





REPORT

OPEN ACCESS



The human antibody sequence space and structural design of the V, J regions, and CDRH3 with Rosetta

Samuel Schmitz ^{a,b}, Emily A. Schmitz ^{b,c}, James E. Crowe Jr ^{d,e,f}, and Jens Meiler ^{a,b,g}

^aDepartment of Chemistry, Vanderbilt University, Nashville, Tennessee, United States; ^bCenter of Structural Biology, Vanderbilt University, Nashville, Tennessee, United States; ^cDepartment of Molecular Physiology and Biophysics, Vanderbilt University, School of Medicine, Nashville, Tennessee, United States; ^dVanderbilt Vaccine Center, Vanderbilt University Medical Center, Nashville, Tennessee, United States; ^eDepartment of Pathology, Microbiology, and Immunology, Vanderbilt University Medical Center, Nashville, Tennessee, United States; ^fDepartments of Pediatrics, Vanderbilt University Medical Center, Nashville, Tennessee, United States; ^gInstitute for Drug Discovery, University Leipzig Medical School, Leipzig, Germany

ABSTRACT

The human adaptive immune response enables the targeting of epitopes on pathogens with high specificity. Infection with a pathogen induces somatic hyper-mutation and B-cell selection processes that govern the shape and diversity of the antibody sequence landscape. To date, even the largest immunome repertoires of adaptive immune receptors acquired by next-generation sequencing cannot fully capture the vast antibody sequence space of a single individual, which is estimated to be at least 10^{12} potential sequences. Degeneracy of the genetic code means that the number of possible nucleotide triplets (64) is greater than the number of canonical amino acids (20), resulting in some amino acids being encoded by multiple triplets and different amino acids sharing the same nucleotide in 1 or 2 positions in the triplet. We hypothesize that the degeneracy of the genetic code can be used to statistically model an enlarged space of human antibody amino acid sequences, accommodating for the discrepancy between the observed and the hypothesized antibody sequence space. Facilitated by Bayesian statistics and immunome repertoire clustering, we calculated amino acid probabilities from single nucleotide frequencies to infer a human amino acid sequence space that is used to design human-like antibodies with Rosetta. We show that antibodies designed with our restraints are on average up to 16.6% more human-like in the V and J regions compared to the Rosetta designs produced without constraints. The human-likeness of the heavy-chain CDR3 region (CDRH3) could be increased for 8 of 27 antibodies compared to Rosetta designs with a similar number of mutations and could be successfully applied on *Mus musculus* antibodies to demonstrate humanization.

ARTICLE HISTORY

Received 25 August 2021
Revised 05 April 2022
Accepted 14 April 2022

KEYWORDS



Human-likeness; antibody design; rosetta; immunome repertoire; biostatistics; Humanization ABBREVIATIONS


Introduction

As of 2019, over 570 antibody drugs are in development with a substantial increase in late-stage antibody development over the past decade.^{1,2} Historically, antibody reagents were generated using cells from an animal source such as rabbit,³ chicken,⁴ and more prominently murine model organisms.^{5,6} The disadvantage of using antibodies with non-human origin is the elicitation of anti-drug-antibodies (ADA) in human patients.^{6,7} High titers of ADAs can block the drug's antigen-binding site or cause faster depletion of the antibody drugs in the bloodstream, which usually results in reduced efficacy of the antibody drug.⁸ The reasons for the ADA response in patients are multi-faceted, but a common cause is sequence patterns that are foreign to the human system.⁹ This observation prompted the invention of humanization techniques, which yield engineered antibodies with non-human sequences interspersed among human-derived antibody segments.^{10,11} Here, we introduce a method based on the human-likeness (HL) assessment method IgReconstruct,¹² and expand upon it to support the structural design of human-like antibodies. A possible application of our method is supported during the

early development process of antibody biologics that appear human-like. It also may be useful to simulate a possible personal immune response using structural design and the sequenced antibody space of a human individual.

Access to large quantities of observed human antibodies sequences, the so-called adaptive immune receptor repertoires, is essential for HL assessment.^{12–16} Next-generation sequencing (NGS) of peripheral blood samples has given insight into the diversity of human adaptive immune receptor repertoires, sometimes referred to as B-cell immunomes.^{17–19} Despite the high diversity, a small sequence overlap between individual blood donors exists.^{18,19} The major mechanism of antibody diversification is the somatic recombination of variable (V), diversity (D), and joining (J) germline gene segments. The human immune system has approximately 123–129 heavy chain variable genes (IGHV), 27 diversity genes (IGHD), and 9 joining genes (IGHJ) available for use in this process. Light chain genes are grouped into kappa (chromosome 2) and lambda (chromosome 22) genes with 40–76 (IGKV), 73–74 (IGLV) variable genes, 5 (IGKJ), and 1 (IGLJ) joining gene.²⁰ The antibody germline genes contribute to antibody diversity,

CONTACT Jens Meiler  jens@meilerlab.org  Department of Chemistry, Vanderbilt University, 7330 Stevenson Center, Station B, 351822, Nashville, TN 37235; Institute for Drug Discovery, Pharmazie, Leipzig University, Brüderstraße 34, 04103 Leipzig, Leipzig, Germany

 Supplemental data for this article can be accessed on the [publisher's website](#)

© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

with the recombination events alone producing a diversity of 10^6 sequences.²¹ The addition or deletion of single nucleotides in the junctions between the variable, diversity, and joining genes (V-D, V-J, or D-J), and somatic hyper-mutation further increase the antibody diversity. Higher affinity variants of B-cell receptors are generated via somatic hyper-mutation. During this process, double-stranded DNA breaks lead to the introduction of single-point nucleotide mutations or insertion/deletions, which are introduced by error-prone DNA repair mechanisms.²² The resulting diversity of human antibody repertoires has been estimated to exceed 10^{12} unique B-cell receptors.²¹ It has been shown that the HL of the IgG isotype of antibodies can be modeled by assessing the single nucleotide frequency for each germline gene in the observed antibody space.²³

It is difficult to assess HL for the heavy-chain CDR3 loop (CDRH3) because it comprises junctions that are not derived from the germline genes (non-templated regions). Even though the diversity germline gene can contribute to the assembly of the CDRH3, the alignment of the CDRH3 to a diversity germline gene is often of low confidence. Here, we expand upon HL assessment using single nucleotide frequency profiles to model a human sequence space that is able to describe the CDRH3 sequence. To achieve this, all sequence sections that align to V and J germline genes, as well as CDRH3 regions, are clustered based on sequence similarity. Instead of aligning germline genes to the CDRH3 region to assess HL as previously described,¹² we instead choose the most similar V, J, and CDRH3 cluster center to assess the HL of the CDRH3 region. HL is then calculated for all antibody regions by assessing the observed nucleotide frequencies of sequences in the assigned repertoire cluster.

An important question to answer for HL assessment is to what extent NGS has discovered the human antibody space. To date the largest B-cell sequence databases published from single individuals include approximately 325 million nucleotide sequences from three blood donors.¹⁸ Taken together, modern sequencing methods have explored a combined sequence space of 5×10^8 sequences, which is orders of magnitude smaller than the theoretical maximum sequence space for a single individual (at least 10^{12}). Large antibody sequence repertoires are the result of work for the Human Immunome Project, which aims to comprehensively catalog the human B- and T-cell sequence spaces.²⁴ It could be shown that even though the sequence commonality between human-blood donors is greater than anticipated, the overall overlap in sequences remains small ($\ll 1\%$ of heavy-chain clonotypes).^{18,19} This low commonality is primarily the result of high sequence diversity, and the main cause for antibody diversity is the high variability of the CDRH3 region. To compensate for the small ratio of the observed to the expected antibody space, we mathematically calculate an enlarged human amino acid space from the nucleotide frequencies. We hypothesize that there is additional information in the nucleotide sequences that can inform the antibody space for the reasons that follow below.

The genetic code is degenerate, such that 64 unique nucleotide triplets in the standard translation table encode the 20 canonical amino acids. Thus, some amino acids are encoded by multiple nucleotide triplets and different amino acids share

the same nucleotide in 1 or 2 positions of the nucleotide triplet. HL was previously described as independent single nucleotide observations,¹² suggesting that the antibody maturation process is a stochastic process that mutates single nucleotides independently. We therefore postulate that all single nucleotide frequencies not only inform about the frequency of their encoding amino acid, but also inform the likelihood of observing another amino acid at that position, which is partially encoded by the same nucleotides of a different codon. In this study, we employ Bayesian statistics to model the probability of observing amino acids in human antibodies and postulate that the resulting amino acid frequencies model a larger human sequence space than has been observed, with the potential to suggest probabilities for amino acids that have not directly been observed at certain positions. We demonstrate that amino acid frequencies can then be used to inform computational structural protein design with Rosetta²⁵ to generate antibodies that are antigen-specific and thermodynamically stable, while still maintaining HL.

