Computational methods to engineer antibodies for vaccines and therapeutics

By

Samuel Schmitz

Dissertation

Submitted to the School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Chemistry

February 28, 2022

Nashville, Tennessee

Approved:

Jens Meiler, Ph.D.

James E. Crowe Jr., M.D.

Lars Plate, Ph.D.

Lauren Buchanan, Ph.D.

1 Acknowledgments

I would like to start with an exceptional thank you to Jens Meiler for giving me the once in a lifetime opportunity to experience Vanderbilt and all of the conferences along the way. Without Jens, I may never have left Germany and experienced all that is both interesting and unique about working in America. Jens was instrumental in patiently preparing me for my current career at Moderna, and is the reason I met my wonderful wife. Many thanks to both Jens Meiler and James E. Crowe Jr. for funding my PhD in the antibody interface between their labs. I wish to thank both for mentoring me, providing guidance, and challenging me to become a better scientist through the spirited discussions throughout my time in the PhD.

I would like to further give thanks for the wonderful and intellectually engaging collaborations that I was a part of, where I worked with the teams of both Antje Körner and Anna Kirstein at the University of Leipzig, Germany. I am grateful to have gotten the chance to support insight into the structural basis of clinical phenotypes, and was able to learn about new perspectives and constructively contribute.

Cinque Soto was a pleasant collaborator, supportive with discussions, and shared interesting papers with me during his time at and even after his departure from Vanderbilt. I give thanks to his continued support even after my defense, and look forward to staying in contact in the future. It was also a great experience to have intellectually stimulating discussions with Sergey Ovchinnikov when I met him at various RosettaCONs.

The Master student Moritz Ertelt, whom I worked with in support of his Master's thesis project, was a joyous ball of sunshine and now a life-long friend. I cannot express how wonderful it was to have someone to complain to about life and science. Also thanks to another life-long friend I found in Jens's lab, Francois Berenger, even though he was only in the lab for a short time. It was joy to work in the same office. Many thanks to the continuing updates from Tokyo.

Many thanks to all of my committee members, both those on the current iteration of the committee as well as those who left for greener pastures. Thanks to Jens Meiler, James E. Crowe Jr., Lars Plate, Lauren Buchanen, Cinque Soto, and Terry Lybrand. At last but not least I would like to point out my wife Emily Schmitz, for proofreading many of my papers and this written dissertation, my friends for forcing me to gather sunshine out of the lab on occasion, and finally my family for providing long-distance support and many dog pictures.

Contents

List of Figures

List of Tables

4 Summary

The multi-billion antibody drug market shows substantial growth and has many successfully antibody products since 1975. One of the major challenges to produce antibodies as vaccines and therapeutics is the ability to develop and manufacture them, and adverse effects that can reduce the efficacy of an antibody product or induce serious health concerns (immunogenicity). Thus, the projects of these topics evolve around methods to design antibodies with low immunogenic effects, and a method to predict if an antibody can be expressed, which can ultimately aid in re-engineering an antibody for increased expressability.

The increasing availability of immunome repertoires, that is the antibody sequences from B-Cells obtained from peripheral blood samples from human blood donors, and the increasing number of antibody (co-)crystal structures facilitates the development of methods that combine large sequence repertoires of observed sequences and computational structural design. Antibodies can specifically bind to a wide variety of antigens and body-foreign particles. The wide range of specificity is generated by multiple mechanisms, which include germline gene rearrangements, non-templated junction segments, and somatic hyper-mutation. The resulting human antibody sequence space is thus estimated to be at least in the range of 10^{13} unique antibody sequences. At the same time, even the largest immunome repertoires list just 10^6 unique sequence per human individual. Projects in this thesis deal with the design of antibodies that are more human-like and therefore reduce the likelihood of inducing immunogenic effects.

The motivation of the four main projects in this dissertation is outlined in Chapter 5. Since the developed methods are ultimately to be used in conjunction with the structural protein modeling software Rosetta, the developed methods for this dissertation are set into context of existing protocols relevant to antibody design. The four main method developed are 1) the human-likeness assessment of antibodies using statistics of complete immunome repertoires. 2) the design of proteins using homologous sequence information to retain protein function during protein design. 3) the engineering of human-like antibodies with Rosetta and a probablistic human-like sequence space. 4) the prediction and re-engineering of antibodies for increased expressability.

Chapter 6 describes the human-likeness estimation of antibodies using statistics of complete immunome repertoires. This is achieved by creating nucleotide frequency statistics for each germline gene of an antibody, avoiding the need for pairwise alignment of the database. The statistics can be used to distinguish human from non-human, and engineered antibodies (chimeric or non-human origin). The back-translation allows to create a nucleotide sequence for each amino acid sequence.

The following chapter 7 describes the usage of co-evolutionary information during Rosetta design, leading to designed sequences that are more natural and are much more likely to retain the identity of the wild-type for functional residues. Thus, co-evolutionary information can be understood as a fingerprint for function. In the benchmark of human-likeness, it was observed that the highly variable CDRH3 region remains elusive to human-likeness assessment. To further improve upon this technique it is recommended to ultimately make use of co-evolutionary residue information in antibody lineages. This will allow the grouping of antibodies that undergo similar

maturation pressure for antigen binding.

The used immunome repertoire is not grouped into lineages and the antibodies specificity remains unknown. Consequently, chapter 8 describes the first step towards sequence pattern analysis by combining the human-likeness nucleotide frequencies (Chapter 6 with a clustering approach). In conjunction with Rosetta, it was shown that antibodies designed with clustered human-likeness profiles are more human-like and render the CDRH3 statistics more meaningful, with increased human-likeness in some cases.

Finally, the expressability of antibodies was addressed with a Deep Learning approach. The protein expression in general involves a complex cascade comprising: transcription, translation, folding, post-translational modifications, vesicle transport and secretion. Deep Learning has the potential to recognize sequence patterns responsible for low expression independent of its exact biophysical cause. Chapter 9 describes the expressability prediction for a set of paired Flu antibodies. The re-design with Rosetta increased the probability of predicted expressability in all cases while exhibiting distinct N and C terminal mutational patterns.

All four major projects comprising this dissertation are critically discussed in chapter 10 with its potentials and limitations, and future approaches to improve upon these techniques are suggested. With the main takeaways that human-like antibody engineering may profit from incorporating methods for co-evolutionary analysis and potentially Deep Learning techniques. This may ultimately lead to powerful techniques for computational antibody discovery. Antibody expressability remains an unsolved and complex challenge requiring an integrated Research and Development cycle that integrates and collaborates with experimental antibody expression.

5 Introduction

Selected sections of this chapter have been adapted from (Schoeder, Schmitz et al., 2021).

5.1 The Adaptive Immune Response

The immune system defends the host against infection. Innate immunity comprises the skin barrier, blood chemicals, and immune system cells and serves as a first line of defense. However, it lacks the ability to recognize certain pathogens and to provide the specific protective immunity that prevents reinfection. In contrast, adaptive immunity is able to respond dynamically from highly diverse antigen-specific receptors that enable the immune system to recognize any foreign antigen. In the adaptive immune response, antigen-specific antibodies proliferate and differentiate to eliminate the pathogen (Murphy, Weaver, and Janeway 2017). The human Antibody (Ab) consists of a heavy and light chain, both of which can be divided into a constant and variable region (Fig. 1a). The antigen-specificity is for the most part established in the variable region (Fv) and is therefore main focus of this dissertation.

The ability of the Ab to mature from an unspecific (germline) state and to obtain high binding affinity to specific epitopes arises from the germline gene rearrangement, non-templated nucleotides at the junction between gene segments, and somatic hyper-mutation. Genetically both heavy and light chains can be recombined from different gene loci. Human light chains are differentiated between the chain class λ (chromosome 22) and κ chains (chromosome 2), whereas heavy chain loci are located on chromosome 14 (Fig. The exact sequences of the germline genes can differ between individuals and different ethnic groups, but the ImMunoGeneTics information system (IMGT) system has assembled gene sequences that can be used as reference (Giudicelli, Chaume, and M.-P. Lefranc 2005). To date 556 V, 52 D, and 34 J human germline gene alleles have been cataloged by IMGT/GeneDB and resemble the basis for antibody variability by recombination(Fig. 1b).

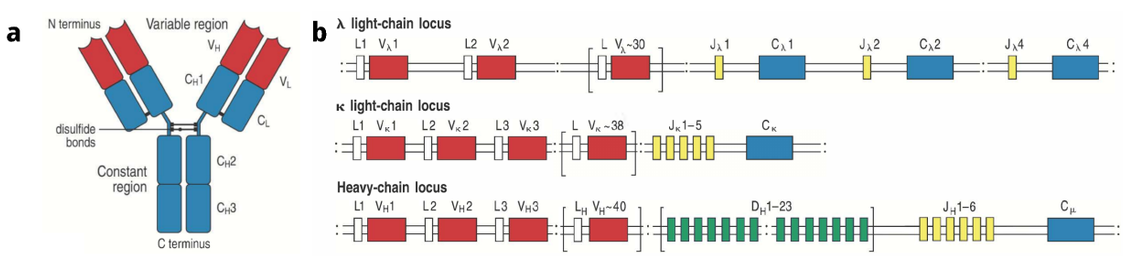

Figure 1: **The high sequence diversity of an antibody is facilitated by gene rearrangements and mutations.** Simplified schematic representation of an antibody molecule. The variable region directly binds the antigen and undergoes B-Cell maturation (a). The germline gene organization of the heavy and light chain loci in the genome. 29-33 κ V light chain loci, 38 λ V light chain loci, and about 40 heavy chain loci (red) across three chromosomes facilitate gene recombination and sequence variability. The highly variable third loop of the heavy chain can be encoded by special D-genes (yellow) (b). (Janeway's Immunobiology page 129-161) (Murphy, Weaver, and Janeway 2017)

Human antibodies consist of a heavy and a light chain, which share a well-conserved constant region (Fc) and framework region (Fr) within the variable region (Fv). Antibody variability is established through the process of recombination of the V, D, and J genes in the creation of the naïve B cell repertoire and by the subsequent somatic hyper-mutation of antibody variable genes in the stimulated B cells during germinal center reactions. Sequence variation and structural variation

of the antibody manifest in the Complementary Determining Region (CDR) as three highly variable loop regions in each heavy and light chain, which facilitates antigen recognition. The sequences of most antibodies are very similar in the Fc and Fr regions if they share the same germline genes. The high variability in the CDR loop regions of the variable domain impedes accurate structure prediction and design of antibodies and has posed a significant challenge in modeling the native conformations of antibody–antigen structures (North, Lehmann, and Dunbrack 2011; Finn et al. 2016).

The segment of the Ab with the greatest sequence variability and therefore potential to differentiate is the third loop of the heavy chain, which is partially based on one or more D gene fragments (B. S. Briney et al. 2012). Whenever the mature Ab sequence is based upon a germline, we speak of it as templated. Identifying the position of the CDR and FR regions is a first crucial step in the characterization of an Ab. Several numbering schemes have been introduced to identify the CDRs of a given antibody from the sequence and to provide a consistent structure-based alignment system (Dondelinger et al. 2018). Prominent numbering schemes include Chothia, Kabat, and AHo numbering schemes, to align CDRs spatially (Chothia et al. 1989; Al-Lazikani, Lesk, and Chothia 1997; Honegger and A. Plückthun 2001). These numbering schemes are either based on antibody sequence alignments (Kabat), the structural superposition of crystal structures (Chothia and Aho). Another commonly used numbering scheme is IMGT numbering, which is derived from the gene assignment (Brochet, M.-P. Lefranc, and Giudicelli 2008; M.-P. Lefranc, Giudicelli, et al. 2015).

Immune repertoire fingerprinting has been developed to group repertoires together that share a common disease state or history. Despite generally high variability of antibody repertoires between individuals, common exposure can lead to a co-evolution of antibody lineages, for example in the case of HIV (Liao et al. 2013; Doria-Rose et al. 2014) or influenza (Krause et al. 2011; Jiang et al. 2013; Joyce et al. 2016). V and J gene distributions as a result of antibody lineages with high specificity can be used to identify a common specificity of immunome repertoires (Sevy, Soto, et al. 2019).

Consequently, a clonotype definition has been developed ('VJ3') that encompasses the V and J germline gene as well as the length of the Heavy Chain Complementary Determining Region 3 (CDRH3) region for the analysis of immune repertoires (Soto, Bombardi, et al. 2019). CDRH3 D germline gene(s) are not considered due to the high sequence variability and the resulting low confidence in germline gene predictions, aggravating the challenging task of a functional characterization of immunome repertoires. Tools that infer germline gene rearrangements like IgBlast (Ye et al. 2013; Soto, Finn, et al. 2020) or MIXCR (Bolotin et al. 2015) provide germline gene rearrangements and the partitioning of the Ab into Fr and CDR domains.

In this work, IgBlastN (Ye et al. 2013) was used to analyze antibody sequences via nucleotide germline gene alignments. Consequently, the Ab partitioning schema of choice is the germline gene based IMGT numbering schema (M.-P. Lefranc, Giudicelli, et al. 2015). Chapter 6 describes the implementation and use cases of a antibody amino-acid partitioning algorithm for protein sequences similar to IgBlastN for nucleotide sequences.

5.2 Combining Next Generation Sequencing And Structural Information

B lymphocytes are a population of cells that express clonally diverse cell surface antibodies. These B-cells of peripheral blood samples can be sequenced on a large scale to assess the sequence space of individuals before, after, or during infections with a pathogen and curated in form of immunome repertoires. With the decreasing costs of Next Generation Sequencing (NGS) over the past decade (Koboldt et al. 2013; Metzker 2010), the availability of Ab sequence databases has consequently increased. NGS immunome repertoires typically comprise T- and B-Cell receptor sequences of the variable region of antibodies. The Adaptive Immune Receptor Repertoire Community (AIRR) facilitates to share these special types of datasets (Rubelt et al. 2017). The platform iReceptor is a portal to access and analyze repertoire repositories across different countries and workgroups in a uniform manner (Corrie et al. 2018). At the time of its publication iReceptor made over 145 million sequences available from 17 studies and 13 research labs organized in 4 remote data repositories (Fig 2a).

Approximately 6800 Ab crystal structures have been deposited at the Protein Databank (PDB) (Berman et al. 2000) often with detailed insight into the binding mode with the corresponding antigen as co-crystal structure (Fig. 2b). Out of these, roughly 1000 Ig molecules are annotated as fully human in the Structural Antibody Database (SAbDab) (Dunbar et al. 2014).

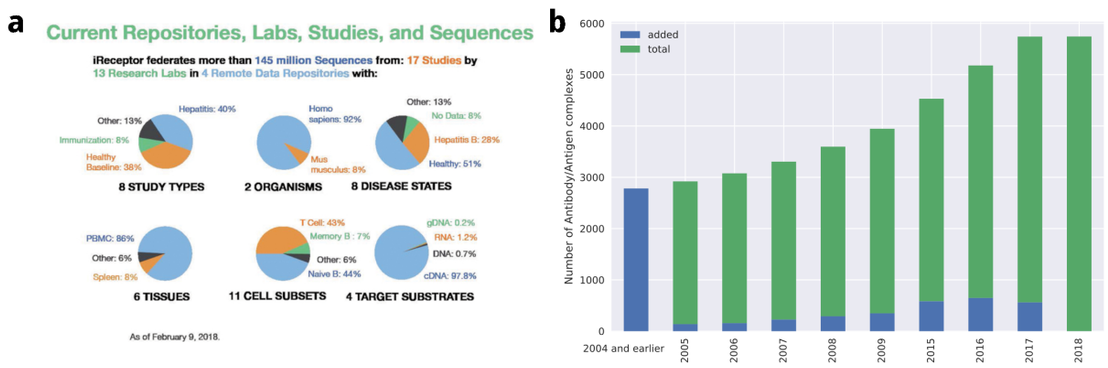

Figure 2: **Availability of antibody sequences and protein structures as of 2018.** AIRR-seq repositories in available in the iReceptor portal allows access of 145 million sequences and growing. The platform is decentralized and accesses data repositories of different work-groups in multiple countries (a). Antibody structure deposits in the Structural Antibody Database (SAbDab) has been steadily growing since 2004 and had approximately 6800 structures available (b).

This dissertation aims to produce new technologies to process large sequence datasets and antibody engineering tools, to inform the computational structural antibody engineering. Here, about 350 million unique nucleotide sequences (Soto, Bombardi, et al. 2019) were used in combination with high-resolution antibody structures from SAbDab (Dunbar et al. 2014) and computational predictions of antibody structures.

5.3 Monoclonal Antibodies for Clinical Use

Antibodies also represent a class of therapeutic proteins, that can routinely be produced in large quantities for either therapeutic use or as vaccines. To date, at least 550 therapeutic monoclonal antibody (mAb)s have been studied in clinical trials and 79 mAbs have been approved by the

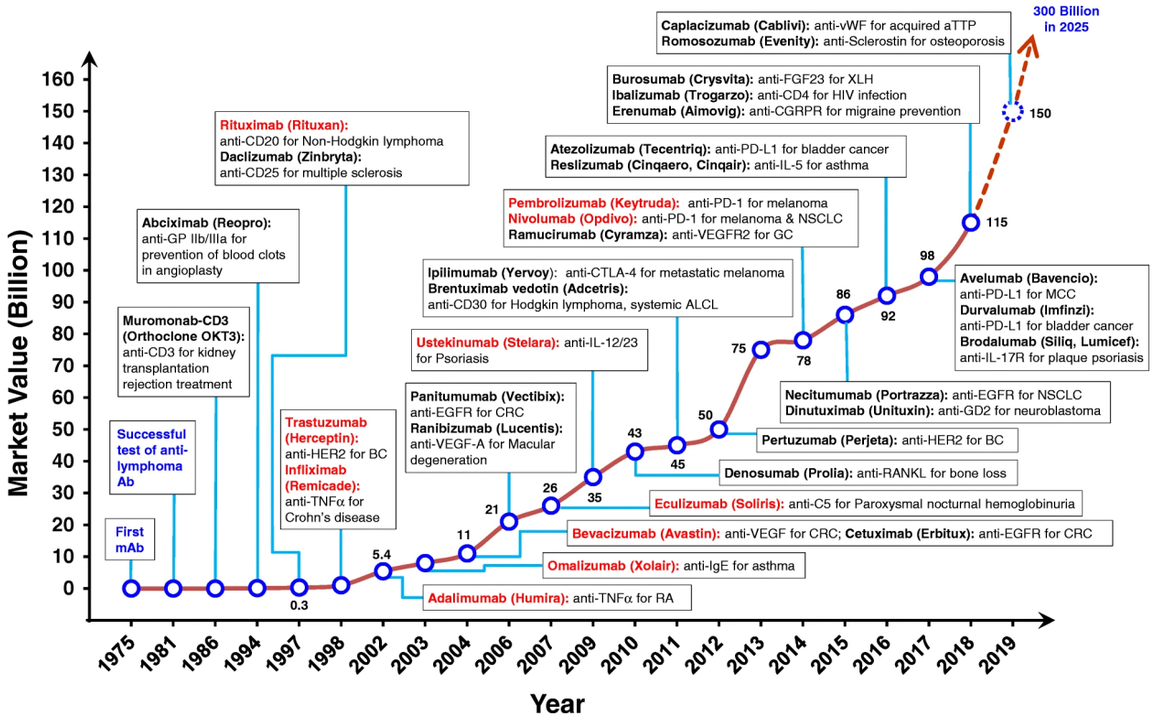

Figure 3: **The antibody drug market shows substantial growth and multiple successfully developed therapeutics since 1975.** The estimated market value of mAb therapeutics for each year. The best selling antibodies in the year 2018 are colored red. The estimated market value has approximately tripled within the past decade (R.-M. Lu et al. 2020).

Federal Food and Drug Administration (FDA) for clinical use (R.-M. Lu et al. 2020; Kaplon and Reichert 2019). The success of various antibody products and the market growth highlight the importance of antibody products for clinical use (Fig. 3).

Advances in sequencing technologies enable the creation of large repertoires containing up to several hundred million unique sequences from one or more donors. The function of antibodies involves specific binding to pathogen specific antigens and immune regulatory roles. The six hypervariable loops, which are usually referred to as the Complementary Determining Region (CDR), determines the specific binding of an antibody to its antigen(s). The antibody maturation mechanism in B-cells enables large sequence diversity of the CDR by germline-gene recombination and somatic hyper-mutation (SHM) of the Fv. This leads to a repertoire diversity of at least 10^{13} antibody sequences capable of binding to a large variety of antigens and pathogens (Wooden and Koff 2018). Antibody repertoires allow the systematic assessment sequence motifs which may be indicators for antigen-specific, differentiated variable regions. For example, the large-scale comparison of Fv repertoires from different donors gave insight into the extreme variability of sequences that are for the most part specific for one person ("private") with little sequence overlap between individuals (Soto, Bombardi, et al. 2019).

NGS repertoire research can be invaluable for antibody discovery and immunologic research, however mAbs for clinical use must fulfill additional requirements. First, the mAb as an industrial product must be capable of being produced in unphysiologically high titers (Mathias et al. 2020). Second, the mAb must exhibit broad acceptance by the human immune system to avoid adverse effects and/or compromised efficacy (Ducancel and Muller 2012). The fact that many Ab repertoires do not distinguish between Ig receptors secreted by the B-Cell and Ig which are not secreted, and

therefore rendered inactive, does aggravate this challenge (Soto, Bombardi, et al. 2019; B. Briney et al. 2019; DeWitt et al. 2016).

This dissertation describes novel tools to assess and improve the qualification of an Ab for clinical use, like the design of more human-like antibodies by informing the structural design with immunome repertoire statistics.

5.4 Manufacturability and Immunogenicity of antibodies

Therapeutic antibodies and vaccines are a class of mAb which are tight and specific binders, inhibitors, or steric blockers. Engineering of non-natural Ab is hereby standard practice to create biologicals that meet these requirements (Spiess, Zhai, and Paul J. Carter 2015; Jost and Andreas Plückthun 2014). Antibodies do not necessarily experience evolutionary pressure for high yields in mammalian cell lines and often show low product titer (Johari et al. 2015). These difficult to express (DTE) Ab can potentially halt product development in late stages due to low product titers (Pybus, Dean, et al. 2014). The developability of an product depends on the ability to recognize and work around DTE candidates. In this dissertation modern Deep Learning (DL) methods are applied to predict the expressability of antibodies. In combination with computational structural design, biologicals are re-engineered to improve the predicted expressability while retaining its biophysical properties.

Additional risk assessment of early-stage mAb biologics development includes screening for sequence liabilities that r compromise the developability and manufacturability (Jarasch et al. 2015). Clinical-stage mAb are exposed to heat and different pH values that can occur during the manufacturing to evaluate performance during long-term storage and assess the risk of modification of the biophysical properties. The CDR is especially susceptible to chemical modification (CM), which includes deamidation and isomerization processes that can disrupt the binding mode to the antigen (X. Lu et al. 2019). Over 200 ptm are known and can affect stability, and efficacy of the Ab product (Amann et al. 2019). Glycosylation in the constant region (Fc) for example is a PTM that plays a role in mitogenicity (Bolt et al. 1993), and therapeutic efficacy (Mimura et al. 2018; Chen et al. 2017) by regulating the antibody's immunobiologic downstream effects. Simultaneously, N-linked glycosylation of the Fv can sterically prohibit antigen binding, or influence immunogenicity (Waldmann 2019).

The immunogenic response to antibody products can be caused by CM, PTM, or aggregation results in an anti-drug antibody (ADA) response. ADA includes human anti-mouse antibody (HAMA) (Schroff et al. 1985), or human anti-chimeric antibody (HACA) (Afif et al. 2010) upon artificial engineering can impact patient safety as well as pharmacokinetic properties and ultimately limit drug efficacy. Surprisingly even fully-human Ab can result in the human anti-human antibody (HAHA) response (Nechansky 2010a). The natural immune response against engineered antibodies is an effect of the sequence being very dissimilar to sequences generated by the adaptive human immune response. Consequently humanization techniques were developed (Tsurushita, Hinton, and Kumar 2005) that has reduced the immunogenicity of engineered Abs (Hwang and Foote

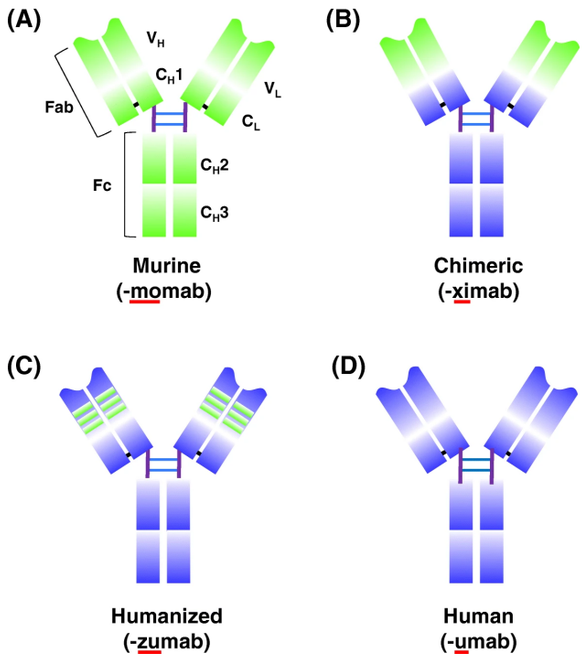

Figure 4: **Antibody humanization from fully antibodies (green) to fully human antibodies (blue).** Fully murine Ab (a), murine Fv of chimeric Ab (b), humanized Ab with murine CDR loops (c), and fully human mAb (d). mAbs are often annotated with the International Nonproprietary Name (INN) nomenclature (-momab, -ximab, -zumab, -and umab) which uses suffixes to roughly indicate the type of humanization effort.

2005). One of the more widespread humanization technique for mAb products is CDR grafting, where the murine Fv is transplanted onto human framework sequences (Maloney et al. 1997), or only transplanting murine CDR-loops while keeping the Fv framework regions human (Queen et al. 1989). Until 2017, the different degrees of human-likeness was captured by International Nonproprietary Name (INN) naming schema which indicates the level of humanization of an mAb. For example Adalimumab indicates a human antibody (Fig. 4). However, the naming system has undergone repeated reviews and its details have been changed since 2017 (Mayrhofer and Kunert 2021).

