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; Centre for Individualised Infection Medicine (CIIM), Hannover, Germany; Cluster of Excellence RESIST (EXC 2155), Hannover Medical School, Hannover, Germany*

Abstract

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. 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. It is suitable for large simulations of affinity maturation. We perform *in silico* immunizations via germinal center simulations and show that our model can explain complex phenomena like recognition of the same epitope by 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 antigen 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, affinity maturation, antigen-antibody interaction, simulation

[*]philippe.robert@ens-lyon.org
[†]Current address: Clinical Immunology Department, Oslo University, Oslo, Norway
[‡]mmh@theoretical-biology.de

Introduction

Receptor-ligand recognition is at the basis of biological function, and defense by the adaptive immune system. Fast computational tools to undersand evolution or mutation imprint under affinity of binding are needed to understand and exploit the specificity of interactions.

Highly mutating pathogens like HIV, hepatitis C virus or influenza are poorly targeted by vaccines, and elicited antibodies do not necessarily protect again the next mutated strain. Strikingly, a few individuals naturally develop broadly neutralizing antibodies (bnAbs) against a large range of strains [1–5]. Injecting these bnAbs is protective against future infections in certain contexts [6, 7], demonstrating their therapeutic potential.

In view of billions of possible antibody sequences, engineering antibodies with target specificity, desired binding landscape, and no self-reactivity, is a timely and complex challenge. Meeting the requirements for the induction of bnAbs *in vivo*, especially in humans or primates, would support better vaccine design and the development of new bnAbs that are functional but not self-reactive. Antibody responses and ultimately generation of bnAb happens 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. Consequently, the affinity of the antibodies increases over time, a process called affinity maturation (AM).

The GC's response to single well-defined antigens has been studied *in vivo* [8]. 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 [9–15]. 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) sequences 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 [16]. 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 GC B-cells [17], potentially changing the immunodominance landscape of the response [18].

Recent mathematical models have accounted for the quantitative properties of AM towards multiple antigens and compared different vaccine strategies such as simultaneous cocktail immunizations or sequential vaccinations [19]. Each model comes with its own abstract encoding of binding affinities [14, 20–26]. We have previously reported [19] that these representations carry different levels of cross-reactivity. Only some can incorporate differential epitope accessibility, 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 necessity to be manually added to abstract models. Several prediction tools simulate protein folding thermodynamics, sometimes with the knowledge of known antibody–antigen structures [27]. Only a few antibody crystal structures are available so far. Full prediction tools like Rosetta [28], have reached powerful binding accuracy at the expense of few hundreds of computational hours for a single calculation.

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 a single reaction [29]. Therefore, the current available prediction techniques for structural antigen–antibody binding are unable to achieve a single GC simulation in feasible time.

Here, we develop a new hybrid model of antigen–antibody folding that simulates the best folding of the complementary determining region 3 (CDR3) around a predefined antigen structure on a 3D lattice. Real AA sequences are used, up to 12 AAs for the CDR loop, and with complex antigen structures up to 1000 AAs for the antigen. The interaction between neighboring AAs is based on experimentally measured potentials [30]. The hybrid model can derive binding energies in computational times suitable for GC simulations (a few minutes for 1000 affinity calculations). 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 an *in silico* model [9] with this structural antigen representation. The model showed physiological GC dynamics and proper AM. We show that the use of cocktails of similar antigens generated favorable conditions for cross-reactivity. 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.

We validated the presented method and package on the physiological relevance of somatic hypermutation in GCs, showing it is suited for physiological scale receptor-ligand interactions.

Results

A fast computational model for lattice-based antigen–antibody binding. Calculations of antibody-antigen binding energy were so far either simplified to an unphysiological degree or so expensive in computational time that they could not be used in real world problems involving many interactions. Here, we describe a fast algorithm that still captures the important physiological properties.

GC simulations need access to affinity changes by somatic hypermutations. Given a predefined antigen structure (the 'ligand'), binding energy and affinity have to be computed for a large number of mutated binding regions of the antibody (the 'receptor'), (Figure 1A). We defined a regular lattice, on which residues can occupy predefined positions on a grid (Figure 1B). Real AA sequences are used,

consecutive AAs are on neighboring points and only one AA is allowed on each grid point. A protein structure is represented as a list of moves, namely *Straight, Up, Down, Left, Right* (Figure 1B). All covalent bonds are equal and are neglected. Neighboring but non-covalently linked AAs are assumed to contribute to the binding strength. Their binding energy has been previously estimated from structural databases for each pair of AAs [30]. For two interacting proteins, the binding energy E_{bind} is estimated as the energetic sum of individual bonds between the two proteins. The total energy E_{tot} of one protein is the energetic sum of interacting and intrinsic bonds, thus, including intrinsic bonds as a stabilization factor of the protein structure (see Methods). We assume that the antigen has a stable conformation, which is not changed by the bound antibody. Instead, the CDR3 loop of the antibody folds onto it. For each CDR3 sequence, a folding has to be generated and the folding structure around the antigen with lowest total energy has to be found (Figure 1C).

Because of the exponentially high number of possible foldings, it is unrealistic to compute all foldings for each CDR3 sequence at every time-step in a dynamic computer simulation of GCs. We developed an optimized algorithm that pre-computes only the interesting possible foldings of CDR3 loops of realistic length L between 7 and 14 AAs. Starting from a predefined antigen structure, a recursive algorithm enumerates all possible foldings of the not yet specified CDR3 sequence, with the constraint of touching the antigen at least n=4 times (Figure 1C). This considerably reduces the amount of structures to be enumerated and stored in the memory. 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 the binding energy of the most stable structure, i.e., the structure with optimal total energy.

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 reflect 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 E_{bind} and the total energy E_{tot}. 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.

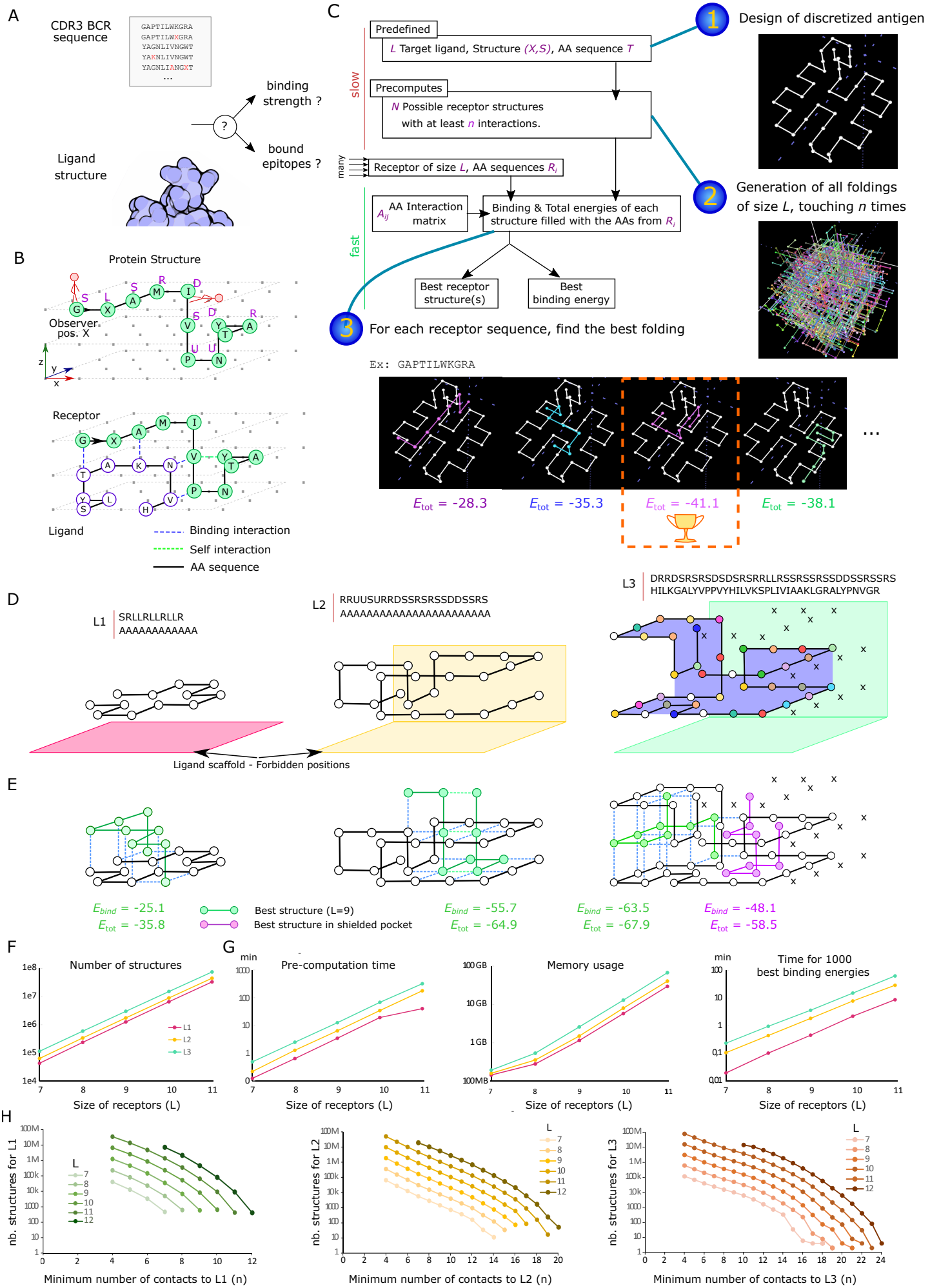
A
CDR3 BCR sequence
GAPTILWKGRA
GAPTILWXGRA
YAGNLIVNGWT
YAKNLIVNGWT
YAGNLIANGXT
...
binding strength ?
bound epitopes ?
Ligand structure
B
Protein Structure
Observer pos. X
Receptor
Ligand
Binding interaction
Self interaction
AA sequence
C
Predefined
L Target ligand, Structure (X,S), AA sequence T
Precomputes
N Possible receptor structures with at least n interactions.
slow
many
Receptor of size L, AA sequences R_i
AA Interaction matrix
Binding & Total energies of each structure filled with the AAs from R_i
fast
Best receptor structure(s)
Best 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_tot = -28.3
E_tot = -35.3
E_tot = -41.1
E_tot = -38.1
D
L1
SRLLRLLRLLR
AAAAAAAAAAAA
L2
RRUUSURRDSSRSRSSDDSSRS
AAAAAAAAAAAAAAAAAAAAAAA
L3
DRRDSRSRSDSDSRSRRLLRSSRSSRSSDDSSRSSRS
HILKGALYVPPVYHILVKSPLIVIAAKLGRALYPNVGR
Ligand scaffold - Forbidden positions
E
E_bind = -25.1
E_tot = -35.8
Best structure (L=9)
Best structure in shielded pocket
E_bind = -55.7
E_tot = -64.9
E_bind = -63.5
E_tot = -67.9
E_bind = -48.1
E_tot = -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, like 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 (blue). The binding energy is the sum of interacting bonds, and the total energy of the receptor additionally includes its intrinsic bonds. **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) For any specified receptor sequence, a binding energy to the ligand E_{bind} and total energy E_{tot} are derived from each structure of this subset. Then, the 'best binding energy' (average binding energy over all structure(s) with minimum total energy) is returned (see Methods). All equally optimal structures are 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 structures (green for L1, L2; purple for reaching the pocket of L3) are shown with the associated binding energy E_{bind} and total energy E_{tot} in kT. **F.** Number of structures for each ligand, depending on the length of receptors L (in AAs), when at least $n = 4$ contacts are required. **G.** Time and memory requirements for pre-computing the structures and then calculating the best energy of 1000 receptor sequences. **H.** Total number of sequences as a function of the minimum requested number of contacts n.

