A 3D structural affinity model for multi-epitope *in silico* germinal center simulations.

Philippe A. Robert*[1] and Michael Meyer-Hermann†[1,2]

[1]*Department of Systems Immunology and Braunschweig Integrated Centre of Systems Biology, Helmholtz Centre for Infection Research, Braunschweig, Germany,* [2]*Institute for Biochemistry, Biotechnology and Bioinformatics, Technische Universität Braunschweig, Braunschweig, Germany*

Abstract

Antibody-based immunotherapies require the tedious identification and development of antibodies with specific properties. In particular, vaccine development for mutating pathogens is challenged by their fast evolution, the complexity of immunodominance, and the heterogeneous immune history of individuals. Mathematical models are critical for predicting successful vaccine conditions or designing potent antibodies. Existing models are limited by their abstract and poorly structural representations of antigen epitopes. Here, we propose a structural lattice-based model for antibody–antigen affinity. An efficient algorithm is given that predicts the best binding structure of an antibody's amino acid sequence around an antigen with shortened computational time. It is suitable for large simulations of affinity maturation. This structural representation contains key physiological properties, such as affinity jumps and cross-reactivity, and successfully reflects the topology of antigen epitopes, such as pockets and shielded residues. We perform *in silico* immunizations via germinal center simulations and show that our model can explain complex phenomena like recognition of the same epitope from unrelated clones. We show that the use of cocktails of similar epitopes promotes the development of cross-reactive antibodies. This model opens a new avenue for optimizing multivalent vaccines with combined cocktails or sequential immunizations, and to reveal reasons for vaccine success or failure on a structural basis.

Index terms— vaccine design, germinal centers, structural affinity, antibody discovery, combinatorial enumeration

*philippe.robert@ens-lyon.org
†mmh@theoretical-biology.de

Introduction

In view of the billions of possible antibody sequences and their therapeutic potential, engineering antibodies with target specificity is a timely and complex challenge. Antibody discovery goes beyond screening existing antibody sequences by also adding carefully designed mutations, with the aim of achieving the desired binding landscape and pharmaceutical biochemical properties, and no self-reactivity. Engineering antibodies from antibodies raised *in vivo* is safer, as *in vitro* tests for self-reactivity, eventually in non-human species, can hide unpredicted deadly cross-reactivity to human self-antigens [1].

Highly mutating pathogens like HIV, hepatitis C virus or influenza are poorly targeted by vaccines. Consequently, antibodies elicited during infection or vaccination do not necessarily protect again the next mutation of the strain. Strikingly, during such infections, a few individuals naturally develop broadly neutralizing antibodies (bnAbs) that bind to a large range of strains [2–6]. Injecting these bnAbs has been shown to protect against future infections in certain contexts [7, 8], showing their therapeutic potential. Meeting the requirements for the induction of bnAbs *in vivo*, especially in humans or primates, would support better vaccine design and enhance the discovery of new bnAbs that are functional but not self-reactive as a basis for further engineering. For this purpose, computational modeling is an attractive approach. To understand broad neutralization, one needs to look at the cellular basis of antibody responses. These happen in anatomical structures called germinal centers (GCs), where B-cells selectively mutate their B-cell receptors (BCRs) through somatic hypermutation, which are later secreted as antibodies. High-affinity B-cells are selected for survival and proliferation at the expense of low-affinity B cells. As a consequence, the affinity of the antibodies increases over time, a process called affinity maturation (AM).

The GC's response to single well-defined antigens has long been studied *in vivo* (reviewed in [9]). Many predictive mathematical models have been developed that use abstract antigen–antibody affinities in a probabilistic manner: a mutation is an improvement or decrease in affinity to the single target antigen [10–16]. However, at the scale of multiple or complex antigens, new layers of complexity arise that are not covered by these models. Firstly, mutating pathogens evolve multiple antigenic variants that differ in their accessible epitopes, their frequency and the amino acid (AA) sequence of their accessible sites. Secondly, during a chronic HIV infection, the epitopes recognized inside the GCs evolve over time and different B-cell clones expand successively towards different epitopes [17]. Therefore, the immunogenicity of antigens (i.e., their capacity to mount an immune response) is very complex and might evolve between epitopes of the same or different antigens over time. Thirdly, it has been reported that antibodies produced by GC-derived plasma cells can diffuse back to the GCs themselves and compete for epitope binding of the GC's B-cells [18], potentially changing the immunodominance landscape of the response [19].

Several recent mathematical models have been developed to account for the quantitative properties of AM towards multiple antigens, either fixed or evolving. The models compare the efficiency of different vaccine approaches such as simultaneous cocktail immunizations or sequential vaccinations with the same or different antigens [20]. The models need antibody–antigen affinity representations that account for properties like cross-reactivity. In particular, an antibody with high affinity to an epitope should still have a correlated affinity to a related epitope but a less correlated affinity to a more mutated or unrelated epitope. Interestingly, each published model comes with its own abstract encoding for antibody–antigen interactions (reviewed in [20]), including (i) an abstract N-dimensional shape-space model [15]; (ii) the size of the longest substrings compared between proteins taken from a finite alphabet [21]; (iii) binary proteins with a predefined accessibility profile for each clone [22]; and (iv) binary proteins with location-dependent affinity in the facing residues, from predefined affinity distributions [23]; (v) abstract arbitrary matching between epitopes and binary receptors [24] and (vi) a structural representation of cubic folded lattice proteins [25].

We have previously reported [20] that these representations carry different levels of cross-reactivity. Only some can incorporate key mutations or shielding, and it is not easy to translate a real antigen or structure into any of them. A more structural encoding of antibody–antigen affinity would overcome these limitations, and the physiological properties would naturally emerge without the need to be manually added to abstract models.

Several prediction tools have been developed based on protein folding thermodynamics and the knowledge of known antibody–antigen structures (reviewed in [26]). Sadly, only a few antibody crystal structures have been generated. Therefore, these tools can only use very limited datasets or threading on known substructures. Full prediction tools like Rosetta [27], on the other hand, have reached powerful binding accuracy at the expense of few hundreds of computational hours for a single calculation.

At the scale of GC simulation, one GC typically contains 10,000 cells at its peak. Through mutations and high turnover between proliferation and death, a GC will typically explore 10^5 to 10^6 mutations during its lifespan [28]. Therefore, the current available prediction techniques for structural antigen–antibody binding are unable to achieve a single GC simulation in feasible time, not mentioning the hundreds of GCs produced during vaccination in humans.

Here, we develop a new hybrid model of antigen–antibody folding that simulates the best folding of one antibody loop, the complementary determining region 3 (CDR3), around a predefined antigen structure on a 3D lattice. Real AA sequences are used, up to 11 AAs for the CDR loop, and with complex antigen structures up to 200 AAs for the antigen. The interaction between neighboring AAs is based on experimentally measured potential [29]. Thanks to the rapid combinatorial enumeration of only the foldings with a minimum number of contact points with the antigen, we show that a binding energy can be derived in feasible computational time for GC simulations (a few minutes for 1000

affinity calculations). We show that key properties naturally arise from this representation, such as polyreactivity, cross-reactivity, accessibility and shielding effects, and mutations inducing affinity jumps. As a proof of concept, we adapted and ran GC simulations from the *in silico* model developed in [10] with our structural antigen representation. The model showed physiological GC dynamics and proper AM. We show that the use of cocktails of very similar antigens generated favorable conditions for cross-reactivity. Therefore, GC simulations combined with the 3D affinity model presented here, are suitable for testing vaccine strategies and predicting therapeutic methods to modulate immunodominance in realistic computational time.

Results

A fast computational model for lattice-based antigen–antibody binding. A major challenge of *in silico* GC simulations is to access the affinity change gained from somatic hypermutations in a very short computational time. Given a predefined antigen structure (the 'ligand'), we aim to compute the binding energy and derive an affinity for large numbers of mutated binding regions of the antibody (the 'receptor'), (Figure 1A).

To generate a realistic but fast model for antigen–antibody binding, we developed a lattice model of protein folding, inspired by previous work [25], where residues can only occupy predefined positions on a grid (Figure 1B). Real AA sequences are used, consecutive AAs should be on neighboring points and only one AA per grid point is allowed (hard-core repulsion). Conveniently, a protein structure can therefore be named as a list of moves, namely *Straight, Up, Down, Left, Right*, such as those a plane pilot would follow in 3D (Figure 1B). To derive binding energies, all covalent bounds are considered equal (bounds have the same length and only 90° or 180° angles) and are therefore neglected. Instead, neighboring but non-covalently linked AAs are considered to interact. Their binding energy has been previously estimated from structural databases for each pair of AAs [29]. For two interacting proteins, the binding energy I can be estimated as the energetic sum of individual bonds between the two proteins, whereas the total energy E of one protein will be the energetic sum of interacting and intrinsic bonds, in order to include intrinsic binding as a stabilization factor of the protein structure (see Methods). We do not yet talk about affinity, because this will depend on other physiological parameters, including temperature, pH, and the rest of the antibody. Affinity will be introduced below. We assume that the antigen has a stable conformation and is not significantly impacted by the antibody binding; instead, the CDR3 loop of the antibody folds onto it. Therefore, for each CDR3 sequence, a folding has to be performed and the folding structure around the antigen with lowest total energy has to be found (Figure 1B).

Because of the exponentially high number of possible foldings, it would be unrealistic to compute all foldings for each CDR3 sequence at each time-step in a dynamic computer simulation of GCs. There-

fore, we developed an optimized algorithm that pre-computes only the 'interesting' possible foldings of receptors of realistic length L (typically between 7 and 14 AAs). More precisely (see Figure 1C), starting from a predefined antigen structure, a recursive algorithm enumerates all possible foldings the receptors could adopt, with the constraint of 'touching' the antigen at least n times (we used a minimum of four interactions). This considerably reduces the amount of structures to be enumerated and stored in the memory. The structures are stored independently of the CDR3 sequence, which is not considered yet. For each CDR3 sequence to be evaluated, all the previously enumerated foldings are taken one by one and filled with the AA sequence of the CDR3, and the binding and total energies are calculated. As a result, the best binding energy is given by the foldings with optimal total energy, and the statistical binding energy can also be derived by weighting each folding by a Boltzmann factor (see Methods). The statistical and best binding energies are very similar (Figure S1) such that the best folding is representative of the ensemble of foldings (frozen state).

As an example, we show three typical antigen structures (Figure 1D): one simple accessible and flat antigen (L1), one antigen with an accessible tail and a hidden pocket (L2), and an antigen with both an accessible hook and a pocket, and additional inaccessible shielded positions, marked X (L3). We assume that positions below the epitope (and on the side of L2 and L3) are inaccessible to receptors, to model the antigen scaffold (shown as filled planes). For a randomly chosen CDR3, the best binding structures are shown in Figure 1E, together with the binding energy I and the total energy E (the lower the better). As expected, the best receptor structure around L2 is located in the accessible pocket. Similarly, the best receptor structure around L3 binds the accessible hook. For comparison, the best structure that would bind the shielded pocket of L3 is shown in purple (Figure 1E) and has much worse binding energy.