Sequence liability screening for PTM, CM, and aggregation therefore increases the likelihood of successful development of an mAb into a product that is manufacturable and has been adopted in product development pipelines (Y. Xu et al. 2013). The variety of factors that influence manufacturability, developability, and immunogenicity requires different methodological approaches.

This dissertation addresses these factors by statistically measuring human-likeness for the first time by assessing nucleotide statistics from complete immunome repertoires. A Deep Learning (DL) based expressability method was developed to predict if an antibody can be expressed and to re-engineer antibodies for increased expression rates.

5.5 Computational Approaches for antibody design

5.5.1 Rosetta protein design

The Rosetta protein design software package (Leaver-Fay, Tyka, et al. 2011) employs the Monte Carlo (MC) simulated-annealing (Xiangqian Hu, Beratan, and W. Yang 2009) algorithm for heuristic sampling of the sequence as well as conformational space for a protein. The fundamental concept behind Rosetta's protein-design algorithm is the *packer*. The packer builds new amino acid side-chains onto a protein scaffold by evaluating a set of rotamers at each position (Ponder and Richards 1987). The large degree of freedom within the protein renders this problem computationally highly expensive (NP-hard) and can not be solved be enumerating all solutions exhaustively (Pierce and

Winfree 2002). Instead, MC simulated annealing approach attempts to find sub-optimal to optimal sequence spaces (B. Kuhlman and D. Baker 2000). The semi-random walk over the sequence and conformational space is evaluated by Rosetta using a scoring function until the solution converges to a (local) minimum. Thus, the non-deterministic behavior of this design approach demands the generation of a variety of decoys to be inspected and filtered by the user with biophysical expert knowledge (Kufareva and Abagyan 2012).

The Rosetta scoring function has historically derived statistical potentials (Simons, Kooperberg, et al. 1997) that describe residue-pair interactions from the PDB (Berman et al. 2000). This early version of the scoring function was purely knowledge-based and did not handle amino-acid side-chain conformations explicitly. Improvements to the scoring function were made by adding physics-based potentials, like van der Waals interactions, or hydrogen bonding terms (Simons, Ruczinski, et al. 1999). The addition of rotamer libraries, a Lennard-Jones solvation model (Neria, Fischer, and Karplus 1996), and electrostatic considerations for hydrogen bonds (Kortemme, Morozov, and David Baker 2003) facilitated the first all-atom energy function (B. Kuhlman and D. Baker 2000). The modern Rosetta Energy Function 2015 (REF15) computes the free energy of a bio-molecule's conformation as a linear combination of its weighted individual terms (Alford et al. 2017) (Eq. 1). The Rosetta score ΔE_{total} is the sum of its individual physics and knowledge-based potentials (E_i) as functions of degree of freedom (Θ) and residue types (aa). The contribution of each term to the final score is carefully fine-tuned by a sets of weights w_i, allowing to design native-like backbone torsion angles (Renfrew, Butterfoss, and Brian Kuhlman 2008).

$$\Delta E_{total} = \sum_i w_i Ei(\Theta, aa_i) \tag{1}$$

REF15 incorporates a set of 19 weighted energy terms, one of which allows for sequence design. The Rosetta amino acid reference energies ΔG_i^{ref} facilitate sequence design on a protein conformation. ΔG_i^{ref} was optimized empirically to maximize the native sequence recovery and allows for estimation of the free energy difference between the folded and unfolded state. Thus, reference energies help to estimate for the energetic change of a mutation (Alford et al. 2017; Jain, Cerutti, and McCammon 2006).

Rosetta is designed to be easily extended by knowledge based potentials and has experimentally been used to combine the primarily thermostabilizing scoring function potentials with additional weighted terms ('constraints'). Adding experimental data constraints to the energy function, for example, further improved *de novo* structure prediction with paramagnetic constraints (Kuenze et al. 2019), electron-electron resonance spectroscopy decay traces (Del Alamo et al. 2020), or information of co-evolving residue pairs (Ovchinnikov, D. E. Kim, et al. 2016).

In this dissertation, novel scoring terms were developed to demonstrate that a) the human-likeness of antibodies can be improved by adding immunome repertoire sequence constraints and b) co-evolutionary information can be used for conservative protein design in order to retain protein function, and c) a Deep Learning (DL) based expression prediction can guide the mutational space towards a greater likelihood of Ab expressability.

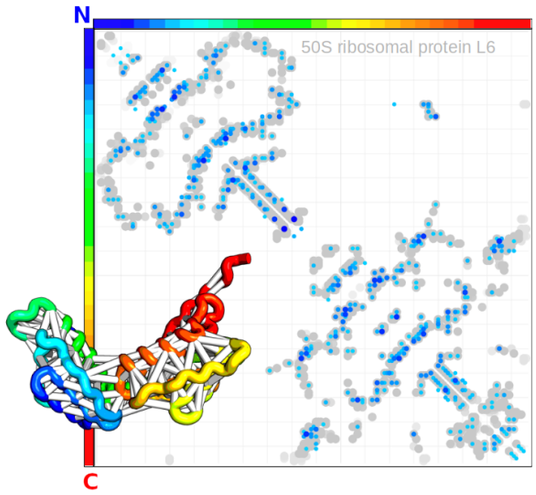

Figure 5: **Co-evolution of protein residue pairs reflect their spacial contacts and ultimately the 3D-structure.** Predicted contacts (blue), X-ray contacts (gray). Co-evolution of residue pairs (couplings) can be inferred from homologous sequences. Couplings resemble a network of evolutionary restraints across the protein that arise from structure, function, and protein dynamics. In the case of antibodies, coupling analysis therefore may be able to inform about antibody specificity and reveal binding patterns specific for an antigen and may ultimately support antibody discovery from large immunome repertoires (Figure by Sergey Ovchinnikov)

5.6 The role of residue-pair co-evolution in protein design

The protein sequence of a protein arises from stability, structural, functional, constraints. Evolutionary related, 'homologous' proteins accumulate sequence patterns specific for protein family and function. Tools have been developed that allow for extracting co-evolving residue pairs (couplings) that are characteristic for a set of homologous sequences (Morcos et al. 2011; Marks, Colwell, et al. 2011; Ekeberg et al. 2013; D. T. Jones et al. 2012; Kamisetty, Ovchinnikov, and David Baker 2013; Ovchinnikov, Kamisetty, and David Baker 2014). Methods that extract couplings decompose the chaining of co-evolving residues (residue A evolves with B, B evolves with C and thus, A evolves with C), resulting in pairwise terms that do not have to be in direct contact with each other. The majority of couplings however ($> 91\%$) are within 5-15 Å in at least one homologous structure and thus can be considered in physical contact with each other (Anishchenko et al. 2017). Consequently, the incorporation of co-evolutionary information into Rosetta as additional scoring term during *de novo* structure prediction has led to a substantial improvement of the computational models (Ovchinnikov, D. E. Kim, et al. 2016). Figure 5 visualizes the contact map of the 50S ribosomal subunit as a matrix, where each cell represents the correlation strength between two residues within the protein. The inferred couplings of the 50S ribosomal subunit (blue) match closely with the residue-pair distances of its corresponding X-ray structure.

Non-local couplings have been proposed to be of phylogenetic origin (Wollenberg and Atchley 2000), results of codon usage (Jacob, Unger, and Horovitz 2015), or allosteric interaction networks (Süel et al. 2003). In this dissertation we demonstrate, that non-local couplings can inform the Rosetta design process in favor of function and protein interaction that may support the design of highly specific antibodies without the requirement of fully studying the binding mode and antigen. The hypothesis of leveraging couplings for protein function and dynamics is supported by previous studies that used co-evolutionary analysis for protein-protein complex prediction (Burger and Nimwegen 2008; Hopf et al. 2014; Ovchinnikov, Kamisetty, and David Baker 2014), interaction partners (Bitbol et al. 2016), and modeling of conformational changes (Dago et al. 2012; Schug et al. 2009).

In this dissertation, we will demonstrate that co-evolutionary analysis can be applied to structural design with Rosetta to conserve the functional knowledge of a protein without its explicit

knowledge. Immune repertoires yet can not be systematically be functionally annotated and discovering functional antibodies remains challenging (DeWitt et al. 2016; Soto, Bombardi, et al. 2019; B. Briney et al. 2019). Immunome repertoire analysis may benefit from elaborate co-evolutionary analysis of antibody lineages that share functionally relevant sequence patterns. As a first step into this direction, we extract sequence patterns from large repertoires by applying an efficient clustering technique capable of processing billions of sequences and demonstrate that antibodies designed with sequence constraints of appropriate clusters increase their similarity to the repertoire. This is an indication that sequence sub-populations can be extracted, which support a certain antibody structure and therefore likely function and sequence patterns.

5.6.1 Rosetta methods for antibody design

Rosetta has been successfully applied on a variety of biological design questions, including protein design (Leaver-Fay, Jacak, et al. 2011), *de novo* protein folding (Brian Kuhlman et al. 2003; Rohl et al. 2004), peptide design and docking (Raveh et al. 2011), enzyme design (Richter et al. 2011), and small-molecule docking (Nguyen et al. 2013). A number of protocols are specifically tailored towards antibody design. To place the methods developed for this dissertation in the context of existing approaches, Rosetta immunogen design protocols that are applicable on human antibodies are described here.

Antibody Structure Prediction. In protein structure prediction, two major approaches are used: (1) *de novo* folding in the absence of a structural reference or template and (2) comparative modeling, which takes advantage of the availability of a structurally similar template to build a target model (Brian J. Bender et al. 2016). Given the large number of experimental antibody structures deposited in the PDB and the conserved immunoglobulin (Ig) fold, the large number of homology templates provides little to no need for de novo folding of the complete Fv domain. This makes antibodies ideal targets for comparative modeling approaches. However, the true challenge of antibody structure prediction lies in the correct orientation and fold of the CDRs, as all further scientific questions concerning antigen binding depend on the accuracy of the modeled loop conformations. Excluding HCDR3, five of the six loops usually fall into canonical clusters as defined by North et al., which can greatly simplify structure prediction (North, Lehmann, and Dunbrack 2011; Adolf-Bryfogle, Q. Xu, et al. 2015). Here, we will review three available protocols for antibody structure prediction from sequence in Rosetta: RosettaAntibody, AbPredict, and RosettaCM.

The RosettaAntibody application uses a three-step protocol for modeling the variable domain from sequence (compare Figure 6A): (1) template selection for the framework and the five canonical loops, (2) grafting of selected templates into a preliminary model, and (3) HCDR3 de novo loop modeling while simultaneously optimizing the VH–VL interface orientation (Weitzner, Jeliazkov, et al. 2017; Weitzner, Kuroda, et al. 2014; Sivasubramanian, Sircar, et al. 2009). For template selection, a BLAST sequence search matches the parsed sequence to a modified copy of the PyIgClassify database provided as part of Rosetta to assign both the Fv template and CDR conformations. This assignment can be checked with the identify_cdr_clusters application in

Rosetta such that any mismatches or other poor assignments within the template selection can be manually modified (Weitzner, Jeliazkov, et al. 2017; Sivasubramanian, Sircar, et al. 2009). As a next step, the initial VH–VL orientation is diversified by sampling VH–VL orientations from the BLAST list based on light–heavy orientational coordinates (LHOC), a metric that combines the VL and VH opening angles, the packing angle between the VH and VL domains, and the interdomain distance (Marze, Lyskov, and Gray 2016). Somatic hypermutation at the interface results in multiple angles between VL and VH even from sequences derived from the same germline genes such that a small difference in VH–VL distance and orientation may result in a drastic change in the CDR placement. This modulation of chain interface relationships has been investigated recently by Cisneros et al., who found VH–VL interface residues were reverted to the germline sequence, which resulted in significant loss of affinity, and indicated that the rigidification of the VH–VL interface, which will determine its orientation, is a major driver for affinity maturation (Cisneros et al. 2019). RosettaAntibody selects 10 different framework matches as starting structures for loop grafting. The selected template loops are superimposed on the framework based on two overlapping residues and optimized through a cycle of minimizations, random torsional sampling and cyclic coordinate descent (CCD) (Wang, Bradley, and David Baker 2007; Canutescu and Dunbrack 2003). Subsequently, HCDR3 conformations are modeled with the next-generation kinematic loop closure (KIC) algorithm in a low-resolution step (Stein and Kortemme 2013). The full model is then refined in full atom mode, with the VH–VL orientation reoptimized with rigid-body docking, (Gray et al. 2003) and the model is subsequently refined with an additional high-resolution step of next-generation KIC, residue side chain packing, and minimization (Weitzner, Jeliazkov, et al. 2017; Sivasubramanian, Sircar, et al. 2009).

Accurate modeling of the target antibody with RosettaAntibody relies on the availability of templates in the database that are highly similar in sequence to the antibody target. Most of the antibody structures determined so far are either human- or mouse-derived. Given the variability of the species-specific germline repertoire, such as the varying number of V genes or the different structural features represented, modeling of non-human or non-murine antibodies may be problematic due to the lack of appropriate templates. Therefore, when antibodies from other species are being modeled, it may be advisable to either curate a custom database or provide selected templates manually.

RosettaAntibody participated in both the 2011 and 2014 antibody modeling assessments (AMAs) (Almagro, Beavers, et al. 2011; Almagro, Teplyakov, et al. 2014). RosettaAntibody performed well overall on the basis of MolProbity scores and loop Cα RMSDs in AMA I (Almagro, Beavers, et al. 2011). In AMA II, RosettaAntibody was compared to six other software suites on a set of 11 unpublished antibody structures. It predicted 42 of 55 non-HCDR3 loops with an accuracy of better than 1 Åand generated the best HCDR3 model for 4 of 11 antibody structures from the other six competing methods (Weitzner, Kuroda, et al. 2014; Almagro, Teplyakov, et al. 2014). Subsequent analysis of the AMA II results identified some areas in the protocol that had weakened its performance: the lack of good loop templates, the inaccurate modeling of the HCDR3 due to

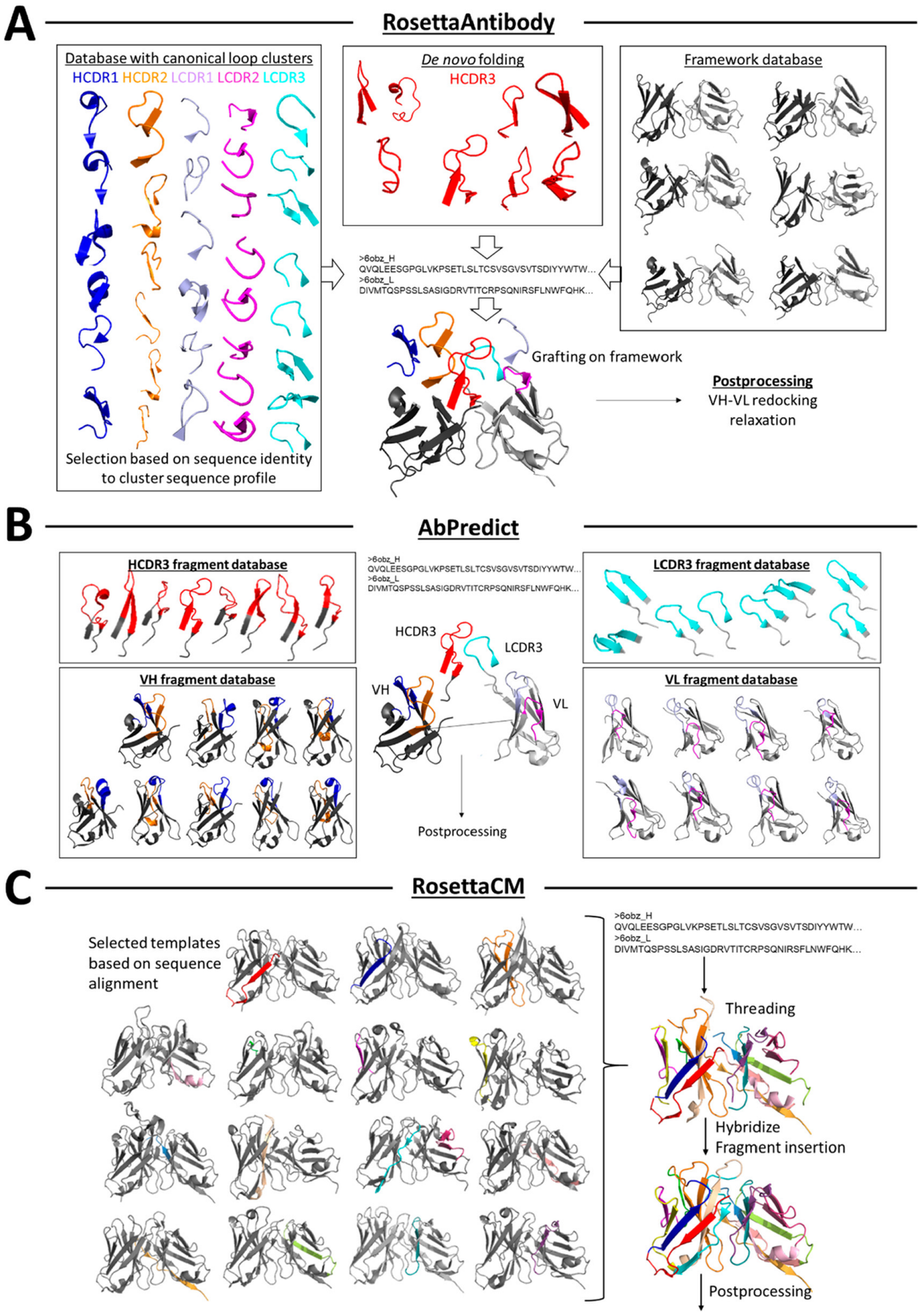

Figure 6: **Methods in Rosetta for antibody structure prediction.** (A) Schematic workflow of the RosettaAntibody application, in which HCDR1–2 and LCDR1–3 are modeled from templates in the loop database, and HCDR3 is de novo folded and grafted on a selected framework. (B) Schematic of the AbPredict protocol, which assembles an antibody from templates in four fragment databases, containing VL, LCDR3, VH, and HCDR3 templates. (C) Schematic overview of RosettaCM, which creates models by threading and hybridization of template structures based on user-provided sequence alignments.

limitations in the loop modeling protocols, and the wrong orientation of the VH–VL interfaces (Weitzner, Kuroda, et al. 2014). All of these issues were addressed in the present RosettaAntibody protocol, which samples a variety of VH–VL starting structures60 and incorporates next-generation KIC with HCDR3 conformational constraints (Weitzner, Jeliazkov, et al. 2017; Weitzner and Gray 2017). The problem of missing starting structures, which prevents accurate sampling of rare CDR loop conformations, can be further improved only when more structural data are deposited in the PDB that are continuously integrated into PyIgClassify and the RosettaAntibody database (Weitzner, Jeliazkov, et al. 2017; Adolf-Bryfogle, Q. Xu, et al. 2015).

A similar approach that combines antibody structural templates in another way has been implemented in the AbPredict protocol (compare Figure 6B) (Norn, G. Lapidoth, and Fleishman 2017). AbPredict selects low-energy combinations of backbone fragments derived from experimentally determined structures of antibodies in the PDB (Norn, G. Lapidoth, and Fleishman 2017). The template antibodies are segmented into four parts: (1) heavy chain CDR3, (2) light chain CDR3, and (3 and 4) heavy and light chain V gene regions each containing CDR1 and CDR2 and the framework as defined by the conserved core disulfide in the variable region. Additionally, AbPredict considers the rigid-body orientation between VL and VH, which is represented by the spatial distance of the disulfide's cysteine residues to L88 and H92 (Kabat numbering). Briefly, a database of randomly recombined backbone fragments and rigid-body orientations with the target sequence length is created. After the target sequence has been threaded on a random starting conformation, a Monte Carlo search that samples backbone fragments from the curated database, repacks side chains, and minimizes the whole structure is executed, which is output as scFv (Norn, G. Lapidoth, and Fleishman 2017; G. D. Lapidoth et al. 2015).

AbPredict has been benchmarked using the AMA II antibody set and compared to the methods presented therein. It performed in the upper third of all compared methods and showed beneficial performance in the prediction of the HCDR3 stem and the rigid-body orientation (Norn, G. Lapidoth, and Fleishman 2017).

Because AbPredict draws from an antibody template database provided as part of Rosetta, the representation of rare CDR loop length combinations is again a potential limitation, especially because AbPredict requires that target and template length match. A protocol capture is included within Rosetta.

Although antibody-tailored homology modeling protocols like RosettaAntibody can take advantage of knowledge-derived features of antibody structure, Rosetta's general multitemplate homology modeling protocol, RosettaCM can also be used (Figure 6C) (Song et al. 2013). RosettaCM might be advantageous in specific cases, especially if the antibody structure shows noncanonical structure elements such as unusual loop lengths or conformations, which would not be available in the antibody template databases. Using the DetailedControls option, RosettaCM can be employed to model only specific ranges of peptide sequences within a protein, for example, just one CDR. A similar approach was used to model G protein-coupled receptor loop regions with great accuracy (Brian Joseph Bender et al. 2019).

Overall, for most antibody structure prediction tasks, a good starting point is to employ RosettaAntibody as described in the tutorial section. Depending on specific features of the target antibody such as template availability or unusual loop length, models may need further refinement. In this case, the user can consider using only selected templates or perform a partial remodeling with RosettaCM. It is advisable to run smaller test runs with only a few output models in the beginning and monitor the outcome for reasonable modeling performance by looking at the total_score, a metric for predicted protein stability, which should be negative. In production runs, up to 10000 models should be created, depending on the complexity of the modeling task and the specific requirements of the protocol. Using metrics such as the total_score and Cα RMSD, the performance of the modeling run and the quality of the models can be assessed. This can also be used to compare the modeling performance of different protocols.

This work makes use of a RosettaCM (Song et al. 2013) based multi-template homology modelling protocol for antibodies (Kodali et al. 2021). Structural models were used to demonstrate that the predicted likelihood of expression can be increased via re-design in conjunction with a novel scoring term.

HCDR3 Structure Prediction. Structure prediction of HCDR3 has been challenging to date due to its high length and conformational diversity. Although half of HCDR3 loops are shorter than 16 residues, HCDR3 has been described to adopt loop sizes far longer, up to 32 residues, and even longer outliers have been described (IMGT nomenclature) (North, Lehmann, and Dunbrack 2011). The mean HCDR3 loop length has been determined to be 16 residues (B. S. Briney et al. 2012). Ultralong HCDR3 loops ($\geqslant$ 28 amino acids) have been described as necessary for the neutralization of disease states such as HIV or malaria, (Pancera et al. 2010; Henderson et al. 2007; McLellan et al. 2011) making the accurate modeling of long HCDR3 loops increasingly important for the structure prediction of therapeutically relevant antibodies.

Canonical loop clustering fails in the case of HCDR3 due to its high degree of diversity. PyIgClassify lists HCDR3 up to lengths of 5-9 residues, which are more restrained in their structural diversity, but structural clusters are not defined for longer HCDR3 lengths (Adolf-Bryfogle, Q. Xu, et al. 2015). However, the HCDR3 "torso" region, encompassing the first three (T1-T3) and the last four residues (T4-T7) of HCDR3 (based on the IMGT numbering scheme), can be classified into "kinked" ("bulged") or "extended" ("non-bulged") (Morea et al. 1998; Shirai, Kidera, and Nakamura 1996). The kinked conformation is predominant in antibodies, although structure prediction software rarely samples this conformation type (Finn et al. 2016; Weitzner and Gray 2017). In the past, sequence-based approaches have been employed to make a distinction between the kinked and extended conformation, relying on the presence of an arginine or lysine in the second position and an aspartic acid in the second to last position of the HCDR3 loop to classify an antibody as having a kinked conformation (North, Lehmann, and Dunbrack 2011; Morea et al. 1998). Although these amino acids are present in a large number of kinked conformations, they fail to cover the entirety of existing kinks (Finn et al. 2016). Therefore, other metrics for describing the kink conformation have been introduced and used as penalties during loop modeling. In an independent protocol to

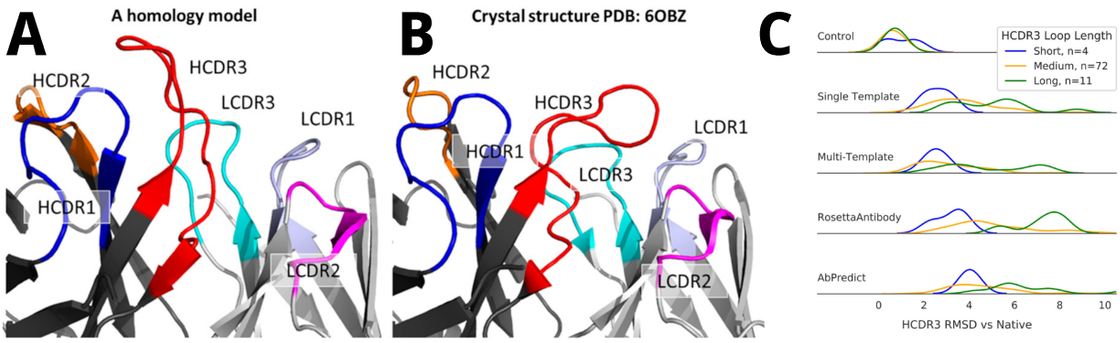

Figure 7: **Incorrect long HCDR3 loop structure prediction.** (A) Model of FluA-20 created with RosettaCM. HCDR1 and HCDR2 are predicted very well; however, the HCDR3 loop has an incorrect conformation that will impair future studies using this model. (B) Experimental structure of FluA-20 for comparison. (C) CDRH3 RMSD for 5 different Rosetta structure prediction methods. Short represents HCDR3 loop lengths of 1-6 amino acids (AA), medium of lengths 7 to 13 AA, and long of lengths 14 to 20 AA

more accurately model near-native HCDR3 loop conformations, Finn et al. described in greater detail the range of dihedral angles present in the torso region and identified a set of rules to guide kinked conformation sampling (Finn et al. 2016). The dihedral angle restraints are defined by the ψ angle at the sixth torso residue (T6) and are added as a Rosetta constraint file using a circular harmonic scoring function that penalizes the incorrect torso residues.