For larger CDR3 sequences, up to 75 millions folding structures exist (Figure 1F). The pre-computation time still did not exceed a few hours (Figure 1G) and has to be performed and saved only once. Thousands of CDR3 can be evaluated for best binding energy to these antigens in less than one hour on a single CPU (Figure 1G), implying that it becomes possible to simulate a full GC within a few hours. By increasing n, the required minimum number contacts to the antigen, calculations are further speeded up (Figure 1H). 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. Thus, the structural model for antibody–antigen binding has the capacity to model complex antigens with an efficiency suitable for GC simulations.

Structural properties of the model.

Antibody specificity is reflected in the model. Thousands of random receptor sequences of length 7 to 11 AAs were sampled. The best binding energy of the receptors varied according to a broad Gaussian distribution in the range of 20, 40 and 50 kT for L1, L2 and L3, respectively (Figure 2A), thus, generating different binding landscapes for each sequence. Despite millions of possible good folding patterns, there are receptors still not matching this antigen. Thus, the calculated best binding energy contains structural information about the receptor sequences. The best 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.

Steric hindrance of epitope access. The impact of the antigen structure onto the binding energies was investigated on a bigger version of ligand L1 with a similar flat shape. It induced a global increase in binding (Figure 2B), due to more available positions, especially corners where three AAs 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 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 ligand L7 (Figure 2C). The distribution of best binding energies remained similar because the receptor structures accessing the pocket were not favored. This shows that the model is a suitable tool for studying the effect of steric hindrance of epitopes such as HIV GP120 protein [31].

Antigen AA complexity. Changing the AA composition of L1 from alanine only to diverse AAs (L5) (Figure 2B) induced a strong improvement of the best binding energy. If an antigen contains only one type of AAs, receptors do 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. All receptor sequences for L6 displayed best binding energies between -90 and -45 kT (Figure 2C). We tested all possible point mutations for eight sequences of length 9 with different best binding energies to L6 (Figure 2D). Mutations could always do both, increase or decrease the best binding energy, even when starting from sequences with very high (green) or very low (purple) energy. Large energetic shifts of ± 15 kT were possible by single mutations.

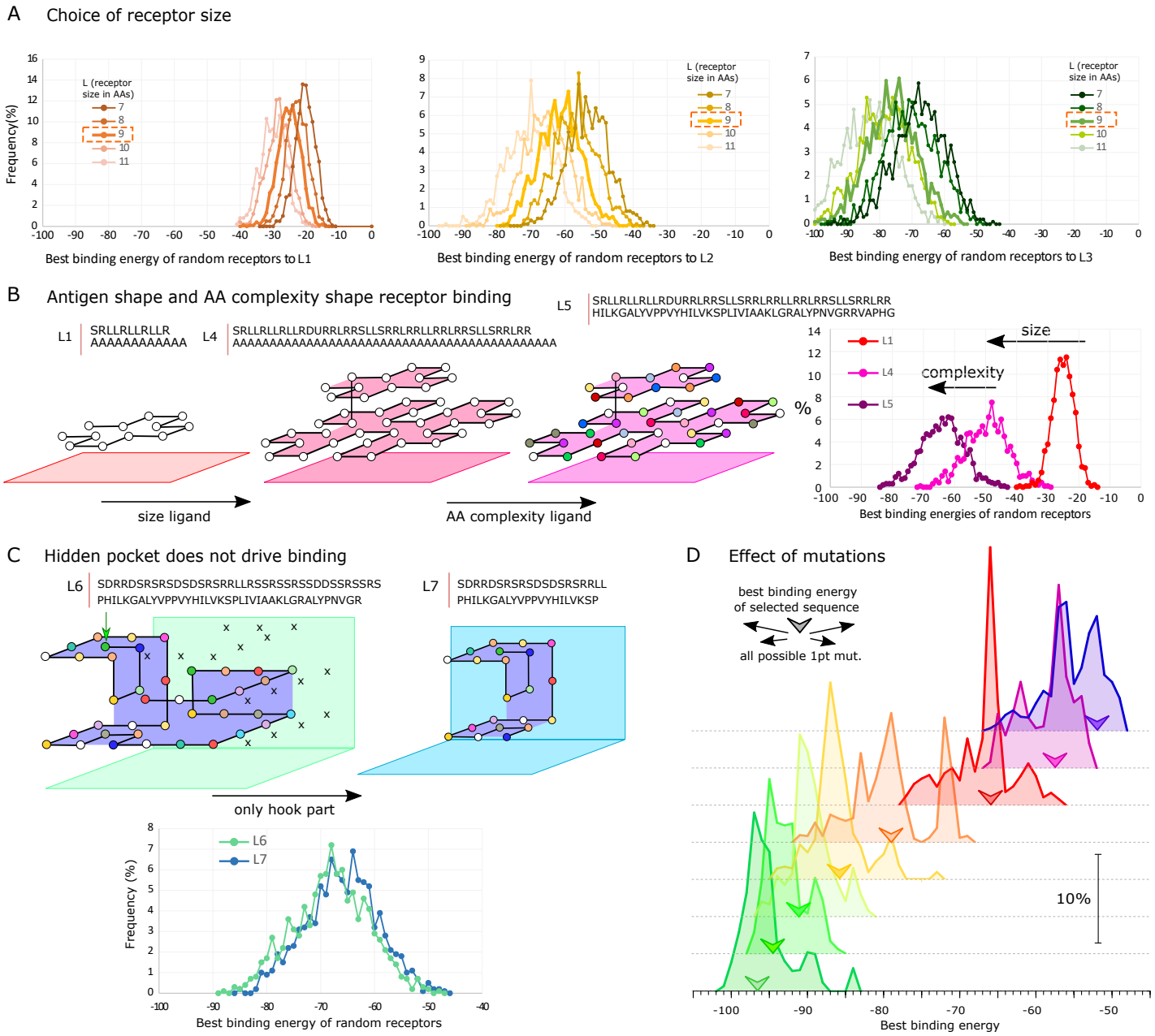

Figure 2: **Distributions of receptor binding energies.** The best binding energy of 2500 randomly sampled receptors **(A)** of length 7 to 11 to ligands L1, L2 and L3; **(B)** of length 9 to ligands L1, L4 and L5; and **(C)** to L6 and L7. The structure of the ligands L2 and L3 are as depicted in Figure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

GC simulations show affinity jumps and clonal 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 the structural binding model into an agent-based model for GC dynamics, where B-cells and T-cells are explicitly modelled to move, interact, be selected, proliferate, die or exit in 3D space [9, 32] (see Methods). The best binding energies derived from the structural binding model were mapped to an affinity (see Methods, Eq. (6)). In the simulation, the affinity between the BCR and the antigen is identical to the probability of capturing the antigen, which is needed later to survive T-cell selection.

With suitable values for these two parameters (see Methods), realistic GC dynamics and AM for a single complex target antigen L6 were found (Figure 3A). The number of B-cells over time in single GCs expands first, reaches a maximum at day 8, followed by slow decay. Starting randomly from cells with at least 0.0001 affinity, the average affinity of B-cells reached 0.3 to 1 after 21 days. The average number of mutations of cells leaving GCs was around six mutations, as estimated in [33]. The high affinity receptor sequences from five GCs were different (Figure 3B). In contrast, in most abstract affinity models the high affinity sequence is unique [19].

Next, we analyzed all receptor sequences generated by mutations during 21 days of a GC reaction. We retrieved all (around $70,000$) mutations that occurred in selected or dead cells and found that more than one GC founder cell achieved high affinity (Figure 3C), thus, pointing to parallel AM of multiple founder cells *in silico*. This is in accordance with the observation that some GCs are dominated by single clones while others display coexistence of multiple founder clones with high affinity [32].

To characterize the mutation history of high-affinity cells back to the founder cell, we selected the mutation cluster of a winning or losing founder cell in Figure 3D. The former showed an explosive expansion of a sequence into many mutated daughter sequences with high affinity. More bursts (hubs in the trees) happened within both the winning and losing clones (Figure 3C). The history of affinity changes during AM leading to high-affinity sequences revealed that increases and decreases of affinity happened, typically by 2— to 10—fold, before expansion of the high affinity sequence (see founder with affinity 0.009 in Figure 3E). In the simulations, not exclusively advantageous sequences are selected, which well reflects the stochastic nature of selection observed in [32, 34].