The computational load of this structural representation is shown in Figure 1F-H. For each of the three antigens, the number of possible folding structures (Figure 1F) and computation times (Figure 1G) are shown, depending on the length of the CDR3 considered (7–11 AAs). For larger CDR3 sequences, millions of folding structures exist (up to 75 million), but the pre-computation time still does not exceed a few hours and has to be performed and saved only once. Regarding calculation time, thousands of CDR3 can be evaluated for binding energy to these antigens in less than one hour on a single CPU, implying that it becomes possible to simulate a full GC within a few hours.

As the folding structures are selected to bind at least four points on the antigen, we also show that higher thresholds can improve the speed further (Figure 1H). The number of structures as a function of CDR3 length and the threshold number of contacts is shown separately for each ligand. For instance, for ligand L3, receptors of length 11 lead to 75 million structures with four contacts, which can be brought down to only 5.6 million by requiring nine contacts. In summary, we present a structural model for antibody–antigen binding that can model complex antigens and is efficient enough for GC simulations.

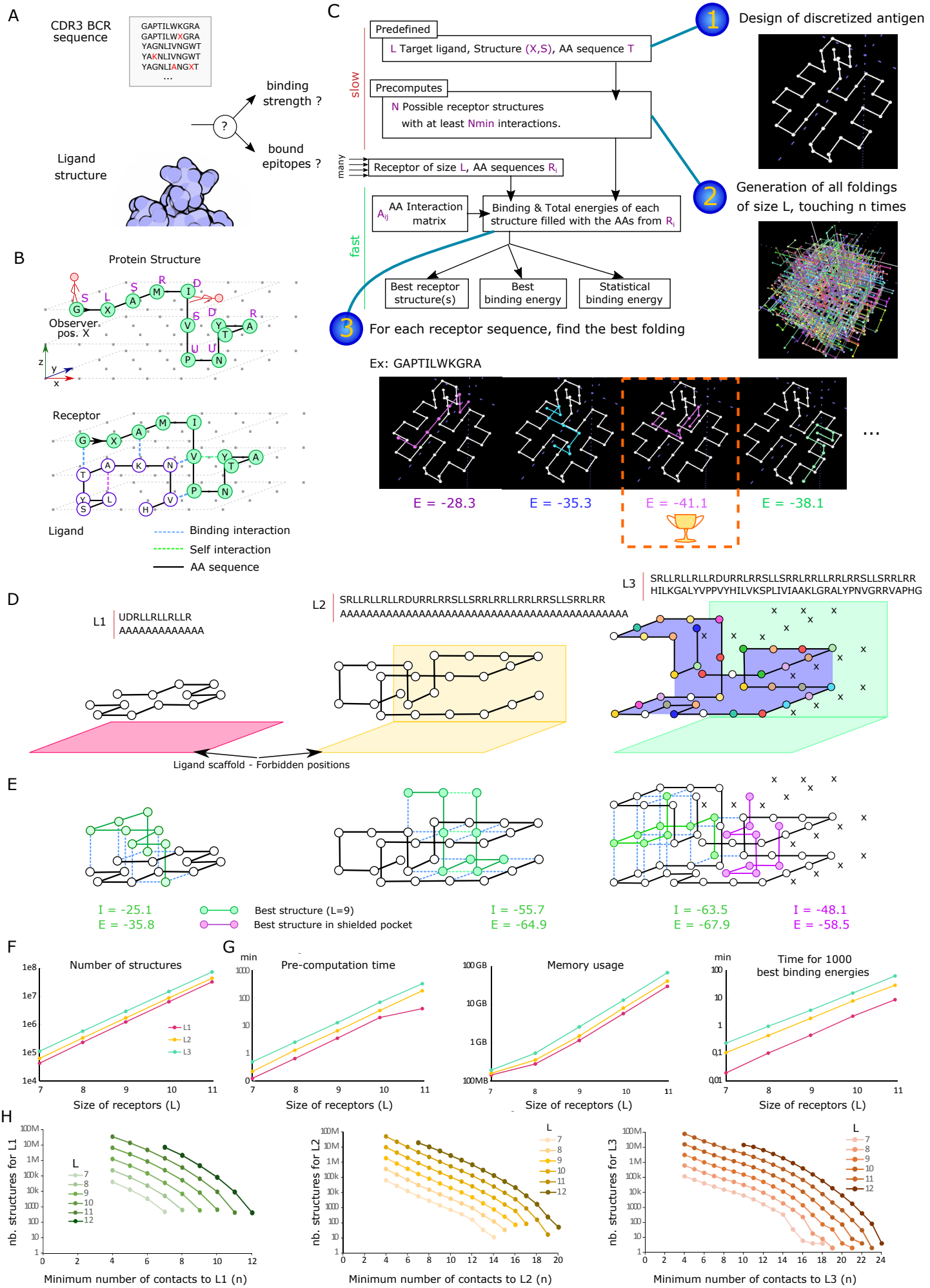
A
CDR3 BCR sequence
GAPTILWKGRA
GAPTILWXGRA
YAGNLIVNGWT
YAKNLIVNGWT
YAGNLIANGXT
...
binding strength ?
bound epitopes ?
Ligand structure
B
Protein Structure
Observer pos. X
Receptor
Ligand
Binding interaction
Self interaction
AA sequence
C
Predefined
L Target ligand, Structure (X,S), AA sequence T
Precomputes
N Possible receptor structures with at least Nmin interactions.
slow
many
Receptor of size L, AA sequences Ri
AA Interaction matrix
Aij
Binding & Total energies of each structure filled with the AAs from Ri
fast
Best receptor structure(s)
Best binding energy
Statistical binding energy
1 Design of discretized antigen
2 Generation of all foldings of size L, touching n times
3 For each receptor sequence, find the best folding
Ex: GAPTILWKGRA
E = -28.3
E = -35.3
E = -41.1
E = -38.1
...
D
L1
UDRLLRLLRLLR
AAAAAAAAAAAAA
L2
SRLLRLLRLLRDURRLRRSLLSRRLRRLLRRLRRSLLSRRLRR
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
L3
SRLLRLLRLLRDURRLRRSLLSRRLRRLLRRLRRSLLSRRLRR
HILKGALYVPPVYHILVKSPLIVIAAKLGRALYPNVGRRVAPHG
Ligand scaffold - Forbidden positions
E
I = -25.1
E = -35.8
Best structure (L=9)
Best structure in shielded pocket
I = -55.7
E = -64.9
I = -63.5
E = -67.9
I = -48.1
E = -58.5
F
Number of structures
L1
L2
L3
Size of receptors (L)
G
Pre-computation time
Memory usage
Time for 1000 best binding energies
Size of receptors (L)
H
nb. structures for L1
nb. structures for L2
nb. structures for L3
Minimum number of contacts to L1 (n)
Minimum number of contacts to L2 (n)
Minimum number of contacts to L3 (n)

Figure 1: **Fast computation of antibody–antigen structural binding on a 3D lattice. A.** Formulation of the problem: given an antigen, generate a binding landscape of thousands of mutated antibody binding regions (focusing on the CDR3 loop). **B.** Proteins (antigen and antibody binding region) are represented as a path in a 3D lattice, written as: S: *go straight*, L: *turn left*, R: *turn right*, D: *move down* and U: *move up*. The sequence of AAs is shown as a black line. Interactions happen between neighboring AAs that are non-covalently linked, either inside a protein (green) or between the two (black). The binding energy is the sum of interacting bonds, and the total energy of the receptor additionally includes its intrinsic bounds. **C.** Workflow for the efficient computation of ligand–receptor binding energies. (1) A ligand is predefined with its structure and AAs in the lattice. (2) The set of possible receptor folding structures of predefined length L, with a minimum number of contact points n to the ligand is pre-computed and stored. (3) Every time a receptor sequence is given, only this subset of structures will be used one by one with this receptor sequence, and tested for binding energy I and total energy E to the ligand. Either the 'best binding energy' (average binding energy of the structure(s) with minimum total energy) or the 'statistical binding energy' (weighted binding energies of all structures according to a Boltzmann factor) is returned (see Methods). All equally optimal structures can also be retrieved. **D.** Example of possible ligands: a simple one filled with alanine only (L1), a ligand with a tail and pocket (L2), and a ligand with diverse AAs, a hook, and additional positions blocked to the receptor (shielded residues, 'X') that cover the pocket, making it less accessible (L3). Alanine is shown in white; other residues are colored. **E.** For a randomly chosen receptor sequence of length 9, the best folding structure is shown in green; the best one reaching the pocket for L3 is shown in purple. The best binding energy I and total energy E are shown for each ligand, in kT units. **F.** The number of structures for each ligand, depending on the length of receptors L (in AAs), when at least four contacts are required. **G.** Time and memory requirements for pre-computing the structures and then calculating the best and statistical energies of 1000 receptor sequences. **H.** Total number of sequences as a function of the minimum requested number of contacts n (threshold). Longer receptors can be considered for larger thresholds.

Structural properties of the model. We assessed the distribution of binding energy for large numbers of receptor sequences and checked whether the properties of antigen binding are physiological.

Varying receptor binding energies. We first asked if the model could reproduce differing ranges of binding energies for different receptor sequences. Thousands of random receptor sequences of length 7 to 11 AAs were sampled; their binding energy to each ligand (L1, L2 and L3), as previously introduced in Figure 1E, is shown in (Figure 2A). The binding energy of the receptors varied according to a broad Gaussian distribution in the range of 20, 40 and 50 kT units for L1, L2 and L3, respectively. It shows that although millions of structures are tested for each receptor sequence, the model successfully generates different binding landscapes for each sequence. In other words, a 'bad receptor' for this antigen does not match any good folding pattern among millions of possible ones. Therefore, the calculated binding energy contains structural information about the receptor sequences. The binding energy of receptors increased linearly with their length (Figure 2A) because they have more options for binding and a bigger folding ensemble. Therefore, we do not recommend comparing energies between receptors of different lengths, unless a correction coefficient is added.

Antigen structure shapes binding. The impact of the antigen structure onto the binding energies is investigated in the model. We created a 'bigger' version of ligand L1 with a similar flat shape. It induced a global increase in binding (Figure 2B), because it opened more positions, especially corners with three AAs that are accessible from the same point. We therefore suggest comparing ligands of a similar size. In the case of L3, we first created a more physiological variant called L6 where the hole in the pocket was filled (green arrow), to avoid receptors going through it. We then monitored the contribution of the shielded pocket by removing it in a new ligand L7 (Figure 2C). Interestingly, it did not significantly change the distribution of the best binding energies because the receptor structures accessing it were not favored. It shows that the model is a suitable tool for studying the effect of steric hindrance of epitopes such as HIV GP120 protein [30].

