1101/sqb.1989.054.01.033.

[129] P. A. Barthelemy, H. Raab, B. A. Appleton, C. J. Bond, P. Wu, C. Wiesmann, S. S. Sidhu, Comprehensive analysis of the factors contributing to the stability and solubility of autonomous human VH domains, Journal of Biological Chemistry 283 (6) (2008) 3639–3654. doi:10.1074/jbc.M708536200.

[130] A. Wörn, A. Plückthun, Different equilibrium stability behavior of scFv fragments: Identification, classification, and improvement by protein engineering, Biochemistry 38 (27) (1999) 8739–8750. doi:10.1021/bi9902079.

[131] C. J. Bond, J. C. Marsters, S. S. Sidhu, Contributions of CDR3 to VHH domain stability and the design of monobody scaffolds for naive antibody libraries, Journal of Molecular Biology 332 (3) (2003) 643–655. doi:10.1016/S0022-2836(03)00967-7.

[132] C. Hamer-Casterman Atarchouch,T, S. Muyldermans, G. Robinson, C. Hamers, E. Bajyana, N. Bendahman, R. Hamilton, Naturally occurring antibodies devoid of light chains, Nature 363 (6428) (1998) 446–448.

[133] S. Boyer, Analyse statistique de la sélection dans des banques minimalistes de protéines, Ph.D. thesis, Université de Grenoble (2015).

[134] B. J. DeKosky, T. Kojima, A. Rodin, W. Charab, G. C. Ippolito, A. D. Ellington, G. Georgiou, In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire, Nature Medicine 21 (1) (2015) 86–91. doi:10.1038/nm.3743.

[135] A. K. Mishra, R. A. Mariuzza, Insights into the structural basis of antibody affinity maturation from next-generation sequencing, Frontiers in Immunology 9 (2018). doi:10.3389/fimmu.2018.00117.

[136] A. Nourmohammad, J. Otwinowski, M. Łuksza, T. Mora, A. M. Walczak, T. Leitner, Fierce Selection and Interference in B-Cell Repertoire Response to Chronic HIV-1, Molecular Biology and Evolution 36 (10) (2019) 2184–2194. doi:10.1093/molbev/msz143.

[137] M. Bonsignori, T. Zhou, Z. Sheng, L. Chen, F. Gao, M. G. Joyce, G. Ozorowski, G. Y. Chuang, C. A. Schramm, K. Wiehe, S. M. Alam, T. Bradley, M. A. Gladden, K. K. Hwang, S. Iyengar, A. Kumar, X. Lu, K. Luo, M. C. Mangiapani, R. J. Parks, H. Song, P. Acharya, R. T. Bailer, A. Cao, A. Druz, I. S. Georgiev, Y. D. Kwon, M. K. Louder, B. Zhang, A. Zheng, B. J. Hill, R. Kong, C. Soto, J. C. Mullikin, D. C. Douek, D. C. Montefiori, M. A. Moody, G. M. Shaw, B. H. Hahn, G. Kelsoe, P. T. Hraber, B. T. Korber, S. D. Boyd, A. Z. Fire, T. B. Kepler, L. Shapiro, A. B. Ward, J. R. Mascola, H. X. Liao, P. D. Kwong, B. F. Haynes, Maturation Pathway from Germline to Broad HIV-1 Neutralizer of a CD4-Mimic Antibody, Cell 165 (2) (2016) 449–463. doi:10.1016/j.cell.2016.02.022.

[138] S. F. Levy, J. R. Blundell, S. Venkataram, D. A. Petrov, D. S. Fisher, G. Sherlock, Quantitative evolutionary dynamics using high-resolution lineage tracking, Nature 519 (7542) (2015) 181–186. doi:10.1038/nature14279.

[139] H. Mouquet, Antibody B cell responses in HIV-1 infection, Trends in Immunology 35 (11) (2014) 549–561. doi:10.1016/j.it.2014.08.007.

[140] J. R. Mascola, B. F. Haynes, HIV-1 neutralizing antibodies: understanding nature's pathways, Immunological Reviews 254 (2013) 225–244. doi:10.1111/imr.12075.