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, single point mutations inducing a 74—fold increase in affinity were observed, similar to key mutations like the acquisition of L33 in antibodies against ovalbumin [35]. Thus, the possibility of key mutations is captured by our model.

The comparison of the history of winning founder cells from different GC simulations (Figure 3G) illustrates how different mutation patterns can be. All of them show clonal bursts to rather different degrees. Few reports have studied single cell sequences inside a single GC. Interestingly, these bursts have been observed in a subset of GCs [32] in response to protein antigens. This indicates that our structural model reflects several non-trivial physiological properties of GCs.

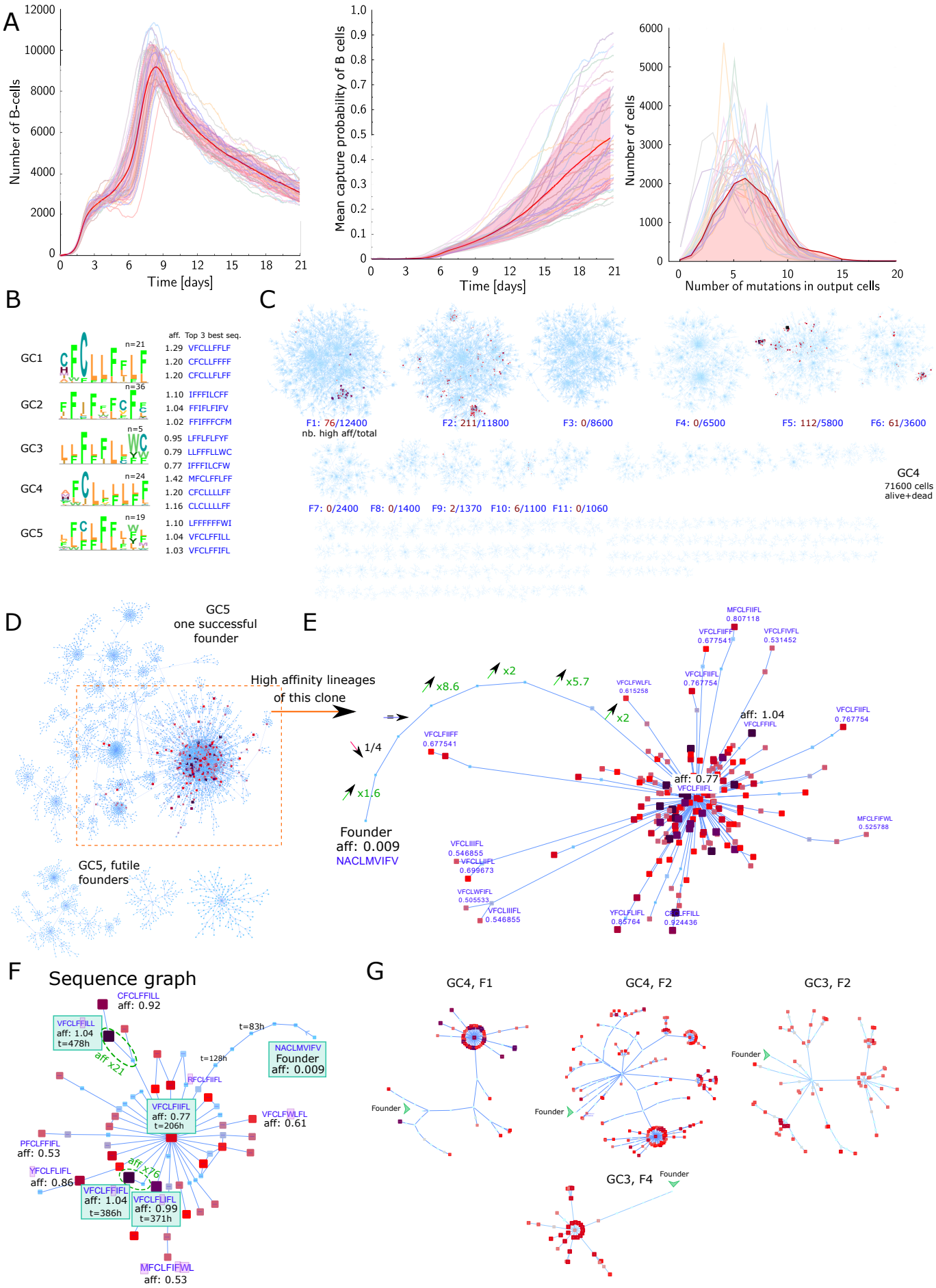
A
Number of B-cells
Time [days]
Mean capture probability of B cells
Time [days]
Number of cells
Number of mutations in output cells
B
GC1
GC2
GC3
GC4
GC5
aff. Top 3 best seq.
1.29 VFCLLFFLF
1.20 CFCLLFFFF
1.20 CFCLLFLFF
1.10 IFFFILCFF
1.04 FFIFLFIFV
1.02 FFIFFFCFM
0.95 LFFLFLFYF
0.79 LLFFFLLWC
0.77 IFFFILCFW
1.42 MFCLFFLFF
1.20 CFCLLLLFF
1.16 CLCLLLLFF
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
aff: 1.04
aff: 0.77
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 x76
VFCLFFIFL
aff: 1.04
t=386h
VFCLFLIFL
aff: 0.99
t=371h
MFCLFIFWL
aff: 0.53
G
GC4, F1
GC4, F2
GC3, F2
GC3, F4
Founder

Figure 3: **Full scale *in silico* GC simulations and mutation histories of high-affinity cells. A.** Dynamics of 25 GC simulations against antigen L6, without glycans, and with receptors of size 9, starting from random BCR sequences with at least 0.0001 affinity. The number of GC B-cells (left), the mean affinity of GC B-cells (middle) and the distribution of mutations per GC output cell (right). **B.** The best three sequences and consensus of BCR sequences with very high affinity (> 0.75) from 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. Each panel represents all the progeny sequences of one founder sequence (i.e., one clone). The panels are displayed starting from the founders with most progeny. High-affinity cells (> 0.5) are shown in dark red and are quantified beneath each founder dynasty. **D.** Details of one founder progeny where a big burst could be observed, and the progeny of futile founders that did not reach high affinity. **E.** Lineage history of high-affinity sequences (> 0.5) of the successful founder in D, including the successive affinity increases with each mutation from the founder sequence to the burst. **F.** Same as E, except that identical sequences are merged. The time of first appearance of the sequence and selected large affinity jumps are indicated. **G.** Progeny of high-affinity cells back to the founder cell for different clones in different GC simulations. In all network representations, node size and color gradually represent affinity: light blue for affinities < 0.3, red at 0.6, purple at 0.9, and black for ≥ 1.0. The position of nodes is determined automatically by the 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 same structure (L6), but more or less sequence similarity (Figure 4A). Interestingly, the GC dynamics and AM were not impacted by the addition of the second epitope variant. For similar epitopes, the simulated GCs could be classified in four groups according to cell affinity to both epitopes at day 11 (Figure 4B). In some GCs, highly cross-reactive cells emerged (GC4), in others the GCs matured to one antigen only (GC2). In most cases, cross-reactive cells were enriched compared with the cells recognizing only one epitope (GC1,3). Therefore, high variability could be observed between GCs and the overall affinity to the best antigen did not reflect the success to generate high cross-reactivity.

Increasing the number of similar epitopes to four did not induce major changes in the GC characteristics (Figure 4C). However, B-cells developed high reactivity to two but not to all antigens (Figure 4D). In order to quantify this effect, we calculated the average B-cell affinity to each epitope for individual GCs with two or four similar epitopes, starting from the best recognized epitope in each GC (Figure 4E). The concomitant use of four similar antigens induced a slightly higher affinity to the two best recognized epitopes as compared to GCs with only two similar antigens. Further, we assessed the average cross-reactivity of each cell for each GC, measured as the minimum affinity to the two best recognized epitopes (Figure 4F). The average cross-reactivity was slightly increased by the combination of four similar epitopes instead of two. Therefore, adding more similar epitopes is not harmful for cross-reactivity but rather supports better recognition of the epitopes.

The model shows that using combination of similar epitopes enhances the chance of generating cross-reactive antibodies, even if less GC founder cell recognizes each epitope.

We believe this will support the design of better vaccines towards poly-reactivity, conserved recognition or broad neutralization.