Antigen AA complexity shapes binding. Changing the AA composition of L1 from alanine only to diverse AAs (L5) (Figure 2B) led to a strong increase in best binding energy. Intuitively, if an antigen contains only one type of AAs, receptors will not benefit from different folding structures.

Point Mutations. In the GC, somatic hypermutation generates point mutations that lead to increased or decreased best binding energies. We tested all possible point mutations along eight selected receptor sequences of length 9 with different binding energies to ligand L6 (Figure 2D). Interestingly, mutations could always both increase and decrease the binding energy, even when starting from sequences with very high (green) or very low (purple) energy. All 2500 randomly sampled receptor sequences for L6 displayed binding energies of -100 to -50 kT units (Figure2A). Point mutations from the selected receptor sequences could increase or decrease by 15 kT units, showing that point mutations can cause huge jumps in binding energy.

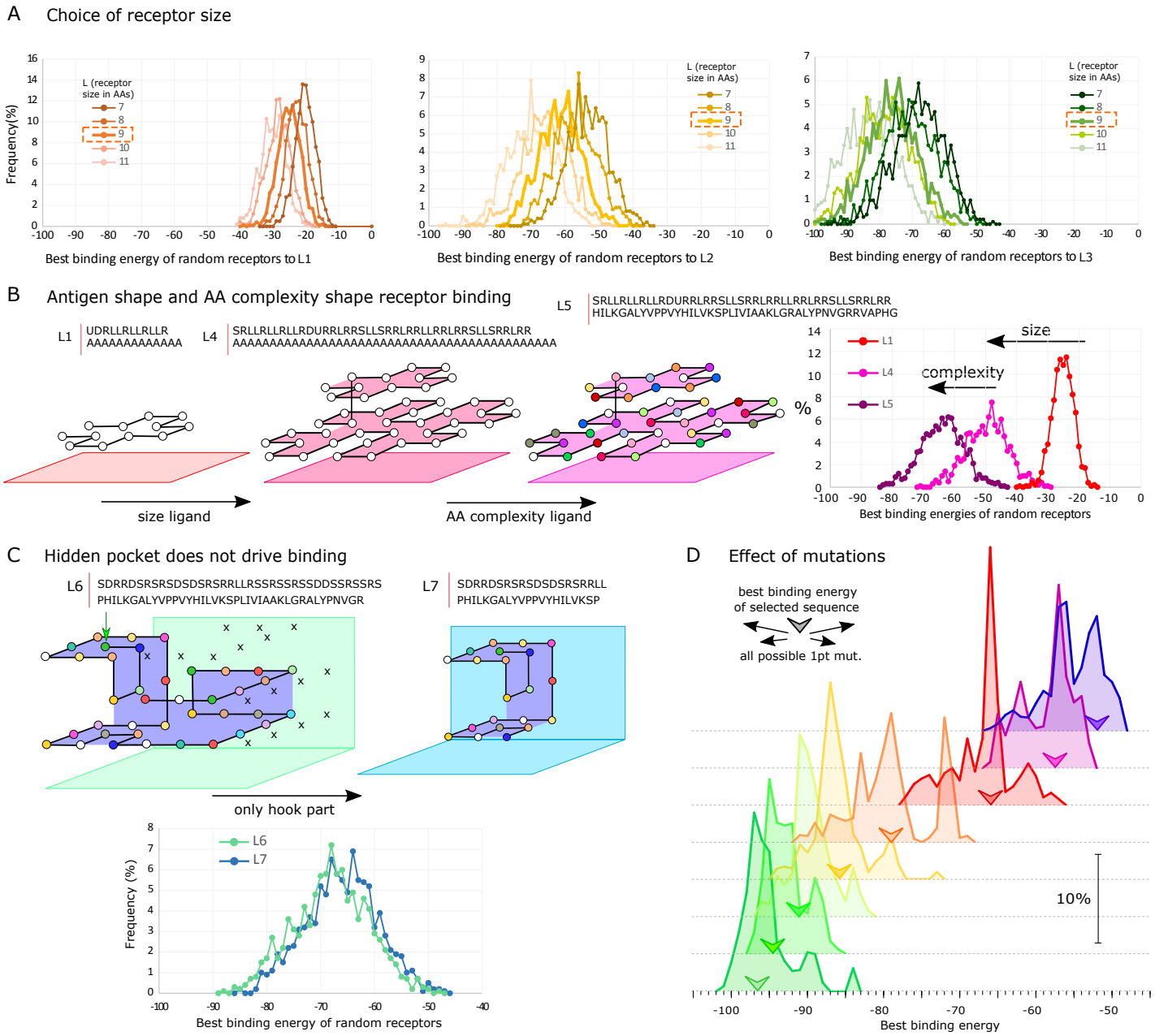

Figure 2: **Distributions of receptor binding energies.** The binding energy of 2500 randomly sampled receptors **(A)** of length 7 to 11 to ligands L1, L2 and L3; **(B)** of length 9 to ligands L1, L4 and L5; and **(C)** to L6 and L7. The structure of the ligands is as depicted in Figure 1D. **D.** Histogram of the effect of all possible single point mutations on the binding energy, for eight selected receptor sequences with different binding energies for L6 are depicted as histograms.

GC simulations show affinity jumps and mutation bursts. In order to predict AM and ultimately the efficiency of multivalent vaccines from the structural properties of the antigens, the next step was to simulate full-scale GC reactions. We incorporated our structural binding model into an agent-based model for GC dynamics in 3D where B-cells and T-cells are explicitly modelled to move, interact, be selected, proliferate, die or exit [10, 31] (see Methods for details). In the GC, the affinity between the BCR and the antigen determines the probability of capturing the antigen, which is needed later to survive T-cell selection. We used two parameters to transform the binding energy into a phenomenological affinity (i.e., the probability of capturing the antigen in the simulation): (1) the lowest (i.e., the best) binding energy is translated into a maximal probability of antigen capture (saturation limit). Therefore, an affinity of 1.0 corresponds to a very low binding energy. Higher affinity values are possible but will not increase antigen capture. (2) A scaling coefficient is added to include the contribution of the remaining antibodies (see Methods for details).

We found suitable values for these two parameters that accounted for realistic GC dynamics and AM. The dynamics of GC simulations when we used a single complex target antigen with diverse AAs (L6) is shown in (Figure 3A). The number of B-cells over time in single GCs shows an initial expansion as expected, followed by a burst and slow decay. Starting randomly from cells with at least 0.0001 affinity (see Methods), the average affinity of cells inside the GCs easily reached 0.3 to 1 after 21 days. The average number of mutations of cells leaving GCs was around six mutations, as estimated in [32]. The best receptor sequences of five GCs are shown in Figure 3B, together with a consensus between receptors of very high affinity (> 0.75). Each GC could mount different receptors with very high affinity, showing that many different receptor sequences could reach high affinity, in contrast to most abstract affinity models where the high affinity sequence is unique [20].

An important biological property of GCs is the possibility of one clone dominating a GC [31] and the possibility of huge affinity jumps by single mutations. As these mutations are rare, it would be impossible to detect them by chance among random receptors and mutations, and we decided to study the mutation landscape in depth in five *in silico* GC simulations (Figure 3C-E). We analyzed all the receptor sequences generated by mutations inside a single GC over the 21 days of the reaction, including dead cells. Therefore, all mutations that occurred were retrieved from the simulation. Around $70,000$ mutations happened; the forest of mutations is shown in (Figure 3C). Each tree (cluster) represents one founder cell sequence and all its progeny. We distinguished the sequences with high affinity (> 0.5) in red and observed that more than one founder cell achieved high affinity. It was observed *in vivo* that GCs exist that are not dominated by a single clone, in which multiple founder clones can co-exist and reach high affinity [31]. Here, we could observe the parallel AM of multiple clones *in silico*.

To further characterize the mutation profile of a successful journey from a founder cell to a high-affinity cell, we selected the mutation cluster of a winning or losing founder cell in Figure 3D. The clone

of the winning founder cell showed the explosive expansion of a sequence into many mutated daughter sequences with high affinity. Interestingly, other bursts (hubs in the trees) could be observed in both the winning and losing clones (Figure 3C). By focusing only on high-affinity sequences and their parents up to the founder cell (Figure 3E) we could track the history of affinity changes during AM leading to those sequences only. Interestingly, from a founder cell with an affinity of 0.009, both increases and decreases of affinity happened, typically by 2— to 10—fold, until reaching the sequence with an affinity of 0.77, showing high expansion. Furthermore, some high-affinity sequences were generated from non-beneficial mutations of the bursting sequence. Therefore, in the simulations, not exclusively advantageous sequences are selected, which reflects the stochastic nature of selection observed in [31].

In the tree representation, the same sequence can occur in different nodes with a different order of mutations. We simplified the graph by merging identical sequences around the bursting cell (Figure 3F). Strikingly, point mutations reaching a 74—fold increase in affinity could be observed. Specific mutations that confer an affinity jump have been described *in vivo*, like the acquisition of L33 in antibodies against ovalbumin [33], and these are positively selected in GCs. Thus, the possibility of key mutations is captured by our model. However, this does not necessarily imply that we actually have reproducible key mutations at a specific position because our receptor sequences can not easily be aligned. Indeed, the consensus sequences in Figure 3B seem to have shifted or to mirror each other. This kind of ambiguity is inherent in the sequence model presented here.

Finally, we show the history of high-affinity cells from different GC simulations in Figure 3G, similar to those depicted in Figure 3E, but for different winning founder cells from two other GC simulations. This illustrates how different mutation patterns can be. All of them show expansion bursts (hubs) to different degrees. Few reports have studied single cell sequences inside a single GC. Interestingly, these bursts have been observed in a subset of GCs [31] in response to protein antigens, and to a greater extent in GCs dominated by a single clone. This indicates that our structural model reflects several non-trivial physiological properties of GCs and is adequately designed for simulating AM in GCs.