In RosettaAntibody, this limitation was overcome by integrating a structurally derived filter based on the kink definition by Shirai et al (Shirai, Kidera, and Nakamura 1996). so that bulged conformations are enriched (Weitzner, Kuroda, et al. 2014). To refine the definition of a "kinked" HCDR3 loop, Weitzner et al. integrated the conformation bias constraint to increase the likelihood of sampling native-like geometries of the last two C-terminal dihedral angles of HCDR3 plus the following framework residue's dihedral angle (as defined by the Chothia numbering scheme) (Weitzner and Gray 2017). Weitzner et al. hypothesized that the kink increases the degree of HCDR3 structural diversity by disrupting the propagation of β-strand pairing. Such a trend was also observed for proteins from other families where similar kinks occur in ligand recognition sites (Weitzner, Dunbrack, and Gray 2015).

Homology modeling of influenza hemagglutinin protein-specific human monoclonal antibody FluA-20 provides an illustrative example of challenging HCDR3 loop modeling (Figure 7A-B). While the best scoring homology model created with a RosettaCM protocol had accurate HCDR1 and HCDR2 predictions, the HCDR3 tip is flipped compared to the crystal structure. Structure prediction methods have difficulty with FluA-20 due to its 18-residue HCDR3 loop. The rules and protocols that Finn et al. and Weitzner et al. provide are a good starting point to improve native-like HCDR3 placement despite its noncanonical conformation. Accurate prediction of all CDR loops, especially HCDR3, from an antibody modeling protocol is paramount in obtaining biologically relevant results in downstream protocols, such as antibody–antigen docking.

The RosettaCM based homology modeling protocol used in this dissertation (Kodali et al. 2021) outperformes the single-template RosettaCM protocol for long CDRH3 loops.

Antibody–Antigen Docking. The structural study of antibody–antigen complexes is crucial for the understanding of antibody–antigen interactions, guides optimization and design approaches

of both docking partners, and ultimately helps develop new antibody-based therapies. Prediction of antibody–antigen complexes with computational protein–protein docking is of particular interest in investigating antibody function, as high-resolution experimental models of antibody–antigen complexes are rare due to the difficulty of co-crystallization. While more and more antibody–antigen complexes are now becoming available through the use of cryo-EM, the experimental data may not fully support atomic-level accuracy in all regions.

In Rosetta, a general protocol called RosettaDock can be employed for rigid-body docking with full backbone flexibility of two interacting proteins (Chaudhury, Berrondo, et al. 2011; Chaudhury and Gray 2008; Gray et al. 2003). This protocol was reviewed previously by Bender et al (Brian J. Bender et al. 2016). and will be discussed only briefly here. A low-resolution docking step, where docking poses are identified by rigid-body movements about the surface of the binding partner(s) (namely rotation and translation moves), is followed by a high-resolution step in full atom mode with fine-grained docking moves and side chain optimization stages (Gray et al. 2003). RosettaDock requires as input a structure of both docking partners, optimally with a user-defined starting point. However, RosettaDock also can perform a global docking step to identify low-energy docking poses (Gray et al. 2003; Chaudhury, Berrondo, et al. 2011).

SnugDock is an antibody- and antigen-specific extension of the RosettaDock protocol that is especially useful for docking homology modeling-derived antibody structures. SnugDock incorporates antibody-specific moves to overcome limitations of homology model-based inaccuracy in rigid-body docking that were observed in docking challenges (Weitzner, Jeliazkov, et al. 2017; Sircar and Gray 2010). Specifically, SnugDock adds a refinement step for HCDR2 and HCDR3 loops after low-resolution docking, allowing for greater loop backbone sampling with small, shear, and CCD moves. During the high-resolution phase, explicit sampling of the rigid-body VH–VL orientation and HCDR2 and HCDR3 conformations is achieved by CDR minimization, and loop backbone perturbation accompanied by additional small, shear, or CCD moves. SnugDock also can be combined with EnsembleDock, providing a database of input models for a higher diversity of starting structures (Weitzner, Jeliazkov, et al. 2017; Sivasubramanian, Sircar, et al. 2009; Sircar and Gray 2010). SnugDock (together with EnsembleDock) has been benchmarked on a set of 11 antibody–antigen complex structures, resulting in four medium and seven acceptable ratings using the critical assessment of prediction of interactions (CAPRI) criteria (Sircar and Gray 2010). SnugDock performed significantly better than did the standard RosettaDock protocol. However, SnugDock can also overfit, closing voids and constructing unnaturally tight interfaces (Sircar and Gray 2010). A protocol capture for SnugDock has been published by Weitzner et al (Weitzner, Jeliazkov, et al. 2017).

Generally, a docking approach will greatly benefit from including experimentally obtained restraints, which can be used to limit the conformational space to relevant structures. Examples of such experimentally derived restraints are alanine or site-directed mutagenesis, hydrogen–deuterium exchange mass spectrometry (HDX) or also HDX-NMR, NMR chemical shift perturbations, low-resolution cryo-EM, and chemical cross-linking data (Sivasubramanian, Chao, et al. 2006; Thorn-

burg et al. 2013). In the presence of a low-resolution EM map, however, it can be very difficult to dock an antibody in the right orientation, and a combination of structural methods may be necessary to obtain a high-confidence antibody–antigen complex model (Thornburg et al. 2013). Both SnugDock and RosettaDock are compatible with a wide variety of general constraints and filters in Rosetta. The general performance of a docking attempt can be assessed by calculating the interface energy for the created models, and also the Cα RMSD, for example, to the best scoring model. In many cases, some kind of experimental or knowledge-derived restraints are available that can also guide model selection, either manually or using filters in Rosetta. As docking normally has a high number of degrees of freedom, it is advisable to sample a high number of models when performing production runs for thoroughly sampling the conformational landscape (e.g., 10000, depending on the complexity of the problem).

In this dissertation, co-crystal structures were used for antibody design. To estimate the effect of mutations on the binding affinity to the antigen, Rosetta was used to calculate the Rosetta interface energy normalized by the size of the binding interface.

Antibody Design. Where structure prediction seeks to identify the optimal three-dimensional protein fold for a particular one-dimensional amino acid sequence, protein design seeks to find potential amino acid sequences that can maintain at least one previously determined, stable three-dimensional protein structure. Therefore, in contrast to antibody structure prediction and docking, where an antibody of fixed sequence is considered, antibody design modifies the sequence of an antibody to improve antibody affinity, specificity, and breadth, guided by knowledge-based sampling strategies.

Single-State Design. Single-state design protocols focus on the optimization of the binding affinity of a single antibody to a specific antigen. Such an approach can be used either to improve an already existing interaction or to create a new interaction for a nonbinding antibody–antigen pair. This refinement of an antibody sequence can be seen as a computational analogy to the natural affinity maturation process (Willis, B. S. Briney, et al. 2013). Somatic hypermutation introduces changes in sequence in the highly variable CDR regions during clonal expansion, leading to a high adaption to the presented antigen and to the expression of the tightest binder in a plasma cell. Rosetta on the contrary samples random mutations, using its energy function and Monte Carlo sampling to differentiate between beneficial and destabilizing mutations. While such a design process can proceed naïvely, naturally occurring patterns can be used as knowledge-based restraints to restrict the sequence search space.

Sequence design in the presence of an antigen can be performed by a very basic design algorithm in Rosetta, focusing the design to amino acids within the antibody–antigen interface. An example for this procedure is given in Bender et al (Brian J. Bender et al. 2016). First, a Python script is used to identify residues that are within a distance of specified residues that define the antibody–antigen protein interface. Subsequently, these interface residues are listed in a so-called "resfile", or a space-delimited file that designates designable residues, labeled by their residue number and chain identification, and to what entities, e.g., amino acid side chains, each residue may

be designed. In essence, the resfile controls which residue side chains can be mutated through design, repositioned through repacking, or kept rigid during design. Because interface design includes more than one protein, it is important to consider which side of the interface should be "mutated"; typically, it is desired to optimize the binding interface of the antibody through design while maintaining the antigen-binding interface. Therefore, it is most common to specify the residues within the antibody's interface as designable residues, while the antigen interface residues are limited to repacking to accommodate amino acid changes in the interface (Figure 8A).

The Rosetta design protocol optimizes the sequence on the basis of the overall energy of the complex, including the internal energy of the antibody and antigen, rather than the binding energy specifically. The resulting binding energy can be evaluated afterward by using InterfaceAnalyzer. Ideally, the binding affinity increases or decreases in value, while the overall energy (as a measurement for stability) does remain relatively constant. These criteria provide an initial filter to select models for further evaluation. More rigorous analysis, however, should evaluate each proposed mutation independently for its contribution to the total energy and binding energy in relation to the native model. A notable application using a similar protocol and analysis was the redesign of PG9, a human monoclonal antibody targeting the HIV envelope glycoprotein, where a RosettaDesign variant displayed increased potency and neutralization breadth (Willis, Sapparapu, et al. 2015).

This method is generally applicable to protein–protein interactions, and as such, it does not use any information about the natural sequence profiles for antibodies. Furthermore, its ability to sample backbone conformations is limited, which in turn limits accurate prediction of residues critical for forming antibody–antigen interaction. To circumvent such a limitation, it may be advisable to run the protocol on an ensemble of pregenerated starting conformations, or to integrate a backrub step, (Smith and Kortemme 2008) which will introduce greater backbone conformational flexibility.

Multi-state design (MSD) is a popular approach to inform Rosetta about protein flexibility. In this dissertation, we benchmarked our design approach that incorporates co-evolutionary information against MSD since both, ensemble of structures and evolutionary information allows the design of structures that favor protein flexibility. We show, that design using co-evolutionary information is capable of producing more natural protein sequences and retains residues that have been discovered to be functionally relevant in previous studies.

RosettaAntibodyDesign (RAbD). RosettaAntibodyDesign (RAbD) is capable of both de novo antibody design from a nonbinding antibody and also affinity maturation of an already existing antibody. It classifies the antibody into regions, including framework, the five canonical loops, and the HCDR3 loop, similar to the methodology in RosettaAntibody. Additionally, it can also redesign the DE loop, or H/LCDR4, as reported by Lehmann et al. for anti-EGFR scFv antibodies (Lehmann et al. 2015). RAbD starts from an assembled antibody–antigen complex and allows for both sequence and graft design based on the canonical clusters described by North et al.: (North, Lehmann, and Dunbrack 2011) GraftDesign exchanges a whole CDR for another from the canonical cluster database, and SequenceDesign optimizes the sequence on the basis of the canonical cluster

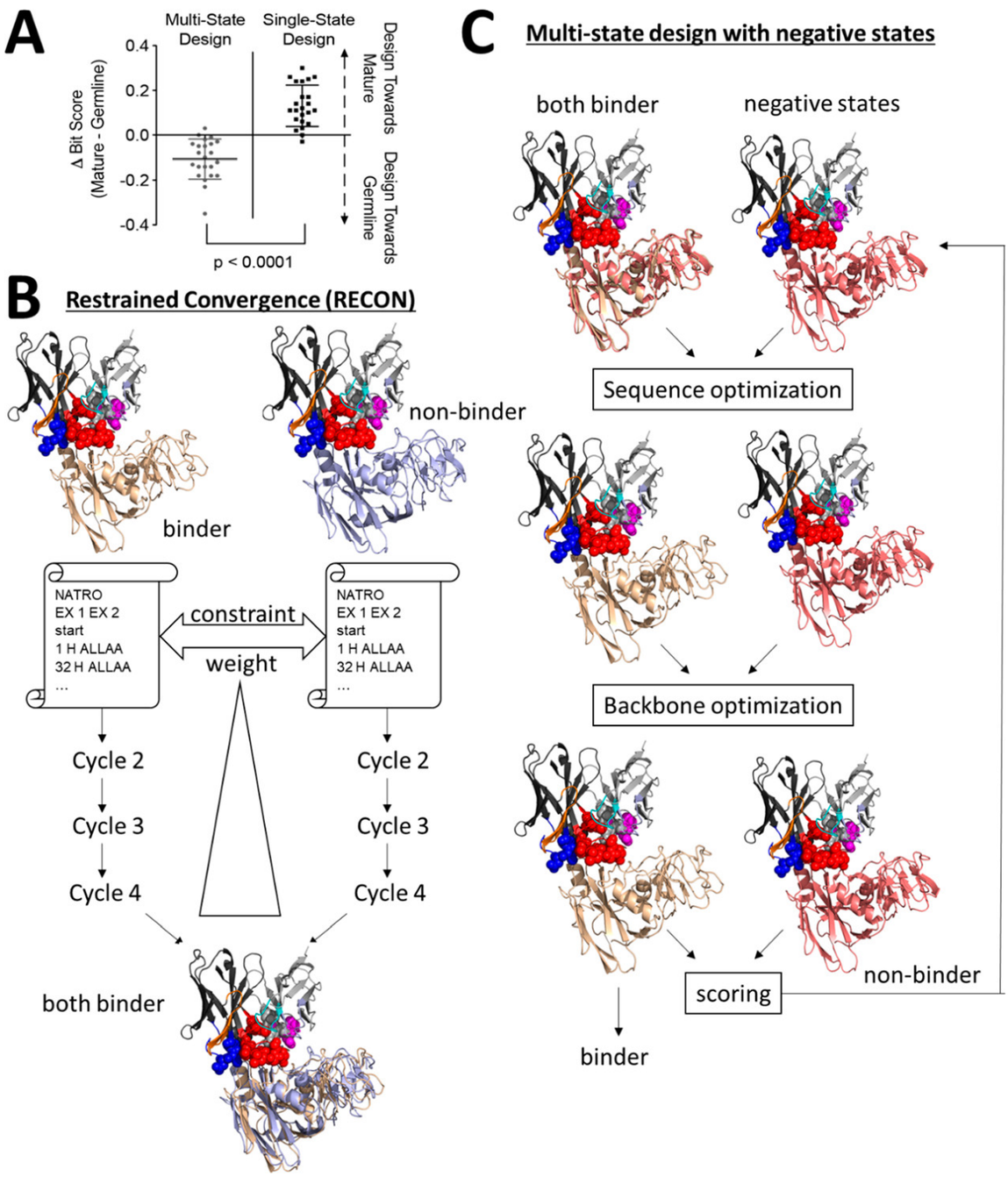

Figure 8: **Overview of multistate design protocols in Rosetta.** (A) Multistate design reverts the antibody sequence back to the germline sequence, while single-state design approximates affinity maturation (figure reproduced under CC-BY from ref (101)). (B) The RECON protocol (REstrained CONvergance) expedites discovery of states that bind multiple targets faster than traditional MSD algorithms because of its independent search of sequence space. (C) Design with negative states performs selectivity design against binders.

sequence profiles. The protocol is highly tunable by using a CDR instruction file, which allows users to include and exclude clusters, loop length, or PDB entries on the basis of the user's preferences. An example for this can be found in the tutorial section in the Supporting Information. Briefly, RAbD consists of an outer loop, which performs the graft design if enabled, and then passes the structure to an inner loop of sequence design, side chain repacking, CDR minimization, and optional integrated docking with epitope and paratope constraints. The structure is energy-minimized through the use of cluster-based CDR dihedral constraints and uses the Metropolis Monte Carlo criterion in the inner and outer loop for optimization. The default cycle number is set to 25 outer loops and one inner loop. The RAbD Metropolis Monte Carlo criterion can be set to the total energy (the protein stability score) or can be set to look specifically at the interface energy (corresponding to the computational binding affinity) using the integrated InterfaceAnalyzer methodology, as described above (Adolf-Bryfogle, Kalyuzhniy, et al. 2018).

RAbD therefore samples through all experimentally observed antibody conformations of different lengths and their corresponding sequence and structure space, allowing the design of loops with different lengths if desired. The protocol was benchmarked on a set of about 60 antibody–antigen complex structures and tested in two experimental antibody design cases, where it improved binding affinities for both antibodies.

AbDesign. AbDesign relies on backbone fragment recombination from experimental structures of antibodies deposited in the PDB, mimicking V(D)J recombination and allowing more native-like packing between the heavy and light chain than other antibody design protocols (G. D. Lapidoth et al. 2015; Baran et al. 2017). In short, AbDesign first predicts candidate apo structures of an antibody and, then following antibody docking, optimizes the antibody-binding interface against the target antigen. Like AbPredict, each heavy and light chain is segmented into one segment containing the CDR1, CDR2, and framework region (resembling the part of the protein encoded by the V gene) and another containing CDR3. Conformational representatives of each of the four segments are selected from precomputed Rosetta databases containing backbone segment torsion and sequence profiles. The selected segments are inserted, or grafted, onto the template scaffold by being subjected to CCD moves (Canutescu and Dunbrack 2003) using dihedral and coordinate constraints. Afterward, the individual antibody segments are scored against the original segment, and if the difference is <1 Åacross all segments, the predicted antibody model is kept for design. In addition, the antibody sequence is optimized on the basis of conformation-dependent position-specific scoring matrices (PSSMs) for each segment cluster, thereby combining knowledge-based sequence space with backbone plasticity. Following sequence and backbone optimization, the pool of generated models is docked onto the target antigen using low-resolution docking. This is followed by a last design step. It is important to note that the sequence constraint is less strict for residues in the antigen interface, thereby encouraging a high degree of sequence variability for the optimization of the binding energy upon design, whereas the more conserved framework regions have stricter constraints, to encourage selection of naturally occurring sequences to maintain stability. Using a many-valued fuzzy-logic approach (Warszawski, Netzer, et al. 2014) in the final selection, antibodies

are chosen on the basis of stability (total energy), binding energy, buried surface area, packing between the heavy and light chain, (Sheffler and David Baker 2009) and shape complementarity (Lawrence and Colman 1993) between the antibody and antigen (G. D. Lapidoth et al. 2015).

AbDesign was benchmarked on a set of nine antibody–antigen complexes and evaluated on sequence recapitulation and interface side chain rigidity (G. D. Lapidoth et al. 2015). Furthermore, AbDesign was used for two de novo designs of scFv in combination with yeast display and error-prone PCR in five consecutive cycles over which the protocol was adapted to its final version. Major modifications were necessary, however, because the first designs expressed poorly, which was attributed to cavities in designs, unpaired buried charges, and the loss of long-range hydrogen bonds (Baran et al. 2017). Even with the two successfully predicted scFv models, crystallization of the models as Fabs (notably without antigen) revealed structural differences between the experimentally determined models and the AbDesign models, especially in HCDR3 and HCDR1 (Baran et al. 2017).

RAbD is one of the most feature-rich Rosetta antibody design protocols. In contrast to AbDesign, it allows the integration of sequence profiles. These profiles are sourced from a few thousand clustered PDB structures. In this dissertation, a new protocol to incorporate sequence statistics derived from complete immunome repertoires was developed that ultimately will awill th design of the challenging CDRH3 region.

Balancing between Sampling and Stability. The protocols presented above represent multiple options to design antibodies in Rosetta. The optimal choice of protocol depends on the design task. The more changes are made to the native antibody that initially is expressible and capable of being crystallized, the less likely are the designs to be expressed and stable (Baran et al. 2017; Sevy, N. C. Wu, et al. 2019). Like in affinity maturation, however, it is often necessary during protein design to sample a broad sequence and conformation space to identify the optimal combination of antibody sequence and structure to achieve both high specificity and binding affinity for a target antigen. This can require sampling beyond energy barriers that confine the native antibody's sequence and structure space to a local energy minimum, and in such cases, protocols that provide a means for more extensive sampling may be superior to more conservative approaches that limit the sampling space to a local energy minimum. In general, if the goal is to improve the binding affinity of an antibody within an already determined antibody–antigen complex, it is generally advisable to begin with a more conservative approach. Otherwise, it is often a good idea to use more than one protocol and to compare results for convergence onto the same sampling space. Even after cross-checking multiple approaches, it may be necessary to alter the chosen protocol to account for problems like expressibility or solubility. However, to overcome energy maxima in the conformational landscape, it might be necessary to sample more thoroughly, and in these cases, protocols with more sampling can be superior compared to more conservative approaches. Upon comparison or establishment of protocols, a smaller size of models can be sampled and evaluated for chosen parameters, which could include the interface energy as metric for predicted binding affinity, but also sequence similarity, type of newly created interactions, or other knowledge-derived metrics depending on the complexity and the specific questions of the design task. The number of

models that should be created for a design task can vary quite heavily depending on the number of positions to design and the protocol used. Generally, more output models will be needed for less conservative approaches.

The Rosetta designs generated in the studies of this dissertation were compared to already published, similar Rosetta protocols or to unrestrained Rosetta design runs, without additional scoring terms (control). The control group acts as baseline the estimate the rate of improvement when a novel scoring term was added.

Multistate Design (MSD). While single-state design considers just a single antibody or antibody–antigen structure, MSD protocols provide a wide platform for addressing several types of higher-complexity design problems. Most commonly, MSD encompasses the design of one antibody in the presence of more than one antigen. The goal can be to optimize the breadth of the antibody to bind multiple antigens, find an antibody that can bind to multiple conformations of a single antigen, or to optimize the selectivity of the antibody through negative design against a subgroup of antigens.

For all three of these possibilities, protocols have been developed in Rosetta and used in the field of antibody design.

Broadly neutralizing antibodies (bnAbs) have proven to be a powerful therapeutic tool. A highly optimized antibody is at risk of losing its binding affinity when small changes in the antigen's amino acid composition occur, whereas bnAb maintains its ability to bind to antigens from multiple strains, subtypes, or even species. The bnAb therefore is more likely to provide protection for a longer period of time. Such breadth is normally mediated through limited but tight binding to conserved residues that are functionally less susceptible to antigenic drift.

One classical MSD task includes designing an antibody initially known to bind to a single antigen to optimize its sequence to form multiple novel binding interactions with one or more antigens. The Rosetta MSD design protocol using the REstrained CONvergence (RECON, 8B) algorithm was originally developed to perform such a task to increase antibody breadth by constraining the sampled sequence space to adopt multiple (binding) conformations (Sevy, Jacobs, et al. 2015; Sevy, N. C. Wu, et al. 2019). Broad antigen recognition, or polyspecificity, may be linked to germline antibody sequences; it has been hypothesized that naïve germline antibodies exhibit greater conformational flexibility, which enables polyspecificity (Babor and Kortemme 2009). Interestingly, using RECON MSD to design the sequence space of a single antibody when in complex with a set of antigens reverted an antibody's sequence back toward its germline gene sequence (Figure 8A). Conversely, using single-state design-introduced mutations will make the difference from the germline gene sequence greater (Willis, B. S. Briney, et al. 2013).

Design of polyspecificity requires that the antibody of interest be spatially aligned with all antigens for which a common binding motif should be found, which comprise the antibody's intended targets, and that a common antibody-binding interface be the subject of design. For RECON MSD, the antibody interface of interest is based on a known antibody–antigen complex structure, such that any novel binding interfaces are based on the superimposition of target antigens to the known

antibody–antigen complex. RECON MSD is novel with respect to other MSD protocols in that rather than treating design as a combinatorial problem, it reduces the design of a large conformation space by treating each structure, or state, included in the design as a separate design problem, thus making RECON MSD very efficient. More specifically, design sampling identifies the lowest-relative free energy sequence for each single conformation but will accept a redesigned sequence only if the sequence has the lowest average energy across all states. RECON MSD assumes that the native sequence is close to the sequence that is ideal for conformational flexibility or polyspecificity and encourages the selection by using a convergence restraint to favor the selection of native sequences. Convergence is further encouraged by using multiple rounds (typically four rounds) of design. To converge on a common sequence, a sequence similarity restraint is introduced. The restraint is kept small in early rounds of design to sample a broad sequence and conformational space specific to each antigen and ramped up in later rounds of design to find convergence over multiple antigens. In the case in which selection of a sequence does not converge for a designed position, the last step in the protocol forces a selection based on the lowest fitness over all sampled amino acids for nonconverging positions. In the end, this sequence convergence encouraged through restraints is hypothesized to find minima in the energy landscape more rapidly (Figure 8B). The independent sequence search allows trajectories to adopt sequences that are favorable in one state but might not be in another state, which in contrast to classic MSD algorithms prevents the exclusion of these intermediate states. Thus, the encouraged convergence bypasses high-energy states. RECON was benchmarked in comparison to the traditional Rosetta MSD, where it showed improved performance to recapitulate evolutionary sequence profiles, a metric chosen to represent polyspecificity (Sevy, Jacobs, et al. 2015). RECON was further refactored to run in parallel on separate processors using message passing interface (MPI) communication, which enables massive parallel design against a large number of antigens (Sevy, N. C. Wu, et al. 2019). It was applied to design broad influenza hemagglutinin H1 antibodies based on the C05–H3 complex structure (Ekiert et al. 2012) and could propose mutations that showed an enhanced breadth against additional virus strains, including a strain with a known escape mutation (Sevy, N. C. Wu, et al. 2019). In this work, criteria that yield greater success in design were identified. For example, a high drop of energy for some antigens, especially the antigen that is bound by the antibody in the original complex structure, indicates nonfavorable mutations (Sevy, N. C. Wu, et al. 2019). Mutations that establish new hydrogen bonds, relieve clashes with the antigen, or create more van der Waals interactions are favorable. To increase the sampling space, the protocol can be combined with backrub moves, which creates a backbone ensemble and enables the sampling of a larger sequence space (Sevy, Jacobs, et al. 2015). Generally, the evaluation criteria are similar to a single design task, however, considering only such amino acid changes that improve predicted binding affinity (e.g., interface energy) for all multistate design targets while not compromising protein stability (total_score). An example protocol for multistate design with the RECON design protocol can be found in the Supporting Information.

The BROAD (BReadth Optimization for Antibody Design) algorithm has been developed to

enhance MSD performance further than RECON MSD. The RECON protocol becomes computationally expensive when designing antibodies against large panels of antigens, or many different conformations of a protein. BROAD includes support-vector machines to classify antibody binders versus nonbinders and optimizes breadth through the use of integer linear programming. This method is very fast and can be applied to large sets of antigens (e.g., a large panel of different viral strains). The method has been tested computationally, but the protocol has not yet been applied to an experimental application (Sevy, Panda, et al. 2018).