[141] B. F. Haynes, D. R. Burton, Developing an HIV vaccine, Science 355 (6330) (2017) 1129–1130. doi:10.1126/science.aan0662.

[142] L. E. McCoy, D. R. Burton, Identification and specificity of broadly neutralizing antibodies against HIV, Immunological Reviews 275 (1) (2017) 11–20. doi:10.1111/imr.12484.

[143] D. R. Burton, L. Hangartner, Broadly Neutralizing Antibodies to HIV and Their Role in Vaccine Design, Annual Review of Immunology 34 (1) (2016) 635–659. doi:10.1146/annurev-immunol-041015-055515.

[144] I. S. Georgiev, R. S. Rudicell, K. O. Saunders, W. Shi, T. Kirys, K. McKee, S. O'Dell, G.-Y. Chuang, Z.-Y. Yang, G. Ofek, M. Connors, J. R. Mascola, G. J. Nabel, P. D. Kwong, Antibodies VRC01 and 10E8 Neutralize HIV-1 with High Breadth and Potency Even with Ig-Framework Regions Substantially Reverted to Germline, The Journal of Immunology 192 (3) (2014) 1100–1106. doi:10.4049/jimmunol.1302515.

[145] L. M. Walker, M. Huber, K. J. Doores, E. Falkowska, R. Pejchal, J. P. Julien, S. K. Wang, A. Ramos, P. Y. Chan-Hui, M. Moyle, J. L. Mitcham, P. W. Hammond, O. A. Olsen, P. Phung, S. Fling, C. H. Wong, S. Phogat, T. Wrin, M. D. Simek, W. C. Koff, I. A. Wilson, D. R. Burton, P. Poignard, Broad neutralization coverage of HIV by multiple highly potent antibodies, Nature 477 (7365) (2011) 466–470. doi:10.1038/nature10373.

[146] J. F. Scheid, H. Mouquet, N. Feldhahn, M. S. Seaman, K. Velinzon, J. Pietzsch, R. G. Ott, R. M. Anthony, H. Zebroski, A. Hurley, A. Phogat, B. Chakrabarti, Y. Li, M. Connors, F. Pereyra, B. D. Walker, H. Wardemann, D. Ho, R. T. Wyatt, J. R. Mascola, J. V. Ravetch, M. C. Nussenzweig, Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals, Nature 458 (7238) (2009) 636–640. doi:10.1038/nature07930.

[147] 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. doi:10.1126/science.1225416.

[148] G. Villain, System for the generation of controlled affinity maturation trajectories of antibody fragments, Ph.D. thesis, Université de Paris (12 2019).

[149] N. E. B. Inc., Molecular cloning: Technical guide, https://www.neb.com/-/media/nebus/files/brochures/cloning_tech_guide.pdf?rev=5e4ee766c39f49e08fe1a378c4cbd2e0, accessed: June 30, 2020.

[150] R. M. De Wildt, C. R. Mundy, B. D. Gorick, I. M. Tomlinson, Antibody arrays for high-throughput screening of antibody-antigen interactions, Nature Biotechnology 18 (9) (2000) 989–994. doi:10.1038/79494.

[151] A. Griffiths, S. Williams, O. Hartley, I. Tomlinson, P. Waterhouse, W. Crosby, R. Kontermann, P. Jones, N. Low, T. Allison, Isolation of high affinity human antibodies directly from large synthetic repertoires., The EMBO Journal 13 (14) (1994) 3245–3260. doi:10.1002/j.1460-2075.1994.tb06626.x.

[152] G. Smith, Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface, Science 228 (4705) (1985) 1315–1317.

[153] A. D. Keefe, J. W. Szostak, Functional proteins from a random-sequence library, Nature 410 (6829) (2001) 715–718. arXiv:20, doi:10.1038/35070613.

[154] G. P. Smith, V. A. Petrenko, Phage display, Chemical Reviews 97 (2) (1997) 391–410. doi:10.1021/cr960065d.