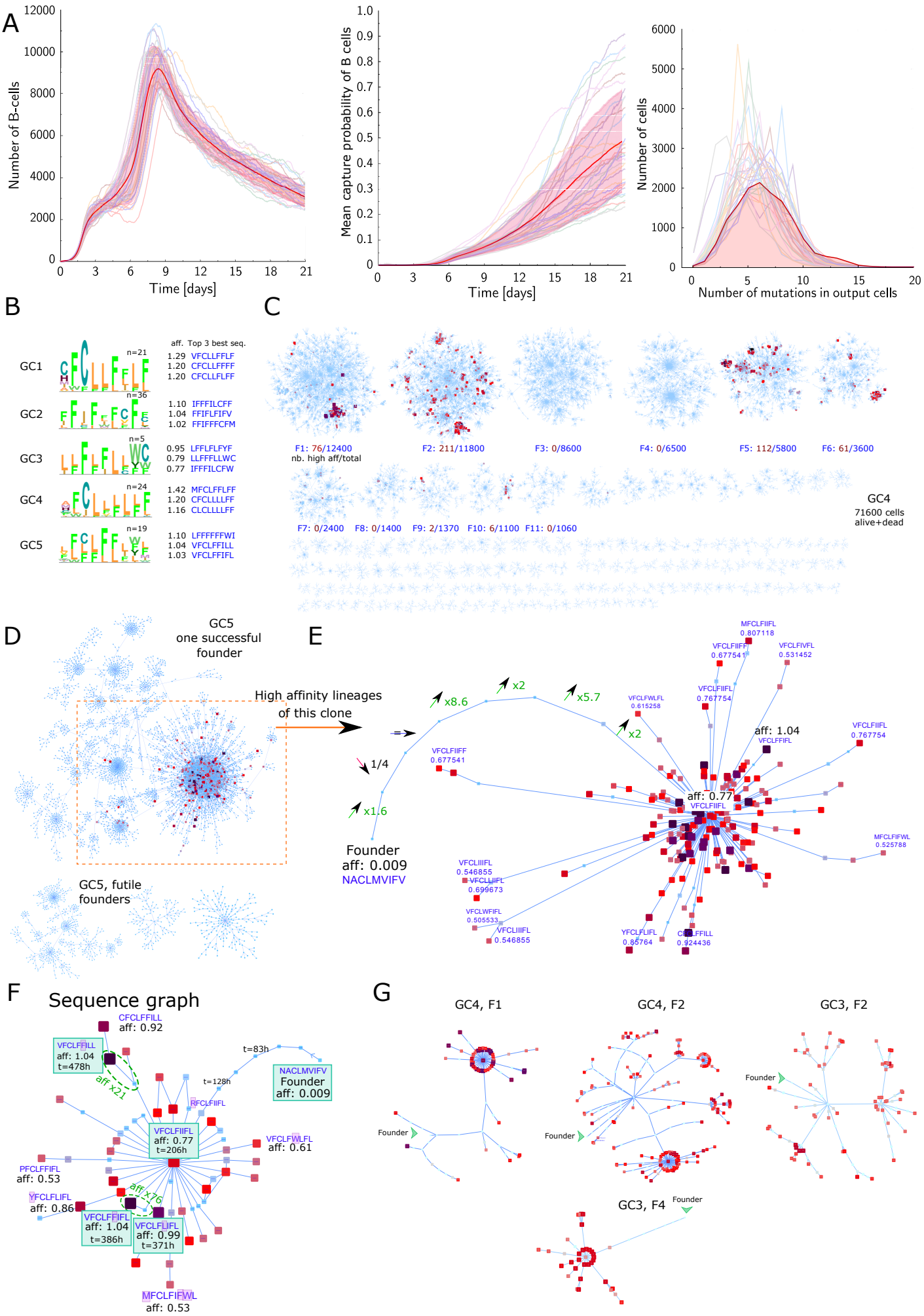
A
Number of B-cells
Time [days]
Mean capture probability of B cells
Time [days]
Number of cells
Number of mutations in output cells
B
aff. Top 3 best seq.
GC1
n=21
1.29 VFCLLFFLF
1.20 CFCLLFFFF
1.20 CFCLLFLFF
GC2
n=36
1.10 IFFFILCFF
1.04 FFIFLFIFV
1.02 FFIFFFCFM
GC3
n=5
0.95 LFFLFLFYF
0.79 LLFFFLLWC
0.77 IFFFILCFW
GC4
n=24
1.42 MFCLFFLFF
1.20 CFCLLLLFF
1.16 CLCLLLLFF
GC5
n=19
1.10 LFFFFFFWI
1.04 VFCLFFILL
1.03 VFCLFFIFL
C
F1: 76/12400
nb. high aff/total
F2: 211/11800
F3: 0/8600
F4: 0/6500
F5: 112/5800
F6: 61/3600
F7: 0/2400
F8: 0/1400
F9: 2/1370
F10: 6/1100
F11: 0/1060
GC4
71600 cells
alive+dead
D
GC5
one successful
founder
High affinity lineages
of this clone
GC5, futile
founders
E
x8.6
x2
x5.7
x2
1/4
x1.6
Founder
aff: 0.009
NACLMVIFV
VFCLFIIFF
0.677541
VFCLFWLFL
0.615258
VFCLFIIFL
0.767754
MFCLFIIFL
0.807118
VFCLFIIFF
0.677541
VFCLFIVFL
0.531452
aff: 1.04
VFCLFFIFL
VFCLFIIFL
0.767754
aff: 0.77
VFCLFIIFL
MFCLFIFWL
0.525788
VFCLIIIFL
0.546855
VFCLIIIFL
0.699673
VFCLWFIFL
0.505533
VFCLIIIFL
0.546855
YFCLFLIFL
0.85764
CFCLFFILL
0.924436
F
Sequence graph
CFCLFFILL
aff: 0.92
VFCLFFILL
aff: 1.04
t=478h
aff x21
t=83h
t=128h
NACLMVIFV
Founder
aff: 0.009
RFCLFIIFL
VFCLFIIFL
aff: 0.77
t=206h
VFCLFWLFL
aff: 0.61
PFCLFFIFL
aff: 0.53
YFCLFLIFL
aff: 0.86
aff x16
VFCLFFIFL
aff: 1.04
t=386h
VFCLFLIFL
aff: 0.99
t=371h
MFCLFIFWL
aff: 0.53
G
GC4, F1
Founder
GC4, F2
Founder
GC3, F2
Founder
GC3, F4
Founder

Figure 3: **Full scale *in silico* GC simulations and mutation histories of high-affinity cells with structural affinities.** **A.** Dynamics of 25 GC simulations against antigen L6, with receptors of size 9, starting from random BCR sequences with affinity 0.0001 at least. Volume (number of B-cells) of GCs over time (left), the mean affinity of B-cells in the GC over time (middle) and the distribution of mutations in cells that left the GC during the first 21 days (right). **B.** The best three sequences and consensus of BCR sequences with very high affinity (> 0.75) for five GC simulations **C.** Network (forest) representing all the mutation histories of the cells (alive or dead) that existed during a single GC simulation. Nodes are sequences and the edges connect a mutated sequence in daughter cells to the mother sequence. The same sequence can arise from different cells and can be found in multiple nodes. One cluster (tree) represents all the progeny sequences of one founder sequence (i.e., one clone). The clusters are displayed starting from the founders with the most progeny. High-affinity cells (affinity > 0.5) are shown in dark red and are quantified beneath each founder dynasty. **D.** Details of one founder progeny where a big burst (hub) could be observed. As a comparison, the progeny of futile founders that did not reach high affinity is shown. **E.** Lineage history of high-affinity sequences (> 0.5). This is a sub-network of D by keeping only high affinity cells and their parents up to the founder sequence. The successive affinity increases with each mutation are shown from the founder sequence to the burst. **F.** To understand the history of mutations in Panel E, the graph of sequences is shown for high-affinity cells and ancestors, except they are merged if they share the same sequence. Affinity jumps can be observed up to 74—fold and the paths did not always increase in affinity. **G.** Progeny of high-affinity cells back to the founder cell for different clones in different GC simulations, illustrating different shapes of the mutation graph, all showing hubs with different degrees. In all network representations, node size and color represent the affinity: light blue for affinities less than 0.3, then color gradients follow, with red at affinity 0.6, purple at affinity 0.9 and black for the 1.0 and higher. The position of nodes is determined automatically by a visualization algorithm, Prefuse Force Directed Layout, that segregates components of the network in 2D, allowing to see the hubs.

Antigen cocktails with high similarity promote cross-reactivity. Next, we simulated GCs with two epitope variants with more or less similarity (Figure 4A). Using related epitopes that differed by two point mutations (similar epitopes) promoted higher affinity and diversity while slightly increasing the GC dynamics. An opposite effect was observed with less similar epitopes (four point mutations), with less affinity and diversity over time, and smaller GCs. For similar epitopes, the affinity of individual cells to both epitopes on Day 11 exhibited four different characteristics (Figure 4B). In some cases (GC4), highly cross-reactive cells emerged, although GC2 matured only to one antigen. In most cases, cross-reactive cells were enriched compared with the cells recognizing only one epitope.

We compared the dynamics of GC simulations and the affinity of individual cells in the case of a cocktail of four similar epitopes (Figure 4C, D). There was no significant difference between two and four unrelated epitopes, and cells could emerge with high cross-reactivity to two antigens but not all. In order to quantify this effect, we calculated the average affinity to each epitope from the best recognized to the least, for individual GCs, in the case of two or four similar epitopes (Figure 4E). Again, using two or four epitopes showed similar results, with a recognized driving epitope and the others showing less recognition and B-cell cross-reactivity. Therefore, the combination of similar epitopes is favorable for cross-reactivity and adding bigger cocktails of similar epitopes is not harmful in this regard. At the scale of many GCs, each GC had a different driving antigen and an immunization cocktail therefore induces different combinations of cross-reactivity in each GC, promoting the notion of stochastic diversity of GC reactions [31].

The model shows that using very similar epitopes supports the emergence of more cross-reactive antibodies. We believe this will support the design of better vaccines towards poly-reactivity, conserved recognition or broad neutralization.