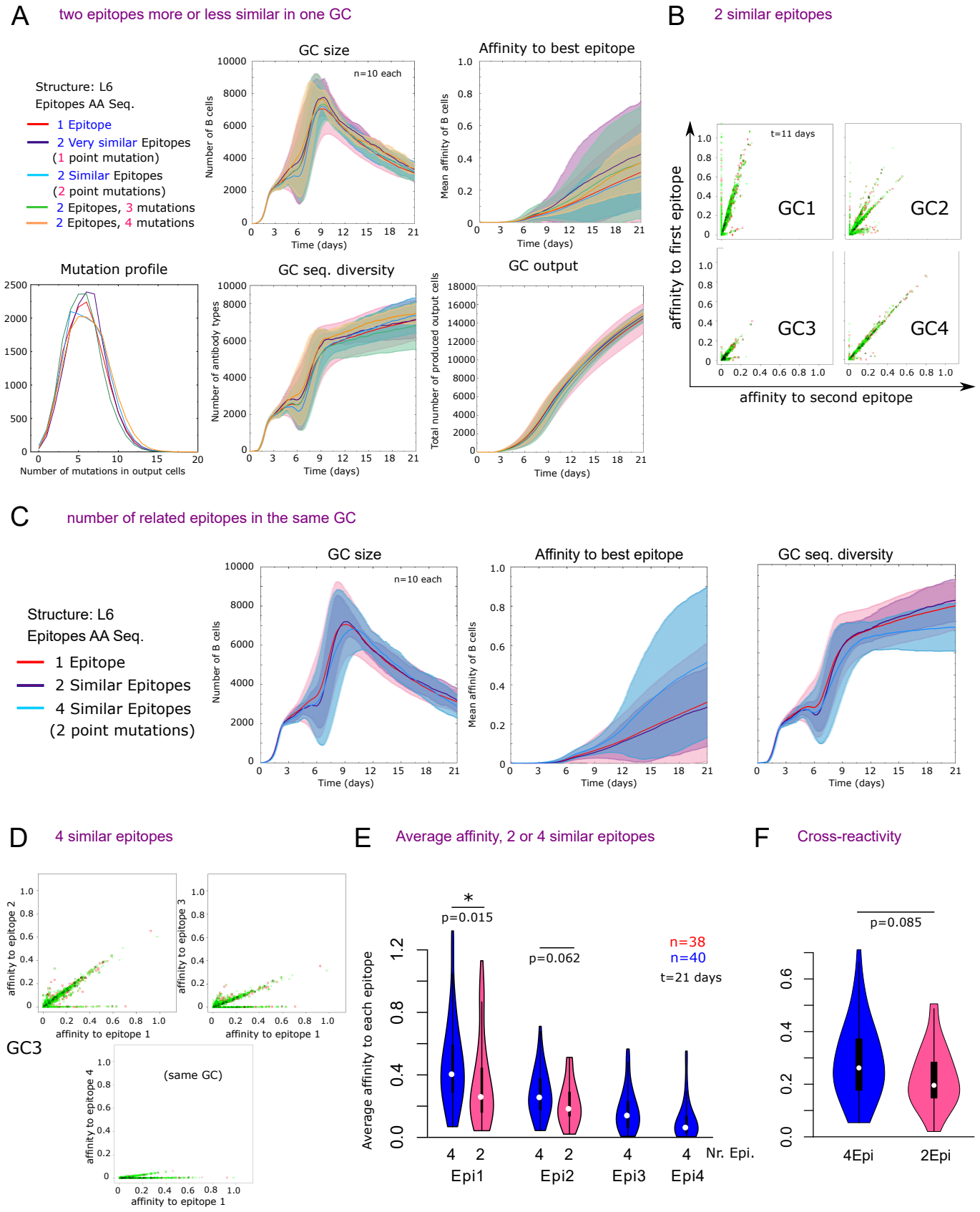

Figure 4: **Dynamics and cross-reactivity of GCs with two or four antigens. A.** Dynamics of GC and affinity maturation for two epitopes variants of the antigen structure L6 that differ by one or two point mutations. **B.** Affinity of single cells in four different GC simulations to two similar epitopes, which differ by two mutations, at day 11. **C.** GC dynamics with two or four similar epitopes (differing by two mutations). **D.** Affinity of single cells in a single GC to pairs of epitopes at day 11. **E.** Average affinity to each epitope of single GCs with two or four similar epitopes. The epitopes are sorted from highest (Epi1) to lowest (Epi2 or 4) average affinity in each GC. **F.** Average cross-reactivity of single GCs raised with two or four similar epitopes, defined as the minimum affinity to the two best recognized epitopes, for each cell.

Discussion

A realistic representation of antibody–antigen binding that can be simulated efficiently on regular computers was presented. For the first time, an affinity directly derived from a structural *in vivo*-like folding was used in full scale GC simulations of AM. The mutation history and AM can directly be linked to the structural properties of the antibodies–antigen binding *in silico*. This representation can encompass the complex topology of real antigens, and simulate vaccination with either one antigen with complex epitopes or cocktails of many related epitopes. The structural affinity model can be easily transferred to other applications like thymic selection.

Several assumptions were made to achieve efficient computational time. The antigens and receptors were discretized on a 3D grid lattice. Instead of enumerating receptor structures and docking them, we perform both simultaneously, allowing to discard huge amounts of possible structures. We do not pretend to simulate the exact affinity of antigen–antibody interactions: Real affinities will not be captured by simulation of the folding of a single CDR loop around a pre-folded antigen but is influence by many other surrounding factors.

However, critical features and behavior of antibody–antigen recognition are well captured in our system: The binding energy of each individual AA pair is taken from an experimentally derived energy potential [30]. Many non-trivial properties of antibody–antigen binding were successfully reproduced, like less easy recognition of hidden pockets and the emergence of many unrelated antibody sequences binding an antigen with high affinity. AM was correctly simulated at the scale of a complex 3D GC reaction. Realistic dynamics of individual GC cell numbers and affinity could be obtained. A substantial variation could be observed between simulated GCs, ranging from monoclonal bursts to polyclonal co-existence of clones, which is in line with the stochastic nature of GCs [32]. Each GC produced a different antibody sequence with high affinity, and some GCs performed better than others. The mutation path to high-affinity cells was not stringent and allowed for intermediate reductions of affinity. Along the path to high-affinity, typical affinity jumps of 2- to 10-fold and up to 74-fold were observed, which is associated with key-mutations in real GCs [36]. These results together provide evidence that the structure-related affinity landscape is sufficiently complex for studying structural implications of antibody development.

Thanks to new methods for enumerating and clustering possible peptide backbone structures [37], it will become possible to adapt the present algorithm to more realistic space conformations by discarding improper angles using rotamer libraries on finer lattices [38], and using approximate scoring functions for receptor-ligand affinities, as for protein docking. The use of real Amino Acids will allow to include out of frame and SHM hotspots into nucleotide sequences.

Several studies have used mutation trees of B-cell repertoires as readout of selection signatures [39]. The robustness of those methods may be improved by generating affinity matured repertoire benchmarks [40] with structurally related antigens and proposing the appropriate size and time-points of sequencing datasets to be analyzed [41].

Vaccine development is hindered by the hierarchical immunodominance of some epitopes over others, on the same antigen [42]. Some epitopes may be accessible and easily targetable, while others are hidden in pockets [43]. By reflecting accessible, shielded and hidden epitopes, the structural affinity model is suitable to study causes of immunodominance.

Immunodominance can also evolve over time, as observed during HIV vaccination [44]. Antibody feedback has been proposed to contribute to hiding immunodominant epitopes [18]. Our model can easily be extended to account for antibody feedback: The B-cell antibody sequence and an earlier produced antibody can compete to bind to the full antigen using mass action kinetics at equilibrium provided their target structures overlap, although such antagonism might be more complex [45]. In this setting, time-shifting of immunodominance will emerge naturally.

Vaccines against highly mutating pathogens like the hepatitis C virus, Dengue fever, influenza or HIV need to raise antibodies against many strains at the same time, ultimately towards the holy grail of broad neutralization. For instance, HIV envelope accessible regions are highly variable, while the conserved core pocket is harder to target [43]. The accessible immunodominant region diverts the immune system into recognizing regions that can escape easily without impact on viral function. Broad neutralization is induced by antibodies recognizing the conserved region like the CD4 binding site of HIV Env protein or influenza Haemaglutinin, thus binding most strains (conserved cross-recognition). But also cross-reactivity of antibodies that recognizes slightly mutated regions (promiscuous cross-reactivity) [46] or completely unrelated epitopes (poly-reactivity) [47] on various strains have the potential of broad neutralization. It is obvious that the best vaccination strategy depends on the antigen structure and the associated kind of cross-reactivity.

Simulations have the potential to inform the best vaccination strategy. Earlier works have simulated cross-recognition of conserved region with abstract affinities with a penalty on affinity for recognition of a shielding pocket [22]. Here, we have shown that GC simulations can be performed with antigens of variable and controllable structure and mutation landscape. This enables predictions of vaccine efficacy with cocktails of antibodies with any structural or mutation relationship, not only in the case of conserved recognition.

Further, by simulating waves of mutant viruses over a long period of time, it has been proposed [48] that the breadth of recognition by Tfh cells in GCs can determine the emergence of broadly neutralizing antibodies. More generally, the structural affinity model could be used to simulate the co-evolution of virus protein and GC B- and T-cells, as was performed at the repertoire level [21]. Altogether,

the presented model paves the ground for computer simulations of AM and conditions that favor the development of bnAbs, thus, 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 inspired from previous single-protein lattice models [24, 49]. Successive AAs occupy neighboring positions in the grid (Figure 1A). Only a single AA can occupy the same position. Starting from a grid point, a protein structure is represented as a sequence of moves on the grid, 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 (Figure 5A,B).

For each structure, a unique 'absolute structure' is defined by (1) a starting position and (2) a string built on the alphabet [SUDLRB], where B can only be the first letter. The 'relative structure' is defined as the group of absolute structures identical by translation or symmetry (Figure 5C). This is achieved by forcing the first move to be 'S' and the first turn to be 'U', thereby removing the 'B'. Relative structures follow the regular expression: 'S* | S[S]*U[SUDLR]*' and do not need a start position on the grid.

The use of relative structures has several advantages. It allows to enumerate and manipulate possible structures by generation of random strings according to the regular expression. The representation is compact (an alphabet of five letters) and collisions (two AAs at the same position) among three successive AAs are avoided by removal of 'B'. Rotations around a covalent bond are performed by changing a single letter, inducing a turn of the observer coordinates and propagating changes of the turn names according to the new coordinates, but without computing the new 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 found by fusing the sequences and performing turns of 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 neighbor positions. Interactions can be within a protein (folding) or between two proteins (binding). To distinguish the notions 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:

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

The ith grid positions of a structure S is denoted by S_i, irrespective of the type of AA at this position. The operator $\text{Touch}(S_j,K_k)$ returns 1 if the residues S_j and K_k are non-covalent neighbors. $A(R_j,G_k)$ is the interaction potential between the residues types given by [30]. 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:

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

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

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

Energies are negative and calculated in kT units. Strong binding implies low energies.