Vaccine Design through Thermostabilization. A major challenge of vaccine design is the flexibility and instability of immunogenic proteins. For now, computational generation of novel epitope-presenting proteins, such as with the methods described above, requires several rounds of testing and optimization both computationally and experimentally. Another approach is to simply stabilize a protein of interest (Goldenzweig, Goldsmith, et al. 2016). In the latter, the needs of thermostabilizing a protein structure and maintaining its function were achieved through amino acid changes guided by information about the protein sequence's evolutionary diversity. The rationale is that evolution does not allow for destabilizing mutations as those would render the protein inactive. The protocol represents the evolutionary diversity with a PSSM, which it uses to sample possible mutations. The effect of a mutation on the stability of the protein and its interactions are evaluated by a $\Delta\Delta$G calculation in Rosetta. Stabilization is achieved by a combinatorial search of groups of amino acid changes that can have an additive effect on protein stability. This protocol was benchmarked on multiple proteins to predict known stabilizing mutations without choosing known destabilizing mutations (Goldenzweig, Goldsmith, et al. 2016). Additionally, the protocol was tested for thermostabilization of human acetylcholinesterase (hAChE), which is usually expressed in eukaryotic cells and could be obtained in large amounts in *Escherichia coli* expression. Of the five chosen designs that had 17–67 mutations in total, four maintained activity while having higher deactivation temperatures (Goldenzweig, Goldsmith, et al. 2016). The protocol is available as a Web server, called the Protein Repair One Stop Shop (PROSS, http://pross.weizmann.ac.il).

This approach also has been applied to a vaccine design project, namely, the thermostabilization of Plasmodium falciparum reticulocyte-binding protein homologue 5, a relevant target for malaria vaccine development. In total, 18 mutations were introduced and yielded a design that was expressed in *E. coli* and showed higher stability, while maintaining its immunogenicity. An experimentally determined structure proved that the design was very similar to the original protein (Campeotto et al. 2017).

In this study, a novel method was benchmarked that restraints the design with co-evolutionary information. It has been shown that the evolutionary design protocol outperforms classic sequence profiles (PSSM)s and is likely to improve stability and conserved function.

5.7 Difficult-to-express (DTE) antibodies

Historically, the foundation of monoclonal antibody production was made 1975 by the immortalization of B-lymphocytes (Köhler and Milstein 1975). Briefly, antibody-secreting B-cells extracted

from a model system (e.g. mouse) are fused with immortal myeloma cancer cells induced virally or chemically. The hybrid cell line (hybridoma) can be used to identify, characterize and produce monoclonal antibodies as a result of a murine immune response. The technology is still popular and has been used with recent successes in recent mAb discovery using murine hybridoma for diagnostics and research (Aguiar et al. 2016; Parray et al. 2020) and was awarded with the nobel prize in physiology and medicine of 1984 (Leavy 2016).

The method evolved after hybridoma instability issues and human anti-mouse antibody (HAMA) antibody responses in patients, reducing the antibody efficacy as therapeutic and inducing adverse effects (immunogenicity). During the 1990s chimeric antibodies were developed with 1) a human Fc region and mouse Fv regions to reduce the HAMA response and 2) humanization strategies were employed to remove T-cell epitopes in the Fv.

Humanization, also referred to as reshaping, complementary determining region (CDR)-grafting, veneering, resurfacing, specificity-determining residue (SDR)-transfer, or DeImmunization, include strategies to reduce the immunogenicity of antibodies of non-human origin. The design of the humanized antibody sequence is critical for reproducing the affinity, specificity, and function of the original molecule while minimizing HAMA responses elicited in patients. A natural strategy is to keep the engineered antibody human from the very beginning of the design phase which may circumvent the biggest challenges faced in late-stage humanization processes. In this dissertation, a method was developed that allows the structural affinity maturation of antibodies using a human germline gene restraints.

Recombinant cell lines large-scale for mAb expression include CHO, NS0, Sp2/0, HEK-293, and PER.C6. The vast majority of approximately 70% of presently industrially produced proteins is conducted in Chinese ovary hamster (CHO) cell lines (Jayapal et al. 2007). Modern fed-batch cultivation processes using CHO cell lines are able to produce monoclonal antibodies in the range of multiple grams per liter (Kunert and Reinhart 2016). Protein synthesis is mediated by a complex process that involves tightly regulated and balanced network of steps involving different cellular compartments (Alberts et al. 2017). For an antibody product to be expressed and secreted, the journey begins with ribosomal synthesis in the endoplasmatic reticulum (ER). The first regulatory lever for the expression rate is the nucleotide sequence codon usage itself. Codon usage is specific for the production system and directly affects the efficiency of messenger RNA transcription (Z. Zhou et al. 2016). Folding in the ER lumen is facilitated by specific protein chaperones that belong to the heat-shock protein family such as the 70kDa binding immunoglobulin protein (BIP), calnexin/calreticulin of the leptin protein family, and peptidyl-prolyl isomerases (Braakman and Hebert 2013; Ellgaard and Helenius 2003). Cystine form disulfide bridges between two residues add additional rigidity and support the proteins tertiary structure. Disulfide bridges are covalent bonds and formed by isomerases after the folding process (Appenzeller-Herzog 2011).

Correctly folded proteins are transported to the Golgi apparatus, a part of the intracellular vesicular transportation system, where post-translational modifications (PTM) occur before the protein is transported into vesicles for secretion. Misfolded proteins are degraded in the proteosome

as part of the ER-associated degradation pathway (Xudong Wu and Rapoport 2018).

With pitfalls on many levels of protein expression, starting from transcription, to folding, to vesicular transport and PTM, the complexity of the challenge to optimize protein expression requires substantial experimental data-collection. In case of a lack of an appropriate data source, a model cannot be created that describes protein expression in sufficient detail. In this dissertation, the Deep-Learning architecture Long short-term memory (LSTM) was employed to extract relevant sequence patterns that influence expressability by one or more unknown bio-physiological effects, automatically from a limited amount of data. An additional Rosetta energy term is then developed to support the design of antibodies with increased expression rates.

5.7.1 Engineering of human-like antibodies

Methods for detecting human-likeness in antibody amino acid sequences support the screening and engineering of antibodies with immunogenic effects, tend to reduce the efficacy of Abs in a clinical setting. The H-Score method to estimate human-likeness developed by Abhinadan et al. in 2007 was based on pairwise sequence identity calculations (Abhinandan and Martin 2007). The method evolved by replacing pairwise sequence calculations with Basic Local Alignment Search Tool (BLAST) databases. The resulting T20 score was also derived from a dataset of about 38,700 sequences (Gao et al. 2013). To take germline gene family specificity of immunogenic effects into account, the germline gene aware G-Score was developed (Thullier et al. 2010). Seeliger et al (Seeliger 2013). demonstrated the usefulness of a heuristic scoring function to increase human-likeness and reduce immunogenic effects. The heuristic scoring function is capable of suggesting mutations to reduce immunogenicity and increase human-likeness based on a pairwise probabilistic model.

The Human String Content (HSC) is an alternative method to decrease immunogenic effects by increasing the germline similarity to 9-mer fragments of germline genes in order to reduce the class II MHC binding affinity (Lazar et al. 2007). The HSC has successfully been combined with structure-based antibody design to produce humanized antibodies with high affinity (Choi et al. 2015). The methods H-Score, T20 and the heuristic scoring function have been developed from small amino acid sequence datasets of several thousand sequences. Recent advances in deep-learning methods enabled Wollacott et al. to precisely capture human-likeness of antibody sequences using a Long short-term memory (LSTM) model trained on 25,000 sequences (Wollacott et al. 2019). Human-likeness scores are usually derived from small datasets, and are primarily concerned with the question of how to separate human from non-human antibodies instead of developing a sequence model that explains how an Ab can emerge from a repertoire.

Computational assessment of Human-likeness has first been described as an alignment of several hundred amino-acid Ab sequences (Abhinandan and Martin 2007). The alignment human, or murine sequences allows for statistical assessment of the frequency of each amino acid type at each position and can be used as a distinct species specific antibody profile. This technique has been evolved to be more scalable on larger sequence sets (approximately 10,000) (Gao et al. 2013; Seeliger

2013).

The Rosetta Antibody design protocol (RAbD) (Adolf-Bryfogle, Kalyuzhniy, et al. 2018) allows for inclusion of sequence restraints from human antibody sequences available as structures deposited in the Protein Databank (PDB) (Berman et al. 2000), and from conformational loop clusters (Adolf-Bryfogle, Q. Xu, et al. 2015). The limitation of this approach is, that the number of available human antibodies ranges at the time of writing between 1,000 and 2,0000 unique antibodies. The sequence profiles developed in this dissertation are based on complete immunome repertoires and further are expanded using probabilistic modeling of an amino acid sequence space

6 Human-likeness of antibody biologics determined by back-translation and comparison with large antibody variable gene repertoires

This chapter has been published under (Schmitz et al., 2020).

6.1 Introduction

Antibodies (Abs) bind to epitopes on the surface of microbial pathogens like bacteria and viruses. Abs are produced by B lymphocytes that use genetic mechanisms to increase sequence diversity of the expressed repertoire. These genetic mechanisms include recombination of variable (V), diversity (D), and joining (J) gene segments as well as enzymatic modification and addition of non-templated (N) or palindromic (P) nucleotides in the V-D, D-J and V-J junction regions (Jung and Alt 2004). The variable domain of an antibody is encoded by the three genes (V, D, and J) for heavy chain sequences, and two genes (V, and J) for light chain sequences. The variable domain can further be divided into framework regions (FR) and complementarity determining regions (CDR). The introduction of somatic mutations in the variable domains occurs in recombined genes during the secondary immune responses (Jung, Giallourakis, et al. 2006). The resulting sequence space of the combined set of naïve and mature sequences of the V domain in an individual organism depends on general characteristics of the Ab genes for a species and on the prior experience of the individual including pathogen exposures. We previously determined the immunome (adaptive immunome receptor repertoire) comprising Ab sequences for three healthy human blood donors using very deep next-generation sequencing (NGS) (Soto, Bombardi, et al. 2019). The Ab sequences of this dataset either cover the full variable domain or start midway into the FR region.

The analysis of human Ab sequences usually comprises the partitioning into V, D, and J gene-encoded domains, and the determination of the FR and CDR as well as somatic mutations. Various computational tools are available to assign inferred genes and domains to portions of Ab sequences by making species-specific germline gene calls (Ye et al. 2013; Bolotin et al. 2015; Russ, Ho, and Longo 2015; Xihao Hu et al. 2018; Brochet, M.-P. Lefranc, and Giudicelli 2008; Gaëta et al. 2007). Germline genes also may vary in individuals and ethnic subgroups, potentially biasing the maturation process in ways that may be of clinical relevance (Brovkina et al. 2018). The increasing availability of large immunome datasets (Soto, Bombardi, et al. 2019; DeWitt et al. 2016; B. Briney et al. 2019; Corrie et al. 2018; Kovaltsuk et al. 2018) was leveraged to create a position- and gene-specific scoring matrix (PGSSM) for datasets in order to describe the human Ab sequence space. For this study we used the sequencing dataset from the Soto et al (Soto, Bombardi, et al. 2019). dataset composed of the antibody sequencing from the blood compartment of three healthy human donors. The PGSSMs were derived from this dataset and consisted of 326 million unique antibody sequences. The PGSSM was used to model the single nucleotide frequencies (SNFs) per position in the germline gene, allowing us the estimation of similarity of an Ab sequence to a given immunome repertoire collection. SNFs can arise from different sources such as: allelic differences, hypermutation, or sequencing errors. The method developed in this study attempts to capture

frequencies caused by hypermutations by grouping all SNFs to their respective germline gene. The size of immune repertoire dataset ensures that any errors that arise from sequencing are minimized.

Our PGSSMs are germline gene-specific (Sheng et al. 2017) for templated regions, and length-dependent for the heavy chain complementarity-determining region three (CDRH3). This approach allows us to model SNFs that exclude insertions, but include non-templated (N) and palindromic (P) nucleotide additions that bracket the CDR3. This feature enables us to derive the nucleotide sequence that maximizes the nucleotide frequencies in the PGSSM model so that the resulting nucleotide has a high human likeness. In this study, we attributed each optimized nucleotide sequence with a score for the variable (V) and joining (J) domain ($\mathrm{PGSSM_{VJ}}$) and characterized the properties of the $\mathrm{PGSSM_{VJ}}$. We show that the $\mathrm{PGSSM_{VJ}}$ represents a similarity measure between an amino acid sequence and a given immune repertoire. Thus, the $\mathrm{PGSSM_{VJ}}$ could in principle be used to engineer an antibody sequence to make it more human-like in the future (Olimpieri, Marcatili, and Tramontano 2015).

Methods for detecting human-likeness in antibody amino acid sequences support the screening and engineering of antibodies with immunogenic effects, which tend to reduce the efficacy of Abs in a clinical setting. The H-Score method to estimate human-likeness developed by Abhinadan et al. in 2007 was based on pairwise sequence identity calculations (Abhinandan and Martin 2007). The method evolved by replacing pairwise sequence calculations with Basic Local Alignment Search Tool (BLAST) databases. The resulting T20 score was also derived from a dataset of about 38,700 sequences (Gao et al. 2013). To take germline gene family specificity of immunogenic effects into account, the germline gene aware G-Score was developed (Thullier et al. 2010). Seeliger et al (Seeliger 2013). demonstrated the usefulness of a heuristic scoring function to increase human-likeness and reduce immunogenic effects. The heuristic scoring function is capable of suggesting mutations to reduce immunogenicity and increase human-likeness based on a pairwise probabilistic model.

The Human String Content (HSC) is an alternative method to decrease immunogenic effects by increasing the germline similarity to 9-mer fragments of germline genes in order to reduce the class II MHC binding affinity (Lazar et al. 2007). The HSC has successfully been combined with structure-based antibody design to produce humanized antibodies with high affinity (Choi et al. 2015). The methods H-Score, T20 and the heuristic scoring function have been developed from small amino acid sequence datasets of several thousand sequences. Recent advances of deep-learning methods enabled Wollacott et al. to precisely capture human-likeness of antibody sequences using a Long-Short-Term-Memory (LSTM) model trained on 25,000 sequences (Wollacott et al. 2019). Human likeness scores are usually derived from small datasets, and are primarily concerned with the question of how to separate human from non-human antibodies instead of developing a sequence model that explains how an Ab can emerge from a repertoire.

In this study, we developed the algorithm IgReconstruct, which draws conclusions about Ab human-likeness that are distinctly different from other methods. Firstly, our method is based on single nucleotide frequencies. Secondly, to estimate the similarity of a target Ab amino acid

sequence to a given repertoire, a germline gene rearrangement tailored to the nucleotide frequency observations made in the repertoire is generated. Thirdly, the target Ab amino acid sequence is back-translated to the nucleotide sequence to allow a fine-grained comparison with the observed immune repertoire nucleotide frequencies. IgReconstruct scales well with large repertoires consisting of hundreds of millions of sequences, and will be useful for computational antibody engineering.

6.2 Results

We calculated position- and gene-specific PGSSM matrices (Figure 11.1) from a publicly available human immunome repertoire of 326 million antibody Ab sequences (Soto, Bombardi, et al. 2019). The PGSSM matrices encode the observed single nucleotide frequencies in the repertoire. The PGSSM matrices were used to calculate the PGSSM_{VJ} score (Figure 9, Equation 2) for any given antibody sequence, which essentially represents the similarity of a given antibody sequence to the immunome repertoire. We then curated a set of in total of 181,355 GenBank (Benson et al. 2013) sequences from 20 different species (see Material and Methods for a sequence breakdown by species). To measure the performance of our PGSSM method with an independent dataset, we used the GenBank sequences and estimated the similarity to the human immunome repertoire of 326 million naturally occurring antibody Ab sequences.

Human Likeness was assessed by calculating the Z-Score of the PGSSM_{VJ} score (Equation 3), for which we used the distribution of PGSSM_{VJ} scores of human GenBank sequences as reference. As expected, human GenBank antibody sequences were most similar to the antibody sequences in our human immunome repertoire.

We demonstrated that our statistical PGSSM model captures a human-like antibody sequence space by recovering the human-like nucleotide sequences. We further were able to calculate a score of the V and J gene-encoded regions to quantify the similarity of an antibody sequence to a given immunome repertoire. The PGSSM_{VJ} score is the average of SNFs in the V and J gene-encoded region of the optimized sequence (Equation 2). We successfully used the score to distinguish between human, non-human, and engineered antibodies. We assessed the scores for 475 antibodies in clinical trials or approved by the U.S. Food and Drug Administration (FDA), indicating a high level of human likeness, but distinguishable difference from natural human antibody sequences.

6.2.1 Processing of immune repertoire data and counting SNFs in V, D, J gene-encoded, and CDR3

Our NGS sequence dataset was annotated with IgBLASTn results comprising germline gene alignments (Figure 9, A1). We only considered Ab sequences without sequencing ambiguity that contain nonstandard nucleotide letters. A collection of 196,755,218 heavy chain and 128,815,779 light chain sequences was used to create PGSSMs (325,570,997 in total). The dataset was processed with IgBLASTn and inferred germline gene alignments were assigned. We generated a full-length PGSSM for each of the 287 VH, 79 VK, 72 VL, 37 D, 13 JH, 9 JK, and 9 JL germline gene alleles. In-frame (+open reading frame (ORF)) germline reference sequences that are pre-annotated with CDR and

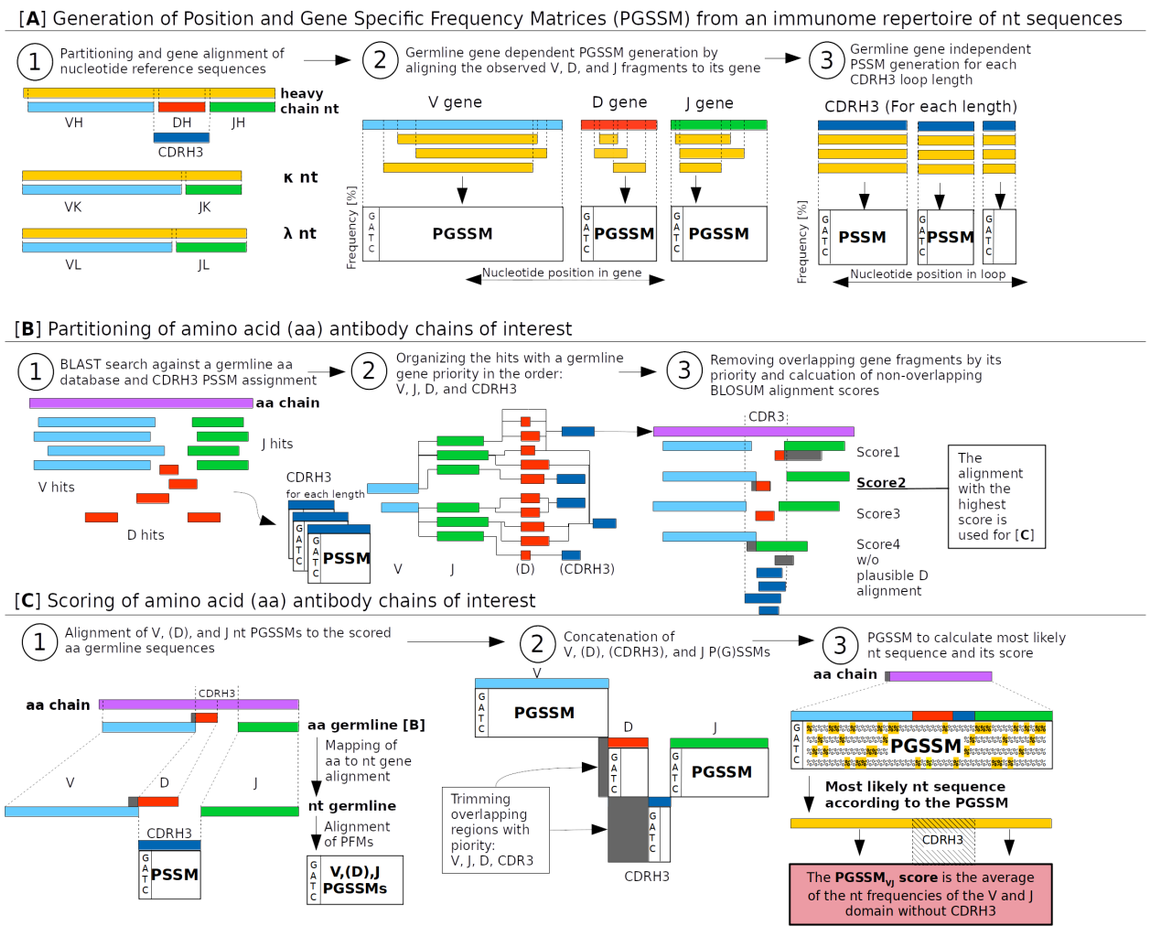

Figure 9: **Flowchart of scoring Ab sequences with IgReconstruct.** The algorithm can be divided into three tasks (a-c) with three steps (1–3) in each task. (a) The IgReconstruct algorithm starts with the generation of Position and Gene Specific Scoring Matrices (PGSSM) for the variable (V, light blue bars), diversity (D, red bars), joining (J, green bars) and CDR3 (dark blue bars) regions of the Ab nucleotide sequence (yellow bars). In this study, nucleotide sequences were obtained from a large immunome repertoire dataset. (b) For a given amino acid Ab sequence (purple bars), the V, D, and J germline gene rearrangement is determined from the alignment to the PGSSMs by creating a hierarchic tree of aligned nucleotide PSSMs. (c) The highest scoring rearrangement then is mapped to germline gene-dependent V, D, J and germline gene independent CDR3 PSSMs. The resulting nucleotide model is used to determine a back-translation which maximizes the observed nucleotide frequencies in the repertoire. The V and J regions of back-translated sequence is then scored ($PGSSM_{VJ}$) after the observed nucleotide frequencies in the repertoire

FR start positions were pulled from IMGT/GENE-DB (Giudicelli, Chaume, and M.-P. Lefranc 2005). Each of the matrices ultimately contains the frequency of observed G, A, T or C nucleotides for each position in each human germline gene (SNF). Here, we defined the CDR3 sequence as the sequence that starts with the first untemplated position after the V germline gene-encoded alignment and stops one position before the first J germline gene-encoded residue. For each observed heavy chain CDR3 loop (CDRH3) length, we created a germline gene independent PGSSM.

6.2.2 Calculation of PGSSMs from single nucleotide counts

To generate the PGSSMs, we first counted nucleotide observations in each germline gene as well as CDR3 loops. We extracted the V, D, and J gene alignments for each sequence as well as the untemplated region of the CDR3 loops. For some light chains and heavy chain sequences with high mutation frequency, no unambiguous D gene assignment was possible, whereas V, and J alignments are present for all analyzed sequences. Here, we refer to this D gene segment uncertainty with (D).

IgBLASTn generates alignments that contain in some cases overlaps of a few residues between V, (D), and J genes. In this case, we prioritized the alignments in the following descending order: V, J, (D). Each column of a PGSSM matrix corresponds to a nucleotide position in a germline gene. We then incremented either the G, A, T, C or gap cell in each aligned column of the PGSSM, avoiding double counts caused by gene overlaps (Figure 9, A2). We converted the observed counts into frequencies for each column after adding one pseudo-count to each cell, which resembles the SNFs. In addition to germline gene dependent V, D, and J PGSSMs, we generated germline gene independent CDR3 PGSSMs for each observed loop-length in the same manner (Figure 9, A3).

6.2.3 BLAST database generation and searches for creating a plausible amino acid germline gene rearrangement

In order to construct a PGSSM for a given amino acid target Ab sequence, we create a germline gene rearrangement as the first step (Figure 9, B1). For this purpose, we translated all human nucleotide germline genes using the reference sequences in the ImMunoGeneTics information system® (IMGT) database (Giudicelli, Chaume, and M.-P. Lefranc 2005) in all reading frames, allowing non-productive sequences, and generated separate BLAST databases (Stephen F. Altschul et al. 2009) containing V, D, and J genes while not distinguishing between heavy, kappa, or lambda chains. For each target Ab amino acid sequence, our algorithm conducts three independent BLAST searches with e-value thresholds of 20 (V), 100 (D), or 50 (J). The number of alignments was limited to 3 (V), 100 (D), or 10 (J). Word sizes were 4 (V), 2 (D), or 3 (J). BLAST hits were discarded if a stop codon was observed in the aligned region or if a corresponding PGSSM was not available. The length and position of the CDR3 is defined by the V, and J germline gene alignments. For each combination of V, and J BLAST hits, we assigned its distinct CDRH3 PGSSM, which is solely chosen by the length of the non-templated part of the CDRH3.

6.2.4 Assignment of a plausible V(D)J rearrangement for an amino acid target sequence

Our algorithm chooses a plausible V(D)J rearrangement for an amino acid sequence by scoring the combinations of BLAST hits. First, we create a V-J-D-CDRH3 tree hierarchy in the form of a nested data structure for each possible V(D)J alignment (Figure 9, B2). We prevented incorrect alignments from being added to the tree, such as D alignments that were not overlapping with the CDR3, and J alignments not overlapping with the FR4 region. Both regions were calculated for each V germline gene dynamically following the IMGT Unique Numbering scheme, (M.-P. Lefranc, Pommié, Ruiz, et al. 2003; M. P. Lefranc 1997) which encodes the positions of FR and CDR as fixed positions in gapped germline genes. The pattern [WF]GXG in the J gene-encoded region marks the end of the CDR3. We also ensured the rearrangements were consistent regarding chain type (heavy, kappa, or lambda).

Second, to choose a final V(D)J rearrangement from the tree, we rescored all recombinations of V, (D), and J alignments after trimming all overlapping regions (Figure 9, B3). We calculated the

BLOSUM62 scores for each alignment after pruning the aligned region from overlaps. Overlapping alignments were trimmed or kept with the following descending priority: V, J, D. For example, a D gene alignment overlapping with N residues of a J gene alignment shortens the scoring area of the D gene alignment by N residues. The remaining V(D)J recombinations then were sorted after summing the scores of the individual alignments. We discarded all rearrangements but the one with the highest score. This process does not require D germline gene alignments, since BLAST D germline genes could not be aligned in about 50% of all cases.

It is important to point out, that the germline gene rearrangement tree is individually generated for each antibody and depends on the unique SNF of the repertoire. A rearrangement in the tree is preferred if a compatible and optional CDRH3 PSSM has been found. A CDRH3 PSSM is compatible if it can bridge the distance between the last aligned V residue and the first J residue. Hence, the chosen V, J, D, CDRH3 rearrangement is dependent on observed CDRH3 lengths in the repertoire.