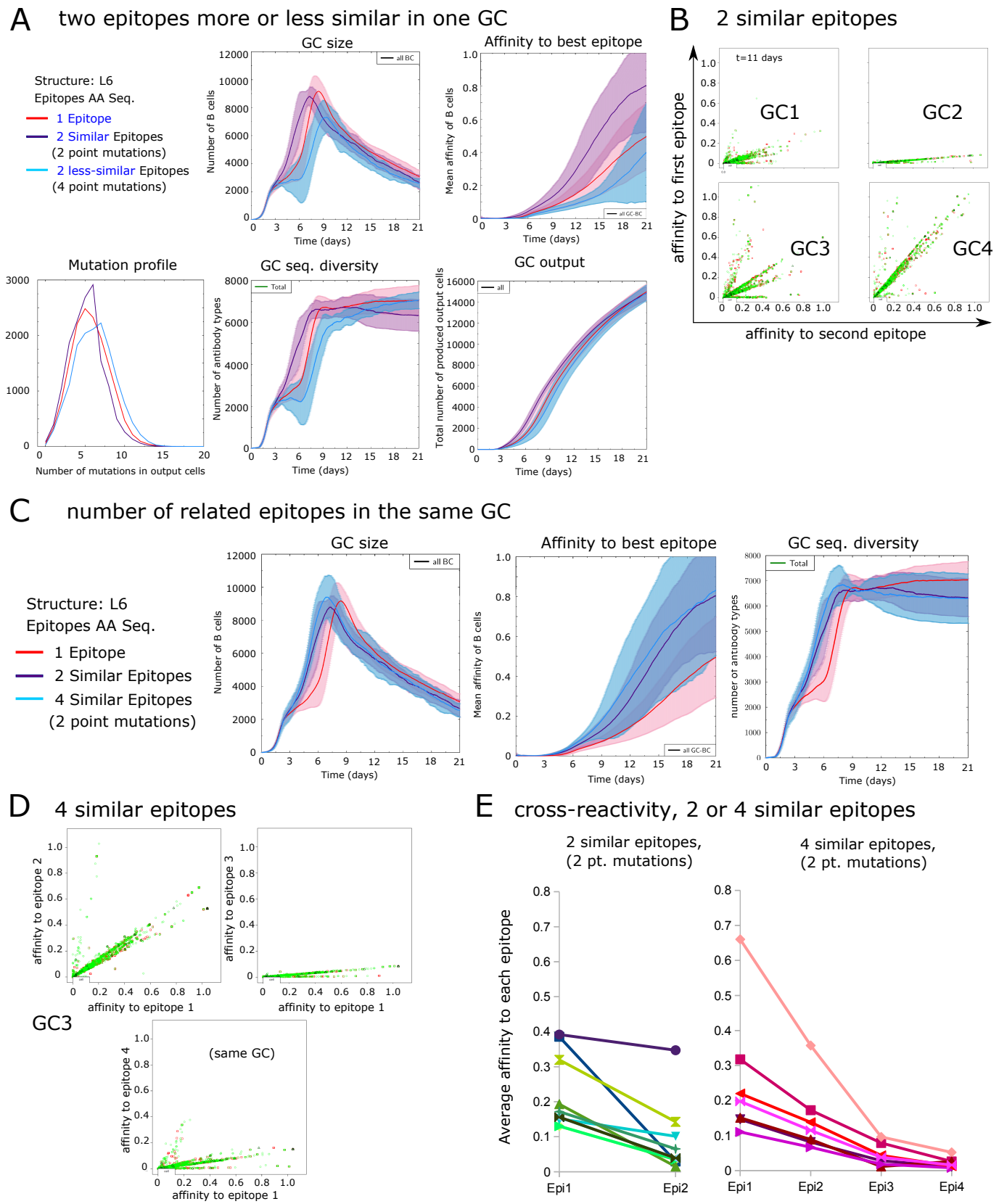

Figure 4: **The degree of similarity between epitopes impacts GC cross-reactivity.** **A.** Dynamics of GC and affinity maturation for two epitopes variants of the antigen structure L6 that differ by two or four point mutations. **B.** Affinity of single cells in four different GC simulations to two similar epitopes, which differ by two point mutations, on day 11. **C.** GC dynamics with two or four similar antigens (also differing by two mutations). **D.** Detail of single affinities in a single GC between each combination of epitopes (that all differ by two mutations) on day 11. **E.** Average affinity to each epitope of single GCs with two or four similar epitopes. Each line represents one GC. The epitopes are shown from the one with maximal average affinity in each GC (Epi 1), to the one with least average affinity (Epi 2 or 4). The order of epitopes differs between the GCs, implying that the GCs did not always recognize the same epitope best.

Discussion

We provide a realistic representation of antibody–antigen binding that can be simulated efficiently on regular computers while retaining important physiological features. The fields of vaccination and antibody design critically need such computational tools. The available tools either resolve the structural folding of antibodies or the dynamics of AM. Here, for the first time, we combine the two facets: an affinity that directly comes from a structural folding but also successfully simulating AM at full GC scale. Therefore, the mutation history and AM can be linked directly to the structural properties of the antibodies–antigen binding *in silico*. This representation can not only encompass the complex topology of real antigens but it can also simulate vaccination with either one antigen with complex epitopes or many related antigens as cocktails. Therefore, according to the antigen's structure, this technique enables the prediction of the structural component of epitope immunodominance and to predict the best combinations of antigens or epitopes for vaccine design based on their structural properties.

Several assumptions were made to achieve efficient computational time. First, the antigens and receptors were discretized on a 3D grid lattice. We do not pretend to simulate the exact affinity of antigen–antibody interactions, for example, because we simulate the folding of only one CDR loop around a pre-folded antigen. However, we believe the features and behavior of antibody–antigen recognition are well captured in our system for the following reasons:

The binding energy of each individual AA pair is taken from an experimentally derived energy potential [29]. We showed that many non-trivial properties of antibody–antigen binding are successfully reproduced. Hidden pockets are harder to recognize (Figure 1E), showing that the antigen's structure impacts the affinity. Cross-reactivity became evident by many unrelated antibody sequences that bound an antigen with high affinity (Figure 3B). In addition, we needed to show that correct AM behavior can be simulated at the scale of a complex 3D GC reaction. To this end, an agent-based model for cellular interactions in the GC, originally developed from an abstract mutation landscape (shape space), was reprogrammed to use structural affinities. We showed realistic dynamics of individual GC cell numbers and affinity (Figure 3A). As in the real world, the 'best' possible antibody sequence against an antigen was unknown to both us and the GC simulations. Interestingly, one GC could systematically produce sequences on its own with affinities much higher than those randomly selected.

A substantial variation could be observed between GCs. Each GC produced a different antibody sequence with high affinity, and some GCs performed better than others. In addition, during one GC simulation, around $70,000$ mutations were tested (including cells that died during the complete GC reaction) and the forest of mutations revealed some very interesting features (Figure 3C-G). First, the path to high-affinity cells was not always achieved by increasing affinity: sometimes, a combination of a mutation decreasing and then increasing their affinity led to higher affinity. Second, very different

affinity jumps could be observed. Along the path to becoming a high-affinity cell, typical jumps of 2x to 10x could be observed, and sometimes a mutation would create or disrupt affinity by up to 74-fold, which is associated with key-mutations in real GCs.

Therefore, we think that our structural affinities model, thanks to the complex structure-related affinity landscape, is a suitable tool for studying and comparing the structures of mutation trees within patients, and to possibly compare driving selection forces under pathological conditions.

Implications for vaccine design. The challenges for vaccine development come in different types. First, immunizing a subject with an antigen does not guarantee the desired epitope recognition. Some epitopes will become immunodominant, meaning the immune response successfully targets them and specific antibodies are raised against them, whereas other ones are weak and are not targeted. This gives a hierarchy of immunodominance among the different epitopes of the same antigen. Immunodominance can arise from different levels of accessibility: some epitopes are accessible and easily targetable, but other ones are hidden in pockets. Our model is designed with the aim of being able to represent the antigen's topology in a fair way on the lattice to mimic the core properties of the antigen. We have also shown that hidden or shielded pockets can easily be mimicked.

Immunodominance can also evolve over time, as observed during HIV vaccination, where easily accessible epitopes produce an early response and then other epitopes become targeted [34]. One potential reason for this time shift could be via antibody feedback, where previously produced antibodies come back and hide the earlier epitopes [19, 34]. Our model can easily be extended to account for antibody feedback in the following way: the B-cell antibody sequence and an earlier antibody compete to bind to the full antigen. If they occupy distinct positions in space (no overlap), their binding energy is not impacted. However, if they do, the mass action chemical kinetics at equilibrium can correct for the probability of binding of the B-cell antibody or BCR.

When the pathogen evolves at a high number of mutations and develops Tier 2 strains, like the hepatitis C virus, Dengue fever, influenza or HIV, successful vaccines need to produce antibodies against many strains at the same time, ultimately towards the golden grail of broad neutralization. Different cases have to be separated and this will impact the vaccine strategy, which is also related to the antigen's structure. For instance, highly mutating viruses sometimes harbor highly variable regions that are very accessible at their surface, while a conserved, more functional part is inside a core pocket and is harder to target. Therefore, the accessible immunodominant region tricks the immune system into recognizing regions that can escape easily. The first case of broad neutralization comes when an antibody recognizes the conserved region, thus binding most strains (conserved recognition). In this case the immune system succeeds in targeting the hidden epitope and the antibody does not need to be cross-reactive. Other cases of broad neutralization arise when the same antibody recognizes regions that are slightly

mutated in many strains, in which case, the antibody is cross-reactive to many related epitopes with a similar structure and only few mutations (promiscuous cross-reactivity) (for example, see [35] for the structural properties of bnAbs against HIV). It has also been reported that antibodies that recognize completely unrelated epitopes could have some advantages in the fight against HIV (poly-reactivity) [36]. It is obvious that the vaccination regimen for all three broad neutralization scenarii might be substantially different and will strongly depend on the antigen's structure. Earlier works have simulated cross-recognition with abstract affinities, where recognizing a shielding pocket is associated with a penalty on affinity [23]. Here, we have shown that GC simulations can be performed with multiple antigens whose structure and mutations can be decided and controlled, thus allowing the prediction of vaccine efficiency with cocktails of antibodies with any structural or mutation relationship. This paves the ground to assess which combinations of antigens are successful for achieving broad neutralization in many cases, providing an advance in the field of *in silico* design of antibody-related immunotherapies.

Methods

3D-lattice representation of proteins. Proteins structures are represented on a 3D Euclidean grid, where successive AAs should occupy neighboring positions in the grid (Figure 1A). Hard-core repulsion is assumed, meaning that two AAs cannot occupy the same position. Starting from a point in space, a protein structure is represented as a sequence of moves in the lattice, namely *straight* (S), *up* (U), *down* (D), *left* (L) or *right* (R). The first move can also be backwards (B). From predefined observer coordinates, each move is made relative to the previous observer coordinates and turns the observer in a new direction (except for a straight move). The rules for changing the observer coordinates are shown in Figure 5A together with an example (Figure 5B).

Therefore, one structure can be stored as (1) a starting position and (2) a string built on the alphabet [SUDLRB], where B can only be the first letter (named 'Absolute Representation' here). This unambiguously describes any correct protein structure (respecting hard-core repulsion) in a very compact form. Furthermore, the group of structures that are similar by translation or symmetry can be represented by forcing the first move to be 'S' and the first turn to be 'U', (thereby removing the 'B'). This is called the 'Relative Representation' of a structure (Figure 5C). Relative structures follow the regular expression: 'S* | S[S]*U[SUDLR]*' and does not need a starting position in space.

The use of relative structures has several advantages. It allows to enumerate and manipulate the possible structures by generating random strings according to the regular expression. Their representation is compact (an alphabet of five letters) and obvious collisions (two AAs at the same position) among three successive AAs are directly avoided by forbidding 'Backwards' inside a structure. Rotations can be performed around one covalent bond just by changing one letter and propagating a turn

in the observer coordinates, without computing the spatial positions of the AAs. It is also easy to fuse two proteins into a longer one. All possible structures resulting from the fusion, according to different rotations, can be done by fusing the sequences and propagating turns in the observer coordinates. It has to be noted that the same protein structure can be described from both ends, and the sequence of a relative structure can be easily reversed to describe it by starting from the other end.