The computational structural design package Rosetta allows sequence design of proteins. Rosetta evaluates protein conformations and their sequences with its scoring function. The Rosetta scoring function comprises the weighted sum of physical, and knowledge-based potentials.²⁶ The scoring function can be extended by adding additional weighted restraints. This approach is commonly used to bias the protein design to include experimental observations such as alanine or site-directed mutagenesis, hydrogen-deuterium exchange mass spectrometry (HDX) or HDX-NMR, NMR chemical shift perturbations, low-resolution cryo-EM, and chemical cross-linking data.^{27,28} In this study, we re-design human antibody structures with our Bayesian human sequence profiles for increased HL. To benchmark our method, we chose 27 human antibody structures from structures deposited in the Structural Antibody Database (SabDab).²⁹ Choosing human antibodies provides us with the HL of human antibody sequences, which serves as a reference for benchmark purposes. Antibodies designed without human restraints are expected to decrease in their HL and exhibit reduced wild-type (WT) sequence identity. Thus, Rosetta-designed antibody sequences created with our amino acid frequency restraints were evaluated for HL, and sequence identity to the human WT antibody, and compared to Rosetta-designed antibodies without restraints. We hypothesize that the sequence recovery rate of designs using HL profiles should increase if our Bayesian model indeed resembles a human sequence space. We expect the Bayesian sequence space to be larger compared to the observed antibody space. We use Rosetta to narrow down the sequence space, and create antibody sequences that are suited for the antibody/antigen complex. Our method suggests a way to create novel antibodies with Rosetta that are more human-like, or to re-design existing antibodies for increased HL.

Results

The IgReconstruct method assesses HL via single nucleotide frequency statistics from immunome repertoires,¹² and has been compared with 10 similar approaches.³⁰ In this study, we extend IgReconstruct to improve its ability to assess the HL

of the heavy-chain CDR3 region (CDRH3). IgReconstruct assigns observed frequency statistics to the antibody germline genes.

The germline gene-centric approach cannot be applied to the CDRH3 since its genes either cannot or can only be partially assigned. Instead, heavy and light-chain sequences are divided into their V, J, and CDRH3 regions and are clustered. This enables us to assign nucleotide frequencies by using the sequence of the cluster center instead of a germline gene, which allows us to apply our algorithms not only to the V and J regions but also to the CDRH3.

The following sections describe how we model the Bayesian antibody space, and our clustering algorithm as an extension to IgReconstruct, followed by the results of our Rosetta design benchmark.

Calculation of the Bayesian antibody space

The proposed method calculates amino acid probabilities for an antibody from single nucleotide frequencies. The frequency profiles are assessed from large immunome repertoires of ~325 million unique sequences (Figure 1a). IgReconstruct¹² is a method developed to assess HL as nucleotide frequency profiles for each germline gene and position, and for each CDRH3 regions (length dependent) (Figure 1b). The method we used to expand upon this approach by creating clustered frequency profiles for genes and for CDRH3 regions is described below. The frequency profiles for V, J, and heavy-chain CDR3 regions (CDRH3) were then combined into position-specific frequency matrices. The combined frequency profile spans the variable region of an antibody and is mapped onto the structure (Figure 1c, Supplement 2 and 3).

The high diversity of the CDRH3 gives rise to the low commonality between human immunome repertoires.^{18,19} The observed antibody space used in this study (approximately 325 million sequences from three healthy human blood donors) is small compared to the estimated antibody diversity

of $10^{12,21}$. This study therefore develops Bayesian statistics to model amino acid frequencies from the observed nucleotide frequencies (Equation 1 in Methods section). Here, it is assumed that all positions of the antibody variable region have the potential to mutate to any canonical amino acid via somatic hyper-mutation. It is also assumed, that the nucleotide distribution observed in the immunome repertoire of 325 million sequences is representative of the human antibody space. Thus, the Bayesian statistics can be simplified by assuming that the *a priori* probability $p(aa)$ to observe each amino acid at each position is 1.

Different amino acids are encoded by a different number of triplets. Equations 2–3 (see Methods section) take the number of different triplets ($trpl$) that encode for a specific amino acid into account as a normalization parameter. For each amino acid probability $p(aa|trpl)$ with a given distribution of nucleotide frequencies ($trpl$), a substitution score s_{ij} is calculated. The substitution score represents statistical significance of the calculated frequencies for each position and will be used as a Rosetta restraint for HL antibody design. The calculation of the substitution score (Equation 4 (see Methods section), Figure 1e) has been adapted from the description for PSI-BLAST.^{31–33} The lambda parameter of s_{ij} is a scaling factor and is optimized for each nucleotide profile to correlate with the change of HL when amino acid i is replaced by j (see Supplement 1, Section 4). The tables of amino acid frequencies and substitution scores are then converted into a PSI-BLAST compatible ASCII file that can be parsed by Rosetta for further design (Figure 1d).

Extending the nucleotide HL metric with a clustering algorithm

Unlike the framework region of the antibody variable region, which is templated by germline genes, the highly variable CDRH3 region is either non-templated or has low confidence diversity (D) gene alignments. This compromises our approach

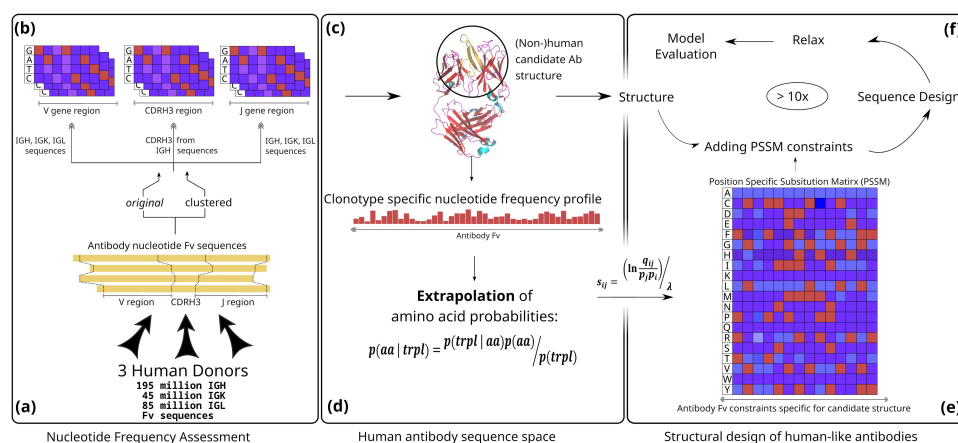


Figure 1. From immunome repertoire processing, to Bayesian statistics of an amino acid space created from single nucleotide frequencies, to structural human-like antibody design. First, V, CDRH3, and J regions are extracted from heavy and light chain sequences in the immunome repertoires to (a) generate position specific nucleotide frequency profiles for each region either by grouping the sequence regions by germline gene and CDRH3 length as described in our previous work (original). Or, each region is clustered by sequence identity resulting in multiple (clustered) PSSMs per germline gene and CDRH3 length (b). The structure of an antibody that is to be re-designed for increased human-likeness (c) is assigned a unique nucleotide frequency profile. Bayesian statistics provide amino acid frequencies for the antibody inferred from the assigned nucleotide frequency profile (d), and is converted into amino acid substitution scores (e). The substitution scores are used as restraints for structural design with Rosetta.

to assess HL via germline gene-specific nucleotide frequencies. Consequently, the CDRH3 was excluded for HL calculations in our previous study.¹²

To enable CDRH3 HL assessment, we extended the position-specific substitution matrix (PSSM) generation method by implementing a basic clustering approach capable of processing large datasets quickly. We expect that clustered PSSMs are enriched with functionally related sequences that can offer a distinct HL signal characteristic for specific CDRH3 lengths and conformations. Clusters are created based on nucleotide sequence identity and are represented as frequency profiles (clustered PSSM). The cluster center is the sequence that can be generated from the most frequent nucleotides observed in the PSSM and is not necessarily a sequence directly observed in the repertoire.

The clustering method, which can be subdivided to four steps, took place while iterating once over our immunome repertoire of approximately 325 million unpaired heavy and light chain human antibody sequences. The first cluster is initialized with the first random sequence encountered (Figure 2a). Every other sequence was either added to any of the existing cluster(s), or was added to a new cluster based on the sequence identity of the cluster center (Figure 2b). Here, the cluster center is the sequence that can be generated by picking the most frequently observed nucleotide at each position of the V, CDRH3, or J PSSM. Distance cutoffs for sequence identity vary for V, J, and CDRH3 domain due to the distinct sequence diversity of the regions. For V and J regions, a sequence identity of 90% was used, whereas the CDRH3 clusters had a sequence identity cutoff of 30%. Selection of the sequence identity cutoff was based on consideration of the sequence diversity of the V, J, and CDRH3 regions, number of final clusters, and their size. The higher the sequence identity cutoff, the more, and thus smaller, clusters are created. Since the V and J regions are more conserved, higher cutoffs were applied to these regions. A smaller cutoff was chosen for the much more diverse CDRH3 region. Here, we set the requirement that each cluster must contain at least 100 sequences in order to ensure sufficient numbers for creation of position-specific nucleotide frequencies for HL assessment.