6.2.5 Creation of the final PGSSM model and scoring of an amino acid target sequence

We used the V(D)J rearrangement chosen earlier and mapped the aligned amino acids corresponding to V, (D) or J genes to their nucleotide counterparts. In addition, we assigned one CDR3 PSSM depending on the length of the loop (Figure 9, C1). We concatenated each V, (D), J and (CDR3) PGSSM such that overlapping parts were discarded. We again respected the domain priority in the descending order V, J, D, CDR3 (Figure 9, C2). Despite the important role of the CDRH3 PSSM for back-translation as well as scoring of the germline gene rearrangement, we chose to not include the untemplated CDRH3 region in the score calculation for two reasons. Firstly, the germline D gene and CDRH3 PSSMs cannot always be assigned. Success depends on the chain type and the availability of CDRH3 PSSMs of a certain length, i.e., the CDRH3 must be observed in the repertoire. Secondly, the CDRH3 PSSM contains all CDRH3 loops of 128,815,779 heavy chain sequences, solely grouped by length. As a result, we do not expect predictive capabilities to the PSSM regarding human-likeness (Figure 11.1b), even though it supports the generation of a back-translated sequence in this region (Figure 11.1a).

We therefore restricted calculation of the PGSSM score to V and J PGSSMs, whereas residues without assigned V or J PGSSM remain unscored (Equation 2). Mann-Whitney statistics were used to assess the significance between $\mathrm{PGSSM_{VJ}}$ scores of human, non-human Abs and Ab drugs.

To assess the human likeness of the $\mathrm{PGSSM_{VJ}}$ score, we calculated the Z-Score using mean and standard deviation of $\mathrm{PGSSM_{VJ}}$ scores obtained for all human GenBank antibody sequences separated by heavy or light chain type (Equation 3).

6.2.6 Strategy to reconstruct nucleotide sequences from Ab amino acid sequences

The concatenated nucleotide PGSSM (Figure 9, C2 and Figure 11.1) aligned and cropped to fit the amino acid target sequence was used to calculate the $\mathrm{PGSSM_{VJ}}$ score. Naturally, this approach

also can deduce a nucleotide sequence that maximizes the SNFs (Figure 9, C3). Such a nucleotide back-translation is codon-optimized and exhibits the highest possible similarity to the PGSSM and its underlying immune repertoire data. Creating an optimized nucleotide sequence eliminates a potential sequence bias of reported nucleotide sequence and increases the robustness of our method in scenarios where only amino acid sequences are available. This situation occurs frequently in artificial computational protein Ab design in which typically the design process is performed without regard to germline gene rearrangements or nucleotide sequences (Adolf-Bryfogle, Kalyuzhniy, et al. 2018; Sircar, E. T. Kim, and Gray 2009). The generation of our nucleotide sequence comprises two steps. First, we interrogated for each amino acid the aligned nucleotide PGSSM and chose the triplet with the smallest hamming distance to the wild-type germline gene. For the untemplated CDRH3, we skipped this step. Second, if multiple triplets after step one are available, we chose the triplet, which maximizes the cumulative SNF.

Figure 9 depicts the complete strategy from amino acid Ab target sequence to nucleotide reconstruction. This method presents per-nucleotide frequency statistics for almost the complete Ab variable domain, including the junction areas of the CDR3 loop and the loop itself. The few exceptions to this assignment are N and C termini without alignments, short light chain junctions, or residues encoded by insertions in the templated regions. Figure 11.1 shows the complete PGSSM rearrangement of the heavy chain with GenBank accession number EU620063.

6.2.7 The $\mathrm{PGSSM_{VJ}}$ acts as a human likeness score in the context of immunomes from healthy humans

We calculated the $\mathrm{PGSSM_{VJ}}$ (Equation 2) for all reconstructed nucleotide sequences in the context of three human healthy immunome repertoires (Figure 10b). The scores for human heavy and light sequences were significantly higher with 93.6% $\pm$ 3.5% (heavy chain) and 93.7 $\pm$ 2.9% (light chain), respectively, than the scores for other species.

The non-human primates Callithrix jacchus (91.1 $\pm$ 2.2%/90.9 $\pm$ 2.9%), Chlorocebus sabaeus (89.1 $\pm$ 2.4%/91.5 $\pm$ 2.7%) and Macaca fascicularis (89.2 $\pm$ 2.4%/91.7 $\pm$ 2.1%) scored significantly lower with P values from a Mann-Whitney test $\ll$ 10-7. The lowest scoring species include Gallus gallus (Red junglefowl) and Salmo salar (Atlantic salmon) with 78.6 $\pm$ 1.9%/82.0 $\pm$ 1.5% and 79.3% $\pm$ 3.7%/N.A (heavy chain/light chain). The lower bound of $\mathrm{PGSSM_{VJ}}$ as well as sequence recovery is constrained by the chance to guess nucleotides of a fixed amino acid sequence correctly, which is approximately 73.68% (Appendix). Scores around the value of 73.68% are strong indicators for sequence alterations such as engineered sequences.

6.2.8 The $\mathrm{PGSSM_{VJ}}$ score can be used to identify engineered and atypical antibodies

Some sequences of the species Homo sapiens are outliers in that they score significantly lower than the 95% confidence interval. For Abs annotated with Mus musculus, a number of high-scoring outliers outside the 95% confidence interval occurred (Figure 11.1b). These findings can be attributed to engineered or other non-natural Abs. For the case of Mus musculus, sequences often

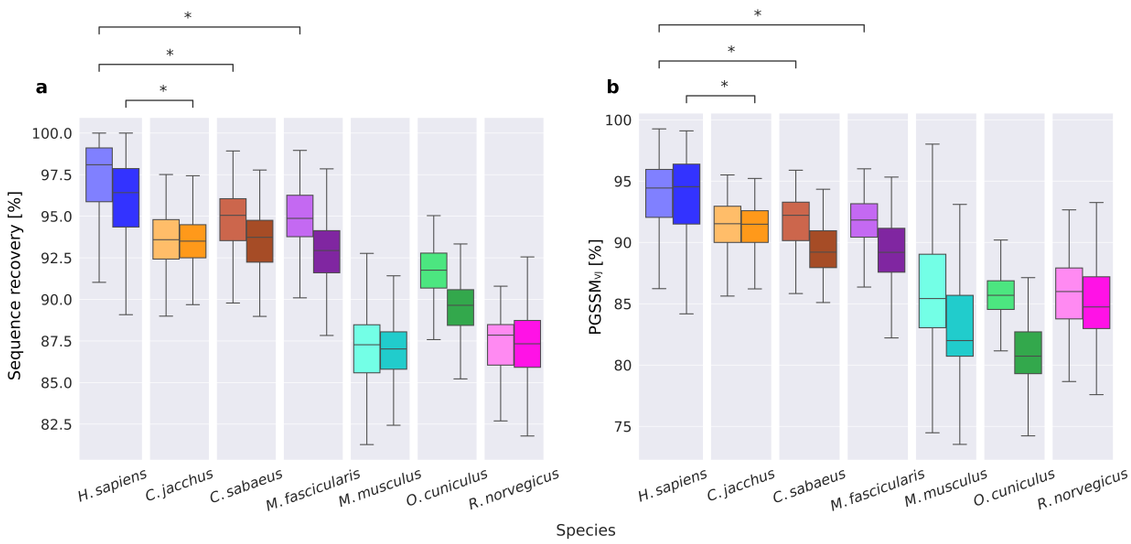

Figure 10: **Native nucleotide sequence recovery and PGSSM$_{VJ}$ score for Ab sequences taken from GenBank.** Amino acid sequences were downloaded from GenBank (Benson et al. 2013) and then back-translated to nucleotide sequences using IgReconstruct. (a) The sequence recovery rate after back-translation with IgReconstruct is highest for human (H. sapiens) sequences when compared to that for sequences from non-human primates (C. jacchus, C. sabaeus, M. fascicularis), mouse (M. musculus), rat (R. norvegicus) or rabbit (O. cuniculus). (b) The PGSSM$_{VJ}$ score for the same set of back-translated nucleotide sequences also scores highest for amino acid sequences derived from humans. Light colors (left bar in each subplot) represent light chain sequences, dark colors (right bar in each subplot) represent heavy chain sequences. A Mann-Whitney test shows statistically significant ($\star$, p $\ll$ 10-7) recovery rates and scores for human sequences compared to the other species.

can be associated to studies involving transgenic mice with human Ab loci (Sok et al. 2016; Longo et al. 2017; Suárez et al. 2006; M. Tian et al. 2016; Protopapadakis et al. 2005).

A large number of low scoring human sequences are annotated with patents related to engineering and or animal Ab sources (US20050002930A1, JP2007524605A, EP2150565A2) often directed to human cancer and immune disorder treatments (JP2009221224A, EP2150565A2, WO2005063299A3, WO2004085474A2) like prostate cancer (WO0173032A2, JP2003528591A), or patents evolving in the vicinity of anti-human Abs (WO2005067477A3). Another possible explanation for the low scoring GenBank entries are their annotations designating them as unpublished or having incomplete publication records (e.g., GenBank IDs: EU620060, FW576479, DQ187727). Our observations match previously reported concerns of incorrectly annotated Abs (Martin and Rees 2016).

Heavy chain/light chain sequences of structures from the Protein Database (PDB) (Berman et al. 2000) with IDs 1GAF (79.9%/86.3%), 1AXS (80%/83.9%), 1BBJ (81.9%/84.7%) 4UOK (88.0%/82.8%), and 4UOM (80.7%/90.0%) were scored. These PDBs were reported previously as incorrectly annotated with human origin (Martin and Rees 2016). The low PGSSM$_{VJ}$ scores ($<1\sigma$ of GenBank sequences assigned as human) also underlines the probable non-human origin of all heavy chains and most light chains.

One shotgun sequenced human light chain of the transcriptome with ORF expressed sequence tags described in 2000 (Dias Neto et al. 2000) exhibits two insertions and a region of five deletions, dropping the sequence score to 77.16%. Other examples for sequences with presumably human

background but atypical mutation patterns are broadly neutralizing HIV Abs (Xueling Wu, T. Zhou, et al. 2011; Liao et al. 2013) like VRC01 and its derivatives that occurred after long-term lineage evolution (Xueling Wu, Zhang, et al. 2015). These highly matured Abs can indicate sensitivity to the progress in sequencing methods. Low-scoring HIV mAbs may highlight the challenge for the human system to generate the right combination of rare mutations against the highly variable sequences of HIV envelope protein (Bhatti, Usman, and Kandi 2016).

Another example of Abs with rare mutations are fetal lymphocyte progenitors, (Kolar et al. 2004) highly mutated Abs of tonsillar IgD-cells, (Seifert et al. 2009) or expanded multiple sclerosis associated lineages in immortalized B cells (Fraussen et al. 2013). Some of these Abs are related to tissue location or to autoimmune diseases, and might therefore not be typical of Abs found circulating in the peripheral blood, which is the current context of our Ab analysis method.

6.2.9 The $\mathrm{PGSSM_{VJ}}$ score correlates with the phylogenetic distance to human V germline genes

We further interrogated the $\mathrm{PGSSM_{VJ}}$ properties and estimated their correlation with the phylogenetic distances between human and non-human species. The phylogenetic distance was calculated as the sum of the branch length between the two closest germline genes of the same class (heavy, kappa, lambda) of two species. We calculated a phylogenetic tree between the available IMGT reference germline sequences. Nucleotide frequencies in V and J gene-encoded domains are on average low in number and guide the overall sequence space of a species. This germline gene preference of nucleotides is directly captured in the PGSSM frequencies and ultimately in the $\mathrm{PGSSM_{VJ}}$ score.

The average $\mathrm{PGSSM_{VJ}}$ score for all studied sequences is plotted against the phylogenetic distance from the assigned human V gene to its closest V gene of the organism of origin separately for heavy chain (Figure 11a) and light chain V genes (Figure 11b). GenBank sequences of the species Mus musculus are frequently the subject of lineage evolution and of engineering studies, and such sequences exhibit highly artificial mutation patterns, which causes a low correlation between phylogenetic distance and score. We therefore separated Mus musculus sequences and highlighted these in red color. The correlation of heavy chains remains less affected due to the higher number of datapoints.

Single nucleotide frequencies in Abs roughly recapitulate phylogenetic distances. One can thus use the $\mathrm{PGSSM_{VJ}}$ to confirm or question the Ab species annotation. The $\mathrm{PGSSM_{VJ}}$ therefore can be used as a measure of the degree of recombinant engineering with known phylogenetic relations.

6.2.10 $\mathrm{PGSSM_{VJ}}$ allows for the recovery of nucleotide sequences for human Abs

We performed a nucleotide sequence recovery benchmark to demonstrate that triplet independent observations of single nucleotide frequencies can approximate the human Ab sequence space. 181,335 GenBank sequences of 20 different species were translated with IgBLASTn (Ye et al. 2013). The nucleotide sequence was optimized by maximizing the $\mathrm{PGSSM_{VJ}}$ score.

Back-translation recovery rates peak for human sequences, with an average heavy and light chain

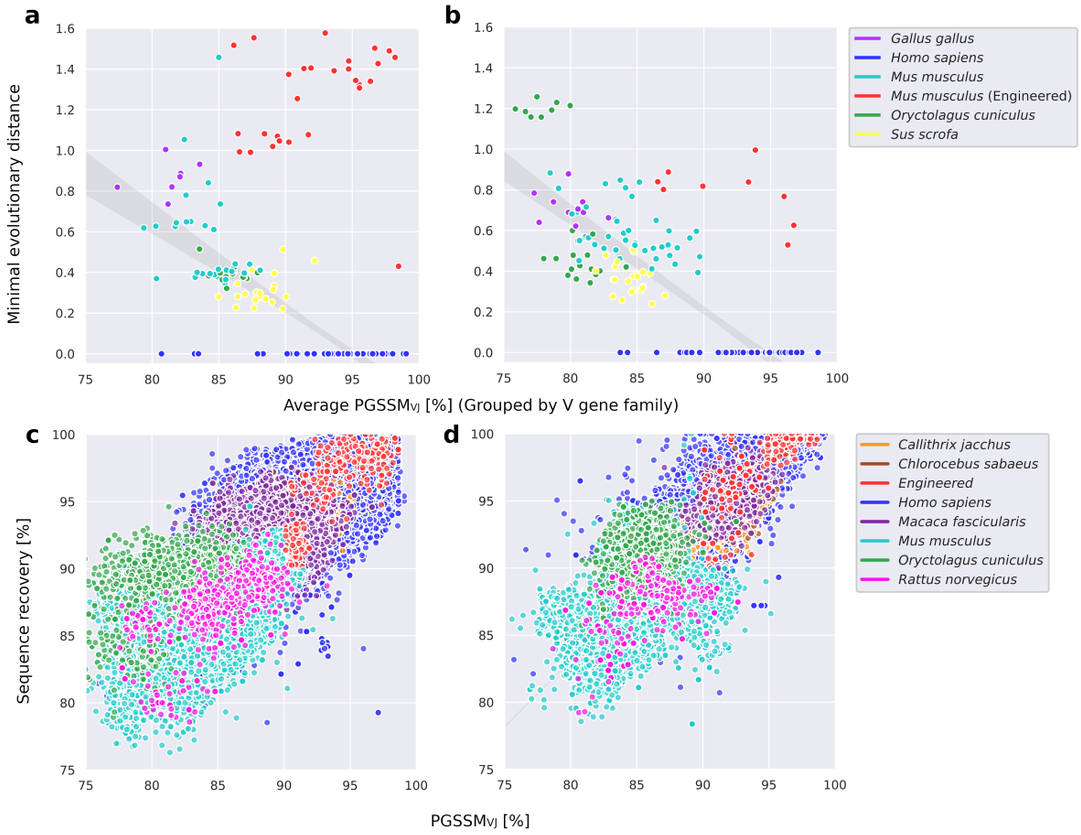

Figure 11: **The $\mathrm{PGSSM_{VJ}}$ score approximates the evolutionary distance from human immunoglobulin germline genes to immunoglobulin germline genes belonging to 20 species.** Amino acid sequences were downloaded from GenBank (Benson et al. 2013) and then back-translated to nucleotide sequences using IgReconstruct. (a) The average $\mathrm{PGSSM_{VJ}}$ scores for heavy chain Ab sequences or (b) light chain Ab sequences are plotted against the phylogenetic distance from the assigned human germline gene using IgReconstruct (see Methods section for details). The $\mathrm{PGSSM_{VJ}}$ scores correlate with the phylogenetic distance with a Spearman rank correlation coefficient of ρ = -0.83 (P = 2e-41, $\alpha = 0.01$) for heavy chain Ab sequences and ρ = -0.83 (P = 2e-37, $\alpha = 0.01$) for light chains Ab sequences. (c) Sequence recovery between native heavy chain sequences and back-translated nucleotide sequences, made using IgReconstruct, gave a Spearman rank correlation coefficient of $\rho = 0.92$ (P = 0, $\alpha = 0.01$). (d) Sequence recovery between native light chain sequences and back-translated nucleotide sequences using IgReconstruct gave a Mann-Whitney correlation coefficient of $\rho = 0.86$ (P = 0, $\alpha = 0.01$). Mouse (M. musculus) Abs engineered to be human-like are colored red (top right corner of subplot a and b)

recovery of 95.9 ± 2.6% or 97.2 ± 2.8%, respectively (Figure 10a, Figure 11.1a). As expected, when we leveraged the human $\mathrm{PGSSM_{VJ}}$ score to determine the most likely human nucleotide sequence for Abs of different species, correct nucleotide identification dropped, labeling these Abs as non-human. For non-human primates, recovery rates were Callithrix jacchus (93.5 ± 1.5%/93.3 ± 2.2%), Chlorocebus sabaeus (93.4 ± 1.9%/94.5 ± 2.7%) and Macaca fascicularis (92.8 ± 1.9%/94.7 ± 1.9%). The lowest scoring species included Gallus (Red junglefowl) and Salmo salar (Atlantic salmon) with heavy/light chain scores as low as 82.7 ± 1.1%/82.9 ± 1.4% and 82.6 ± 2.2%/N.A. A comparison of $\mathrm{PGSSM_{VJ}}$ scores with sequence recovery rates (Figure 10) shows striking similarity, suggesting that the $\mathrm{PGSSM_{VJ}}$ score is a predictor of sequence recovery. Figure 11.1 depicts the similarity of sequence recovery (a) with $\mathrm{PGSSM_{VJ}}$ score (b) for all 20 species.

6.2.11 The sequence recovery frequency strongly correlates with the $PGSSM_{VJ}$

A third property of $PGSSM_{VJ}$ is the ability to estimate the nucleotide sequence recovery rate. We calculated the correlation between average nucleotide mutation frequency ($PGSSM_{VJ}$ score) with the sequence identities determined in our sequence recovery benchmark. The recovered sequence is of importance to determine the minimal distance to its context for Ab-dataset comparisons. With a Mann-Whitney correlation coefficient of R = 0.92, P = 0 for heavy chains (Figure 11c) and R = 0.86, P = 0 for light chains (Figure 11d), the $PGSSM_{VJ}$ is approximately the sequence recovery rate for human sequences ± 5%.

6.2.12 Ab therapeutics in context of the Ab repertoire of healthy humans

We used 475 unique Abs that are either approved by the U.S. FDA or are in clinical trials (Jain, Sun, et al. 2017; Poiron 2021). All biologics were either annotated with the INN designations (Parren, Paul J Carter, and Andreas Plückthun 2017) HU, ZU, XI, and XIZU as reported by Jain et al (Jain, Sun, et al. 2017). or annotated with Human, Humanized, Chimeric, and Mouse in case of antibodies taken from IMGT/mAb-DB (Poiron 2021). For this study, we chose appropriate labels for HU (Human), ZU (Humanized), XI (Chimeric), and XIZU (Humanized Chimeric Hybrid) to match the designations used in IMGT/mAb-DB. The sequences were treated the same way independent from its labeling in the algorithm. We investigated the Ab sequences in the context of our three individual immunome repertoires and in the context of one large merged repertoire. For Z-Score calculation, mean and standard deviation (σ) from GenBank sequences (Figure 10b) were used (Equation 3).

We compared the Z-Score of $PGSSM_{VJ}$ either grouped by clinical stage (Figure 14) or source subsystem, which indicates the origin and type of engineering of the biologics (Figure 13) (Parren, Paul J Carter, and Andreas Plückthun 2017). Drugs with a human source scored highly similar to GenBank sequences (Z-Score around 0), followed by humanized, chimeric and murine Abs. This trend was consistent for both drug datasets processed. Scores of sequences from mice still score in a similar range of GenBank Mus musculus sequences. This finding shows that antibody sequences from IMGT/mAb-DB with a murine background remain distinguishable from biologics with human origin. On the other hand, humanized and chimeric sequences populate a scoring range closer to human and non-human primate sequences. Pooling drugs by their clinical status shows that drugs in Phase 2 to 3 clinical trials and approved Abs have an average Z-score of -0.56 ± 1.05 (Phase 2), -0.77 ± 1.35 (Phase 3), and -1.18 ± 1.45 (Approved). On average, human drugs appear human-like with a Z-Score greater than -2, caused by the high number of human (57) and humanized (68) drugs compared to 13 chimeric. The low number of available sequences aggravates the challenge to draw reliable conclusions. The $PGSSM_{VJ}$ indicates that there is a non-human sequence space compatible with the human system. However, we hereby choose a Z-Score cutoff of -2 or greater to roughly group the majority of clinical stage antibodies (Figure 14, horizontal red line). For our next experiment, we used this cutoff to distinguish between biologics/human antibodies, and non-human antibodies.

To further investigate the role of public and private repertoires on the eligibility of Abs as drugs, we calculated $PGSSM_{VJ}$ scores using each of the three individual immunome repertoires. The majority of staged antibodies exhibit a cutoff of -2 or greater (Figure 14). Hence, we roughly defined any of the three scores as human-like as long as the Z-Score of the $PGSSM_{VJ}$ was greater or equal to -2. Figure 15 depicts the number of human-like scores for non-human (orange), human GenBank Abs (blue), and biologics (green), separated by light chains (a) and heavy chains (b). We observed high agreement between the three scores for human and therapeutic Abs. We also observed high agreement rates between all three repertoires, including 70.0% of all biologics and 92.3% of all human GenBank heavy chain sequences and 81.8% of all biologics and 94.6% of all human GenBank light chain sequences. In contrast only 8.8% light chain and 8.8% heavy chain sequences of biologics and 1.3% of light chain biologics and 2.6% of heavy chain human GenBank sequences were scored as non-human in all three cases.

6.2.13 Performance and robustness

The initial release of our algorithm requires amino acid Ab sequences that cover at least a fraction of the V and J gene-encoded region, which can be successfully aligned via BLAST. The algorithm then places optional D PGSSMs as well germline gene CDR3 loop PGSSMs in the appropriate locations if available. Templated regions as well CDR3 junctions are modeled statistically; insertions are represented in the statistical SNF model as gaps.

We compared the germline gene families with the top five germline gene families assigned by IgBLASTp, the IgBLAST tool for protein sequences (Table 1). Our method reliably assigns germline V genes to our sequences when IgBLASTp is taken as reference.

6.2.14 Output

We provide a webservice called IgReconstruct (http://meilerlab.org/index.php/servers/IgReconstruct), which takes amino acid sequences of Ab variable domain in FASTA format as input. The output is presented graphically in a downloadable PDF file (Figure 12), and a spreadsheet with equivalent machine-readable information. The PDF report presents the query amino acid sequence aligned to its reconstructed nucleotide sequence, V, (D), and J germline gene alignments. The germline gene alignments indicate sequence identity with a dot and residue type replacements with a one-letter code. The variable region is annotated in the form of branches for the predicted IMGT-CDR1-3. V(D)J domains are colored blue, red, and green and match the colors used in the IgReconstruct flowchart (Figure 9). In case of overlapping alignments, the region is colored according to the hierarchy of the rearrangement tree.

6.3 Discussion

We have shown that statistics of SNFs of the variable region using large human immunome repertoires are capable of modeling the human Ab sequence space by predicting nucleotide sequences from amino acid sequences (Figure 10). With more and more large NGS nucleotide sequence

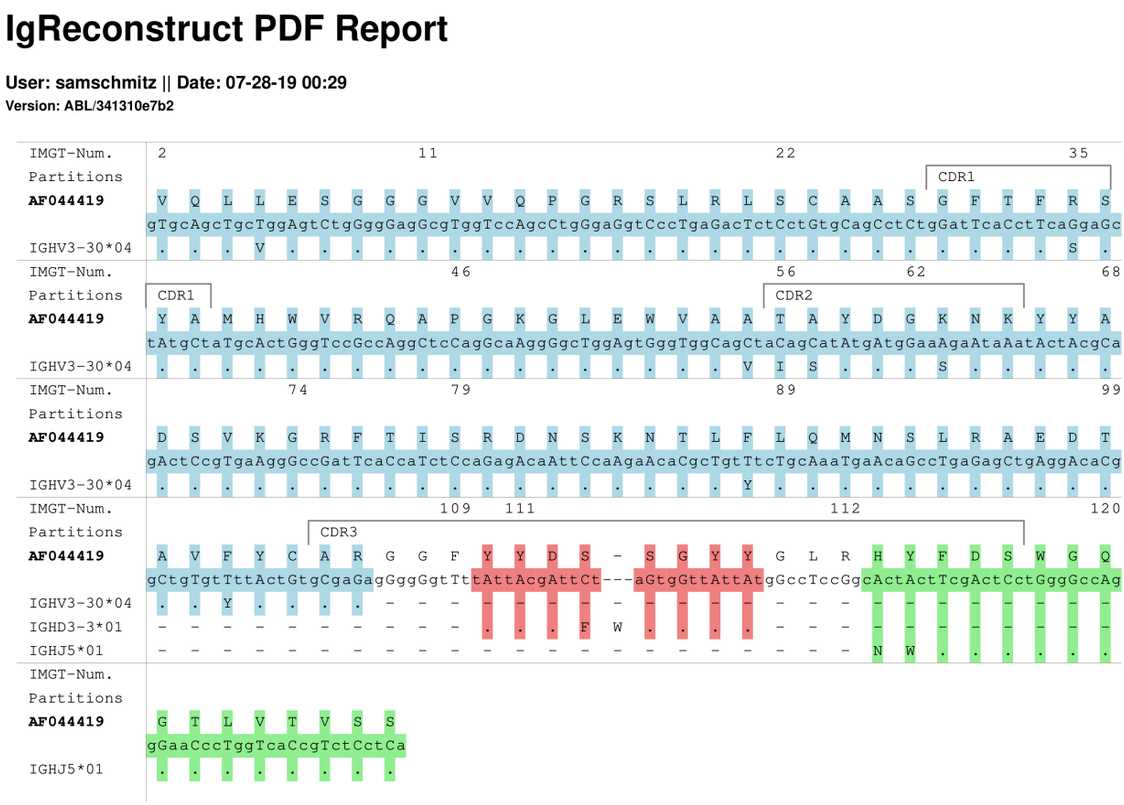

Figure 12: **Alignment report generated by IgReconstruct. An example alignment report for the human heavy chain Ab sequence with the GenBank accession number AF044419.** Reports generated by IgReconstruct provide information on the query amino acid sequence (first row), the back-translation (second row) and alignments to the germline gene sequences (third and following row if applicable). The color code blue (V gene), red (D gene), and green (J gene) refers to the aligned germline PSSMs which were used to create the back-translated sequence. Columns without color are not aligned to a specific germline gene. Dots represent the germline sequence; mutations are shown using the one-letter amino acid code. CDR loops 1 to 3 are inferred based on alignments to the V and J germline genes. The numbers on top of the amino acid sequence was implemented using the IMGT numbering scheme (M. P. Lefranc 1998). Non-templated regions at the V-D and D-J junctions flanking the D gene alignment (red) are covered by the CDRH3 PSSM, but are not visualized in the color scheme. The PDF report gives a quick insight into the nature of the germline gene rearrangement which is used to generate the back-translation and the human-likeness score

datasets becoming publicly available, (Soto, Bombardi, et al. 2019; DeWitt et al. 2016; B. Briney et al. 2019; Corrie et al. 2018; Kovaltsuk et al. 2018) IgReconstruct resembles an approach to link the nucleotide sequence space with resources of Abs where primarily amino acid information is available, like de-novo computational models or structural databases (Berman et al. 2000; Dunbar et al. 2014). Approaches of structural modeling of Abs (Adolf-Bryfogle, Kalyuzhniy, et al. 2018) have been made to include amino acid sequence profiles of V and CDR3. IgReconstruct may pave the way to completely model the germline gene rearrangement of an amino acid sequence at the nucleotide level and provide full access to large-scale human immunome repertoire statistics.