[155] U. Ravn, F. Gueneau, L. Baerlocher, M. Osteras, M. Desmurs, P. Malinge, G. Magistrelli, L. Farinelli, M. H. Kosco-Vilbois, N. Fischer, By-passing in vitro screening - Next generation sequencing technologies applied to antibody display and in silico candidate selection, Nucleic Acids Research 38 (21) (2010). doi:10.1093/nar/gkq789.

[156] E. Dias-Neto, D. N. Nunes, R. J. Giordano, J. Sun, G. H. Botz, K. Yang, J. C. Setubal, R. Pasqualini, W. Arap, Next-generation phage display: Integrating and comparing available molecular tools to enable costeffective high-throughput analysis, PLoS ONE 4 (12) (2009) 1–11. doi:10.1371/journal.pone.0008338.

[157] P. S. Daugherty, G. Chen, B. L. Iverson, G. Georgiou, Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies, Proceedings

of the National Academy of Sciences of the United States of America 97 (5) (2000) 2029–2034. doi:10.1073/pnas.030527597.

[158] C. O. McCoy, T. Bedford, V. N. Minin, P. Bradley, H. Robins, F. A. Matsen, Quantifying evolutionary constraints on B-cell affinity maturation, Philosophical Transactions of the Royal Society B: Biological Sciences 370 (1676) (2015). doi:10.1098/rstb.2014.0244.

[159] H. Gram, L. A. Marconi, C. F. Barbas, T. A. Collet, R. A. Lerner, A. S. Kang, In vitro selection and affinity maturation of antibodies from a naive combinatorial immunoglobulin library., Proceedings of the National Academy of Sciences of the United States of America 89 (8) (1992) 3576–3580. doi:10.1073/pnas.89.8.3576.

[160] F. Kluge, W. L. Staudenbauer, P. H. Hofschneider, Replication of bacteriophage m13: Detachment of the parental dna from the host membrane and transfer to progeny phages, European Journal of Biochemistry 22 (3) (1971) 350–354.

[161] S. W. Smeal, M. A. Schmitt, R. R. Pereira, A. Prasad, J. D. Fisk, Simulation of the M13 life cycle I: Assembly of a genetically-structured deterministic chemical kinetic simulation, Virology 500 (2017) 259–274. doi:10.1016/j.virol.2016.08.017.

[162] G. Smith, Phage Display of Single-Chain Antibody Constructs, Immunology (2002) 1–27.

[163] A. Royant, M. Noirclerc-Savoye, Stabilizing role of glutamic acid 222 in the structure of Enhanced Green Fluorescent Protein, Journal of Structural Biology 174 (2) (2011) 385–390. doi:10.1016/j.jsb.2011.02.004.

[164] X. Shu, N. C. Shaner, C. A. Yarbrough, R. Y. Tsien, S. J. Remington, Novel chromophores and buried charges control color in mFruits, Biochemistry 45 (32) (2006) 9639–9647. doi:10.1021/bi0607731.

[165] J. Shendure, H. Ji, Next-generation DNA sequencing, Nature Biotechnology 26 (10) (2008) 1135–1145. doi:10.1038/nbt1486.

[166] W. L. Matochko, S. Cory Li, S. K. Tang, R. Derda, Prospective identification of parasitic sequences in phage display screens, Nucleic Acids Research 42 (3) (2014) 1784–1798. doi:10.1093/nar/gkt1104.

[167] A. Cohen, Estimating the Mean and Variance of Normal Populations from Singly Truncated and Doubly Truncated Samples, The Annals of Mathematical Statistics 21 (4) (1950) 557–569.

[168] H. A. Levine, M. Nilsen-Hamilton, A mathematical analysis of SELEX, Computational Biology and Chemistry 31 (1) (2007) 11–35. doi:10.1016/j.compbiolchem.2006.10.002.

[169] R. Stoltenburg, C. Reinemann, B. Strehlitz, SELEX - A (r)evolutionary method to generate high-affinity nucleic acid ligands, Biomolecular Engineering 24 (4) (2007) 381–403. doi:10.1016/j.bioeng.2007.06.001.