The cutoffs of 90% (V and J regions) and 30% (CDRH3) led to 14,638 V, 390 J, and 411 CDRH3 clusters. Of all of the clusters, 2,669 V and 116 J clusters are clusters of light-chain sequence regions (lambda and kappa class). We considered the

median sequence population of V and J clusters with 263 and 291 sequences, respectively, and the median CDRH3 cluster population with 9,863 sequences sufficiently above the chosen minimum of 100 sequences per clusters. In comparison, a higher CDRH3 cutoff of 50% would result in 142,295 clusters, with the majority strongly underpopulated. Only 18,380 clusters would contain more than 100 sequences.

The Rosetta human-like antibody design protocol

We made one major change to the IgReconstruct method.¹² Instead of relying solely on germline gene alignments to create HL frequency profiles, a cluster of sequences with the greatest sequence identity of the cluster center to the V, J, or CDRH3 region was assigned when creating HL profiles. To create amino acid restraints that can be interpreted by Rosetta,²⁵ we calculated amino acid frequencies from cluster nucleotide frequencies using Bayesian statistics.

In this study, we benchmarked antibodies designed with Rosetta that were created with the Bayesian antibody space, and without any HL restraints. Below, we refer to proteins that were designed with the Rosetta suite as decoys. The antibody space is calculated using PSSMs from our previous study (original), which used one sequence profile per germline gene and CDRH3 length, and clustered PSSMs that comprise multiple sequence profiles per germline gene and CDRH3 length. We then compare the total Rosetta score, predicted binding energy, sequence recovery, and HL between the designs. In the optimal case, the binding energy is not compromised compared to the WT and the HL increases. If the Bayesian amino acid frequencies of clustered immunome repertoires are able to model the human antibody space, we also expect to see increased sequence recovery, since the WT sequence of the designed antibodies are of human origin.

We curated a set of 27 high resolution (better than 2 Å) antibody crystal structures (Supplement 2 and 3) that is: 1) of human origin, and 2) available as a complex bound to its antigen. For each of the antibody structures, we created a Bayesian PSSM for the heavy and light chain separately. PSSM restraints were added to Rosetta in the form of a PSI-Blast formatted ASCII PSSM file.³² During Rosetta design, each mutation is then reevaluated for increased HL by either favoring a mutation (positive substitution score), or disfavoring

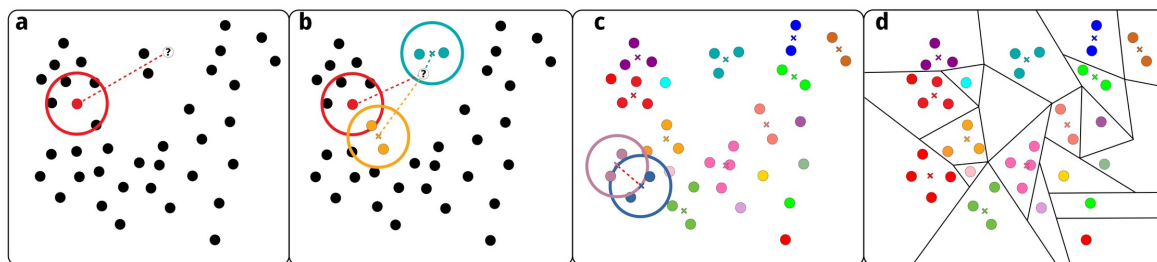


Figure 2. Schematic of fast immunome repertoire clustering. Each V, J, and CDRH3 sequence not assigned to a cluster in the repertoire is represented as a black dot. The arbitrary first sequence (red dot, a) initializes the first cluster (red circle, a). New sequences (shown as a question mark (?)) are processed in random order. The sequence identity compared to the cluster center (cross, b) is used as distance measure (dotted line, b). If a new sequence has an identity smaller than the threshold, it is assigned to an existing cluster (cyan, b), otherwise a new cluster is created until each sequence is assigned. Finally, overlapping clusters with an average cluster distance smaller than the sequence identity threshold were merged (c). Each cluster represents a unique nucleotide PSSM after processing the repertoire once (d).

a mutation (negative substitution score). The substitution scores ultimately guide Rosetta to prefer mutations that are more human-like.

Rosetta restraints must be carefully balanced to not overshadow the scoring terms that evaluate the thermodynamic stability of the protein. To estimate the effect on the protein's stability and the binding of the antibody to its antigen, Rosetta decoys created with HL restraints were compared to decoys without HL restraints (control). To ensure that the difference in the number of mutations between control and designs did not affect the results, decoys were also compared to control designs with a similar number of mutations. Each Rosetta HL design was assigned one control design that matched the V and J sequence identity the closest, and another control design that matched the sequence identity of the CDRH3 region the closest. We refer to this set of control sequences with matching sequence identity as "native group". The native group is assigned to both, the CDRH3 and the combined V/J regions separately. We ensured that a control design with a similar number of mutations exists by limiting the number of mutations with a separate Rosetta design run using the FavorNativeResidue function, using the weights of 1.0, 1.5, 2.0, and 2.5. The native group is used as a reference to calculate the difference of HL between designs, and the next closest control design with a similar number of mutations. **Supplementary Figure 1** demonstrates the close correlation of sequence identities between native, and human-like designs. Thus, for each Rosetta design, a control design that was generated by limiting the mutation rate can be found with a comparable mutation rate.

Rosetta design of human-like antibody structures remain thermodynamically plausible and antigen-specific

To prove that the Rosetta restraints were balanced correctly, the Rosetta energies of both control and human-like decoys were compared with each other. Rosetta Energy Units (REU) are a measure for thermodynamic stability of a protein complex.²⁶ The REU score can be used to compare different protein conformations, and to estimate the effects of mutational changes on thermodynamic stability. The more negative the score, the higher the predicted stability. Here, we compare the total REU scores of the Rosetta decoys, with the REU score of the WT crystal structure. A score smaller than 0 indicates an improvement compared to the WT structure. The more negative the reported results, the greater the improvement of the predicted stability of the protein compared to the WT.

On average, the Rosetta energy was improved during design compared to the relaxed WT structure by -142.8 ± 25.0 (control), -115.6 ± 26.6 (native), -82.7 ± 24.4 (original), or -68.7 ± 22.3 (clustered) REU. REU scores of the decoys that were restrained by the original or clustered PSSMs are more positive compared to the control (e.g., -142.8 ± 25.0 for the control vs. -68.7 ± 22.3 for clustered), which is expected due to the additional restraints added and indicates that a normally unexplored sequence space was sampled. When comparing the control group with the native group, we see a similar trend. This is mainly due to the limited number of mutations in the native group, which gives Rosetta fewer degrees of freedom to

optimize the protein. Overall, the design protocols improved the Rosetta energy compared to the WT energy in all cases (**Figure 3a**). The binding energy (see Methods), normalized by its interface size, retained the original values, suggesting a conserved specific antibody binding to its antigen and no weakened antigen affinity due to biophysical frustrations that may arise from an inappropriately balanced energy term (**Figure 3b**). We conclude that the chosen weights (see Supplement 1, Section 3) for HL restraints can be considered appropriate for the design task.

Improved human WT antibody sequence recovery for the V and J region

We hypothesize that our Bayesian amino acid profiles from clustered nucleotide repertoires can be used to model the human antibody space. As a consequence, it can be expected that antibodies designed with HL restraints explore a human sequence space that is more similar to the WT sequences of the designed structures. Sequence recovery rates of the human WT sequence were measured for the V, J, and CDRH3 regions separately.

When compared to the control group, the heavy-chain sequence recovery increases from $74.5 \pm 6.3\%$ (control) to $84.8 \pm 3.8\%$ (original), or $85.5 \pm 4.6\%$ (clustered). Similarly, the light-chain sequence recovery is increased from $77.1 \pm 7.2\%$ (control) to $85.6 \pm 4.3\%$ (original), or $85.5 \pm 4.6\%$ (clustered) (**Figure 4a**). The HL distribution of heavy and light chains in the control group diverges more compared to the HL Rosetta decoys, likely due to the significantly higher degree of freedom in the heavy chain than in the light chain due to its length and diversity. This effect is reduced for the HL Rosetta decoys since the possible sequence space to explore is more restricted by the Bayesian restraints.

In contrast to the increased sequence recovery of the V and J regions, the CDRH3 sequence recovery does not change significantly ($45.6 \pm 11.1\%$ (control) to $45.0 \pm 13.3\%$ (original)). With a slight decrease of sequence recovery to $40.6 \pm 10.1\%$, the clustered human-like design approach appears to influence the average sequence recovery (**Figure 4b**). We hypothesize that the CDRH3 sequence is a consequence of antibody maturation and differs between individuals too much to be reproducible without access to their sequence repertoires.