We demonstrated that the $\mathrm{PGSSM_{VJ}}$ score, derived from the SNF statistics of an individual Ab, is an appropriate distance measure of a particular chosen Ab to a nucleotide immunome repertoire or arbitrary large set of sequences (context). For this, we fulfilled the requirement to find the minimal distance by suggesting the most probable nucleotide sequence for a given repertoire (context-dependent). The $\mathrm{PGSSM_{VJ}}$ then can be used to estimate the likelihood to observe a context-dependent nucleotide sequence in the dataset. Finally, the $\mathrm{PGSSM_{VJ}}$ strongly correlates

with the phylogenetic distance between human and non-human germline genes (Figure 11). These combined properties allowed us to estimate the similarity of a variable domain to a dataset and to interpret it as a distance value. For example, further studies might conclude that infections like HIV exhibit a greater distance to the human sequence space, which results in less effective immune responses.

A current shortcoming of our method is that our CDRH3 statistics, which include the heavy chain junctions, are only length dependent. As a result, the major domain that diversifies an immunome repertoire (Glanville et al. 2009; Saada et al. 2007; Warren et al. 2011) is merged into relatively small bins, disregarding the sequence similarity and function. As a result, our $\mathrm{PGSSM_{VJ}}$ score is currently exclusively calculated from V and J gene templated regions. We do not anticipate or observe sufficient performance using solely CDRH3 PSSMs to distinguish between non-human, human, and biologics only using CDRH3 sequences due to high variability (Figure 11.1). However, CDRH3 PSSMs can be used to support the back-translation of amino acid sequences to nucleotides (Figure 11.1).

We evaluated Ab sequences from 20 species and were able to distinguish sequence origins between human primates, non-human primates and other species reliably. While doing this, we found that the prior species annotation in deposited sequences often was not reliable. The signal that allows us to distinguish between human vs. non-human persisted while studying the IgReconstruct results of clinical-stage and FDA-approved Abs (Figure 13). A non-human source could reliably be detected in murine, chimeric, humanized chimeric and humanized Abs. Due to the higher count of therapeutic Abs with a human sequence background, the combined population of sequences scores at the lower end of "human-like" (Figure 14). A more comprehensive therapeutic Ab and immunome repertoire relationship might be developed in the future, when our statistical Ab model incorporates a more sophisticated CDRH3 model. The results indicate that there is a non-human sequence space, which is compatible with human biology (i.e., is associated with a manageable frequency of adverse effects). Abs from that space can be used as therapeutics. These sequences remain unlike the repertoire in our study with low human likeness scores, despite humanization efforts. However, the majority of Z-Scores of antibody biologics in clinical phases appears to be -2 or greater (red horizontal line). For our next experiment, we used this cutoff to distinguish between biologics/human antibodies, and non-human antibodies.

Krawczyk et al. used amino acid alignments of variable and CDR regions to show that sequences with high similarity to therapeutic Abs can emerge in the human antibody repertoire, whereas chimeric and humanized antibodies tend to be slightly more dissimilar (Krawczyk et al. 2019). This observation could be reproduced using SNFs mapped onto germline genes instead of amino acid sequence alignments. In addition, a Z-score cutoff of -2 was chosen, which enables us to separate between non-human and human as well as biologics. The ability to separate drugs from non-human antibodies is hypothesized to support antibody drug development in the future.

The human-likeness score in this study is distinctly different from previously published methods, where typically the ability of the separation of real human and non-human sequences was

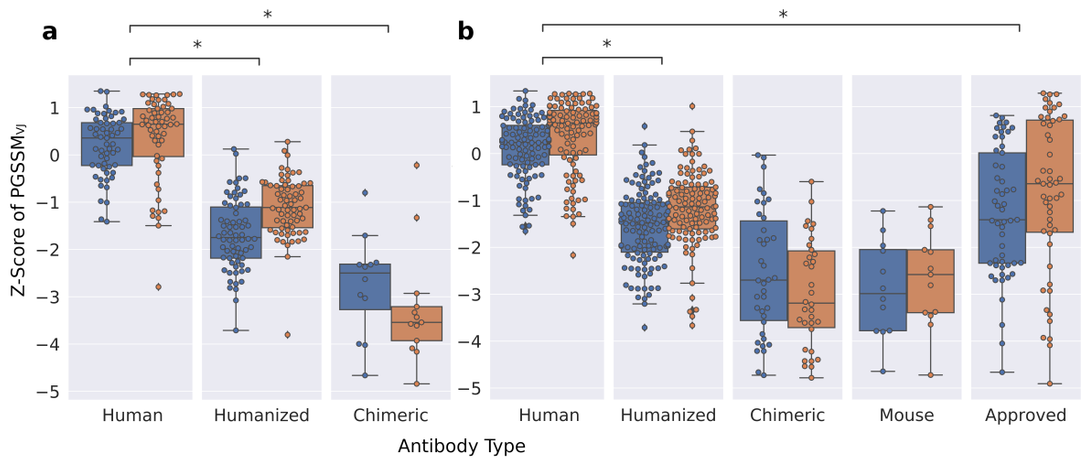

Figure 13: **The $PGSSM_{VJ}$ score ranks human Abs highest when compared to either chimeric or mouse Abs used as biologics.** Ab sequences for biologics were obtained from IMGT/mAb-DB (Poiron 2021) separated by heavy chain (blue) and light chain (orange). All $PGSSM_{VJ}$ scores were transformed into Z-scores and ranked within each group. (a) Biologics analyzed from the Jain et al (Jain, Sun, et al. 2017). study show that human Abs rank highest when compared to either chimeric or mouse Abs. Humanized Abs also rank higher than either chimeric or mouse Abs. (b) Biologics from the IMGT monoclonal Ab database show a similar picture, with human sequences scoring higher than biologics with a non-human origin. Approved Biologics are distinguishable from human antibodies. Mann-Whitney significance tests show statistical significance ($p \ll 10\text{-}7$) and are labeled with a star (*)

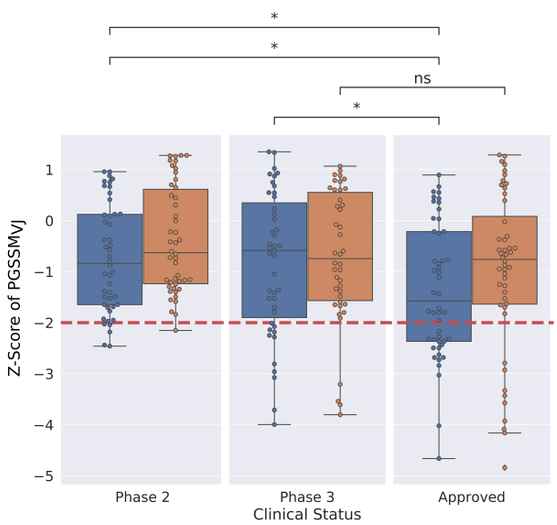

Figure 14: $PGSSM_{VJ}$ **score cannot discriminate between clinical stage and FDA-approved biologics.** The Z-Scores of heavy chains (blue) and light chains (orange) were calculated using the distribution of GenBank sequences annotated as human. $PGSSM_{VJ}$ scores of biologics from Jain et al., (Jain, Sun, et al. 2017) grouped by their clinical phase, show an overall picture of human-like sequences (within one standard deviation of human GenBank sequences) and a smaller population of low scoring sequences. A Mann-Whitney test between clinical trial Phase 2, 3 and FDA-approved Abs revealed no significance (ns) to very weak statistical significance ($p < 5 \times 10\text{-}2$, *)

being maximized. Recent advances in deep-learning have shown excellent classification capabilities (Wollacott et al. 2019). Here, we devised a method that generates a nucleotide frequency model based on repertoire observations, which represents the plausibility that an Ab sequence arises from a particular repertoire. The results of a previous study could be confirmed, which has shown that biologics can be distinguished from human sequences (Krawczyk et al. 2019). On the one hand, this study does not aim to maximize the separation between truly human and non-human sequences, resulting in less clear boundaries between human and, for example, macaque sequences (Thullier et al. 2010). On the other hand, the approach could hypothetically be used to capture the biologically relevant question of the immunogenicity of an Ab, which cannot strictly be answered by separating human from non-human sequences. Consequently, a slightly worse separation performance compared to the deep-learning approach of Wollacott et al. could be observed with an Area

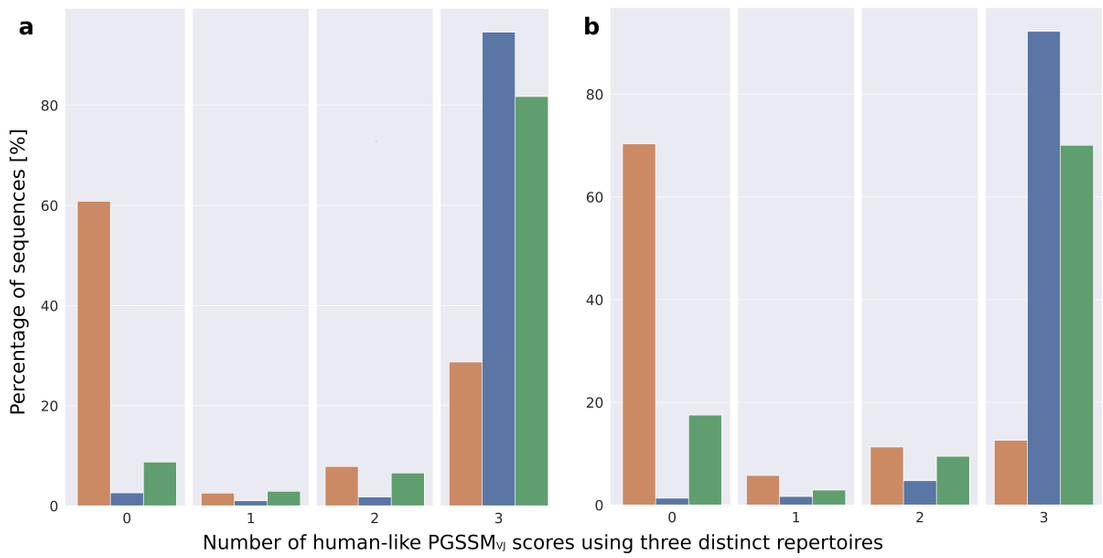

Figure 15: **Scoring medically relevant Abs using sequencing data from three individual human immunome repertoires.** PGSSM$_{VJ}$ scores of biologics (Jain et al.47) versus human and non-human Ab sequences from GenBank. All Ab sequences were scored using sequencing data from three separate immunome repertoires (Soto, Bombardi, et al. 2019). Z-Scores of the PGSSM$_{VJ}$ was calculated using GenBank sequences annotated as human. A binary score was used to indicate if an amino acid sequence was human-like. A score of 1 indicates a human-like sequence with a Z-score of -2 or greater. A score of 0 indicates a non-human-like sequence with a Z-score less than -2 (see human data in Figure 14 for Z-score cutoff value). Each sequence was scored against each repertoire and summed up. Thus, a maximum number of three scores can be achieved for any individual sequence which signifies that the sequence is human-like according to comparison with all three individual repertoires. Using the cutoff of -2 allows to roughly separate between non-human (orange), human (blue) GenBank sequences and biologics (green). In case of light chains (a) the cutoff of -2 classifies a larger amount (30%) of non-human antibodies as human than in the case of heavy chains (12%) (b)

Under the Curve (AUC) of 0.94 compared to 0.97 (Figure S6). At the same time, IgReconstruct is able to leverage the substantial sizes of the largest repertoires with hundreds of million to billion sequences like the Observed Antibody Space (Kovaltsuk et al. 2018) by using nucleotide germline gene rearrangements as reference, as opposed to using smaller datasets in the ranges of ten thousands of sequences of previous methods (Abhinandan and Martin 2007; Gao et al. 2013; Thullier et al. 2010; Seeliger 2013; Wollacott et al. 2019). IgReconstruct provides an alternative to extrapolating sequence landscapes from a small representative set of sequences in favor of leveraging large repertoires to its full extent. The definition of human-likeness in this study is a novel approach with the potential to support Ab engineering and explain immunogenic effects in future studies.

6.4 Materials and methods

We developed the PGSSM method and supplementary tools for repertoire processing in Python-3.7.1. We provide a webserver that generates germline gene rearrangements for amino acid Ab sequences in text or PDF form, and numeric information in a spreadsheet format.

6.4.1 Curation of sequences from three sources

We curated a set of 181,355 Ab sequence from GenBank (Benson et al. 2013). 119,827 heavy chains Ab were from the following species: Bos indicus (5), Bos taurus (1,520), Callithrix jacchus (328),

Camelus dromedarius (388), Canis lupus familiaris (253), Chlorocebus sabaeus (82), Equus caballus (427), Felis catus (94), Gallus gallus (157), Homo sapiens (76,728), Lama glama (499), Macaca fascicularis (3,592), Mus musculus (27,863), Oryctolagus cuniculus (1,253), Ovis aries (1719), Rattus norvegicus (544), Salmo salar (109), Sus scrofa (4029), Vicugna pacos (237). 61,528 light chain sequences were from the species Anas platyrhynchos (298), Bos indicus (191), Bos taurus (353), Callithrix jacchus (874), Camelus dromedarius (32), Canis lupus familiaris (417), Chlorocebus sabaeus (74), Equus caballus (319), Felis catus (76), Gallus (301), Homo sapiens (41,347), Lama glama (15), Macaca fascicularis (673), Mus musculus (13,249), Oryctolagus cuniculus (1,099), Ovis aries (583), Rattus norvegicus (299), and Sus scrofa (1,328). We applied our method on the translated variable domains reported by IgBLASTn. To estimate the performance, we calculated the nucleotide sequence identity of the complete variable region and compared the germline gene families assigned with our method with the results from IgBLASTp for protein sequences.

In addition to GenBank, we used a dataset of 137 Ab drugs (Jain, Sun, et al. 2017) and extracted 382 Abs for clinical use from IMGT/mAb-DB (Poiron 2021). In total, we had sequences for 475 unique Ab drugs available for analysis.

6.4.2 Calculation of $\mathrm{PGSSM_{VJ}}$ scores and assessment of human-likeness

We developed a method that creates position- and gene-dependent scoring matrices for a given immunome repertoire (Figure 9). Our $\mathrm{PGSSM_{VJ}}$ score assesses the similarity of any given amino acid antibody sequence to the repertoire by averaging the observed single nucleotide frequencies over the Ab V and J gene-encoded regions. The single nucleotide frequencies were looked up in the PGSSM matrix that was generated for each antibody individually (Figure 11.1). Equation 2 was used to calculate the similarity score using a specific sequence and PGSSM matrix.

Equation 2 Calculation of the $\mathrm{PGSSM_{VJ}}$ score for the variable and joining region calculated as an average of the observed single nucleotide frequencies

$$\Delta E_{total} = \sum_{i}^{N} PGSSM_{VJ}(resi, resn)/N \tag{2}$$

N:= Sequence Length

resi:= Residue Position i

resn:= Residue type at position i ∈ {G, A, T, C}

The Z-Score of the $\mathrm{PGSSM_{VJ}}$ score was used to estimate the human likeness of an antibody. For Z-Score calculation, we used the average and standard deviation of $PGSSM_{VJ}$ scores we calculated for 76,728 human GenBank sequences (Equation 3). We also defined an antibody as human-like as long as its Z-Score was -2 or greater.

$$Z = (PGSSM_{VJ} - \mu)/\sigma \tag{3}$$

μ: = Mean of $PGSSM_{VJ}$ scores of human GenBank sequences

σ: = Standard deviation of $\mathrm{PGSSM_{VJ}}$ scores of human GenBank sequences

6.4.3 Phylogenetic tree construction and the evolutionary distance of germline genes

To characterize the $PGSSM_{VJ}$ score, we correlated scores of 20 species with the phylogenetic distance to human germline genes. For this purpose, we constructed a phylogenetic tree from the complete set of IMGT reference sequences (M. P. Lefranc 1998) of all species available using the program PhyML (Guindon et al. 2010). For each human V germline gene allele, we calculated the minimal phylogenetic distance to each genus of the same chain class (heavy, lambda, kappa) by summing up the branch lengths of the closest path. We averaged the sequence recovery rate and $PGSSM_{VJ}$ score for each germline gene in the tree.

6.5 Availability

IgReconstruct is available as a webservice, hosted by Meiler Lab with no restrictions for sequence files up to 4 MB. (http://meilerlab.org/index.php/servers/IgReconstruct)

7 Rosetta design with co-evolutionary information retains protein function

This chapter has been published under (Schmitz et al., 2021).

7.1 Introduction

Proteins play a vital role in fundamental processes of life, and their diverse three-dimensional structures allow for highly diverse functions. Computational protein design explores the sequence landscape and side chain conformational space for a given protein backbone to find a residue combination that supports a function. The protein modeling suite Rosetta (Leaver-Fay, Tyka, et al. 2011) has been applied with marked success on various applications (Raveh et al. 2011; Rohl et al. 2004), including protein (Brian Kuhlman et al. 2003) and enzyme design (Richter et al. 2011). A critical element of Rosetta is a scoring function that is fine-tuned to respect knowledge-based statistics and physical approximations. Without additional restraints, this scoring function reflects the thermodynamic stability of one static protein conformation in a distinct environment (Alford et al. 2017).

However, protein function often relies on structural flexibility (Süel et al. 2003), thus multiple Rosetta protocols have been developed to favor sequences which do not only thermostabilize but also account for protein flexibility. Multi-state design (MSD), for example, supports design on multiple protein conformations simultaneously which benefits the design of conformational changes (Sevy, Jacobs, et al. 2015; Leaver-Fay, Jacak, et al. 2011; Löffler et al. 2017). The MSD implementation RECON (Sevy, Jacobs, et al. 2015; Sauer et al. 2020) optimizes in an iterative protocol the individual sequences of the conformational states. Each iteration increases a restraint to converge the individually designed sequences into a single sequence that supports all conformations.

Improving thermodynamic stability or function of a given protein is an important aspect of protein design (Goldenzweig and Fleishman 2018). As protein sequences observed in nature are often close to the optimum (B. Kuhlman and D. Baker 2000), the design of sequences constrained towards native conformations and sequences is a successful strategy. It can be implemented by using sequence profiles (Goldenzweig, Goldsmith, et al. 2016) that mirror the residue occupancy at each position of a backbone and serve as additional constraints on sequence selection. However, as each residue is treated independently, a severe limitation of sequence profile design is the neglection of subtle interdependencies between residue occupancies.

The reasons for these mutual dependencies are often the maintenance of structural stability by compensatory mutations but are also more importantly related to sophisticated functional aspects like information transmission, conformational plasticity, and the binding of ligands or other proteins (Z. Hu et al. 2007; Marks, Hopf, and Sander 2012). Thus, a network of evolutionary constraints may exist in a protein that fine-tunes the occupancy of several pairs of residue-positions. Various methods like GREMLIN (Balakrishnan et al. 2011), plmDCA (Ekeberg et al. 2013), and PSICOV (D. T. Jones et al. 2012) have been developed to identify these constraints, which are also named couplings, to indicate the dependency between the occupancy of residue pairs. In a pioneering

study, co-evolutionary fitness landscapes have been used to design three different stable protein folds with the ability to bind native ligands with high affinity (P. Tian, Louis, et al. 2018).

Pairwise sequence requirements in natural proteins are a consequence of maintaining thermodynamic stability, structural flexibility (plasticity), and other requirements for protein function, such as recognizing interaction partners, catalyzing chemical reactions, and many more. Computational protein design with Rosetta primarily favors thermodynamic stability and is conceptually unaware of couplings required for protein flexibility and/or function. The premise of this study is that this restriction in evolutionary tolerated sequence space is not reflected in Rosetta designed proteins. This leads to design solutions that are thermodynamically stable but might change flexibility or lose function. While custom protocols for a specific design task can circumvent this shortcoming, we wondered about a general approach to maintain native-like couplings in the sequences designed beyond the couplings dictated by thermodynamic stability. For this study, we evaluate a number of computational design protocols in Rosetta: 1) One biased towards the wild-type sequence as a baseline for comparison, 2) Design with a sequence profile, which encodes the sequence space as observed in functional proteins, 3) RECON multi state design, which has the potential to capture couplings critical for protein plasticity, and 4) Constraining co-evolving residues directly in the Rosetta design process.

We hypothesize that incorporating evolutionary constraints in the Rosetta design process will allow us to optimize the sequence across all functionally relevant conformations even for single state design (SSD), including intermediate states that are difficult to obtain experimentally (Bonetti et al. 2016). Thus, we have implemented a novel RosettaScripts (Fleishman et al. 2011) element, the ResCue (residue coupling enhanced) mover, which transforms coupling strengths inferred from a MSA into an energy function bias (restraint). These restraints are generalizable and applicable on different design scenarios that can be addressed with Rosetta. To evaluate our method, we captured two performance metrics: First, we measured the recovery of couplings. Second, we assessed the overall sequence recovery of the full protein and of residues which were reported as functionally relevant in literature. We found that proteins designed with ResCue had significantly higher recovery rates compared with three other state-of-the-art design approaches.

We use native sequence recovery as one of our metrics of design success in order to facilitate comparison with other studies. Although it might appear counter-intuitive to use this measure to assess coupling recovery, we argue that it is a useful metric as increased coupling recovery will imply increased sequence recovery. Our method achieves high recovery rates by conserving networks of co-evolving residue pairs, in contrast to an alternative approach that trivially increases sequence recovery rates by limiting the number of mutations. We show, that our method is superior in recapitulating the wild-type residues especially in functionally active sites compared to other approaches and thus is suitable to retain function during design.

Description	PDB IDs	Resolution [Å]	Length	RMSD [Å]	MSA size
HPPK	1HKA 1Q0N	1.5 1.3	435	0.5	4534
FixJ	1D5W 1DBW	2.3 1.6	126	0.5	38021
RasH	6Q21 4Q21	2.0 2.0	189	0.5	46795
G-protein Arf6	1E0S 2J5X	2.3 2.8	174	1.0	36036
S100A6	1K9P 1K9K	1.9 1.8	90	1.9	14768
Calmodulin	1CKK 1CFD	NMR NMR	148	9.9	13561
LAO Binding protein	2LAO 1LAF	1.9 2.1	238	4.5	23810
Phosphate-binding protein	1QUK 1OIB	1.7 2.4	321	2.9	5898
Thioredoxin reductase	1TDE 1F6M	2.1 3.0	316	6.6	33408
Adenylate kinase	1AKE 4AKE	2.0 2.2	214	7.0	30589

Table 1: Characterization of the ten benchmark proteins ($\text{bench}_{\text{coev}}$) used in this study.

7.2 Results and discussion

7.2.1 Assembling a benchmark $\text{bench}_{\text{coev}}$ of ten proteins representing conformational flexibility

In order to test our hypothesis that co-evolutionary information helps to improve the protein design process, we assembled a benchmark of ten proteins, which we named $\text{bench}_{\text{coev}}$ (Table 1). We chose the proteins based on four criteria. First, two conformations had to be available in the Protein Data Bank (PDB), representing two functionally different states e.g. without or with a bound substrate. Second, we accepted only structures with an experimental resolution of 3 Åor better. Third, for each protein of length N, we confirmed that 10 x N homologous sequences were available in databases, which is a prerequisite for a reliable determination of coupling (Ovchinnikov, Kamisetty, and David Baker 2014). Forth, we preferred proteins that are functionally well studied and understood. We ended up with a diverse set of two calcium binding proteins, two GTP binding proteins, one DNA binding protein, one phosphate binding protein, one enzyme, one bacterial solute binding protein and one protein that is part of an ABC transporter.

7.2.2 The ResCue mover and its energy term

Our method aims at the conservation of co-evolutionary networks during the design of protein sequences (Fig 16). For their identification, we opted for GREMLIN (Ovchinnikov, Kamisetty, and David Baker 2014) that deduces from an MSA of homologous sequences a four-dimensional coupling tensor storing the co-evolutionary scores. The first two dimensions list two residue positions and the last two indicate their amino acid interdependencies for all possible combinations. Large positive values represent a strong coupling and negative values indicate their incompatibility; but most values are close to zero. The tensor allows us to quickly deduce a score (Eq 4) for the strength of the coupling constraint cc(i) for each individual residue i. These scores are then used to constrain sequence design as an add-in to the Rosetta energy function. Note that it is essential to balance carefully between having an efficient constraint but not over-writing the standard energy function, since the designed proteins should fulfill the coupling restraints and be physically realistic.