[170] M. Slattery, T. Riley, P. Liu, N. Abe, P. Gomez-Alcala, I. Dror, T. Zhou, R. Rohs, B. Honig, H. J. Bussemaker, R. S. Mann, Cofactor binding evokes latent differences in DNA binding specificity between hox proteins, Cell 147 (6) (2011) 1270–1282. doi:10.1016/j.cell.2011.10.053.

[171] F. Sun, D. Galas, M. S. Waterman, A mathematical analysis of in vitro molecular selection-amplification, Journal of Molecular Biology 258 (4) (1996) 650–660. doi:10.1006/jmbi.1996.0276.

[172] S. Schulz, A hidden markov model for sequence-dependent analysis of *In Vitro* selection experiments, Ph.D. thesis, École normale supérieure Paris (07 2016).

[173] Z. R. Sailer, M. J. Harms, Detecting high-order epistasis in nonlinear genotype-phenotype maps, Genetics 205 (3) (2017) 1079–1088. doi:10.1534/genetics.116.195214.

[174] R. Adhikary, W. Yu, M. Oda, R. C. Walker, T. Chen, R. L. Stanfield, I. A. Wilson, J. Zimmermann, F. E. Romesberg, Adaptive mutations alter antibody structure and dynamics during affinity maturation, Biochemistry 54 (11) (2015) 2085–2093. doi:10.1021/bi501417q.

[175] M.-N. Papadopoulou, Riboselect: A novel framework for directed evolution of antibodies in vitro, Ph.D. thesis, Sorbonne Université (06 2020).

[176] A. H. Badran, D. R. Liu, In vivo continuous directed evolution, Current Opinion in Chemical Biology 24 (2015) 1–10. doi:10.1016/j.cbpa.2014.09.040.

[177] M. Matysiak, C. Nizak, O. Rivoire, unpublished.

[178] L. A. Brammer, B. Bolduc, J. L. Kass, K. M. Felice, C. J. Noren, M. F. Hall, A target-unrelated peptide in an M13 phage display library traced to an advantageous mutation in the gene II ribosome-binding site, Analytical Biochemistry 373 (1) (2008) 88–98. doi:10.1016/j.ab.2007.10.015.

[179] M. Pavlicev, G. P. Wagner, Coming to Grips with Evolvability, Evolution: Education and Outreach 5 (2) (2012) 231–244. doi:10.1007/s12052-012-0430-1.

[180] J. Villain, Lip-chip: A quick and affordable method for assaying protein thermal stability, Ph.D. thesis, Université de Paris (09 2020).

[181] A. Mayer, V. Balasubramanian, A. M. Walczak, T. Mora, How a well-adapting immune system remembers, Proceedings of the National Academy of Sciences of the United States of America 116 (18) (2019) 8815–8823. arXiv:1806.05753, doi:10.1073/pnas.1812810116.

[182] L. Yan, R. Ravasio, C. Brito, M. Wyart, Architecture and coevolution of allosteric materials, Proceedings of the National Academy of Sciences of the United States of America 114 (10) (2017) 2526–2531. arXiv:1609.03951, doi:10.1073/pnas.1615536114.

[183] E. A. Raleigh, K. Elbing, R. Brent, Selected Topics from Classical Bacterial Genetics, Current Protocols in Molecular Biology 59 (1) (2002) 1–14. doi:10.1002/0471142727.mb0104s59.

[184] K. Elbing, R. Brent, Growth in Liquid Media, Current Protocols in Molecular Biology 59 (1) (2002) 1.2.1–1.2.2. doi:10.1002/0471142727.mb0102s59.

[185] K. Elbing, R. Brent, Growth on Solid Media, Current Protocols in Molecular Biology 59 (1) (2002) 1.3.1–1.3.6.

[186] K. Elbing, R. Brent, Media Preparation and Bacteriological Tools, Current Protocols in Molecular Biology 59 (1) (2002) 1.1.1–1.1.7. doi:10.1002/0471142727.mb0101s59.

[187] C. E. Seidman, K. Struhl, J. Sheen, T. Jessen, Introduction of plasmid DNA into cells, Current Protocols in Molecular Biology 37 (1) (1997) 1.8.1–1.8.10. doi:10.1002/0471142301.nsa01ls11.