Increased human-likeness across the antibody framework region

Similar to the observed increased sequence recovery in the V and J regions of the antibody, we observed a substantial increase of HL. To compare the human-like Rosetta decoys, the control group was scored with both the clustered and the original PSSMs and compared to their respective HL decoys. HL of decoys generated with clustered and original PSSMs were not compared directly with each other due to the different sets of underlying sequences and nucleotide frequency distributions. **Figure 5a** visualizes the HL of the framework regions compared to the control group. While the heavy chain HL of the control group barely differed in their HL of $70.7 \pm 2.8\%$ (original) and $71.3 \pm 2.7\%$ (clustered), the human-like designs both increased substantially to $86.0 \pm 2.8\%$ (original) and

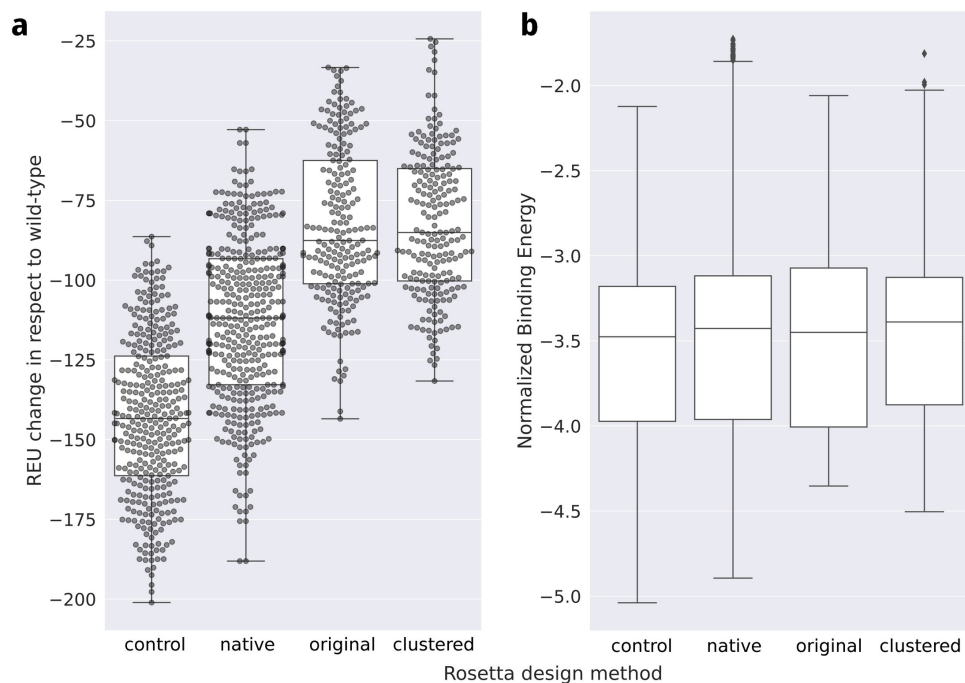


Figure 3. Rosetta energy and binding energy of the human antibody set. The total Rosetta score change relative to the relaxed wild-type score (a) and the interface energy normalized by the interface size (b). Decoys were generated with four protocols, control (Rosetta design without restraints), native (limited number of mutations), original PSSMs (original), and PSSMs of the clustered immunome repertoire (clustered). Overall, the change in Rosetta energy is favorable for all decoys (a), and did not exhibit substantial changes (b), indicating a sufficient balancing of Rosetta restraints.

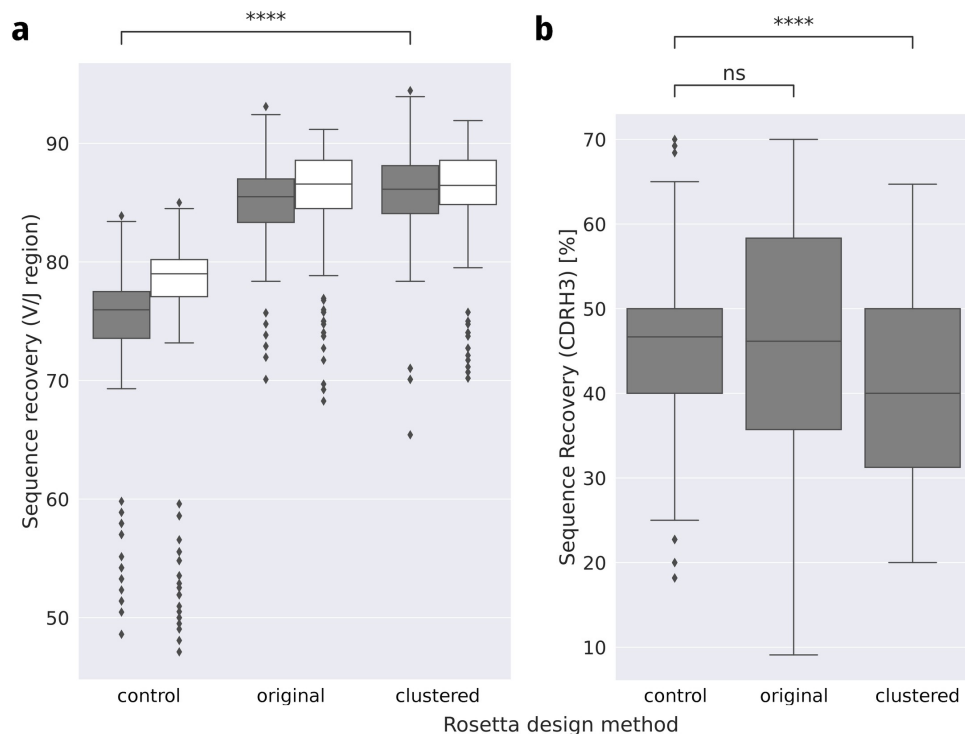


Figure 4. Wild-type sequence recovery rates of the antibody after Rosetta design. The sequence recovery for the V and J regions is increased for Abs designed with Rosetta using clustered and original PSSMs compared to the Abs designed without restraints (control) (a). The sequence recovery of the CDRH3 region does not substantially change when original PSSMs were used but is reduced when clustered PSSMs are applied (b). Heavy chains (gray), and light chains (white). Statistical annotations using the Mann-Whitney significance test (****: $p < 10e-4$; ns: not significant).

$87.9 \pm 2.6\%$ (clustered). Similarly, the light chain HL increased from $71.0 \pm 2.8\%$ to $86.3 \pm 2.2\%$ (original), or from $71.8 \pm 2.8\%$ to $86.8 \pm 2.1\%$ (clustered). The frequency of residues with identical amino acid identities as the aligned germline gene is $65.29 \pm 5.53\%$

(heavy), and $71.12 \pm 5.94\%$ (light) for HL designs compared to the control group with $46.01 \pm 4.95\%$ (heavy) and $52.21 \pm 5.08\%$ (light). WT sequences exhibit the greatest frequency of $82.63 \pm 9.27\%$ (heavy) and $88.55 \pm 8.60\%$ (light).

The HL of the CDR that is templated by V and J germline genes is overall slightly lower and more variable. These residues are more variable, surface exposed, and shorter. The average HL for heavy chains of this part of the CDR is $79.31 \pm 5.07\%$ (clustered) vs. $65.29 \pm 6.39\%$ (control) and $77.31 \pm 5.01\%$ (original) vs. $64.54 \pm 5.74\%$ (control). The templated region of the light chain CDR is on average $78.04 \pm 5.25\%$ (clustered) vs. $60.89 \pm 7.12\%$ (control) and $76.96 \pm 5.31\%$ (original) vs. $60.69 \pm 7.06\%$ (control) (Figure 5b). The CDR was determined using the IMGT antibody numbering.

As previously noted above, the native control group are the control decoys with the highest sequence similarity to the HL antibody design. Thus, the native decoys are the control decoys with a similar number of mutations when compared to HL decoys. When designing an antibody in Rosetta without HL restraints, a decrease of HL is expected as the number of mutations increases. We avoided inflating the performance of our method by comparing the results to a control that is allowed to diverge from the human WT with an unrestricted number of mutations. Instead, the results of our HL design

protocol are also compared to the native group. The native group contains selected control decoys that have a similar number of mutations when compared to our HL decoys.

Compared to the native group, HL Rosetta decoys do not decrease their HL as much as the native control group when compared to the WT HL. In the case of the design with original PSSMs, the HL of one antibody was higher than its WT HL (Figure 5c). In the clustered design scenario (Figure 5d), four antibodies increased their HL compared to the WT. In contrast to our design protocol, all control decoys with a similar sequence identity to the HL designs (“native”) decreased their HL.