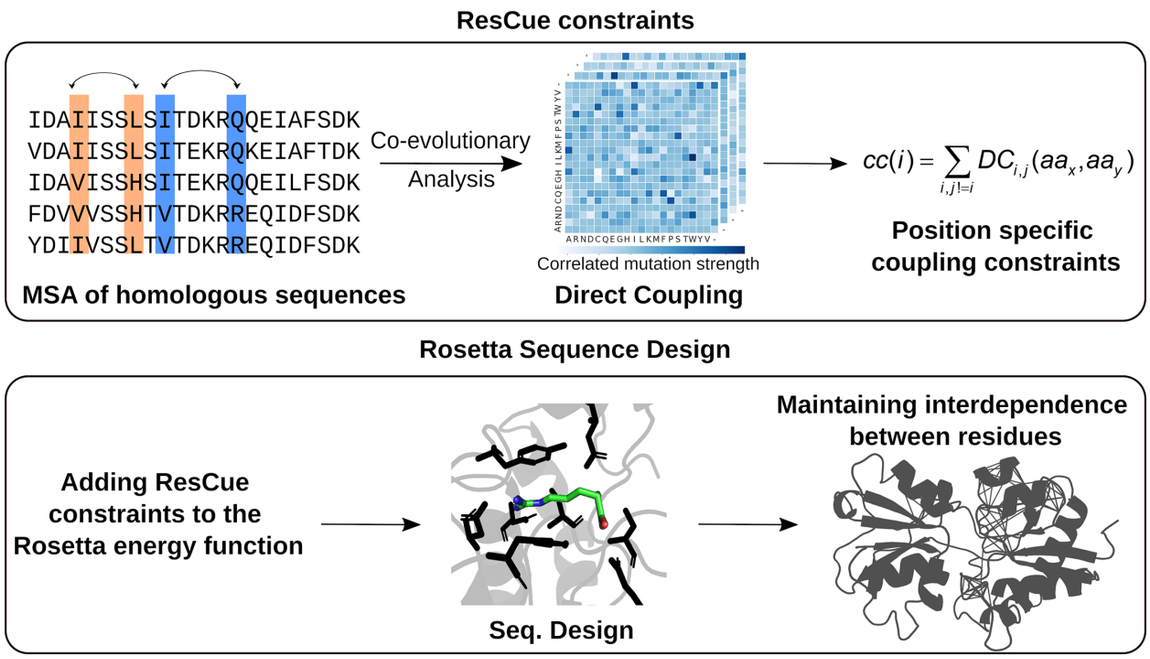

Figure 16: **Basic concept and application of ResCue.** Co-variance scores are deduced from an MSA of homologous sequences and converted to position specific coupling constraints cc(i). The ResCue constraints are then added to the Rosetta energy function to maintain the interdependencies between residues while protein sequence design.

7.2.3 Sophisticated design protocols sample sequences of higher energy

To evaluate our method, we assessed the performance of four established design protocols on the protein benchmark set $\text{bench}_{\text{coev}}$: a) An unmodified, default Rosetta single-state design protocol that served as a reference (RoSSD) (Leaver-Fay, Tyka, et al. 2011), b) RECON MSD design, c) SeqProf Mover (SeqProf), which is sequences profile design using a position-specific scoring matrix (PSSM) (S. F. Altschul and Koonin 1998), d) biased design to prefer the native sequence (FavorNative), and e) our co-evolutionary informed design ResCue. The same MSA was used to derive the PSSMs for SeqProf and the residue-specific coupling constraints cc(i) (Eq 4) utilized with ResCue.

As noted above, it is essential, to balance carefully between having an efficient coupling restraint and designing physically realistic sequences as dictated by the Rosetta energy function. By restraining Rosetta to bias the sampling towards a desired goal, the energy landscape is modified. As a result, when reevaluating the solutions with the unmodified energy function, the energy can and often does increase (get worse). The ResCue protocol (S2 Supplement) was parametrized to produce designs with Rosetta energies comparable to the established SeqProf method and substantially increased coupling recovery. In order to assess the energy increase of the different design approaches, we determined the difference of the Rosetta total energy to the relaxed wild-type structure with the best energy, normalized by protein length. As expected, all design approaches with constraints had significantly higher Rosetta energies compared to the relaxed wild type (Fig 17), (Mann–Whitney U test (MW) p <1.0e-04 for all three comparisons). The differences per residue were -0.15 ± 0.11 REU for single state design, +0.55 ± 0.76 REU for RECON MSD, + 0.28 ± 0.14

REU for the SeqProf design, -0.057 ± 0.15 REU for FavorNative, and +0.13 ± 0.15 REU for our ResCue mover. As the latter value is significantly lower than that of the RECON protocol (MW p = 6.4e-195) and comparable to the SeqProf design, we concluded that our concept of considering evolutionary constraints affects the scoring function less than a well-established MSD approach. The FavorNative design energies remain on average close to the wild-type energies (-0.057 REU). Most likely, for several positions residue choices with similar energies exist and FavorNative helps to select the native residue. As expected, sequence recovery increases substantially with the FavorNative method, which is a trivial result as the correct solution is input into the method. Thus, this approach would not allow the design of new sequences that retain structure, plasticity, and function. Details on the execution of each experiment and the used constraint weights can be found in S2 Supplement.

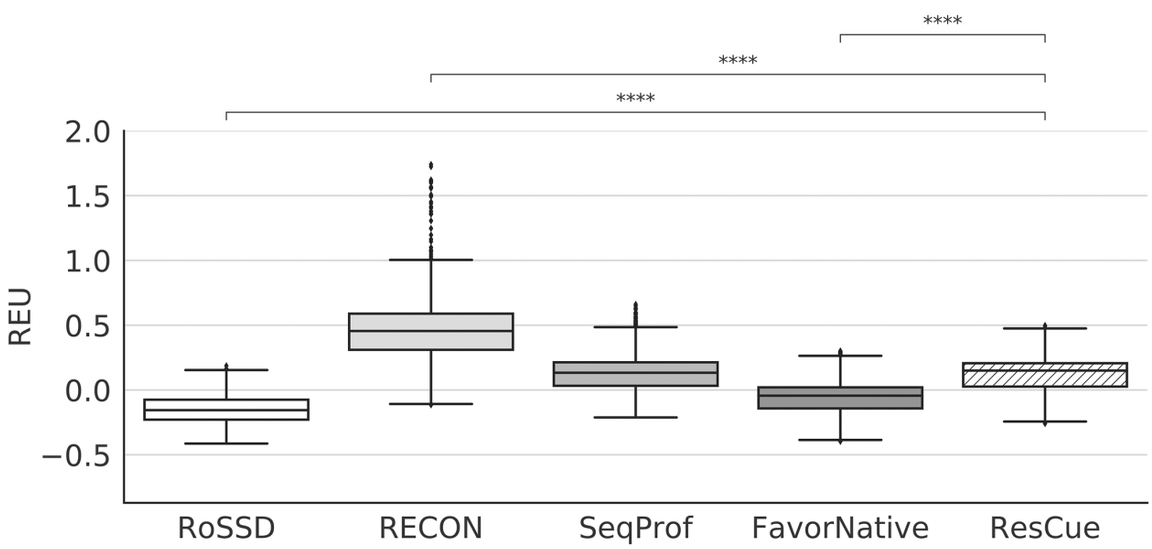

Figure 17: **Distribution of Rosetta total energies for the full benchmark design.** Energies are given in Rosetta energy units (REU) relative to the wild-type on a per residue basis for the five design protocols. For this figure and all subsequent boxplots, the median is indicated as a black line; boxes depict the interquartile range (IQR), whiskers represent 1.5 x IQR. Results of a two-sided Mann-Whitney-Wilcoxon test are indicated as follows: * ≙ 1.00e-02 <p <= 5.00e-02, * ≙ 1.00e-03 <p <= 1.00e-02, * ≙ 1.00e-04 <p <= 1.00e-03, *** ≙ p <= 1.00e-04.

7.2.4 ResCue recovers networks of co-evolving residues

Having shown that our scoring of couplings had no drastic effect on sequence energies, we assessed the conservation of couplings by analyzing for each benchmark prot the designed sequence seqDesign and the native sequences seqNative. We determined the coupling recovery score crs(prot) (Eq 6), which quantifies how well the inferred residue interaction network was maintained. To compute this score, we first calculated for each seqDesign the sum of the corresponding scores cc(i). One can consider the cumulative strength of pairwise couplings as a measure for the selective functional pressure on a particular residue i (Marks, Hopf, and Sander 2012). For normalization, the resulting cs(seqDesign) value (Eq 5) was divided by cs(seqNative). Note that crs(prot) can assume values greater than one.

Fig 18A depicts the crs values of the sequences designed with the five protocols; the standard deviation was ≈ 10% in all cases. The unconstrained Rosetta protocol RoSSD reached an average crs value of 21%. The performance of RECON MSD was slightly better with an average crs value of 25% and SeqProf design gained an average crs value of 28%. FavorNative reached an average crs value of 52% with a higher standard deviation compared to other protocols with 18.3%. In contrast, ResCue reached an average crs value of 109%. Note, that the crs value can be larger than 100% which would suggest that the designed sequences fulfill additionally restraints not found in the native sequences seqNative, but in homologs.

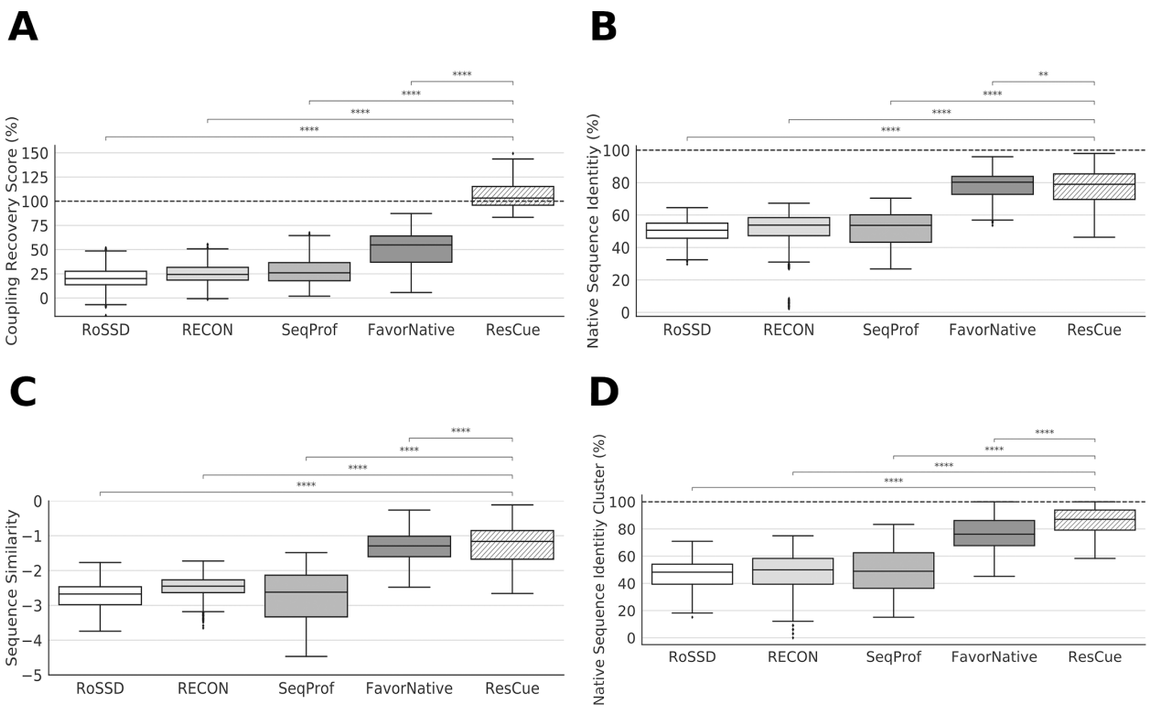

Figure 18: **Performance of four different design approaches.** (A) Coupling recovery scores crs deduced for all designed sequences. The 100% value is indicated by the dashed line. (B) Native sequence recovery nsr of the four design approaches. (C) Sequence similarity seqsim of the full-length sequences. (D) Native sequence recovery nsrCN of residues found in clustered networks. The scores are summarized with boxplots as explained above. Results of a two-sided Mann-Whitney-Wilcoxon test are indicated as follows: * ≙ 1.00e-02 <p <= 5.00e-02, * ≙ 1.00e-03 <p <= 1.00e-02, * ≙ 1.00e-04 <p <= 1.00e-03, *** ≙ p <= 1.00e-04).

These data show that the native Rosetta protocol is not suitable to completely recover the evolutionary constraints for functional connectivity between residues across our ten proteins. Moreover, we expected a better performance of RECON, since multistate design optimizes over both conformations at the same time. This optimization should consider residue couplings that are in spatial proximity in either state. As expected, SeqProf failed to drastically improve the performance as mutual residue-dependencies are ignored. In contrast, the average crs value of ResCue exceeds 100%.

7.2.5 Preserving evolutionary constraints by means of ResCue improves native sequence recovery and sequence similarity

As further quality measures, we determined native sequence recovery nsr(seqDesign) (Eq 7) and sequence similarity seqsim(seqDesign) (Eq 9) values by comparing the designed sequences and the native ones (Fig 18B and 3C). RoSSD reached an average nsr value of 50 ± 6.3% and a seqsim value of -2.72 ± 0.34. For RECON, the nsr and seqsim values were 50 ± 12.2% and -2.50 ± 0.35, respectively. SeqProf design gained an averaged nsr value of 51 ± 10% and a seqsim value of -2.74 ± 0.35. FavorNative reached nsr and seqsim values of 78 ± 9.15% and -1.28 ± 0.5. Note that the FavorNative weights were tuned to approximate the ResCue sequence recovery. Our ResCue design showed a significant increase both in the nsr and the seqsim values, which were 78 ± 11.7% and -1.20 ± 0.61, respectively. Compared to the other design approaches, these improvements were statistically significant ($MWp < 5.0e-04$ for nsr and seqsim).

We wanted to know, whether these protocol-specific performance differences in nsr and crs values affect each individual protein of the benchmark $\text{bench}_{\text{coev}}$ and determined the nsr improvements and the crs improvements. For each of the two conformations of a protein, the nsr value of RoSSD was subtracted from the nsr value reached by each of the three other design protocols, namely RECON, SeqProf, and ResCue. Analogously, the crs values were processed. Thus, a difference greater than zero indicates an improvement compared to RoSSD, whereas a value smaller than zero indicates that the protocol performed worse than the reference.

In Fig 19, these pairs of values are plotted for each protein of $\text{bench}_{\text{coev}}$. RECON showed a slight increase except for the two conformational states of the thioredoxin reductase (two datapoints in the lower left quadrant). Most likely, these results are due to the low resolution of one state, which was 3.0 Å. SeqProf slightly improved sequence identity and coupling recovery values in six of the proteins (thioredoxin reductase, LAO binding protein, phosphate-binding protein, S100A6, FixJ and HPPK) and impaired them in four proteins (RasH, G-protein Arf6, calmodulin and the adenylate kinase). In contrast, ResCue design improved nsr and crs values for the full benchmark set.

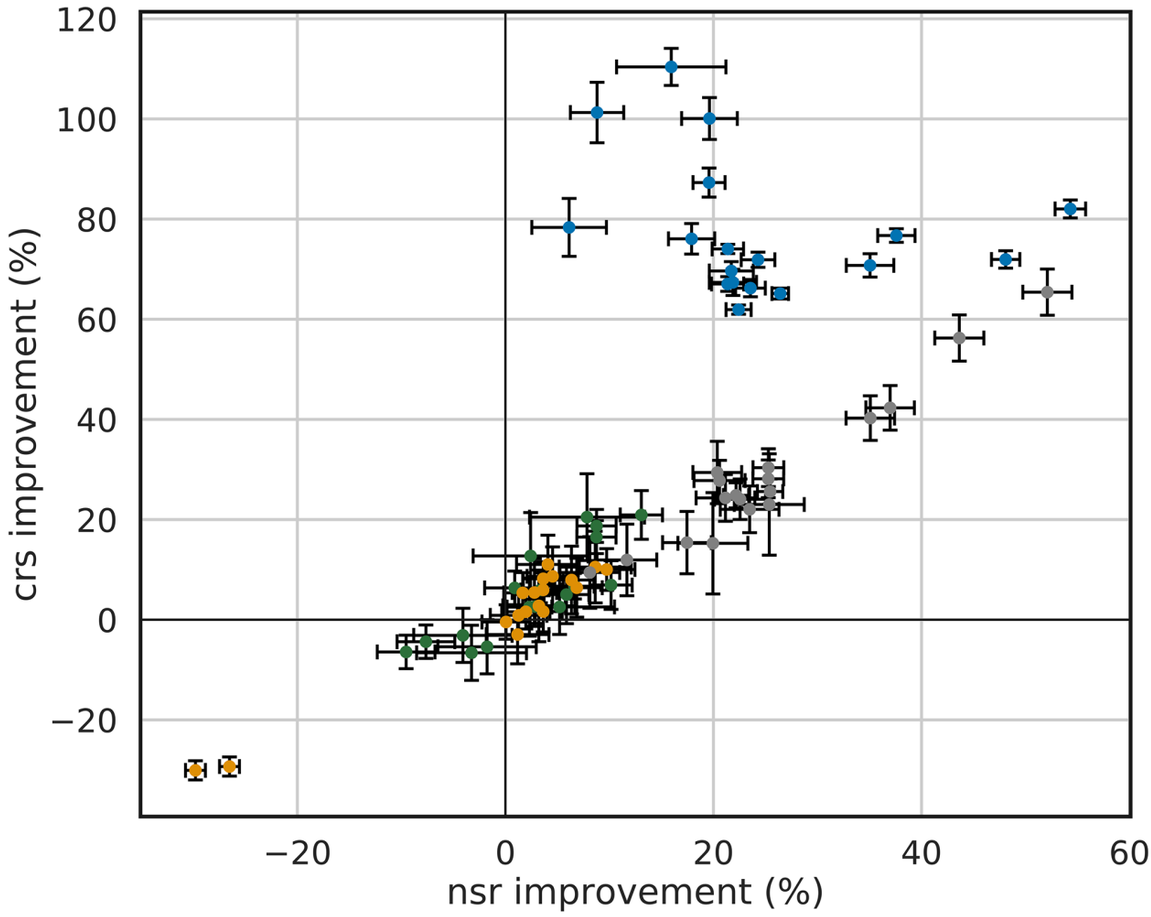

Figure 19: **Improvement of native sequence recovery values and coupling recovery scores for the two conformational states of all ten benchmark proteins.** For each state, the nsr value of the default Rosetta protocol was subtracted from the nsr value reached by one of the other protocols. The crs values were processed analogously. The following color code indicates the design protocols: RECON (orange), SeqProf (green), ResCue (blue), and FavorNative (gray).

Taken together, ResCue directed the design process towards native sequences that facilitate the coupling of residues. Considering these constraints comes at the expense of a relatively moderate energy increase, which however, is not energetically more expensive than maintaining sequence composition by means of sequence profiles.

7.2.6 ResCue recovers functionally relevant residues

Residues involved in evolutionary couplings often form networks (Süel et al. 2003; Jeon et al. 2011; Marino Buslje et al. 2010) and the cc(i) values are a measure for the selective functional pressure on a particular residue (Marks, Hopf, and Sander 2012). To study the most prominent cases, we adopted a previous approach (Jeon et al. 2011) and identified all residues ($res_{cc}^{20}(prot)$) having a coupling constraint cc(i) above the 20th percentile. As expected, these residues were often described as functionally relevant in the literature (see discussion for individual proteins below). We used the residues to compute coupling networks (CN) and determined for the corresponding sets of residues the nsrCN values, which were higher than the global sequence recovery nsr(prot) values. The average nsrCN value of the RoSSD protocol was 46 ± 8.8%, for RECON, SeqProf, and FavorNative design the values were 47.5 ± 14.8%, 50 ± 14.9%, and 49 ± 14.9% respectively.

Our ResCue design reached an average nsrCN value of 87 ± 8.8% (Fig 18D), which is statistically significantly higher than that of the second-best protocol, namely SeqProf (MW p = 1.9 e-22).

To study the coupling networks in detail, we present results for four benchmark proteins, LAO, FixJ, RasH, and calmodulin, which we have chosen for the following reasons: First, the binding site residues of LAO play an essential role in stabilizing both the closed and the open state. Second, FixJ residues are involved in dimerization. Third, RasH uses two highly flexible switch domains to bind GTP. Fourth, residues crucial for peptide binding in calmodulin are only moderately conserved, which complicates efforts to recognize and recapitulate them. Detailed information about the remaining benchmark proteins is provided in S1 Supplement.

7.2.7 The substrate induced conformational change of the lysine-arginine-ornithine binding protein LAO

LAO is a periplasmic protein capable of binding the amino acids L-arginine and L-histidine (Oh et al. 1993). Periplasmic transport systems consist of a substrate-binding protein and a membrane-bound complex that translocates the substrate from the periplasm to the cytoplasm. Following substrate binding, LAO undergoes a conformational change, bringing the two domains into a closed configuration that completely buries the ligand. Recently, residues crucial for substrate binding were identified (Vergara et al. 2020) by performing alanine scanning and categorized into different groups: Main chain binding (D161, S72, R77), guanidino binding (D11), side chain binding (Y14, F52) and water-mediated binding (D30, S70).

Our analysis revealed that $res_{cc}^{20}(LAO)$ the residues form two networks located close to the binding site, a smaller network consisting of seven residues and a more complex one with 35 residues (Figs 20A and 21A). Four of the eight crucial residues (Y14, F52, S70, R77) are part of the more complex network, which highlights that certain combinations of binding site residues enables them to bind the ligand cooperatively. Comparing the structure with and without bound ligand showed that all $res_{cc}^{20}(LAO)$ residues adopt alternative configurations in the open and the closed configuration. Analyzing the sequence space for LAO designs showed that each of the five design approaches was able to recover the native amino acids at position 11 (aspartate) and position 30 (aspartate) (Fig 22A). Both RoSSD design and RECON failed to sample the A77. In contrast, Y14, F52, S70, and S72 were only recovered in ResCue designs. Superimposing the native structure with a ResCue design sampling the correct amino acids illustrates the similarity except for two side-chain conformations (Fig 23A).

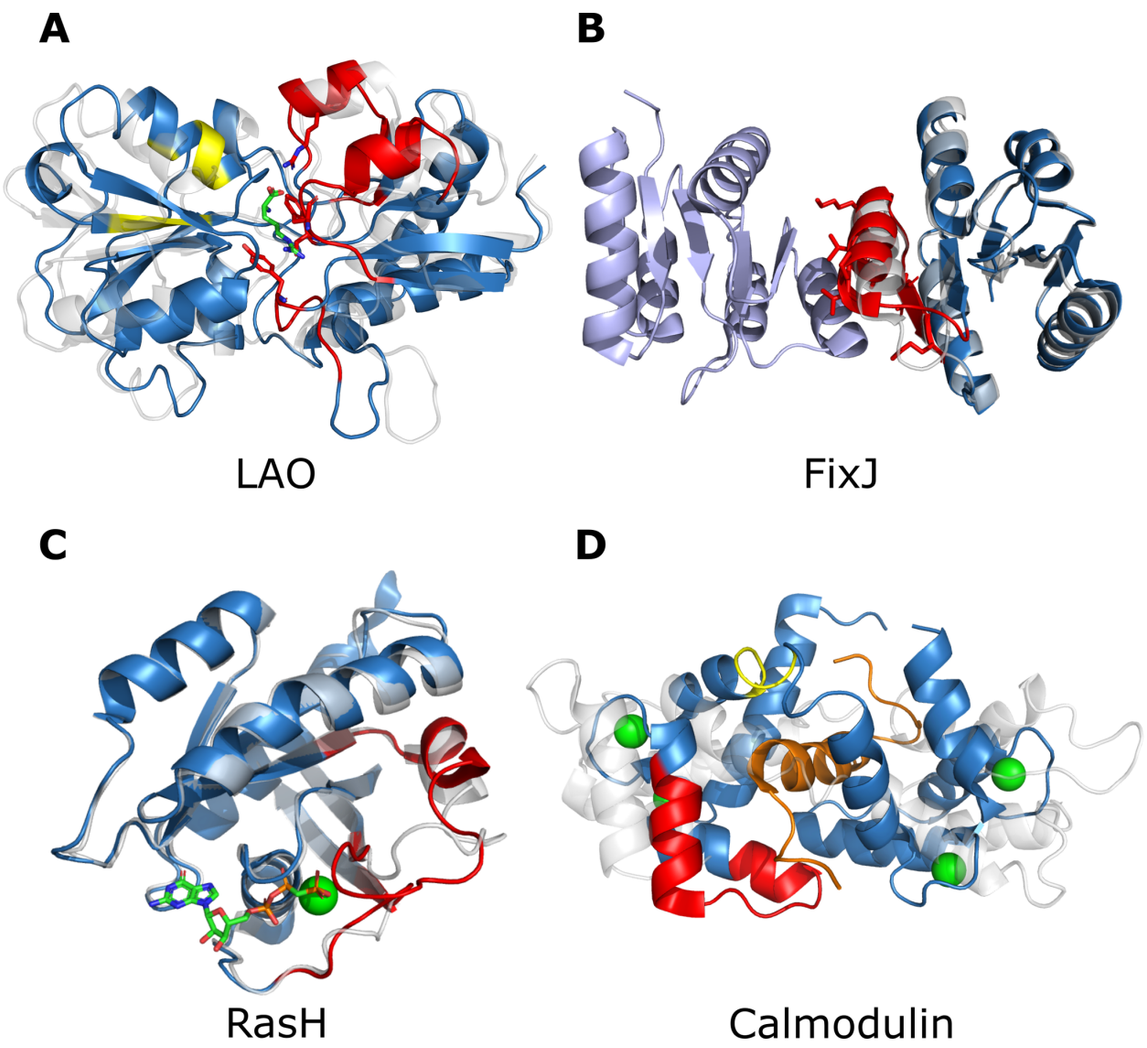

Figure 20: **Localization of highly coupled residues in four benchmark proteins.** Localization of highly coupled residues in four benchmark proteins. (A) Network analysis of highly coupled residues mapped on the structure of LAO (PDB ID: 2LAO (unbound), 1LAF (bound)). Superposition of the unbound (grey) and the bound state (blue). The substrate L-asparagine is shown as sticks. The two networks are highlighted in red and yellow. Residues known to be crucial for substrate binding and belonging to a network are shown as sticks. (B) Network of highly coupled residues (red) mapped on the structure of FixJ (PDB ID: 1DBW (unphosphorylated), 1D5W (phosphorylated)). Superposition of the unphosphorylated (grey), the phosphorylated protein (blue) and a second FixJ (light blue) belonging to the dimer. Residues critical for dimerization are shown as sticks. (C) Network of highly coupled residues (red) mapped on the structure of RasH (PDB ID: 4Q21 (GDP bound), 6Q21 (GTP bound)). Superposition of the GDP bound state (grey) and the GTP bound state (blue). The substrate is shown as sticks. Bound magnesium is depicted as green spheres. (D) Network analysis of highly coupled residues mapped on the structure of calmodulin (PDB ID: 1CFD (without Ca2+), 1CKK (with Ca2+)). Superposition of the unbound (grey) and the bound state (blue). The peptide CaMKK is shown in orange. The two networks are highlighted in red and yellow.