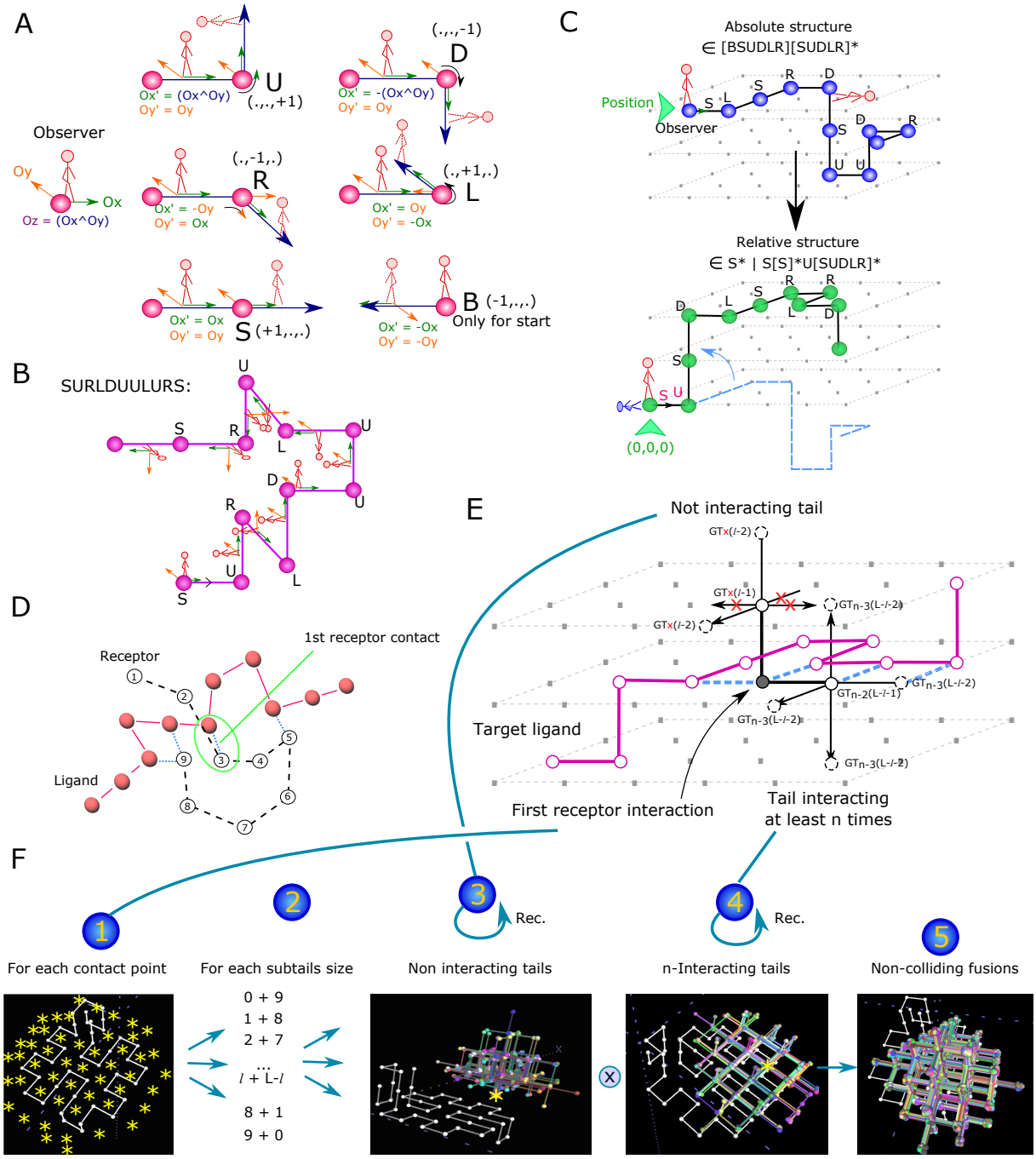

Figure 5: **Representation of proteins and enumeration of receptor foldings. A.** A protein structure is described as a list of 'moves' (see Figure 1B), that are relative to the observer coordinates $(\overrightarrow{Ox}, \overrightarrow{Oy}, \overrightarrow{Oz})$. For each move, the new coordinates are given $(\overrightarrow{Ox'}, \overrightarrow{Oy'})$ and $\overrightarrow{Oz'} = \overrightarrow{Ox'} \wedge \overrightarrow{Oy'}$. **B.** 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 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. Then, one of the two receptor tails does not interact with the ligand. **E.** Recursive rules for enumerating all possible receptors with a specific '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 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 with at least $n(\geq 1)$ interactions with the ligand. Starting from the first receptor residue, its first interaction point with the ligand is localized (Figure 5D). By definition, further interactions with the ligand have to be behind this point.

From a particular empty grid point X that is neighbor of a ligand residue, it is possible to recursively enumerate all receptor structures with X the first contact point to the ligand. All structures without interaction until X and all structures that may interact with the ligand behind position X are enumerated. Next, each pair of starting and finishing structures are checked for collisions with each other and for 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 memory, avoiding excessive recomputing and explaining why memory usage reaches a few GB (Figure 1G).

Note that each starting point X describes a set of mutually exclusive structures if we assume oriented receptor structures (first to last residue). Indeed, every structure, if not oriented, would be enumerated twice, once from each end. Algorithm 2 is still valid in the particular case of $n = 1$ 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 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 concrete receptor sequences. This step leads to a three- to four-fold 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 n 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 n 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, n-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. Given the pre-computed list of possible 3D receptor structures with at least n interactions to the pre-defined ligand, the binding energy and total energy between the two AA sequences is calculated for each structure with the receptor AA sequence. The 'best binding energy' E_{best} is the average binding energy of all optimal structures in terms of 'total energy' E_{tot} (around a small $\epsilon = 0.001$ to account for rounding errors), with

$$E_{\text{best}}\big(R, P(G,K)\big) = \begin{array}{l} \underset{\text{Structures } S}{\text{Average}}\Big(E_{\text{bind}}\left(P(R,S), P(G,K)\right), \text{ for all S with} \\ E_{\text{tot}}\big(P(R,S), P(G,K)\big) < \min\limits_{i=1}^{n_{\text{S}}} \Big(E_{\text{tot}}\big(P(R,S_i), P(G,K)\big)\Big) + \epsilon \Big) \end{array} \tag{4}$$

Self-folding can contribute to stabilize a structure and therefore the best structure is not necessarily the one binding with highest strength, but with lowest total energy. R is the receptor sequence, G the antigen sequence, K the antigen structure, S are the possible enumerated receptor structures, and n_{S} is the number of these structures. If a receptor sequence folds on itself with better energy than the total energy around the ligand, no binding is assumed and the energy is NAN (Not A Number).

Alternatively, a 'statistical energy' can be computed by applying a Boltzmann weight to each protein structure $P(R, S_i)$ according to its total energy:

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

with kT the Boltzmann coefficient and S_i the ith structure. Z is the sum of weights among all possible structures, and is used as normalization coefficient. Many structures could co-exist with good but not optimal total energy, and this may not be described well by the set of optimal structures alone in E_{best}.

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). Structures with $n < 4$ had no effect on Z (not shown) because these come with a small weight. We also computed all the possible self-foldings for different receptor sizes L, and found that the influence of n on Z was negligible (not shown) because the number of possible self-folding structures was small compared to large number of foldings around the ligand. This justifies to discard self-foldings and choose $n = 4$. Higher values of n may be used to increase speed, as shown in Figure 1F, provided that there is still no impact on Z.

Transforming energy into affinity and binding probability. The best binding energy between receptor and ligand was calculated for short receptor structures. 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 other factors. Therefore, an affinity cannot be drawn easily. We define an empirical 're-scaled affinity' a

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

of a receptor AA sequence R, to represent the probability of binding the antigen. $E_{\max}$ is the binding energy, at which the antigen capture probability becomes 1. The binding energy was re-scaled with a coefficient C because the best structures were dominant in the affinity and GC simulations did not develop because of low antigen binding. In Figures 3 and 4 the affinity of L2, L5, and L6 and receptors with $L = 9$ were calculated with: $E_{\max} = -100$ and $C = 2.8$. $E_{\max}$ and C have to be adapted for different 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 [9, 50]. 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 a single B-cell - T-cell interaction for B-cell selection were used [51]. Parameters values are detailed in [50].

The structural affinity model replaced the shape space affinity model in this framework. A mutation is a random change in the AA located at a randomly picked position of the BCR. Selection parameters had to be adapted: Antigen uptake is limited in the model by a 'refractory time' between two attempts of antigen capture. This time was reduced to 3.6 seconds. The simulation time-step had to be lowered to 0.001 hours accordingly.

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 Hill-coefficient from 2.0 to 1.4. This value was also supported in [52]. With a higher slope, cells with intermediate affinity did not expand (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 to at least one epitope, 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 (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. When using multiple epitopes, the same total amount of epitopes in the simulation is used as for single epitope simulations, meaning two epitopes

are used each with half-amount, etc. Each FDC position was initially filled with an equal amount of each epitope. At each position, B-cells can access all of the epitopes simultaneously. The antigen capture probability was determined by the highest affinity to any of the epitopes at this position, and this highest affinity epitope is removed by one epitope unit at this position. Other epitopes are not captured.

Computation times are given for simulations on a single Intel Xeon CPU core (model E5-2690 v3 at 2.6 GHz). The antigen sequences used for 4 are given in Suppl. Table S1.

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 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. 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.

Author contributions

PR and MMH designed the study, MMH programmed the Germinal Center model, PR programmed the structural affinities, MMH and PR analyzed the results and wrote the manuscript.

References

1. Eroshkin, A. M. *et al.* bNAber: database of broadly neutralizing HIV antibodies. *Nucleic Acids Research* **42,** D1133–D1139 (2014).
2. Corti, D. *et al.* Tackling influenza with broadly neutralizing antibodies. *Current Opinion in Virology* **24,** 60–69 (2017).

3. 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).

4. 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).

5. Tan, J. *et al.* A LAIR-1 insertion generates broadly reactive antibodies against malaria variant antigens. *Nature* **529,** 105–109 (2016).

6. 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).

7. 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).