The human-likeness of the CDRH3 benefits from repertoire clustering

The most difficult task of antibody HL assessment and engineering is the highly variable CDRH3 region. We previously introduced with IgReconstruct an HL assessment method based on single nucleotide frequencies of the observed antibody space.¹² The untemplated and diverse character of the CDRH3 requires an alternative approach to address CDRH3

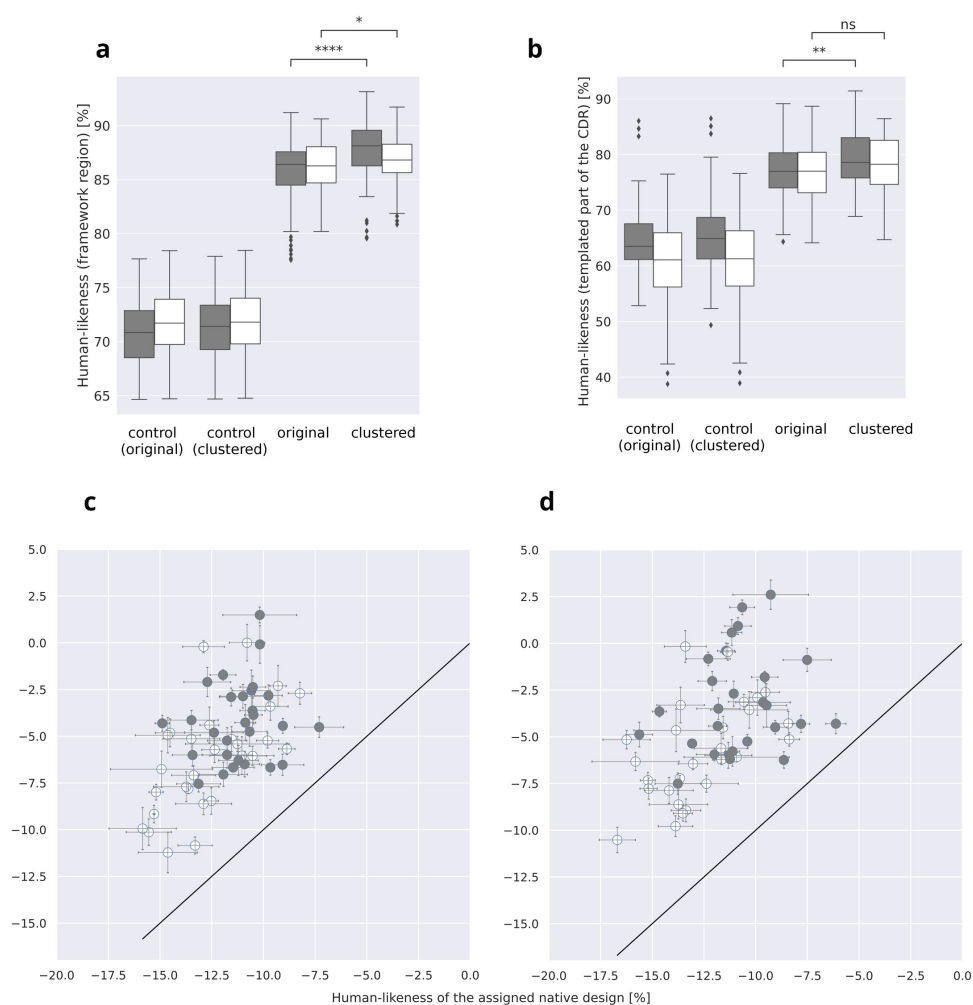


Figure 5. Human likeness of the V and J region, and templated part of the CDR after Rosetta design. The control group was scored using original and clustered PSSMs, and their HL is significantly lower than decoys created with original and clustered PSSMs. HL was assessed for the V and J regions (a), as well as the V and J germline gene templated part of the CDR (b). The change of HL in respect to the native group (unrestrained Rosetta decoys with similar sequence identity to the wild-type used as baseline), shows an improvement of HL in all cases when original (c) and clustered PSSM were used (d). Each data point represents a unique PDB ID. Heavy chains (gray), and light chains (white). Statistical annotations with the Mann-Whitney significance test (****: $p < 10e-4$; *: $1.00e-02 < p \leq 5.00e-02$).

HL. Thus, IgReconstruct was expanded to support repertoire clustering. Instead of germline genes, the sequence of the cluster center was used to assign nucleotide profiles to the CDRH3. Characteristic for the CDRH3 sequence space is its low commonality between human blood donors,^{18,19} and the relatively low number of observed sequences per individual (10^8 per donor versus $> 10^{12}$). To address the difficult task of defining HL for the CDRH3 region, Bayesian statistics were used to infer an enlarged amino acid sequence space from single nucleotide observations.

To assess the performance of our method for the CDRH3 specifically, decoys created with the original or clustered PSSMs were compared to the HL of the native group. Decoys created with and without clustering did not show a substantial change in HL when compared to the control group and the native group (Figure 6a). In contrast, eight antibodies in our benchmark that were designed with clustered PSSMs exhibited a positive change ($> 3.5\%$) of HL compared to their native group (Figure 6b). Structures with an increased HL compared to their natives were 1n0x ($8.0 \pm 0.7\%$), 2yc1 ($3.5 \pm 0.0\%$), 3l5x ($5.0 \pm 0.7\%$), 4hs6 ($4.0 \pm 0.6\%$), 4ioi ($4.3 \pm 1.4\%$), 4j6r ($3.5 \pm 0.8\%$), 5f9o ($6.1 \pm 0.7\%$), or 5xku ($6.5 \pm 0.9\%$). Supplementary Table 1 contains a complete list of changes in HL.

Due to the low shared commonality of CDRH3 sequences between human individuals, and the fact that the antibody repertoires used here were collected from healthy blood donors, we did not expect the PSSMs to carry the information needed to generate mature, highly specific antibody sequences in all 27 cases. Figure 6c visualizes the eight cases with an CDRH3 HL improvement of at least 3.5%. Even though the design approach using original PSSMs may increase the HL slightly, this effect is more pronounced when clustered restraints were used. For interpreting the HL scores, it is important to point out that the maximum possible HL an antibody can achieve is not always 100% and depends on how distinctly sharp the frequency distribution is. Generally speaking, the more diverse the sequence set, the flatter the observed frequency distribution. Here, the HL of the CDRH3 never exceeded 40% for clustered PSSMs (Figure 6b), and was

less than 32.5% for original PSSMs (Figure 6a). We therefore consider 3.5% to be a reasonable cutoff for determination of an improvement in HL.

The Bayesian antibody space can be used to retain and increase human-likeness in antibodies from *Mus musculus*

It was shown that the Bayesian antibody space resembles a human sequence space by comparing the Rosetta decoys with their human wild-type. The design on human antibodies led to observation that clustered PSSMs do not significantly influence the HL of the V and J regions, but lead to a much more significant response for the CDRH3 region. Overall, the design on 22 *Mus musculus* antibody structures (Supplement 3 and Table S2 and S3) can recapitulate these results. In addition, we demonstrate that in the majority of cases, the HL can either be retained or increased. HL is the average of nucleotide frequencies of the respective region of the sequence (V and J, or CDRH3, respectively). We consider a change in HL of 1% or greater significant since this implies that a significant number of residues has changed accordingly. We therefore define the range between -1% and $+1\%$ as no change and $< 1\%$, or $> 1\%$ as a significant change in HL. We categorize the Rosetta decoy (original, clustered, control) with the highest HL to provide an idea of what behavior can be expected for humanization of *Mus musculus* antibodies. The HL of the V and J regions increases for the light chain in 4 (original), and 3 (clustered) cases, whereas in nine cases (original and clustered) we detect no change, 9 (original and clustered) light chains with a decreased HL. The V and J regions of the heavy chain is improved 13 cases (original), or 12 cases (clustered) and do not show a significant change in 5 (original) and 6 (clustered) cases (Figure 7a). The CDRH3 exhibits a more reliable increase in HL in 20 (original) and 15 (clustered) cases, and 4 decoys using clustered PSSMs is reduced (Figure 7b).

To summarize, the HL of *Mus musculus* antibodies can be increased by up to 4.42% (light chain, clustered), 6.02% (heavy chain, clustered), or 9.90% (CDRH3, clustered), compared to

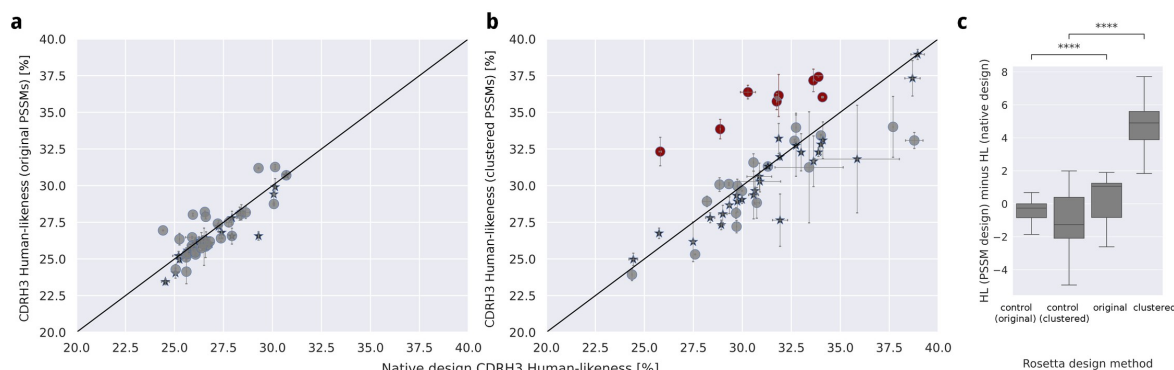


Figure 6. Human-likeness of the CDRH3 compared to Rosetta designs with a similar number of mutations (native). HL antibodies designed with Rosetta using original (a, circles) and clustered (b, circles) PSSMs compared to the HL of control designs (star). Each datapoint corresponds to a unique PDBID. In contrast to the Abs designed with original PSSMs, 8 of 27 PDBs improved their HL with clustered PSSMs when compared to the control and their native group (red). When selecting the eight antibodies with improved CDRH3 human-likeness, the significance of the change becomes visible for HL scores using clustered and original PSSMs. As reference, the control group was scored with both, original and clustered PSSMs (c). Statistical annotations with the Mann-Whitney significance test (****: $p < 10e-4$).