Binding and total energies between two proteins. For two proteins (as in Figure 1), an interaction is defined by a pair of two non-consecutive AAs occupying neigbouring positions. Interactions can be within a protein (folding) or between two proteins (binding). To separate the notion of 'empty' structure and protein (that is a structure with AAs), we introduce the operator $P(R,S)$ that represents the protein of structure S and AA sequence R (both sharing the same length). The binding energy E_{bind} between a receptor protein $P(R,S)$ of length L and an antigen $P(G,K)$ of length L_{G}, is the strength of the interaction between the two proteins and is calculated as the sum of all interactions between them (Equation 1).

$$E_{\text{bind}}\big(P(R,S),P(G,K)\big) = \sum_{k=1}^{L_{\text{G}}}\sum_{j=1}^{L} \text{Touch}(S_j,K_k)A(R_j,G_k) \tag{1}$$

The successive positions in space of a structure S are denoted by S_i for the ith residue (independently of the type of AA at this position). The operator $'\text{Touch}(R_1,R_2)'$ returns 1 if the residues R_1 and R_2 are non-covalent neighbors, and $A(R_1,R_2)$ is the interaction potential between the residues types of R_1 and R_2, given by [29]. The folding energy E_{fold} of a protein $P(R,S)$ of length L, is the sum of (intrinsic) interactions between its own interacting AAs (Equation 2).

$$E_{\text{fold}}\big(P(R,S)\big) = \sum_{j=1}^{L}\sum_{k=1}^{L} \text{Touch}(S_j,S_k)A(R_j,R_k) \tag{2}$$

As we assume the antigen (ligand) has a static conformation, we can also define the 'total energy' E_{tot} of S around the antigen $P(G,K)$ as the sum of its binding and total energies (Equation 3).

$$E_{\text{tot}}\big(P(R,S),P(G,K)\big) = E_{\text{fold}}\big(P(R,S)\big) + E_{\text{bind}}\big(P(R,S),P(G,K)\big) \tag{3}$$

Energies are negative and calculated in kT units; therefore, the most favorable energies are the lowest.

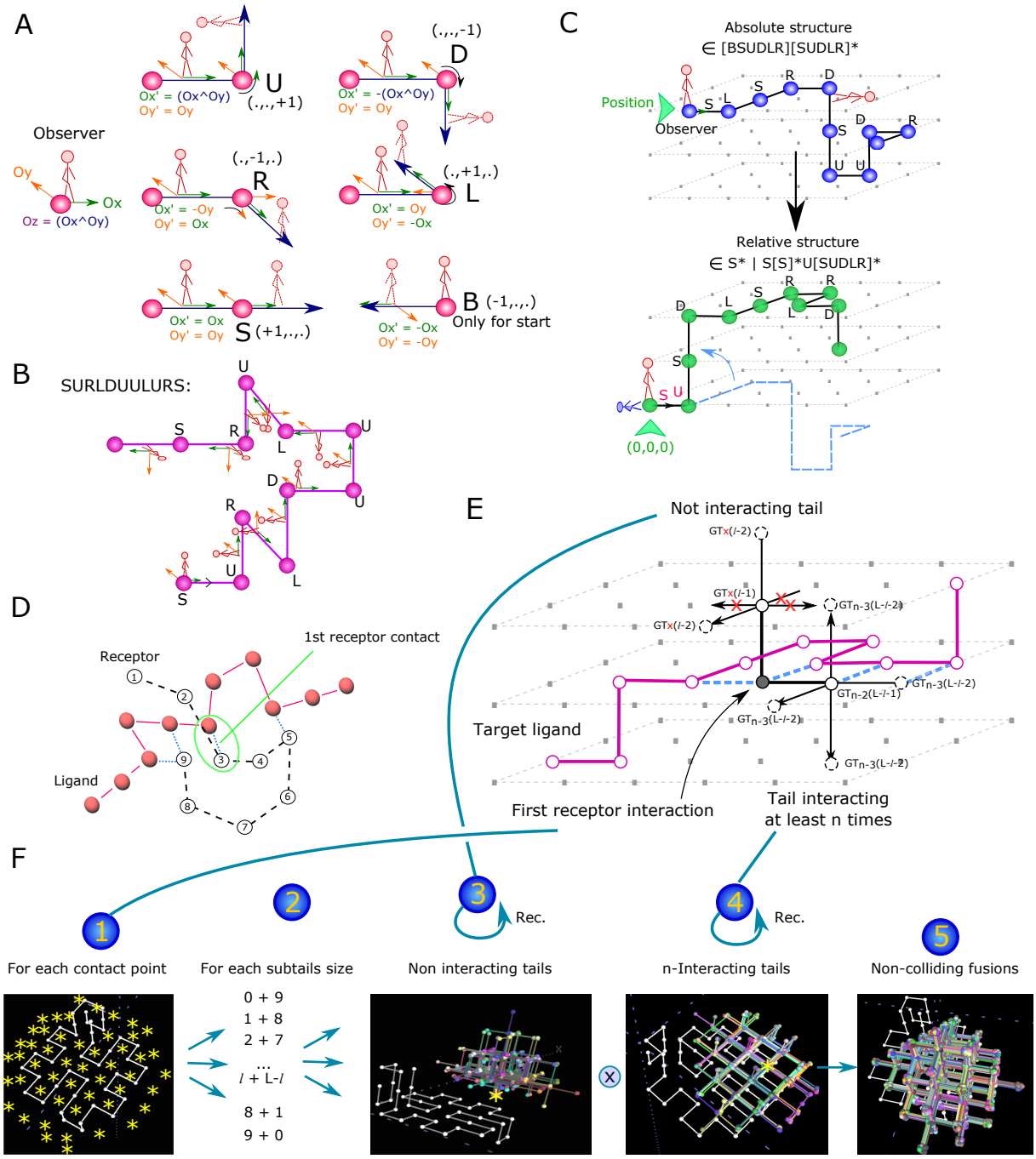

Figure 5: **Representation of proteins and enumeration of receptor foldings. A.** A protein structure is described as a list of 'moves' (see Figure 1B), that are relative to the observer coordinates (Ox, Oy, Oz). For each move, the new coordinates are given (Ox', Oy') and $Oz' = Ox' \wedge Oy'$. **B.** Detailed example of the sequence 'SURLDUULURS' and its structure. **C.** Proteins can be represented as 'Absolute', with a starting position in space, or as 'Relative' after translation and rotation to start with S and be followed by U, thereby representing all structures identical by rotation/translation in the same format. **D.** To enumerate all possible receptors binding a ligand, we define a 'first position of contact' to the ligand. By this definition, one of the receptor tails does not interact with the ligand. **E.** Recursive rules for enumerating all possible receptors with a certain 'first contact position'. All the possible structures for the two tails are enumerated separately, and each combination of structures from both tails is tried as a fusion to get a full, non self-colliding, receptor. The function Generate Tail (GT) of length l from a position P without contacts (red 'x') calls itself from the neighboring positions with length $l-1$; similarly, the function GT of length $L-l$ with at least k contacts calls itself from neighboring positions with length $L-l-1$ and k or fewer minimum contacts, depending on the number of contacts gained. **F.** Graphical illustration of the enumeration of all receptors described in E. For each possible first contact position, for each possible length of each tail with correct sum L, the possible interacting and non-interacting tails are enumerated from this position, concatenated (fused) and kept if they didn't collide.

Combinatorial enumeration of all possible foldings. From a predefined ligand, we explicitly enumerate all the possible 3D structures of the receptor of length L that harbors at least $n(\geq 1)$ interactions with the ligand.

Starting from the first residue of a receptor structure, we identify its first point of contact or interaction with the ligand (Figure 5D). By definition, the receptor residues before this point do not interact with the ligand, although residues on the other tail may do.

For a particular empty grid contact point X in the vicinity of a ligand residue, it is possible to recursively enumerate all receptor structures whose first contact point to the ligand is X. This will involve enumerating all non-interacting structures finishing at position X as the first step, and all structures that may also interact with the ligand, starting from position X. Next, each pair of starting and finishing structures should be checked for collisions with each other and should have the right total length L (Figure 5E).

The recursive algorithm 'GenerateTails' (Algorithm 1) shows how to get a starting or finishing structure, and the algorithm 'GenerateReceptors' (Algorithm 2) shows how to enumerate all possible receptor structures around a ligand, with a minimum number of interactions between receptor and ligand. In order to improve efficiency, the GenerateTails function will be called multiple times from the same point, and the result of each call is stored in the memory, avoiding excessive recomputing and explaining why the memory usage reaches a few GB of memory (Figure 1G).

Note that each starting point X describes a set of mutually exclusive structures if we consider that a receptor structure is oriented (first to last residue). Indeed, every structure (if they are not considered to be oriented) will be enumerated twice, once from each end, because the last contact point from one end is the first contact point from the other end. Algorithm 2 is still valid in the particular case where $n = 1$, because the structures will be enumerated twice in lines 18-23.

Finally, in order to minimize the computational time, a list of structures can be compressed further into a list of binding pairs of positions on the receptor and AAs on the ligand. Some structures produce the same list of binding pairs. By keeping track of which structures have which binding pairs in a dictionary, only the binding pairs need to be stored and evaluated for further exploration of the binding energies of receptor sequences. This step leads to a three- to fourfold increase in computational speed.