8. Victora, G. D. & Nussenzweig, M. C. Germinal centers. *Annual Review of Immunology* **30,** 429–457 (2012).

9. Meyer-Hermann, M. *et al.* A theory of germinal center B cell selection, division, and exit. *Cell Reports* **2,** 162–174 (2012).

10. Zhang, J. & Shakhnovich, E. I. Optimality of mutation and selection in germinal centers. *PLoS Computational Biology* **6,** e1000800 (2010).

11. 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).

12. 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).

13. 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).

14. 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).

15. 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).

16. Gao, F. *et al.* Cooperation of B cell lineages in induction of HIV-1-broadly neutralizing antibodies. *Cell* **158,** 481–491 (2014).

17. Zhang, Y. *et al.* Germinal center B cells govern their own fate via antibody feedback. *The Journal of Experimental Medicine* **210,** 457–464 (2013).

18. Meyer-Hermann, M. Injection of Antibodies against Immunodominant Epitopes Tunes Germinal Centers to Generate Broadly Neutralizing Antibodies. *Cell reports* **29,** 1066–1073 (2019).

19. 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).

20. 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).

21. 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).

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

23. 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).

24. 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).

25. Raoof, S., Heo, M. & Shakhnovich, E. I. A one-shot germinal center model under protein structural stability constraints. *Physical biology* **10,** 025001 (2013).

26. Murugan, R. *et al.* Clonal selection drives protective memory B cell responses in controlled human malaria infection. *Science Immunology* **3,** eaap8029 (2018).

27. 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).

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

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

30. 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).

31. 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).

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

33. 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).

34. Kuraoka, M. *et al.* Complex antigens drive permissive clonal selection in germinal centers. *Immunity* **44,** 542–552 (2016).

35. 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).

36. Mesin, L., Ersching, J. & Victora, G. D. Germinal center B cell dynamics. *Immunity* **45,** 471–482 (2016).

37. Malliavin, T. E., Mucherino, A., Lavor, C. & Liberti, L. Systematic Exploration of Protein Conformational Space Using a Distance Geometry Approach. *Journal of Chemical Information and Modeling* **59,** 4486–4503 (2019).

38. Koliński, A. *et al.* Protein modeling and structure prediction with a reduced representation. *Acta Biochimica Polonica* **51** (2004).

39. Dunn-Walters, D., Townsend, C., Sinclair, E. & Stewart, A. Immunoglobulin gene analysis as a tool for investigating human immune responses. *Immunological reviews* **284,** 132–147 (2018).

40. Davidsen, K. & Matsen IV, F. A. Benchmarking tree and ancestral sequence inference for B cell receptor sequences. *Frontiers in immunology* **9,** 2451 (2018).

41. Uduman, M., Shlomchik, M. J., Vigneault, F., Church, G. M. & Kleinstein, S. H. Integrating B cell lineage information into statistical tests for detecting selection in Ig sequences. *The Journal of Immunology* **192,** 867–874 (2014).

42. Angeletti, D. & Yewdell, J. W. Understanding and manipulating viral immunity: antibody immunodominance enters center stage. *Trends in immunology* **39,** 549–561 (2018).

43. Ward, A. B. & Wilson, I. A. The HIV-1 envelope glycoprotein structure: Nailing down a moving target. *Immunological reviews* **275,** 21–32 (2017).

44. 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).

45. Alanine, D. G. *et al.* Human antibodies that slow erythrocyte invasion potentiate malaria-neutralizing antibodies. *Cell* **178,** 216–228 (2019).

46. 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).

47. 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).

48. De Boer, R. J. & Perelson, A. S. How germinal centers evolve broadly neutralizing antibodies: the breadth of the follicular helper T cell response. *Journal of virology* **91,** e00983–17 (2017).

49. Mann, M., Saunders, R., Smith, C., Backofen, R. & Deane, C. M. Producing high-accuracy lattice models from protein atomic coordinates including side chains. *Advances in bioinformatics* **2012** (2012).

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

51. 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).

52. Binder, S. C. & Meyer-Hermann, M. Implications of intravital imaging of murine germinal centers on the control of B cell selection and division. *Frontiers in immunology* **7,** 593 (2016).

Supplementary material

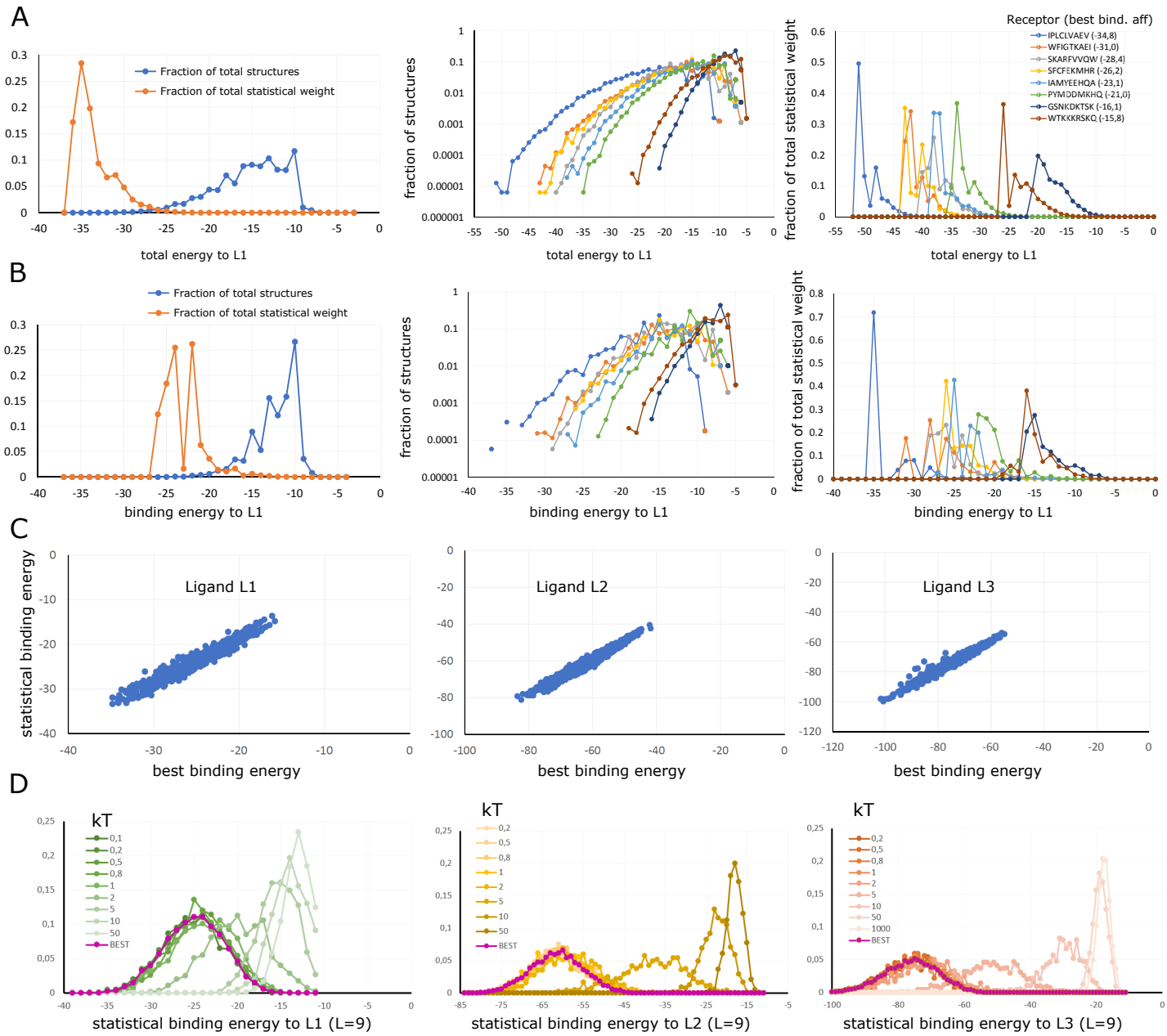

Figure S1: **The best binding energy is a fair approximation of the statistical binding energy. A-B.** Description of the energies of the ensemble of structures around ligand L1, for particular receptor sequences. **A.** Left: Distribution of total energies E_{tot} and statistical weight for possible 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 E_{tot} of the possible structures for selected receptor sequences with different best binding energies, as fraction of the total statistical weight Z (middle) or as fraction of the total number of structures (right). **B.** Same analysis with binding energies E_{bind} 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 with each ligand. Temperature values around $kT = 1$ have no effect on the distributions, meaning that a frozen state is reached, where optimal structures dominate.