5.28% (light chain, original), 6.16% (heavy chain, original), or 5.13% (CDRH3, original). Clustered PSSMs are most effective when applied on the CDRH3.

A complete breakdown of PDB IDs and their HL can be found in Supplementary Tables S2 and S3 for all three design runs (original, clustered, control).

The decoys of decreased HL compared to the WT are likely cases of sub-optimal alignments of the human genes to the non-human antibodies for the V and J regions. In case of the CDRH3, we argue that, like the observation made with the design on human antibodies, some of the CDRH3 regions are not supported by the used human repertoires. This suggests a low likelihood that the blood donors would generate an immune response supporting the structure and specific binding mode. We do not observe such an effect when the original PSSMs are used because it cannot be expected that original CDRH3 PSSMs carry HL information specific for certain binding modes and conformations. We suggest that using an ensemble of PSSMs generated by different germline gene rearrangements has the potential to mitigate this effect for a use-case scenario with practical relevance.

Discussion

Valuable research-grade monoclonal antibodies are often derived from non-human organisms such as mouse, rat, or rabbit and chicken. Humanization techniques are required if such antibodies are developed for clinical use to avoid adverse effects and maintain the efficacy of antibodies when used in the clinic. Here, a method was developed for computational antibody design of IgG antibody isotypes with Rosetta. Even though our findings are exclusive to one antibody isotype, we

suggest that this method can be expanded to all isotypes for which a sufficient large amount of human nucleotide reference sequence are available for creation of the PSSM antibody space.

The observed antibody space of single blood donors (10^8) is magnitudes smaller than the expected diversity of human antibodies ($> 10^{12}$). The main reason for the high diversity and the low commonality^{18,19} of sequence repertoires is the variable CDRH3 region of the antibody, which enables specific binding to a wide variety of antigens. To model HL despite these difficulties, the previously described IgReconstruct¹² method was improved. Amino acid frequency profiles of clustered antibody repertoires were modeled from nucleotide sequences using Bayesian statistics. In this study, we hypothesize that Bayesian statistics are able to infer a larger antibody space by exploiting the degeneracy of the genetic code. The usefulness of this antibody space was demonstrated by improving the HL of the CDRH3 region with Rosetta design for 8 of 27 human co-crystal structures. For the variable and joining segments of the antibody, the HL was reliably improved compared to unrestrained Rosetta designs, suggesting that Rosetta can be employed using our method to either design novel antibodies that are human-like, or to re-design existing antibodies for HL. We also generated human-like decoys of *Mus musculus* antibody structures. We could observe increased HL especially in the CDRH3 of up to 9.9%. The overall similar performance suggests that our restraints can be used for antibody humanization.

It has been shown that human germline genes can be assigned to non-human antibody species,¹² which is the foundation of our method. However, the alignments naturally can be of low quality due to their low HL and low

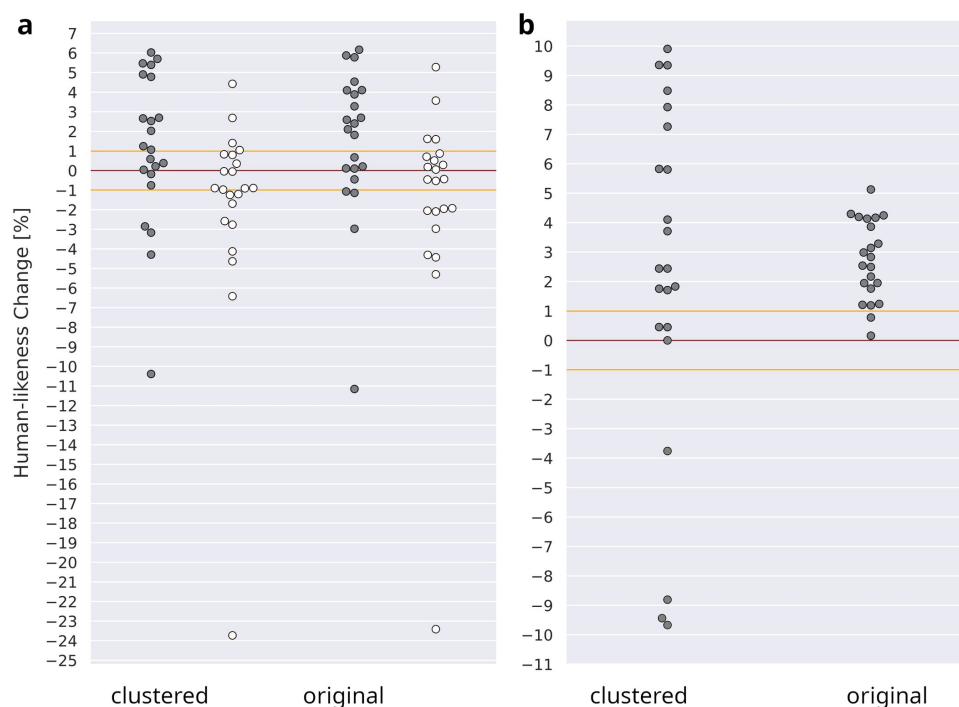


Figure 7. Change of human-likeness for 22 *Mus musculus* antibodies. Visualized is the Rosetta decoy with the highest change in HL relative to the WT for original and clustered PSSMs for heavy chains (gray) and light chains (white) separately. No change (0%) compared to the WT is highlighted with a red line. We consider a change of $\geq 1\%$ significant (Orange lines). HL was calculated separately for the V and J regions combined (a) and the CDRH3 region (b).

germline identity. Consequently, we observed outliers during the HL design of mouse antibodies. To humanize non-human antibodies, decisions must be made on a case-to-case basis, including: 1) which human germline gene combination should be used that optimally supports the (CDRH3) conformation, and 2) which areas of the variable region should be protected from mutations. Evaluation of the humanness may depend on specific project aims and include experimental evidence.

In this study, we suggest that nucleotide frequencies should be exploited to infer amino acid probabilities, instead of assessing amino acid frequencies directly, for three main reasons. First, ambiguities that arise during antibody amino acid characterization can be resolved on the nucleotide sequence level. For example, nucleotide triplets may only be partially aligned to germline genes. Or the same triplet can be assigned to different genes, which may occur in the junctions between V-D, D-J, or V-J gene assignments. The amino acid representation would fail to resolve ambiguities and inaccurately model the frequency statistics. Second, germline gene-dependent nucleotide statistics do not require special handling of frame-shifts. Third, single nucleotide observations can be used in combination with our Bayesian approach to suggest probabilities for amino acids that have not necessarily been observed in immune repertoires.

The low commonality of human CDRH3 sequences between human subjects has been shown before.^{12,19} This finding implies that antibodies specific to the same antigen but derived from different humans, can significantly differ in their sequence. This may explain our observation that clustered PSSMs fail to significantly increase the sequence identity to the WT antibody, since the sequence space is biased by individual repertoires. The CDRH3 HL on the other hand could be increased in 8 of 27 cases. It should be noted that the human blood samples for the repertoires used here were collected from otherwise healthy donors in the US. The individuals did not have exposure histories for all antigens observed in our dataset of 27 co-crystallized antibodies, which target a variety of antigens such as human immunodeficiency virus, hepatitis C virus, auto-antibodies, and dengue virus. In the cases of increased CDRH3 HL and decreased sequence identity to the wild-type, we assume that the immune response of the blood donor(s) would appear differently than the antibody deposited in the PDB.

Further applications may include established Rosetta design protocols that are available, such as RosettaScripts,³⁴ in combination with our PSSMs. The RosettaScript used for this study (see Supplement 1, Section 3) can be considered as a basic single-state affinity maturation protocol when co-crystal structures in complex with the antigen are used, where one conformation is referred to as a single state. Our approach can also be combined with RECON, a multi-state design protocol that can be used to design multi-specific antibodies, or for affinity maturation.^{35,36} Another possible use case is the *de-novo* design with RostetaAntibodyDesign (RabD)³⁷ of both human-like antibodies from an experimental structure of a non-binding antibody, or affinity maturation of an already existing antibody weakly binding antibody while maintaining or increasing HL.

Methods

Generation of single nucleotide frequency profiles

We assign single nucleotide frequency (SNF) profiles that were created from NGS-sequenced immunome repertoires to target amino acid sequences, as described in our previous study.¹² We achieve this by assigning the SNF to a set of reference V and J germline genes, which are then aligned to the amino acid target sequence. The region between the V and J alignments was modeled with CDRH3 PSSMs. Alignments of the human and mouse antibodies are visualized in Supplement 2 and 3. This enables the assessment of HL and the recovery of human like nucleotide sequences. SNF statistics were generated for each germline gene and CDRH3 loop length independently. Here, a similar approach was used, but instead of pooling all sequences depending on germline gene and loop length, the immunome repertoires were clustered based on minimal sequence identity. Separation by sequence identity allowed us to capture the SNF statistics that depend on reading frames or that are unique for antibody lineages. We used SNF profiles with a sequence identity of 50%, 70%, 80%, and 90% for V, D, and J regions and profiles with a CDRH3 loop identity of 16%, 23%, 30%, 37%, 44%, and 50%. We used 196,072,571 heavy chain and 129,095,736 light-chain sequences published by Soto et al.¹⁸ SNF profiles with an identity cutoff of 90% of V and J regions and 30% for the CDRH3 were chosen for all experiments as a compromise between the number of clusters and cluster sizes.