Algorithm 1 Function that generates the list of all possible structures of length l that start from a specific point and direction, and interact at least minInteract times with the protein or that do not touch the ligand.

```
procedure GENERATETAILS(l, firstPos, firstDirection, canInteract?, minNbInteract)
    if l==0 then
        return []                                                    ▷ empty vector
    end if
    if l==1 then
        return [Struct3D(firstPos, firstDirection)]                  ▷ single vector
    end if
    possibleTails ← []
    secondPos ← position(firstPos, firstDirection)
```

```
    for each neighbor N_i of secondPos do
        if (N_i ∉ ligand and !(touch(N_i) and !canInteract))) then
            nbNewContacts ← nbTouchPoints(N_i, ligand)
            nextTails ← generateTails(l-1, minNbInteract - nbNewContacts, canInteract, ligand)
            for each structure S in nextTails do,
                if ! collide(precursorPos, S) then
                    head ← Struct3D(firstPos, firstDirection)
                    possibleTails ← add(fuse(head, S))
                end if
            end for
        end if
    end for
    return possibleTails
end procedure
```

Algorithm 2 Main function to enumerate all receptors structures of length L that interact at least minInteract times with the ligand.

```
procedure GENERATERECEPTORS(L, minNbInteract, ligand)
    possibleReceptors ← []
    set<int> Envelope ← neighbors(ligand)                ▷ Enumerates all positions touching of the ligand
    for each position X_i in Envelope do
                                                         ▷ 1st case: the receptor starts here, can further interact.
        for each neighbor N_i of X_i do
            if Xi ∉ ligand) then
                vector<struct3D> uniTails ← generateTail(L, X_i, N_i, minInteract-1, canInteract=true)
                possibleReceptors ← add(uniTails)
            end if
        end for
                                                         ▷ 2nd case: prepares a list of couples of directions.
        CoupleDirections ← []
        for each direction d_1 ∈ [B, S, U, L, D, R] do
            for each direction d2 > d1 do                ▷ (to avoid repeating a couple)
                N1 ← neighbor(pos, direction d1)
                N2 ← neighbor(pos, direction d2)
                if !touch(N1, ligand) then
                    CoupleDirections ← add (d1, d2)      ▷ d1 tail should not touch
                end if
                if !touch(N2, ligand) then
                    CoupleDirections ← add (d2, d1)      ▷ d2 tail should not touch
                end if
            end for
        end for

        for each (d1, d2) in CoupleDirections do
            for subL from 1 to L do
                tailsRight ← generateTails(L - subL, X_i, d2, minNbInteract-1, canInteract=true)
                if tailsRight ≠ ∅ then
                    tailsLeft ← generateTails(subL, X_i, 0, canInteract=false)
                    for each s1 in tailsRight do
                        for each s2 in structLeft do
                            struct3D combined ← struct3D(X_i, fuse(revert(s1), s2))
                            if properlyFolded(combined) then
                                possibleReceptors ← add(combined)
                            end if
                        end for
                    end for
                end if
            end for
        end for
    end for
    return possibleReceptors
end procedure
```

The best binding energy of a receptor sequence to the ligand. In order to get the best binding energy between two AA sequences, (the ligand sequence with known structure, and the receptor sequence with unknown structure), the list of possible 3D receptor structures is pre-computed for the pre-defined ligand structure, according to the minimum number of interactions n. The binding energy and total energy between the two AA sequences is then calculated for each structure, using the receptor AA sequence in this structure. Two binding energies can be derived. The 'best binding energy' E_{best} is the average binding energy of all optimal structures in terms of 'total energy' E_{tot} (Equation 4). Self-folding can contribute to stabilizing a structure and therefore the best structure is not necessarily the one binding with the highest strength, but with the lowest total energy.

$$E_{\text{best}}\big(R, P(G,K)\big) = \underset{\text{Structures } S}{\text{Average}} \left(\begin{array}{l} E_{\text{bind}}\big(P(R,S), P(G,K)\big), \text{ such that} \\ E_{\text{tot}}\big(P(R,S), P(G,K)\big) = \min\limits_{i=1}^{n_S} \Big(E_{\text{tot}}\big(P(R,S_i), P(G,K)\big)\Big) \end{array} \right) \tag{4}$$

In the formula, R is the receptor sequence, G the antigen sequence and K the antigen structure, S are the possible enumerated receptor structures, and n_S is the number of such structures. If a receptor sequence folds on itself with better energy than the total energy around the ligand, then we assume no binding at all and the energy is NAN (Not A Number).

Alternatively, a 'statistical affinity' can be computed by applying a Boltzmann weight to each protein structure $P(R, S_i)$ according to its total energy (Equation 5). Many conformations could co-exist with adequate total energy, and this may not be described well by the set of optimal structures alone. kT is the Boltzmann coefficient and s_i is the ith structure. Z is the sum of weights among all possible structures, and is used as normalization coefficient.

$$\begin{aligned} E_{\text{stat}}\big(R, P(G,K)\big) &= \frac{1}{Z}\sum_{i=1}^{n_S} E_{\text{bind}}\big(P(R,s_i), P(G,K)\big) \exp\left(-\frac{E_{\text{tot}}\big(P(R,s_i), P(G,K)\big)}{kT}\right) \\ Z &= \sum_{i=1}^{n_S} \exp\left(-\frac{E_{\text{tot}}\big(P(R,s_i), P(G,K)\big)}{kT}\right) \end{aligned} \tag{5}$$

In theory, the Boltzmann distribution should be distributed according to all possible receptor structures, rather than only with respect to a particular minimum number of interactions n, even including those that do not interact with the ligand (self-foldings). Calculating the Z factor with lower $n < 4$ had no effect on Z (not shown) because structures with few contacts always had an infinitesimal small weight. We also computed all the possible self-foldings for different receptor sizes L, and found that the contribution of n to Z was negligible (not shown) because and only a few possible self-folding structures appeared compared with the high number of foldings around the ligand. Altogether, this shows that the choice to discard self-foldings and choose $n = 4$ is a valid assumption. Higher values of n may be used to increase speed, as shown in Figure 1F, provided that they still do not impact Z.

Transforming energy into affinity and binding probability. The binding energy between two folded structures only represents one sliced window of a CDR loop; therefore, the real antibody–antigen affinity needs to include the contribution of the full antibody, with two binding regions and the scaffold. Further, inside a GC, the affinity of a BCR to the antigen translates into capture and internalization of the antigen, a process that depends on many complex factors. Therefore, no direct affinity can easily be drawn. We define an empirical 're-scaled affinity' a of a receptor AA sequence R (Equation 6), to represent the probability of binding to the antigen, based on the previously defined binding energy E_{best} (Equation 4) and where an affinity of 1 represents a chosen very high binding energy E_{max} associated with saturation of the antigen capture probability. We needed to re-scale the energy with a coefficient C because the best structures were dominant in the affinity and GC simulations could not succeed because of low antigen binding.

$$\mathrm{a}\big(R, P(G,K)\big) = \exp\left(-\frac{E_{\text{max}} - E_{\text{best}}\big(R, P(G,K)\big)}{C}\right) \tag{6}$$

In Figures 3 and 4 the affinity of L2, L5, and L6 and receptors with $L = 9$ were calculated with: $E_{\text{max}} = -100$ and $C = 2.8$. E_{max} and C might need to be adapted when changing the receptor length L.

GC simulations. GC simulations were performed on an agent-based model programmed in C++ in the lab, for which the algorithm was described in [10, 37]. The model explicitly simulates the movement and encounters of B- and T-cells, capture of antigen by B-cells, T-cell help, proliferation, recirculation, death and exit from the GC. The model settings with constant inflow of founder cells and only one B-cell - T-cell interaction for T-cell selection were used (as in [38] and with reference parameter values detailed in [37]).

The present structural affinity model was incorporated into this framework, where a mutation is a random change in the AA located at a randomly picked position of the BCR. The mutation landscape was changed as compared with the previous shape space affinity model, and the selection parameters needed to be adapted accordingly. First, the uptake of antigen is defined in the model by a 'refractory time' between two antigen capture events. We needed to take a smaller value (0.001 hours, or 3.6 seconds) compared to [37]. It means, for a maximal affinity, corresponding to a capture probability of 1, an antigen can be captured every 3.6 seconds. The time-step of simulations, previously bigger, needed to be lowered to 0.001 hours to fit the shorter refractory time.

In the model, the individual number of B-cell divisions upon T-cell selection is derived from a Hill-function depending on the amount of captured antigen. We observed that the appearance of high affinity B-cells happened slower with the structural space compared to the shape space. We needed to lower the slope of the Hill function from a Hill-coefficient of 2.0 in [37] to 1.4, and kept the threshold

parameter K equal to 9 antigen uptake events. With a higher slope, cells with intermediate affinity did not manage to expand and the simulated GC population shrinked fast (not shown).

Each founder B-cell entering the GC carries a randomly picked BCR of length $L = 9$ AAs and with a minimum affinity of 0.0001, potentially allowing for an increase in affinity by a factor of 10,000 in a GC reaction. If we lower this 'entry' threshold further, the simulated GCs collapse and the mutations do not reach reasonable affinities in due time (data not shown).

In the model, Follicular Dendritic Cells (FDC) occupy a set of positions in space, and display a certain amount of epitopes at each of position to the B-cells. The spatial distribution of epitopes in the GC was adapted to be compatible with multiple epitopes. When using multiple epitopes, each FDC position was initially filled with an equal amount of each epitope, such that the total amount of antigen is kept the same as for single epitope simulations. B cells with random BCRs enter the GC reaction provided their affinity is higher than 0.0001 for at least one epitope. At each position, B-cells can access all of the epitopes simultaneously. The antigen capture probability was determined by the highest affinity to all epitopes at this position, and this highest affinity epitope is removed by one antigen unit at this position. Other epitopes are not captured.

Graph analysis. Each founder sequence was assigned a unique ID. For every mutation, a new unique ID is assigned to the mutated sequence, even if this sequence already exists in another cell. The mutation history network of one GC is created with sequence IDs as nodes and mutations as edges. It is actually a forest, where each founder dynasty is a tree. The network was analyzed with Cytoscape (Figure 3). To represent the network, the default Prefuse Force directed layout was used, which showed a cluster for each founder in a convenient way. For Figures 3C and G, a cluster was manually selected, the nodes with an affinity of more than 0.5 were selected and their parents were included up to the founder sequence. Again, the Prefuse Force directed layout was used to represent the graph. For Figure 3F, the tree in Figure 3E was selected, the nodes with identical sequences were merged and the network was shown again. The sequence logos (consensus) of Figure 3B were generated using Skylign (skylign.org).

Acknowledgements

This work was supported by the Human Frontier Science Program (RGP0033/ 2015), and a PhD fellowship granted by École Normale Supérieure de Lyon. We thank Victor Greiff, Rahmad Akbar and Gang Zhao for fruitful discussions and suggestions, and Megan Foster (LeafItToMe) for critical reading of the manuscript.