Simulation	Repeats	Combination of antigens	AG name
		Mutable positions (x) x_xx_xxxxx_x_xxxxxx_____x_____	
2AGs, HD=1	1	VHFYKNQNFSPLVWLKFPGSPLIVIAAKLGEALYPNVGR	B1
		VHFYKNQNFSPLVWLKFPGSPLIVIAAKLGAALYPNVGR	B35
	1	CHWHKMRSFMPRVLWMEHNSPLIVIAAKLGVALYPNVGR	H8
		CHWGKMRSFMPRVLWMEHNSPLIVIAAKLGVALYPNVGR	H15
	1	GHQFKTRYILPGVWNIYVHSPLIVIAAKLGWALYPNVGR	I1
		GHQFKTRYILPGVWNIVVHSPLIVIAAKLGWALYPNVGR	I19
	1	QHQDKLWHICPHVCGMFMCSPLIVIAAKLGAALYPNVGR	L4
		QHQDKLWHICPHVVGMFMCSPLIVIAAKLGAALYPNVGR	L18
	1	QHTLKPFVTVPTVCLIHMMSPLIVIAAKLGKALYPNVGR	M11
		QHTLKPFVTVPTVCLDHMMSPLIVIAAKLGKALYPNVGR	M13
	1	QHTLKPFVTVPTVCLDHMMSPLIVIAAKLGKALYPNVGR	M13
		QHTLKPFVTVPTVCLNHMMSPLIVIAAKLGKALYPNVGR	M38
	1	QHDFKFLPDDPNVWAMAFVSPLIVIAAKLGFALYPNVGR	Q15
		QHDLKFLPDDPNVWAMAFVSPLIVIAAKLGFALYPNVGR	Q25
	1	PHILKAALYVPPVYTILVKSPLIVIAAKLGRALYPNVGR	V10
		PHILKTALYVPPVYTILVKSPLIVIAAKLGRALYPNVGR	V43
	1	PHILKGALYVPPVYHILVYSPLIVIAAKLGRALYPNVGR	W0
		PHILKGALYVPPVYHILVDSPLIVIAAKLGRALYPNVGR	W12
	1	PHILKGALNVPPVYHILVKSPLIVIAAKLGRALYPNVGR	W2
		PHILKGALGVPPVYHILVKSPLIVIAAKLGRALYPNVGR	W16
2AGs, HD=2	2	IHLMKVPSREPTVWVMDLASPLIVIAAKLGIALYPNVGR	A0
		IHLMKVPFREPTVWAMDLASPLIVIAAKLGIALYPNVGR	A32
	2	VHFYKNQTFSPLVWLKFPGSPLIVIAAKLGGALYPNVGR	B0
		VHFYKNQNFSPLVWLKFPGSPLIVIAAKLGEALYPNVGR	B1
	2	VHFYKNQTFSPLVWLKFPGSPLIVIAAKLGGALYPNVGR	B0
		VHFYKNQNFSPLVWLKFPGSPLIVIAAKLGAALYPNVGR	B35
	2	LHRNKRVTHIPQVICMQHASPLIVIAAKLGFALYPNVGR	C6
		LHRNKRGTLIPQVICMQHASPLIVIAAKLGFALYPNVGR	C18
	2	CHPVKLWNSKPIVWFAIEESPLIVIAAKLGMALYPNVGR	D0
		CHPVKLWNVKPIVWFDIEESPLIVIAAKLGMALYPNVGR	D3
	2	LHMVKDWFFNPFVIYDKDASPLIVIAAKLGAALYPNVGR	E0
		LHMPKDWFFNPFVIYDCDASPLIVIAAKLGAALYPNVGR	E1
	2	LHMVKDWFFNPFVIYDKDASPLIVIAAKLGAALYPNVGR	E0
		LHMVKDWFGNPFVIYDVDASPLIVIAAKLGAALYPNVGR	E2
	2	HHFGKSCFCTPHVWISFHISPLIVIAAKLGPALYPNVGR	F0
		IHQGKSCFCTPHVWISFHISPLIVIAAKLGPALYPNVGR	F2
	2	EHIYKLFDLWPEVIYKLNNSPLIVIAAKLGWALYPNVGR	H0
		EHIYKLFDLCPEVIKKLNNSPLIVIAAKLGWALYPNVGR	H5
	2	GHQFKKRYILPGVWNIMVHSPLIVIAAKLGWALYPNVGR	I13
		GHQFKTRYILPGVWNIVVHSPLIVIAAKLGWALYPNVGR	I19
	3	GHQFKTRYILPGVWNIYVHSPLIVIAAKLGWALYPNVGR	I1
		GHQFKKRYILPGVWNIMVHSPLIVIAAKLGWALYPNVGR	I13
	3	QHQDKLWHICPRVMGMFMCSPLIVIAAKLGAALYPNVGR	L0
		QHQDKLWHICPHVVGMFMCSPLIVIAAKLGAALYPNVGR	L18
	3	QHTLKHHVTVPTVCLPHMMSPLIVIAAKLGKALYPNVGR	M5
		QHTLKHWVTVPTVCLVHMMSPLIVIAAKLGKALYPNVGR	M15
	3	QHDMKFLPGDPNVWAMAFVSPLIVIAAKLGFALYPNVGR	Q0
		QHDFKFLPDDPNVWAMAFVSPLIVIAAKLGFALYPNVGR	Q15
	3	PHILKGALYVPPVYHILVKSPLIVIAAKLGRALYPNVGR	V0
		PHILKAALYVPPVYTILVKSPLIVIAAKLGRALYPNVGR	V10
	3	PHILKGALYVPPVYHILVYSPLIVIAAKLGRALYPNVGR	W0
		PHILKGALYVPPVYTILVKSPLIVIAAKLGRALYPNVGR	W1
2AGs, HD=3	1	IHLMKVPDREPTVWVMDLASPLIVIAAKLGTALYPNVGR	A6
		IHLMKVPFREPTVWAMDLASPLIVIAAKLGIALYPNVGR	A32
	1	VHFYKNQRLSPLVWLKFPGSPLIVIAAKLGGALYPNVGR	B4
		VHFYKNQNFSPLVWLKFPGSPLIVIAAKLGAALYPNVGR	B35
	1	LHRNKRVTHIPQVICMQHASPLIVIAAKLGFALYPNVGR	C6
		LHRNKDKTHIPWVICMQHASPLIVIAAKLGFALYPNVGR	C11
	1	CHPVKLWNVKPIVWFDIEESPLIVIAAKLGMALYPNVGR	D3
		CHTVKLWNGKPIVWFAIEESPLIVIAAKLGMALYPNVGR	D5
	1	LHMPKDWFFNPFVIYDCDASPLIVIAAKLGAALYPNVGR	E1
		LHMVKDWFGNPFVIYDVDASPLIVIAAKLGAALYPNVGR	E2
	1	HHCGKSCFCTPNVWISFHISPLIVIAAKLGPALYPNVGR	F1
		IHQGKSCFCTPHVWISFHISPLIVIAAKLGPALYPNVGR	F2
	1	EHIYKLFDKWPEVIIKLNNSPLIVIAAKLGWALYPNVGR	H1
		EHIYKLFDLCPEVIKKLNNSPLIVIAAKLGWALYPNVGR	H5
	1	GHQFKIRYILPGVWNINVHSPLIVIAAKLGWALYPNVGR	I3
		GHQFKTRYILPGVWNIVVHSPLIVIAAKLGWALYPNVGR	I19
	1	QHQDKLWHICPHVCGMFMCSPLIVIAAKLGAALYPNVGR	L4
		QHQDKLWHICPRVRWMFMCSPLIVIAAKLGAALYPNVGR	L10
	1	QHTLKHHVTVPTVCLPHMMSPLIVIAAKLGKALYPNVGR	M5
		QHTLKPFVTVPTVCLIHMMSPLIVIAAKLGKALYPNVGR	M11
	1	QHTLKHHVTVPTVCLPHMMSPLIVIAAKLGKALYPNVGR	M5
		QHTLKPFVTVPTVCLDHMMSPLIVIAAKLGKALYPNVGR	M13
	1	MHDMKFLPADPNVWAMAFVSPLIVIAAKLGFALYPNVGR	Q3
		QHDFKFLPDDPNVWAMAFVSPLIVIAAKLGFALYPNVGR	Q15
2AGs_HD=4	1	VHFYKNQNFSPLVWLKFPGSPLIVIAAKLGEALYPNVGR	B1
		VHFSKNQTFSPLVWLKFPRSPLIVIAAKLGGALYPNVGR	B2
	1	LHINKRGTHIPWVICMQTASPLIVIAAKLGFALYPNVGR	C2
		LHRNKDKTHIPWVICMQHASPLIVIAAKLGFALYPNVGR	C11
	1	LHINKRGTHIPWVICMQTASPLIVIAAKLGFALYPNVGR	C2
		LHRNKRGTLIPQVICMQHASPLIVIAAKLGFALYPNVGR	C18
	1	IHPVKLWNSKPIVFFAIEESPLIVIAAKLGMALYPNVGR	D2
		CHTVKLWNGKPIVWFAIEESPLIVIAAKLGMALYPNVGR	D5
	1	LHMPKDWFFNPFVIYDCDASPLIVIAAKLGAALYPNVGR	E1
		LHMVKDLFFNPFVIYHKDASPLIVIAAKLGAALYPNVGR	E4
	1	IHQGKSCFCTPHVWISFHISPLIVIAAKLGPALYPNVGR	F2
		HHFGKSCMCTPHVWISFHISPLIVIAAKLGGALYPNVGR	F3
	1	EHICKLFDLWPEVIYGLNNSPLIVIAAKLGWALYPNVGR	H3
		EHIYKLFDLCPEVIKKLNNSPLIVIAAKLGWALYPNVGR	H5
	1	CHWSKMRSFEPHVLWMEFNSPLIVIAAKLGVALYPNVGR	H1
		CHWSKMNHFMPHVLWMEHNSPLIVIAAKLGVALYPNVGR	H19
	1	QHQCKLWHICPRVMGQFMCSPLIVIAAKLGAALYPNVGR	L1
		QHQDKLWHICPHVCGMFMCSPLIVIAAKLGAALYPNVGR	L4
	1	QHRMKFLEGDPNVWAMAFVSPLIVIAAKLGFALYPNVGR	Q1
		QHDFKFLPDDPNVWAMAFVSPLIVIAAKLGFALYPNVGR	Q15