[188] A. M. Helmenstine, 0.5 m edta solution recipe, https://www.thoughtco.com/0-5m-edta-solution-recipe-608140, accessed: July 1, 2020.

[189] I. M. Tomlinson, G. Walter, P. T. Jones, P. H. Dear, E. L. Sonnhammer, G. Winter, The imprint of somatic hypermutation on the repertoire of human germline V genes, Journal of Molecular Biology 256 (5) (1996) 813–817. doi:10.1006/jmbi.1996.0127.

[190] I. M. Tomlinson, G. Walter, J. D. Marks, M. B. Llewelyn, G. Winter, The repertoire of human germline VH sequences reveals about fifty groups of VH segments with different hypervariable loops, Journal of Molecular Biology 227 (3) (1992) 776–798. doi:10.1016/0022-2836(92)90223-7.

[191] R. Saada, M. Weinberger, G. Shahaf, R. Mehr, Models for antigen receptor gene rearrangement: CDR3 length, Immunology and Cell Biology 85 (4) (2007) 323–332. doi:10.1038/sj.icb.7100055.

[192] 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. doi:10.1158/2326-6066.CIR-14-0029.

[193] L. C. James, Antibody Multispecificity Mediated by Conformational Diversity, Science 299 (5611) (2003) 1362–1367. doi:10.1126/science.1079731.

[194] J. Foote, C. Milstein, Conformational isomerism and the diversity of antibodies, Proceedings of the National Academy of Sciences of the United States of America 91 (22) (1994) 10370–10374. doi:10.1073/pnas.91.22.10370.

[195] G. D. Victora, M. C. Nussenzweig, Germinal Centers, Annual Review of Immunology 30 (1) (2012) 429–457. doi:10.1146/annurev-immunol-020711-075032.

[196] F. Delbos, S. Aoufouchi, A. Faili, J. C. Weill, C. A. Reynaud, DNA polymerase η is the sole contributor of A/T modifications during immunoglobulin gene hypermutation in the mouse, Journal of Experimental Medicine 204 (1) (2007) 17–23. doi:10.1084/jem.20062131.

[197] M. S. Neuberger, R. S. Harris, J. Di Noia, S. K. Petersen-Mahrt, Immunity through DNA deamination, Trends in Biochemical Sciences 28 (6) (2003) 305–312. doi:10.1016/S0968-0004(03)00111-7.

[198] P. Pham, R. Bransteitter, J. Petruska, M. F. Goodman, Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation, Nature 424 (6944) (2003) 103–107. doi:10.1038/nature01760.

[199] J. Bachl, C. Carlson, V. Gray-Schopfer, M. Dessing, C. Olsson, Increased Transcription Levels Induce Higher Mutation Rates in a Hypermutating Cell Line, The Journal of Immunology 166 (8) (2001) 5051–5057. doi:10.4049/jimmunol.166.8.5051.

[200] S. H. Kleinstein, Y. Louzoun, M. J. Shlomchik, Estimating Hypermutation Rates from Clonal Tree Data, The Journal of Immunology 171 (9) (2003) 4639–4649. doi:10.4049/jimmunol.171.9.4639.

[201] A. Burkovitz, I. Sela-Culang, Y. Ofran, Large-scale analysis of somatic hypermutations in antibodies reveals which structural regions, positions and amino acids are modified to improve affinity, FEBS Journal 281 (1) (2014) 306–319. doi:10.1111/febs.12597.

[202] J. Foote, H. N. Eisen, Kinetic and affinity limits on antibodies produced during immune response, Proceedings of the National Academy of Sciences of the United States of America 92 (5) (1995) 1254–1256.

[203] S. Tonegawa, Somatic generation of antibody diversity, Nature 302 (5909) (1983) 575–581. doi:10.1038/302575a0.

[204] J. Foote, C. Milstein, Kinetic maturation of an immune response, Nature 352 (6335) (1991) 530–532.

[205] B. F. Haynes, G. Kelsoe, S. C. Harrison, T. B. Kepler, B-cell-lineage immunogen design in vaccine development with HIV-1 as a case study, Nature Biotechnology 30 (5) (2012) 423–433. doi:10.1038/nbt.2197.