Bayesian approach to model the human amino acid sequence space

We deduced amino acid substitution scores from SNF profiles. We hypothesize that silent mutations and the degeneracy of the genetic code contain additional information that allows us to extrapolate a larger and smoother amino acid sequence space than experimentally determined via NGS sequencing. We developed a Bayesian approach to estimate amino acid probabilities from independent nucleotide triplet observations $p(aa|trpl)$ (Equation 1). We simplified Equation 1 with the assumption that the immunome repertoire is of infinite size and an observation of any amino acid at any position is possible with $p(aa)$ equals 1.0. The denominator $p(trpl)$ is the fraction of observed versus all possible triplet observations for all 20 amino acids.

$$p(aa \vee trpl) = \frac{p(trpl \vee aa)p(aa)}{p(trpl)} \quad (1)$$

The triplet probability for a given amino acid (nominator) and the global triplet probabilities (denominator) was reformulated as a fraction of amino acid pseudo-observations O_{pseudo} and divided by the total number of observations. Working with observations instead of frequencies allows further simplification of the equation. Pseudo-observations were inferred by pooling all encoding triplets together that encode an amino acid together for the first, second, and third position separately. Each nucleotide is counted once, as seen at the example of serine and the unique nucleotides T_{unique} were used to infer the

Table 1. Example of unique nucleotides at each position of the six triplets (T_{unique}) that encode Serine. T_{unique} is used to look up the observed nucleotide frequencies that contribute to a specific amino acid.

	Position 1	Position 2	Position 3
Triplet 1	A	G	T
Triplet 2	A	G	C
Triplet 3	T	C	T
Triplet 4	T	C	C
Triplet 5	T	C	A
Triplet 6	T	C	G
T_{unique}	A, T	G, C	T, C, A, G

triplet frequency for a specific amino acid. Table 1 exemplary shows T_{unique} for serine. The resulting observations are independent of the varying number of triplets that encode an amino acid.

O_{pseudo} is ultimately the sum of SNF observations of all unique nucleotides T_{unique} and resembles the frequency of a specific amino acid.

The probability to observe a specific amino acid resin at position $resi$, given the triplet observations from our SNF profile, is described in equation 2 as $p(resn | trpl)$. Pseudo-observations allow us to determine the greatest common denominator (GCD). The GCD is calculated for all three positions in the triplet.

$$p(resn \vee trpl) = \frac{\sum_{nt} T_{unique}(resn) O_{pseudo}(resi, nt)}{\sum_{aa} \sum_{nt} T_{unique}(aa) O_{pseudo}(resi, nt)} \quad (2)$$

The GCD cancels out, which leads us to our final Equation 3. We expect these amino acid pseudo-observations to approximate the Bayesian human amino acid sequence space.

$$p(aa, resi) = \frac{\sum_{nt} T_{unique}(resn) O_{pseudo}(resi, nt)}{\sum_{aa} \sum_{nt} T_{unique}(aa) O_{pseudo}(resi, nt)} \quad (3)$$

Generation of a position specific substitution matrix

We used the amino acid probabilities calculated in Equation 3 to assemble position-specific frequency matrices for the antibody variable regions. The substitution matrices are deduced from SNF profiles, which are individually generated for each sequence depending on its germline gene rearrangement.¹² We then converted these frequencies into PSI-Blast formatted position-specific substitution matrices for amino acids.³² The method to calculate substitution matrices from probabilities was described for Blast applications.³¹ We adopted the mathematical Equation 4 for substitution score calculation and applied it to each germline gene dependent and CDRH3 loop length-dependent amino acid probability matrix $p(aa, resi)$ (Equation 3).

$$s_{ij} = \frac{\left(\ln \frac{q_{ij}}{p_i p_j} \right)}{\lambda} \quad (4)$$

The target frequency q_{ij} describes the probability of mutating amino acid i to residue j , and the background probabilities for each amino acid i , and j (p_i and p_j). The scaling parameter λ was determined for each probability matrix individually

by optimizing the Spearman correlation between substitution score and nucleotide HL score PFM_{VJ} .¹² We used Powell optimization as optimization function, and cropped the s_{ij} values between -10 and 10 . Cropping ensures that outliers and extreme values turn into forced mutations during Rosetta design. Supplementary Figure 2 visualizes the effect correlation optimization has on the distribution of substitution scores, which leads to a better spread of the values within the allowed range of -10 to 10 .

Design of antibody structures with and without substitution score constraints

Crystal structures were obtained from the Protein Data Bank (PDB),²⁹ removing the solvent and all non-protein objects, as well as and duplicate chains. Rosetta was used for structural sequence design.²⁵ The Fv region of each antibody was designed with and without substitution scores, in apo and holo state if available. Sequence design through amino acid replacements was enabled for each residue in the variable region as long as a HL profile was available. All 20 canonical amino acid were allowed, excluding cysteines. Residues with a cysteine in the WT structure were excluded from design.

To add the constraints to Rosetta we used our PSI-Blast formatted PSSM in combination with the Favor SequenceProfileMover and global scaling, and a weight of 5. All positions in the PSSM without any information about substitution scores (untemplated regions like insertions or the antigen) were filled with zeros. In order to measure the sequence recovery rate, we compared the variable regions of the heavy and light chains only.

Calculation of antibody-antigen binding energies

Binding energies were calculated using Rosetta's interface-analyzer application using the value $dG_{separated}/dSASAx100$, which is the difference of total Rosetta score between the bound and unbound state. The normalization factor in square Angstroms is the solvent-accessible surface area that gets buried in the bound state.

Human-likeness, sequence recovery calculation, and SNF profile generation

HL was calculated as previously described as PFM_{VJ} is a direct measure of observed nucleotide frequencies.¹² HL values were reported for the V and J region as the average of nucleotide frequencies (PFM_{VJ}) and adopted for the CDRH3 region analogously (PFM_{CDRH3}). SNF matrices were generated by from the WT crystal structures using IgReconstruct and the clustered version of the IgReconstruct algorithm. SNF matrices were then used to create the substitution scores/PSSMs as HL restraint.

The sequence recovery was calculated by counting the number of mutations introduced during Rosetta each design run, divided by the total number of residues in the antibody chains. Sequence recovery was calculated for heavy and light chains separately.

Abbreviations

ADA	Anti-Drug-Antibodies
CDR	Complementarity-Determining Region
CDRH3	Heavy chain CDR3 region
GCD	Greatest Common Denominator
HL	Human-likeness
NGS	Next-Generation Sequencing
PSSM	Position Specific Scoring Matrix
REU	Rosetta Energy Units
SNF	Single Nucleotide Frequencies
WT	Wild-Type

Acknowledgments

We thank Prof. Cristina Martina Elisa, and Prof. Clara Schoeder at the University of Leipzig for reviewing the paper for clarity and language.

Disclosure statement

J.E.C. has served as a consultant for Luna Biologics, is a member of the Scientific Advisory Board of Meissa Vaccines and is Founder of IDBiologics. The Crowe laboratory at Vanderbilt University Medical Center has received sponsored research agreements from Takeda Vaccines, IDBiologics and AstraZeneca.

Funding

The work was supported by the National Institute of Health grants U01 AI150739 and R01 AI141661. Jens Meiler is supported by a Humboldt Professorship of the Alexander von Humboldt Foundation in Germany;

ORCID

Samuel Schmitz  <http://orcid.org/0000-0001-5314-6095>
 Emily A. Schmitz  <http://orcid.org/0000-0002-5122-8991>
 James E. Crowe Jr  <http://orcid.org/0000-0002-0049-1079>
 Jens Meiler  <http://orcid.org/0000-0001-8945-193X>

Data availability

The IgReconstruct webservice has been extended to output clustered and original PSSMs which can directly be used in combination with Rosetta scripts (see Supplement 1, Section 3). The IgReconstruct webservice is available at <http://www.meilerlab.org/index.php/servers/IgReconstruct>