Simulation	Repeats	Combination of antigens	AG name
		Mutable positions (x) x_xx_xxxxx_x_xxxxxx_____x_____	
1AG	1	IHLMKVPSREPTVWVMDLASPLIVIAAKLGIALYPNVGR	A0
	1	IHLMKVPDREPTVWVMDLASPLIVIAAKLGTALYPNVGR	A6
	1	IHLMKVPFREPTVWAMDLASPLIVIAAKLGIALYPNVGR	A32
	1	VHFYKNQTFSPLVWLKFPGSPLIVIAAKLGGALYPNVGR	B0
	1	VHFYKNQNFSPLVWLKFPGSPLIVIAAKLGEALYPNVGR	B1
	1	VHFSKNQTFSPLVWLKFPRSPLIVIAAKLGGALYPNVGR	B2
	1	VHFYKNQRLSPLVWLKFPGSPLIVIAAKLGGALYPNVGR	B4
	1	VHFYKNQNFSPLVWLKFPGSPLIVIAAKLGAALYPNVGR	B35
	1	LHINKRGTHIPWVICMQTASPLIVIAAKLGFALYPNVGR	C2
	1	LHRNKRVTHIPQVICMQHASPLIVIAAKLGFALYPNVGR	C6
	1	LHRNKDKTHIPWVICMQHASPLIVIAAKLGFALYPNVGR	C11
	1	LHRNKRGTLIPQVICMQHASPLIVIAAKLGFALYPNVGR	C18
	1	CHPVKLWNSKPIVWFAIEESPLIVIAAKLGMALYPNVGR	D0
	1	IHPVKLWNSKPIVFFAIEESPLIVIAAKLGMALYPNVGR	D2
	1	CHPVKLWNVKPIVWFDIEESPLIVIAAKLGMALYPNVGR	D3
	1	CHTVKLWNGKPIVWFAIEESPLIVIAAKLGMALYPNVGR	D5
	1	LHMVKDWFFNPFVIYDKDASPLIVIAAKLGAALYPNVGR	E0
	1	LHMPKDWFFNPFVIYDCDASPLIVIAAKLGAALYPNVGR	E1
	1	LHMVKDWFGNPFVIYDVDASPLIVIAAKLGAALYPNVGR	E2
	1	LHMVKDLFFNPFVIYHKDASPLIVIAAKLGAALYPNVGR	E4
	1	HHFGKSCFCTPHVWISFHISPLIVIAAKLGPALYPNVGR	F0
	1	HHCGKSCFCTPNVWISFHISPLIVIAAKLGPALYPNVGR	F1
	1	IHQGKSCFCTPHVWISFHISPLIVIAAKLGPALYPNVGR	F2
	1	HHFGKSCMCTPHVWISFHISPLIVIAAKLGGALYPNVGR	F3
	1	EHIYKLFDLWPEVIYKLNNSPLIVIAAKLGWALYPNVGR	H0
	1	EHIYKLFDLCPEVIKKLNNSPLIVIAAKLGWALYPNVGR	H5
	1	EHICKLFDLWPEVIYGLNNSPLIVIAAKLGWALYPNVGR	H3
	1	EHIYKLFDLCPEVIKKLNNSPLIVIAAKLGWALYPNVGR	H5
	1	CHWHKMRSFMPRVLWMEHNSPLIVIAAKLGVALYPNVGR	H8
	1	CHWGKMRSFMPRVLWMEHNSPLIVIAAKLGVALYPNVGR	H15
	1	CHWSKMNHFMPHVLWMEHNSPLIVIAAKLGVALYPNVGR	H19
	1	GHQFKTRYILPGVWNIYVHSPLIVIAAKLGWALYPNVGR	I1
	1	GHQFKKRYILPGVWNIMVHSPLIVIAAKLGWALYPNVGR	I13
	1	GHQFKTRYILPGVWNIVVHSPLIVIAAKLGWALYPNVGR	I19
	1	QHQDKLWHICPRVMGMFMCSPLIVIAAKLGAALYPNVGR	L0
	1	QHQCKLWHICPRVMGQFMCSPLIVIAAKLGAALYPNVGR	L1
	1	QHQDKLWHICPHVCGMFMCSPLIVIAAKLGAALYPNVGR	L4
	1	QHQDKLWHICPRVRWMFMCSPLIVIAAKLGAALYPNVGR	L10
	1	QHQDKLWHICPHVVGMFMCSPLIVIAAKLGAALYPNVGR	L18
	1	QHTLKHHVTVPTVCLPHMMSPLIVIAAKLGKALYPNVGR	M5
	1	QHTLKPFVTVPTVCLDHMMSPLIVIAAKLGKALYPNVGR	M11
	1	QHTLKHWVTVPTVCLVHMMSPLIVIAAKLGKALYPNVGR	M15
	1	QHTLKPFVTVPTVCLNHMMSPLIVIAAKLGKALYPNVGR	M38
	1	QHDMKFLPGDPNVWAMAFVSPLIVIAAKLGFALYPNVGR	Q0
	1	QHRMKFLEGDPNVWAMAFVSPLIVIAAKLGFALYPNVGR	Q1
	1	MHDMKFLPADPNVWAMAFVSPLIVIAAKLGFALYPNVGR	Q3
	1	QHDFKFLPDDPNVWAMAFVSPLIVIAAKLGFALYPNVGR	Q15
	1	QHDLKFLPDDPNVWAMAFVSPLIVIAAKLGFALYPNVGR	Q25
	1	PHILKGALYVPPVYHILVKSPLIVIAAKLGRALYPNVGR	V0
	1	PHILKAALYVPPVYTILVKSPLIVIAAKLGRALYPNVGR	V10
	1	PHILKTALYVPPVYTILVKSPLIVIAAKLGRALYPNVGR	V43
	1	PHILKGALYVPPVYHILVYSPLIVIAAKLGRALYPNVGR	W0
	1	PHILKGALYVPPVYTILVKSPLIVIAAKLGRALYPNVGR	W1
	1	PHILKGALNVPPVYHILVKSPLIVIAAKLGRALYPNVGR	W2
	1	PHILKGALYVPPVYHILVDSPLIVIAAKLGRALYPNVGR	W12
	1	PHILKGALGVPPVYHILVKSPLIVIAAKLGRALYPNVGR	W16
4AGs_HD=2	4	IHLMKVPSREPTVWVMDLASPLIVIAAKLGIALYPNVGR	A0
		IHLMKVPSRTPTVWVMDNASPLIVIAAKLGIALYPNVGR	A2
		IHLMKVPSRPPTVWVMDFASPLIVIAAKLGIALYPNVGR	A31
		IHLMKVPFREPTVWAMDLASPLIVIAAKLGIALYPNVGR	A32
	4	VHFYKNQTFSPLVWLKFPGSPLIVIAAKLGGALYPNVGR	B0
		VHFYKNQNFSPLVWLKFPGSPLIVIAAKLGEALYPNVGR	B1
		VHFYKNQLFSPLVWLKFPGSPLIVIAAKLGDALYPNVGR	B6
		VHFYKNQAFSPLVWLKFPGSPLIVIAAKLGTALYPNVGR	B18
	4	LHRNKRGTHIPWVICMQHASPLIVIAAKLGFALYPNVGR	C0
		LHRNKRVTHIPQVICMQHASPLIVIAAKLGFALYPNVGR	C6
		LHRNKRGTLIPQVICMQHASPLIVIAAKLGFALYPNVGR	C18
		LHREKRGTHIPWVICMFHASPLIVIAAKLGFALYPNVGR	C28
	4	CHPVKLWNSKPIVWFAIEESPLIVIAAKLGMALYPNVGR	D0
		CHPVKLWNSSPTVWFAIEESPLIVIAAKLGMALYPNVGR	D1
		IHPVKLWNSKPIVFFAIEESPLIVIAAKLGMALYPNVGR	D2
		CHPVKLWNVKPIVWFDIEESPLIVIAAKLGMALYPNVGR	D3
	4	LHMVKDWFFNPFVIYDKDASPLIVIAAKLGAALYPNVGR	E0
		LHMPKDWFFNPFVIYDCDASPLIVIAAKLGAALYPNVGR	E1
		LHMVKDWFGNPFVIYDVDASPLIVIAAKLGAALYPNVGR	E2
		FHMVKDWFFNPFVIYDKDCSPLIVIAAKLGAALYPNVGR	E3
	4	HHFGKSCFCTPHVWISFHISPLIVIAAKLGPALYPNVGR	F0
		HHCGKSCFCTPNVWISFHISPLIVIAAKLGPALYPNVGR	F1
		IHQGKSCFCTPHVWISFHISPLIVIAAKLGPALYPNVGR	F2
		HHFGKSCMCTPHVWISFHISPLIVIAAKLGGALYPNVGR	F3
	4	VHMLKMVCFDPEVYATKCFSPLIVIAAKLGWALYPNVGR	G5
		VHMLKMVCDDPEVYAFKCFSPLIVIAAKLGWALYPNVGR	G25
		VHMLKMVCDDPEVYCHKCFSPLIVIAAKLGWALYPNVGR	G6
		VHMLKMVCDDPEVYQDKCFSPLIVIAAKLGWALYPNVGR	G9
	4	EHIYKLFDLWPEVIYKLNNSPLIVIAAKLGWALYPNVGR	H0
		EHIYKLFDKWPEVIIKLNNSPLIVIAAKLGWALYPNVGR	H1
		FHIYKLFDQWPEVIYKLNNSPLIVIAAKLGWALYPNVGR	H4
		GHIYKLFDSWPEVIYKLNNSPLIVIAAKLGWALYPNVGR	H7
	4	YHVVKLGGNYPHVMVPNNLSPLIVIAAKLGIALYPNVGR	K0
		YHVVKLGYNYPHVMVPGNLSPLIVIAAKLGIALYPNVGR	K1
		YHVVKLGYNYPHVMVPNKLSPLIVIAAKLGIALYPNVGR	K27
		YHVHKLGYNYPHVMVPNNLSPLIVIAAKLGIALYPNVGR	K32
	4	QHQDKLWHICPRVMGMFMCSPLIVIAAKLGAALYPNVGR	L0
		QHQCKLWHICPRVMGQFMCSPLIVIAAKLGAALYPNVGR	L1
		QHVDKLWHICPRVMTMFMCSPLIVIAAKLGAALYPNVGR	L2
		QHQDKLWHICPHVVGMFMCSPLIVIAAKLGAALYPNVGR	L18

Table S2: Antigen sequences used for GC simulations. For each group of simulations (left column), the set of antigen sequences used on structure L6 is given, together with the number of replicates of each GC simulation. Briefly, antigen AA sequences that achieve an average best binding among 200 random receptors between the top 1% and 15% best values were selected, to avoid chaotic GC dynamics if the antigen has too low or too high binding, and sequences with 1 to 5 point mutations within this range were kept to be used as combination of antigens with different Hamming Distances (HD).