[206] N. S. Longo, M. S. Sutton, A. R. Shiakolas, J. Guenaga, M. C. Jarosinski, I. S. Georgiev, K. McKee, R. T. Bailer, M. K. Louder, S. O'Dell, M. Connors, R. T. Wyatt, J. R. Mascola, N. A. Doria-Rose, Multiple Antibody Lineages in One Donor Target the Glycan-V3 Supersite of the HIV-1 Envelope Glycoprotein and Display a Preference for Quaternary Binding, Journal of Virology 90 (23) (2016) 10574–10586. doi:10.1128/jvi.01012-16.

[207] J. F. Scheid, H. Mouquet, B. Ueberheide, R. Diskin, F. Klein, T. Y. Oliveira, J. Pietzsch, D. Fenyo, A. Abadir, K. Velinzon, A. Hurley, S. Myung, F. Boulad, P. Poignard, D. R. Burton, F. Pereyra, D. D. Ho, B. D. Walker, M. S. Seaman, P. J. Bjorkman, B. T. Chait, M. C. Nussenzweig, Sequence and Structural Convergence of Broad and Potent HIV Antibodies That Mimic CD4 Binding, Science 333 (6049) (2011) 1633–1637. arXiv:15334406, doi:10.1126/science.1207227.

[208] S. Hoot, A. T. McGuire, K. W. Cohen, R. K. Strong, L. Hangartner, F. Klein, R. Diskin, J. F. Scheid, D. N. Sather, D. R. Burton, L. Stamatatos, Recombinant HIV Envelope Proteins Fail to Engage Germline Versions of Anti-CD4bs bNAbs, PLoS Pathogens 9 (1) (2013). doi:10.1371/journal.ppat.1003106.

[209] D. Sok, U. Laserson, J. Laserson, Y. Liu, F. Vigneault, J. P. Julien, B. Briney, A. Ramos, K. F. Saye, K. Le, A. Mahan, S. Wang, M. Kardar, G. Yaari, L. M. Walker, B. B. Simen, E. P. St. John, P. Y. Chan-Hui, K. Swiderek, S. H. Kleinstein, G. Alter, M. S. Seaman, A. K. Chakraborty, D. Koller, I. A. Wilson, G. M. Church, D. R. Burton, P. Poignard, The Effects of Somatic Hypermutation on Neutralization and Binding in the PGT121 Family of Broadly Neutralizing HIV Antibodies, PLoS Pathogens 9 (11) (2013). doi:10.1371/journal.ppat.1003754.

[210] H. Mouquet, J. F. Scheid, M. J. Zoller, M. Krogsgaard, R. G. Ott, S. Shukair, M. N. Artyomov, J. Pietzsch, M. Connors, F. Pereyra, B. D. Walker, D. D. Ho, P. C. Wilson, M. S. Seaman, H. N. Eisen, A. K. Chakraborty, T. J. Hope, J. V. Ravetch, H. Wardemann, M. C. Nussenzweig, Polyreactivity increases the apparent affinity of anti-HIV antibodies by heteroligation, Nature 467 (7315) (2010) 591–595. doi:10.1038/nature09385.

[211] B. F. Haynes, J. Fleming, E. W. St. Clair, H. Katinger, G. Stiegler, R. Kunert, J. Robinson, R. M. Scearce, K. Plonk, H. F. Staats, T. L. Ortel, H. X. Liao, S. M. Alam, Immunology: Cardiolipin polyspecific autoreactivity in two broadly neutralizing HIV-1 antibodies, Science 308 (5730) (2005) 1906–1908. doi:10.1126/science.1111781.

[212] M. Liu, G. Yang, K. Wiehe, N. I. Nicely, N. A. Vandergrift, W. Rountree, M. Bonsignori, S. M. Alam, J. Gao, B. F. Haynes, G. Kelsoe, Polyreactivity and Autoreactivity among HIV-1 Antibodies, Journal of Virology 89 (1) (2015) 784–798. doi:10.1128/JVI.02378-14.