References

- Kaplon H, Reichert JM. Antibodies to watch in 2019. *MAbs*. 2019;11(2):219–38. doi:10.1080/19420862.2018.1556465. Cited in: PMID: 30516432
- Kaplon H, Reichert JM. Antibodies to watch in 2021. *MAbs*. 2021;13(1):1860476. doi:10.1080/19420862.2020.1860476. Cited in: PMID: 33459118
- Steinberger P, Sutton JK, Rader C, Elia M, Barbas CF. Generation and characterization of a recombinant human CCR5-specific antibody. A phage display approach for rabbit antibody humanization. *J Biol Chem*. 2000;275(46):36073–78. doi:10.1074/jbc.M002765200. Cited in: PMID: 10969070
- Tsurushita N, Park M, Pakabunto K, Ong K, Avdalovic A, Fu H, Jia A, Vásquez M, Kumar S. Humanization of a chicken anti-IL-12 monoclonal antibody. *J Immunol Methods*. 2004;295:9–19. doi:10.1016/j.jim.2004.08.018.
- Gillies SD, Lo KM, Wesolowski J. High-level expression of chimeric antibodies using adapted cDNA variable region cassettes. *J Immunol Methods*. 1989;125(1–2):191–202. doi:10.1016/0022-1759(89)90093-8. Cited in: PMID: 2514231
- Bonwick GA, Cresswell JE, Tyreman AL, Baugh PJ, Williams JH, Smith CJ, Armitage R, Davies DH. Production of murine monoclonal antibodies against sulcofuron and fluclofuron by in vitro immunisation. *J Immunol Methods*. 1996;196(2):163–73. doi:10.1016/0022-1759(96)00098-1. Cited in: PMID: 8841454
- Nechansky A. HAHA – nothing to laugh about. Measuring the immunogenicity (human anti-human antibody response) induced by humanized monoclonal antibodies applying ELISA and SPR technology. *J Pharm Biomed Anal*. 2010;51(1):252–54. doi:10.1016/j.jpba.2009.07.013.
- Holgate RGE, Baker MP. Circumventing immunogenicity in the development of therapeutic antibodies. *IDrugs*. 2009;12(4):233–37. Cited in: PMID: 19350467
- Harding FA, Stickler MM, Razo J, DuBridg RB. The immunogenicity of humanized and fully human antibodies: residual immunogenicity resides in the CDR regions. *MAbs*. 2010;2(3):256–65. doi:10.4161/mabs.2.3.11641. Cited in: PMID: 20400861
- Jones TD, Carter PJ, Plückthun A, Vásquez M, Holgate RGE, Hötzel I, Popplewell AG, Parren PW, Enzelberger M, Rademaker HJ, et al. The INNs and outs of antibody nonproprietary names. *MAbs*. 2016;8(1):1–9. Cited in: PMID: 26716992. doi:10.1080/19420862.2015.1114320.
- Parren PW, Carter PJ, Plückthun A. Changes to international nonproprietary names for antibody therapeutics 2017 and beyond: of mice, men and more. *MAbs*. 2017;9(6):898–906. doi:10.1080/19420862.2017.1341029. Cited in: PMID: 28621572
- Schmitz S, Soto C, Crowe JE Jr, Meiler J. Human-likeness of antibody biologics determined by back-translation and comparison with large antibody variable gene repertoires. *mAbs*. 2020;12(1):1758291. doi:10.1080/19420862.2020.1758291. Cited in: PMID: 32397786
- Wollacott AM, Xue C, Qin Q, Hua J, Bohnuud T, Viswanathan K, Kolachalama VB, Daggett V. Quantifying the nativeness of antibody sequences using long short-term memory networks. *Protein Engineering, Design and Selection*. 2019;32(7):347–54. doi:10.1093/protein/gzz031.
- Seeliger D, Tosatto SCE. Development of scoring functions for antibody sequence assessment and optimization. *PLoS One*. 2013;8(10):e76909. doi:10.1371/journal.pone.0076909. Cited in: PMID: 24204701
- Gao SH, Huang K, Tu H, Adler AS. Monoclonal antibody humaneness score and its applications. *BMC Biotechnol*. 2013;13(1):55. doi:10.1186/1472-6750-13-55. Cited in: PMID: 23826749
- Lazar GA, Desjarlais JR, Jacinto J, Karki S, Hammond PW. A molecular immunology approach to antibody humanization and functional optimization. *Mol Immunol*. 2007;44(8):1986–98. doi:10.1016/j.molimm.2006.09.029. Cited in: PMID: 17079018
- DeWitt WS, Lindau P, Snyder TM, Sherwood AM, Vignali M, Carlson CS, Greenberg PD, Duerkopp N, Emerson RO, Robins HS. A public database of memory and naive B-Cell receptor sequences. *PLOS ONE*. 2016;11(8):e0160853. doi:10.1371/journal.pone.0160853.
- Soto C, Bombardi RG, Branchizio A, Kose N, Matta P, Sevy AM, Sinkovits RS, Gilchuk P, Finn JA, Crowe JE. High frequency of shared clonotypes in human B cell receptor repertoires. *Nature*. 2019;566(7744):398–402. doi:10.1038/s41586-019-0934-8. Cited in: PMID: 30760926
- Briney B, Inderbitzin A, Joyce C, Burton DR. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature*. 2019;566:393–97. Cited in: PMID: 30664748. doi:10.1038/s41586-019-0879-y.
- Giudicelli V, Chaume D, Lefranc M-P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res*. 2005;33:D256–D261. Cited in: PMID: 15608191. doi:10.1093/nar/gki010.

21. Charles A, Janeway J, Travers P, Walport M, Shlomchik MJ. The generation of diversity in immunoglobulins. *Immunobiology: the Immune System in Health and Disease* 5th edition [Internet]. 2001. [cited 2021 Nov 25]
22. Teng G, Papavasiliou FN. Immunoglobulin Somatic Hypermutation. *Annu Rev Genet.* 2007;41(1):107–20. doi:10.1146/annurev.genet.41.110306.130340.
23. Kovaltsuk A, Leem J, Kelm S, Snowden J, Deane CM, Krawczyk K. Observed antibody space: a resource for data mining next-generation sequencing of antibody repertoires. *J Immunol.* 2018;201(8):2502–09. doi:10.4049/jimmunol.1800708. Cited in: PMID: 30217829
24. Wooden SL, Koff WC. The human vaccines project: towards a comprehensive understanding of the human immune response to immunization. *Hum Vaccin Immunother.* 2018;14(9):2214–16. doi:10.1080/21645515.2018.1476813. Cited in: PMID: 29847214
25. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* 2011;487:545–74. Cited in: PMID: 21187238. doi:10.1016/B978-0-12-381270-4.00019-6.
26. Alford RF, Leaver-Fay A, Jeliazkov JR, O’Meara MJ, DiMaio FP, Park H, Shapovalov MV, Renfrew PD, Mulligan VK, Kappel K, et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J Chem Theory Comput.* 2017;13:3031–48. Cited in: PMID: 28430426. doi:10.1021/acs.jctc.7b00125.
27. Thornburg NJ, Nannemann DP, Blum DL, Belser JA, Tumpey TM, Deshpande S, Fritz GA, Sapparapu G, Krause JC, Lee JH, et al. Human antibodies that neutralize respiratory droplet transmissible H5N1 influenza viruses. *J Clin Invest.* 2013;123(10):4405–09. Cited in: PMID: 23999429. doi:10.1172/JCI69377.
28. Sivasubramanian A, Chao G, Pressler HM, Wittrup KD, Gray JJ. Structural model of the mAb 806-EGFR complex using computational docking followed by computational and experimental mutagenesis. *Structure.* 2006;14(3):401–14. doi:10.1016/j.str.2005.11.022. Cited in: PMID: 16531225
29. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, Shi J, Deane CM. SABDab: the structural antibody database. *Nucleic Acids Res.* 2014;42(D1):D1140–1146. doi:10.1093/nar/gkt1043. Cited in: PMID: 24214988
30. Prihoda D, Maamary J, Waight A, Juan V, Fayadat-Dilman L, Svozil D, Bitton DA. BioPhi: a platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *MABS.* 2022;14(1):2020203. doi:10.1080/19420862.2021.2020203. Cited in: PMID: 35133949
31. Altschul SF. Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol.* 1991;219(3):555–65. doi:10.1016/0022-2836(91)90193-a. Cited in: PMID: 2051488
32. Altschul SF, Gertz EM, Agarwala R, Schäffer AA, Yu Y-K. PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Res.* 2009;37(3):815–24. doi:10.1093/nar/gkn981. Cited in: PMID: 19088134
33. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25:3389–402. Cited in: PMID: 9254694. doi:10.1093/nar/25.17.3389.
34. Fleishman SJ, Leaver-Fay A, Corn JE, Strauch E-M, Khare SD, Koga N, Ashworth J, Murphy P, Richter F, Lemmon G, et al. RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PLoS One.* 2011;6(6):e20161. Cited in: PMID: 21731610. doi:10.1371/journal.pone.0020161.
35. Sevy AM, Wu NC, Gilchuk IM, Parrish EH, Burger S, Yousif D, Nagel MBM, Schey KL, Wilson IA, Crowe JE, et al. Multistate design of influenza antibodies improves affinity and breadth against seasonal viruses. *Proc Natl Acad Sci U S A.* 2019;116(5):1597–602. Cited in: PMID: 30642961. doi:10.1073/pnas.1806004116.
36. Sevy AM, Jacobs TM, Crowe JE, Meiler J, Peters B. Design of protein multi-specificity using an independent sequence search reduces the barrier to low energy sequences. *PLoS Comput Biol.* 2015;11(7):e1004300. doi:10.1371/journal.pcbi.1004300. Cited in: PMID: 26147100
37. Adolf-Bryfogle J, Kalyuzhnyi O, Kubitz M, Weitzner BD, Hu X, Adachi Y, Schief WR, Dunbrack RL, Ben-Tal N. RosettaAntibodyDesign (RABD): a general framework for computational antibody design. *PLoS Comput Biol.* 2018;14(4):e1006112. doi:10.1371/journal.pcbi.1006112. Cited in: PMID: 29702641