References

1. Linette, G. P. *et al.* Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood* **122,** 863–871 (2013).
2. Eroshkin, A. M. *et al.* bNAber: database of broadly neutralizing HIV antibodies. *Nucleic Acids Research* **42,** D1133–D1139 (2014).
3. Corti, D. *et al.* Tackling influenza with broadly neutralizing antibodies. *Current Opinion in Virology* **24,** 60–69 (2017).
4. Murira, A., Lapierre, P. & Lamarre, A. Evolution of the Humoral Response during HCV Infection: Theories on the Origin of Broadly Neutralizing Antibodies and Implications for Vaccine Design. *Advances in Immunology* **129,** 55–107 (2016).
5. Chen, E. *et al.* Broadly neutralizing epitopes in the Plasmodium vivax vaccine candidate Duffy Binding Protein. *Proceedings of the National Academy of Sciences* **113,** 6277–6282 (2016).
6. Tan, J. *et al.* A LAIR-1 insertion generates broadly reactive antibodies against malaria variant antigens. *Nature* **529,** 105–109 (2016).
7. Shingai, M. *et al.* Passive transfer of modest titers of potent and broadly neutralizing anti-HIV monoclonal antibodies block SHIV infection in macaques. *Journal of Experimental Medicine* **211,** 2061–2074 (2014).
8. Pegu, A., Hessell, A. J., Mascola, J. R. & Haigwood, N. L. Use of broadly neutralizing antibodies for HIV-1 prevention. *Immunological Reviews* **275,** 296–312 (2017).
9. Victora, G. D. & Nussenzweig, M. C. Germinal centers. *Annual Review of Immunology* **30,** 429–457 (2012).
10. Meyer-Hermann, M. *et al.* A theory of germinal center B cell selection, division, and exit. *Cell Reports* **2,** 162–174 (2012).
11. Zhang, J. & Shakhnovich, E. I. Optimality of mutation and selection in germinal centers. *PLoS Computational Biology* **6,** e1000800 (2010).
12. Oprea, M., Van Nimwegen, E. & Perelson, A. S. Dynamics of one-pass germinal center models: implications for affinity maturation. *Bulletin of mathematical biology* **62,** 121–153 (2000).
13. Keşmir, C. & De Boer, R. J. A spatial model of germinal center reactions: cellular adhesion based sorting of B cells results in efficient affinity maturation. *Journal of Theoretical Biology* **222,** 9–22 (2003).
14. Wang, P., Shih, C.-m., Qi, H. & Lan, Y.-h. A stochastic model of the germinal center integrating local antigen competition, individualistic T–B interactions, and B cell receptor signaling. *The Journal of Immunology* **197,** 1169–1182 (2016).
15. Perelson, A. S. & Oster, G. F. Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *Journal of Theoretical Biology* **81,** 645–670 (1979).
16. Meyer-Hermann, M., Deutsch, A. & Or-Guil, M. Recycling probability and dynamical properties of germinal center reactions. *Journal of Theoretical Biology* **210,** 265–285 (2001).
17. Gao, F. *et al.* Cooperation of B cell lineages in induction of HIV-1-broadly neutralizing antibodies. *Cell* **158,** 481–491 (2014).
18. Zhang, Y. *et al.* Germinal center B cells govern their own fate via antibody feedback. *The Journal of Experimental Medicine* **210,** 457–464 (2013).
19. Meyer-Hermann, M. Injection of antibodies against immunodominant epitopes tunes germinal centres to generate broadly neutralizingred antibodies. *bioRxiv,* 501247 (2018).

20. Robert, P. A., Marschall, A. L. & Meyer-Hermann, M. Induction of broadly neutralizing antibodies in germinal centre simulations. *Current Opinion in Biotechnology* **51,** 137–145 (2018).

21. Luo, S. & Perelson, A. S. Competitive exclusion by autologous antibodies can prevent broad HIV-1 antibodies from arising. *Proceedings of the National Academy of Sciences* **112,** 11654–11659 (2015).

22. Nourmohammad, A., Otwinowski, J. & Plotkin, J. B. Host-pathogen coevolution and the emergence of broadly neutralizing antibodies in chronic infections. *PLoS Genetics* **12,** e1006171 (2016).

23. Wang, S. *et al.* Manipulating the selection forces during affinity maturation to generate cross-reactive HIV antibodies. *Cell* **160,** 785–797 (2015).

24. Kauffman, S. A. & Weinberger, E. D. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology* **141,** 211–245 (1989).

25. Shakhnovich, E. & Gutin, A. Enumeration of all compact conformations of copolymers with random sequence of links. *The Journal of Chemical Physics* **93,** 5967–5971 (1990).

26. Brown, A. J. *et al.* Augmenting adaptive immunity: progress and challenges in the quantitative engineering and analysis of adaptive immune receptor repertoires. *arXiv preprint arXiv:1904.04105* (2019).

27. Song, Y. *et al.* High-resolution comparative modeling with RosettaCM. *Structure* **21,** 1735–1742 (2013).

28. Meyer-Hermann, M. Overcoming the dichotomy of quantity and quality in antibody responses. *The Journal of Immunology* **193,** 5414–5419 (2014).

29. Miyazawa, S. & Jernigan, R. L. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology* **256,** 623–644 (1996).

30. Zhou, J. O., Ton, T., Morriss, J. W., Nguyen, D. & Fera, D. Structural Insights from HIV-Antibody Coevolution and Related Immunization Studies. *AIDS research and human retroviruses* **34,** 760–768 (2018).

31. Tas, J. M. *et al.* Visualizing antibody affinity maturation in germinal centers. *Science,* 1048–1054 (2016).

32. Küppers, R, Zhao, M, Hansmann, M. & Rajewsky, K. Tracing B cell development in human germinal centres by molecular analysis of single cells picked from histological sections. *The EMBO journal* **12,** 4955–4967 (1993).

33. Cumano, A. & Rajewsky, K. Clonal recruitment and somatic mutation in the generation of immunological memory to the hapten NP. *The EMBO journal* **5,** 2459–2468 (1986).

34. Forsell, M. N., Kvastad, L., Sedimbi, S. K., Andersson, J. & Karlsson, M. C. regulation of subunit-specific germinal center B cell responses to the HIV-1 envelope glycoproteins by antibody-Mediated Feedback. *Frontiers in Immunology* **8,** 738 (2017).

35. Pancera, M., Changela, A. & Kwong, P. D. How HIV-1 entry mechanism and broadly neutralizing antibodies guide structure-based vaccine design. *Current Opinion in HIV and AIDS* **12,** 229 (2017).

36. Bournazos, S., Gazumyan, A., Seaman, M. S., Nussenzweig, M. C. & Ravetch, J. V. Bispecific anti-HIV-1 antibodies with enhanced breadth and potency. *Cell* **165,** 1609–1620 (2016).

37. Robert, P. A., Rastogi, A., Binder, S. C. & Meyer-Hermann, M. How to simulate a germinal center, 303–334 (2017).

38. Meyer-Hermann, M., Binder, S., Mesin, L. & Victora, G. D. Computer simulation of multi-colour Brainbow staining and clonal evolution of B cells in germinal centres. *Frontiers in Immunology* **9,** 2020 (2018).

Supplementary material

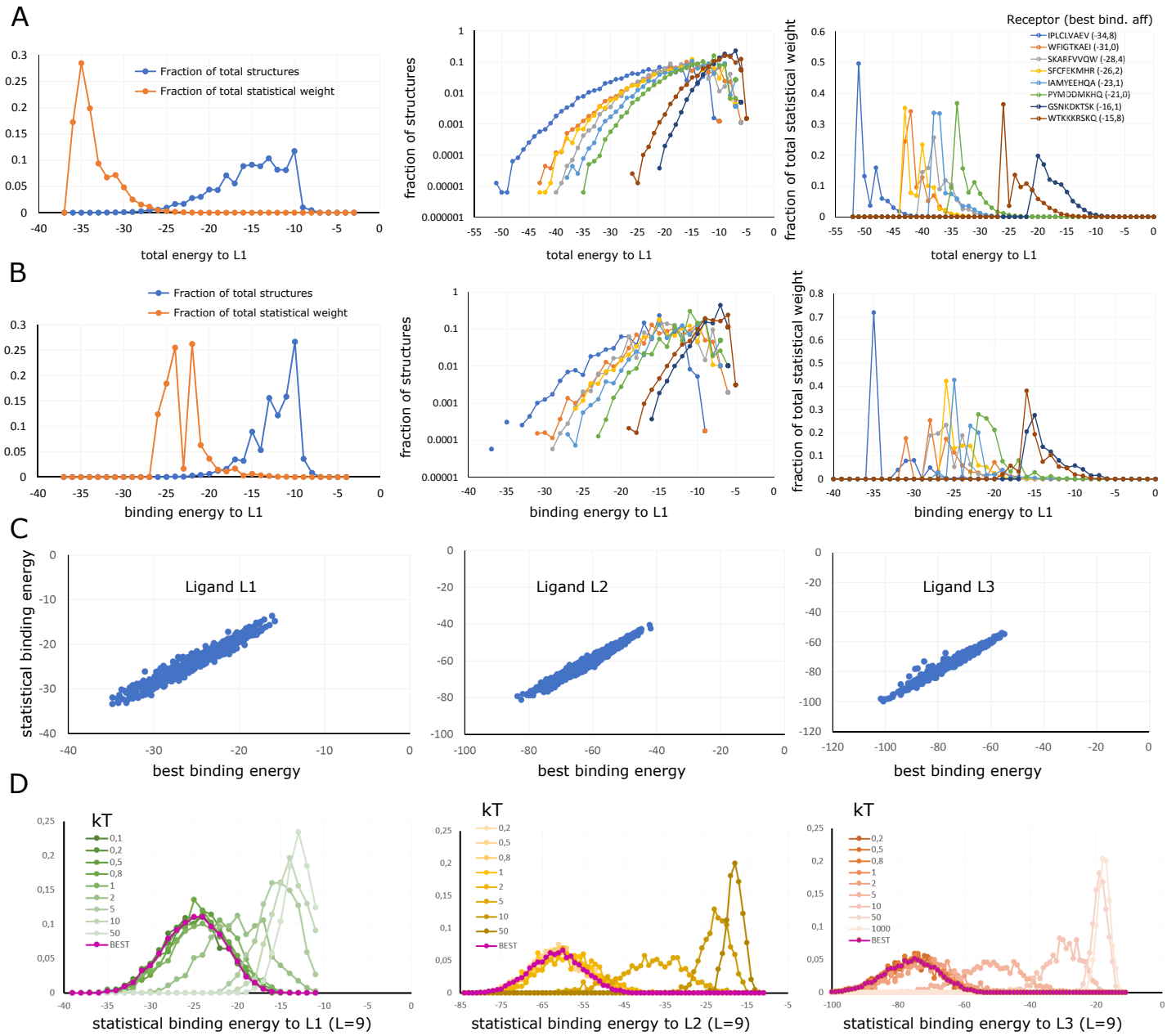

Figure S1: **The best binding energy is a fair approximation of the statistical binding energy. A-B.** Description of the energies of the ensemble of structures around ligand L1, for particular receptor sequences. **A.** Left: Distribution of total energies and statistical weight for possible structures of one receptor sequence. Most structures have higher (non-favorable) total energy, but only the ones with lowest energy have a significant statistical weight. Middle and right: Distribution of the total energy of the possible structures for selected receptor sequences with different best total energies, as fraction of the total statistical weight Z in logarithmic scale (middle) or as fraction of the total number of structures in linear scale (right). **B.** Same analysis with binding energies instead of total energies. **C.** Correlation between the best and statistical binding energy for randomly selected receptor sequences of size 9 against ligands L1, L2 and L3. **D.** Effect of temperature on the distribution of statistical binding energies of randomly picked receptors sequences of size 9 against each ligand. Temperature values around $kT = 1$ have no effect on the distributions, meaning frozen state is reached, where optimal structures dominate.