[213] I. N. Bronshtein, K. A. Semendyayev, Handbook of mathematics, Springer Science & Business Media, 2013.

RÉSUMÉ

Nous caractérisons l'«évoluabilité» des anticorps en combinant des techniques à haut débit en biologie moléculaire, des outils inspirés de physique statistique et les sciences des données, une approche interdisciplinaire déjà implantée dans d'autres contextes biologiques. L'évoluabilité décrit la capacité d'anticorps à évoluer, c'est-à-dire à sélectionner des phénotypes plus favorables sous l'effet de mutations aléatoires. Celle-ci est une propriété essentielle pour la maturation d'affinité qui est un processus évolutif permettant d'augmenter l'affinité des anticorps contre un pathogène donné. Peut-on observer l'évoluabilité ? Peut-on définir un paramètre mathématique qui représente l'évoluabilité ? Peut-on mesurer ce paramètre ? Quels anticorps sont des points de départ prometteurs pour la maturation d'affinité ? Ici, nous étudions l'effet de l'évolution sur l'affinité de liaison en imitant les premières étapes de la maturation d'affinité contre plusieurs cibles antigéniques : Nous sélectionnons l'affinité de liaison dans des banques d'anticorps randomisés sur leurs sites de liaison en utilisant le phage display et le séquençage à haut débit. Nos banques sont construites sur la base d'échafaudages d'anticorps humains possédant des niveaux différents de maturation antérieure contre une cible tierce (VIH). Nous observons des différences importantes dans leurs réponses face à la sélection, 1) au niveau intra-banque avec peu de variants spécifiques à la cible qui dominent tous les autres variants, 2) au niveau inter-banque la banque naïve dominant systématiquement les banques maturées. En utilisant la physique statistique, nous expliquons comment ces hiérarchies dérivent du potentiel sélectif, une composante de l'évoluabilité que nous définissons comme la susceptibilité à la variation et à la sélection. Nous élaborons que les hiérarchies inter- et intra-banques résultent d'une même origine décrite par un paramètre dépendant de la banque et génératif, σ qui encode pour la variance d'énergies de liaison (valeurs sélectives malthusiennes) dans les banques. Curieusement, le potentiel sélectif le plus élevé est observé systématiquement dans la banque basée sur un anticorps naïf ce qui suggère un scénario où les anticorps naïfs auraient été «evolués pour évoluer».

MOTS CLÉS

évolution, évoluabilité, potentiel sélectif, évolution *in vitro*, séquençage à haut débit, anticorps, maturation d'affinité

ABSTRACT

We characterize antibody "evolvability" by combining high-throughput techniques from molecular biology and tools from statistical physics and data science, an interdisciplinary approach already successfully applied in other biological contexts. Evolvability describes the ability of antibodies to evolve, *i.e.* the effect of mutation and selection on their phenotype. It is an essential property for the success of affinity maturation, an accelerated evolutionary process leading to antibodies with improved binding affinity to a given pathogen. Can we observe evolvability? Can we define a mathematical parameter that represents evolvability? Can we measure this parameter? What antibodies are promising starting points for affinity maturation? Here, we study the effect of evolution on binding affinity by mimicking the initial step of affinity maturation against various antigenic targets: We select for binding affinity from libraries of randomized antigen binding sites using phage display and high-throughput sequencing. Our libraries are built around human antibody scaffolds exhibiting different levels of previous maturation against a third-party target (HIV). We observe vast differences in their response to selection, 1) at the intra-library level with few, target-specific variants strongly dominating all others, 2) at the inter-library level with the naïve library systematically dominating mature libraries. Using statistical physics, we argue how these hierarchies are linked to selection potential, a component of evolvability that we define as the susceptibility to variation and selection. We establish that inter- and intra-library differences share a common origin captured by a single, library-dependent, generative parameter σ encoding for the variance of binding energies (Mathusian fitness) within libraries. Interestingly, highest selection potentials are systematically observed in the library based on a naïve antibody, suggesting a scenario of naïve antibodies being "evolved to evolve".

KEYWORDS

evolution, evolvability, selection potential, *in vitro* evolution, high-throughput sequencing, antibody, affinity maturation