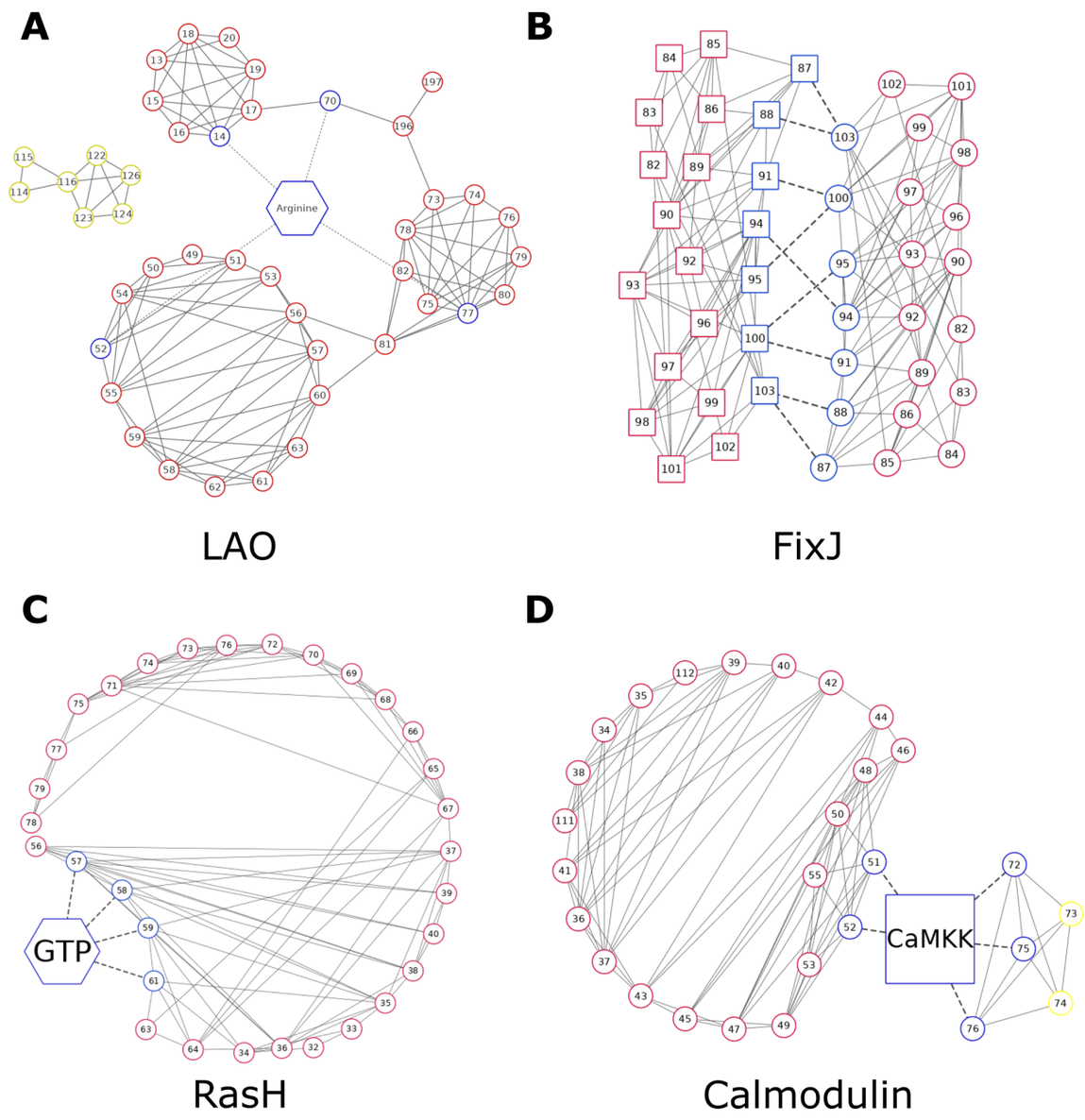

Figure 21: **Representation of residue interaction networks.** Intra-protein couplings are depicted as solid lines and dashed lines indicate substrate-binding residues or inter-protein couplings. (A) LAO possesses two interaction networks (red, yellow). Residues crucial for binding the substrate L-arginine are marked blue. (B) Residue interaction network of FixJ. Residues that are crucial for dimerization are highlighted in blue. Circles/squares distinguish coupled residues belonging to the two protomers of the dimeric complex. (C) Residue interaction network of RasH. Residues that are crucial for GTP hydrolysis are highlighted in blue. (D) Residue interaction networks (red, yellow) of calmodulin. Residues that are crucial to the binding of the peptide CaMKK are highlighted in blue.

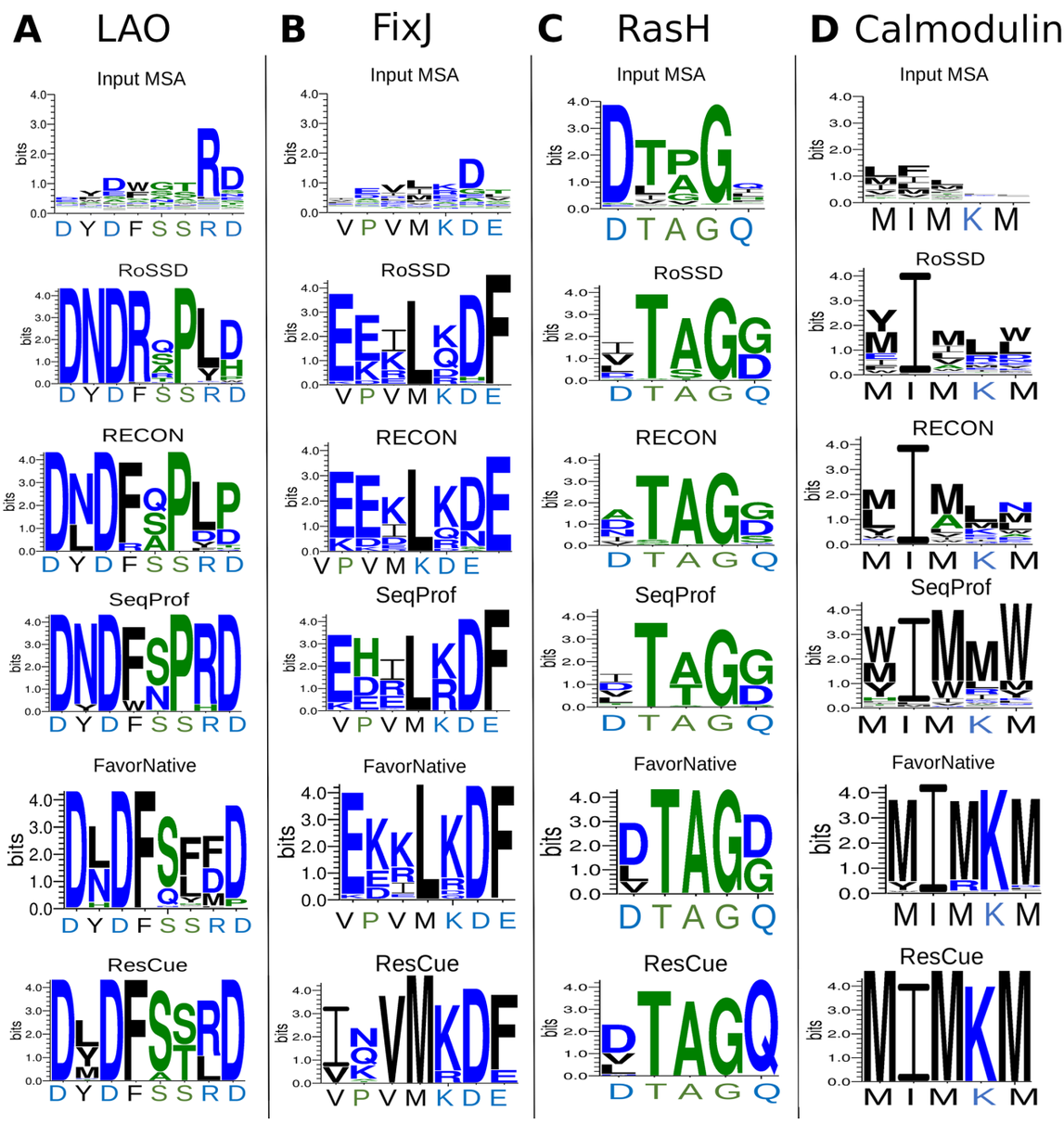

Figure 22: **Sequence logos resulting from five design protocols.** The native sequences are listed below the logos. (A) LAO binding site, eight residues. (B) FixJ dimer interface, seven residues (C) RasH binding site, five residues. (D) calmodulin-binding site, five residues.

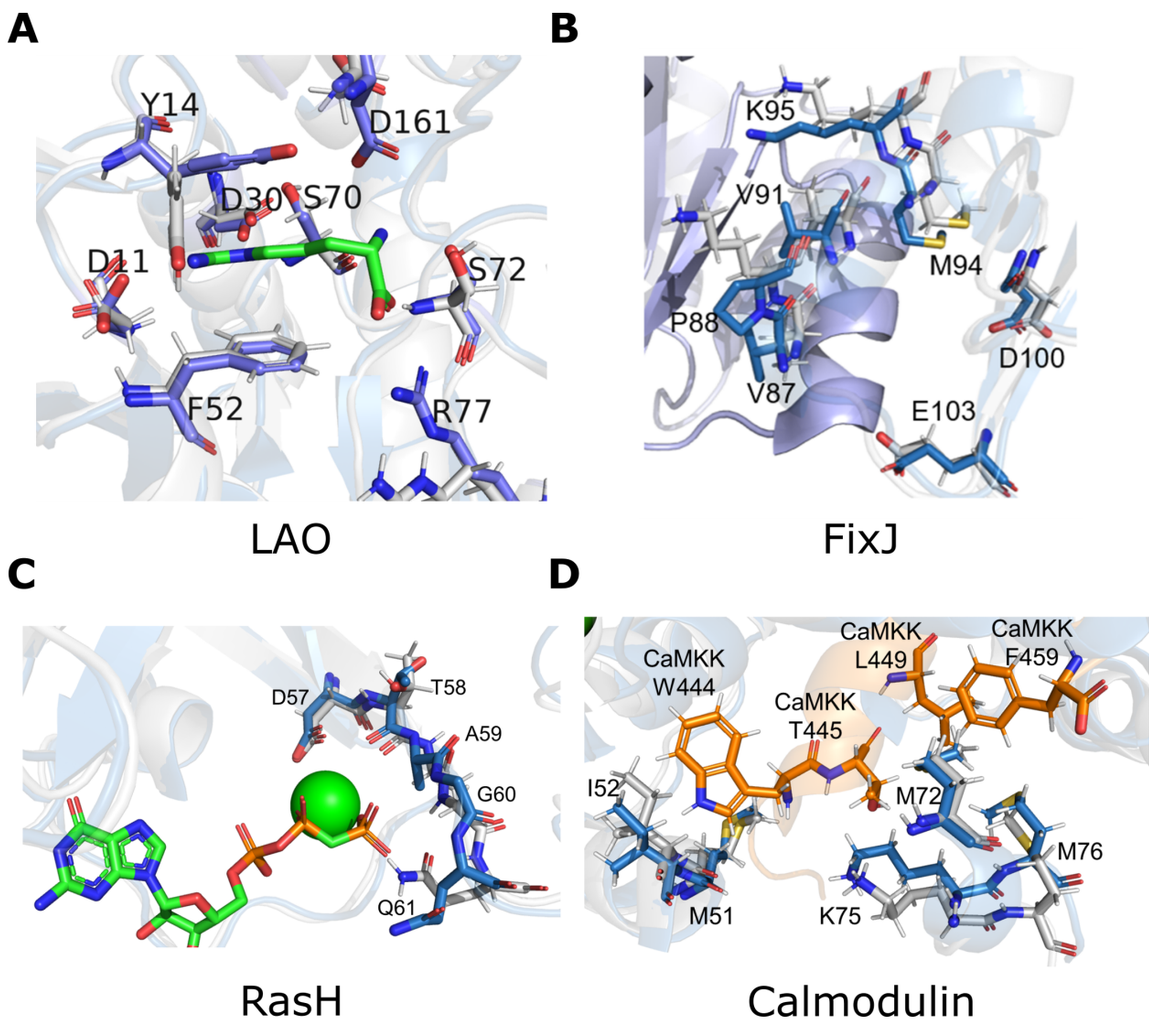

Figure 23: **3D representation of binding sites.** (A) The native structure of LAO in the closed state (PDB ID: 1LAF) is depicted in blue and a protein designed with ResCue is shown in grey. The ligand arginine is shown as green sticks. (B) FixJ in the phosphorylated state (PDB ID: 1D5W) is depicted in blue and a protein designed with ResCue is shown in grey. Residues crucial for dimerization are shown in stick representation. (C) Native structure of RasH (PDB ID: 6Q21) with bound GTP (green sticks) is depicted in blue and a protein designed with ResCue is shown in grey. (D) Native structure of calmodulin (PDB ID: 1CKK) with bound peptide CaMMK (orange sticks). The Ca2+ bound state is depicted in blue and a protein designed with ResCue is shown in grey. For all four protein designs, the ligand was not part of the starting structure.

The coupling strength cs(seq) of the chosen set of residues for each protein was calculated and visualized as a bar plot. Compared to the native sequence, the coupling strength of ResCue was on average 103±50%, followed by FavorNative with 35±55%, SeqProf with 23±58%, RECON with 20±63%, and RoSSD with 4±56% (S4 Supplement). The improvement of ResCue was statistically significant compared to all other design methods (MW p <5.0e-04) The increased sequence recovery of the ResCue protocol can therefore be attributed to the collective interaction of couplings. Full-length sequence logos of the ResCue design are provided in S3 Supplement.

7.2.8 Conformational changes induced by the phosphorylation of the FixJ receiver domain

FixJ is a two-component system crucial for the symbiotic nitrogen fixation in Sinorhizobium meliloti (David et al. 1988; Agron and Helinski 1995). The FixJ receiver domain is arranged in two domains, and phosphorylation of the conserved D54 residue induces the dimerization of the protein (Kahn and Ditta 1991; Da Re, Bertagnoli, et al. 1994; Da Re, Schumacher, et al. 1999; Birck et al. 1999). consists of 20 residues that form an interaction network located at the dimerization interface (Figs 20B and 21B). The network includes the seven residues, which are critical for dimerization, namely V87, P88, V91, M94, K95, D100 and E103 (Gouet et al. 1999). Comparing the sequence logos (Fig 22B) of the five design approaches indicates that only ResCue sampled these seven amino acids correctly and highlights a major advantage of our approach: Since only one protein structure was used during the design phase, all protocols were deprived of the interactions across the dimerization interface, which resulted in sequences probably unable to dimerize. Even SeqProf did not sample critical interface residues, because they are often less conserved (Holinski et al. 2017) and depend on the occupancy of neighboring positions, which induces couplings. Thus, considering co-evolutionary constraints during the design process leads to favorable residue combinations, even without explicit knowledge of restraints related to dimerization (Fig 23B).

7.2.9 RasH switches between two states for signal transduction

RasH is part of a signal transduction crucial for cell growth and differentiation. RasH adopts an 'off' and an 'on' state induced by a substantial conformational change in the so-called switch I region (residues 30–38) and switch II region (residues 60–76) (Milburn et al. 1990). The network connects both regions spanning residues 32–40 and 56–78. The networks highlight how the conformational shift is the product of the subtle interdependencies of protein residues (Figs 20C and 21C). The analysis of sequence logos determined for five GTP binding residues reveals that all four methods recover three residues well (T58, A59, G60) (Fig 22C). The other two (D57, Q61) were only recovered by ResCue. Superimposing the native structure with a model determined for a ResCue design confirms the correct orientation of the side chain residues (Fig 23C).

7.2.10 The conformational switch of the calcium-binding messenger protein calmodulin

Calmodulin is an intermediate calcium-binding messenger protein, playing a critical role in coupling transient Ca2+ influx to events in the cytosol and therefore the calcium signal transduction pathway (Kuboniwa et al. 1995). The protein undergoes a substantial conformational shift to bind the calmodulin-binding domain of the calmodulin-dependent protein kinase kinase (CaMKK) (Osawa et al. 1999). 25 residues (positions 34–53 and positions 72–76) are functionally relevant and the network is located at the N-terminal hydrophobic pocket anchoring T444 of the CaMKK peptide (Figs 20D and 21D). Due to the large number of functionally relevant residues, five exemplary residues were chosen with approximately similar distance; these were M51, I52, M72, K75, M76.

The five residues are part of the coupling network and crucial for the CaMKK binding. The results highlight how our co-evolutionary approach biases the design towards the native amino acids (Fig 22D). Furthermore, sequences designed with co-evolutionary information sample realistic side-chain conformations, even in the absence of CaMKK in the design process (Fig 23D).

7.2.11 Pros and cons of ResCue

In this work, we tested our hypothesis that incorporating co-evolutionary information into protein design helps Rosetta to design stable and functional proteins. For a benchmark consisting of ten difficult cases, native sequence recovery values and sequence similarity values were superior to three alternative design protocols. We imply that considering these evolutionary restraints helps to maintain function by recovering the couplings between residues.

ResCue outperforms SeqProf, because many crucial residues are not prominent in MSAs, and depend on the occupancy of neighboring positions. The design with RECON on only two conformations may underestimate its performance, since it scales well with the amount of different structures available for a given protein (Sauer et al. 2020). At the same time, the choice to use two conformations highlights how our approach is especially useful for proteins with little available additional structural information. Interestingly, ResCue compromises the energy score of the designed sequences less than RECON MSD. Since the score is meant to reflect the thermodynamic stability of the design, impacting it to a high degree would possibly result in a less stable protein. The high native sequence recovery of ResCue, the fact that native sequences are close to optimal to their structure (B. Kuhlman and D. Baker 2000), and the experimental evidence that coupling guided design can generate novel sequences with stability similar to the wild type (Socolich et al. 2005; P. Tian and Best 2017; P. Tian, Louis, et al. 2018) suggests that the slightly increased Rosetta energy is thermodynamically acceptable. A restriction of our method is the need for an MSA consisting of N homologous sequences. However, given that more than 120 million protein sequences are deposited in the UniProt Databank (UniProt Consortium 2019) compared to the 163,141 structures deposited at the PDB (Berman et al. 2000), we suppose that ResCue can be applied to a wide variety of proteins. Our results are in agreement with earlier findings (P. Tian, Louis, et al. 2018; Socolich et al. 2005) indicating that the integration of co-evolutionary information promotes stability and function in protein design.

Primary motivation of ResCue is to demonstrate, that co-evolutionary information can be leveraged for a Rosetta design algorithm that yields more natural sequences while conserving couplings that, at least in some cases, will be critical for plasticity or function, i.e. properties of the protein that are 'hidden' in a single structure/sequence pair. These couplings would encode required properties of the sequence but cannot be derived from a single conformation and in the absence of all binding partners. The Rosetta scoring function will optimize thermodynamic stability (Alford et al. 2017) of a single conformation but miss other aspects of plasticity or function. The altered scoring method with ResCue can be exploited 1) to explore a sequence space more similar to native sequences, 2) for conservative re-engineering while keeping known and unknown functions intact,

or similar 3) to design on one structural conformation while not destabilizing other possible conformations. Alternative design protocols that can indirectly inform the design process about pairwise residue couplings were discussed: These are firstly, the design on multiple conformations (RECON), and secondly, design with a sequence profile. Amongst these, ResCue generates sequences most similar to the wild type and conserves the most plausible functionally important residues and couplings. Nonetheless, future experimental studies are required to confirm the suitability of our approach for the different suggested design scenarios.

It was previously shown that selecting a small number of mutations based on conservation information in sequence alignments can improve expression rates and predict improved protein stability (Goldenzweig, Goldsmith, et al. 2016). This approach however failed to allow mutations in proximity of binding partners and co-factors to prevent activity loss in the first place. By leveraging co-evolutionary information, ResCue is going beyond the task of thermodynamic stabilization and can be exploited to re-design proteins including its functionally relevant sites, even when properties crucial for function are not encoded in one sequence/structure pair.

7.3 Methods

7.3.1 Collection of the benchmark $\text{bench}_{\text{coev}}$

When compiling the benchmark, a major goal was to represent a wide variety of small to massive conformational shifts and proteins of different length N. Proteins were collected that exhibited conformational changes with the criteria that at least two conformations of the protein were known. To prevent discrimination of non-ResCue design protocols caused by low quality protein structures, only structures with an experimental resolution of at least 3Å were accepted. The existence of at least 10 x N non-redundant, homologous sequences was confirmed and the sequences were compiled to an MSA (see below). Sequences were considered redundant if they shared more than 80% sequence identity to the native sequence. All structural models were relaxed by means of Rosetta. For all single state design protocols, both structures were used as starting points and the results were pooled.

7.3.2 GREMLIN-based co-evolution analysis

To analyze co-evolution between residues, multiple sequence alignments (MSA) were created using HHblits (E-value cutoff: 1.0e-10, Iterations: 4) (Remmert et al. 2011; Zimmermann et al. 2018). On average, the MSAs consisted of 24,700 sequences. We omitted sequences that did not cover at least 75% of the original sequence length. Additionally, we removed positions in the MSA with more than 75% gaps. The python version of GREMLIN was used to analyze each MSA and to create a tensor storing covariance values DCi,j(aax, aay) for all possible residue combinations aax, aay at all positions i, j. The coupling strengths DCi,j(aax, aay) from the Markov Random Field (MRF) tensor were used to restrain designs with ResCue. MRF-values were preferred over the derived GREMLIN (pseudo)log-likelihood values (Ovchinnikov, D. E. Kim, et al. 2016) for two reasons: Firstly, (pseudo)log-likelihood values combine coupling strength and amino acid preferences at a

certain position. In the MRF tensor, both values are listed separately. Here, we wanted to focus on coupling strengths. If desired, the user can utilize the already established FavorSequenceProfile term in addition to the ResCue coupling weight to incorporate a sequence conservation term (present in the log-likelihood). Secondly, usage of coupling strengths allows favoring correlations and penalizing anti-correlations. In contrast, the (pseudo)log-likelihood values reflect absolute coupling strengths and thus ignores this information. Eq 4 indicates how DCi,j(aax, aay) values were combined to deduce a coupling constraint cc(i) for each single residue position i:

$$cc(i) = \sum_{i,j!=i} DC_{i,j}(aa_x, aa_y) \tag{4}$$

Here, and in all other formulae, N is the length of the protein. The coupling strength cs(seq) of a given sequence seq was determined by adding the N cc(i) values:

$$cs(seq) = \sum_{i=1}^{N} cc(i) \tag{5}$$

To assess the coupling recovery of a designed protein prot, the coupling recovery score crs(prot) was deduced from the cs values related to the designed and native sequences seqDesign and seqNative:

$$crs(prot) = \frac{cs(seq_{Design})}{cs(seq_{Native})} \tag{6}$$

7.3.3 Assessment of native sequence recovery and sequence similarity

The native sequence recovery nsr(seqDesign) of a sequence seqDesign is the fraction of residues seqDesign[i] that match the corresponding native residues seqNative[i]:

$$nsr(seq_{Design}) = \frac{1}{N} \sum_{i=1}^{N} ident(seq_{Design}(i), seq_{Native}(i) \tag{7}$$

The binary function ident() determines the identity of two residues aa_k and aa_l:

$$ident(aa_k, aa_l) = \begin{cases} 1, & if aa_k == aa_l \\ 0, & else \end{cases} \tag{8}$$

Analogously, sequence similarity seqsim(seqDesign) was computed:

$$seqsim(seq_{Design}) = \frac{1}{N} \sum_{i=1}^{N} BLOSUM62(seq_{Design}(i), seq_{Native}(i)) \tag{9}$$

Scores for the similarity of corresponding residue pairs ware taken from the BLOSUM62 matrix (S. Henikoff and J. G. Henikoff 1992). All computations were performed using Biopython (Cock et al. 2009).

7.3.4 Protein design with ROSETTA

he ROSETTA software suite was used for all different design approaches. RosettaScript XML files and commands can be found in S2 Supplement. Designs with no additional constraints (RoSSD) were performed by one round of fixed backbone rotamer optimization followed by repacking. RECON multistate designs utilized four rounds of fixed backbone design and a convergence step, as described in (Sevy, Jacobs, et al. 2015). For each design, a PSSM was created by means of PSI-BLAST (S. F. Altschul and Koonin 1998). The PSSM was needed for the SeqProf RosettaScripts mover and the MSAprot served GREMLIN to deduce for each design the coupling tensor required for the ResCue Mover. At least 100 designs were generated for each protein in the benchmark and each approach. The resulting designs were scored with the ref 2015 Rosetta energy function.

7.3.5 Network analysis of highly coupled residues

Regions of highly coupled residues were analyzed by using a similar technique, as described in (Jeon et al. 2011). First, for each residue i of the native sequence the cc(i) score (Eq 4) was determined. Then, a sliding window (window size of ten, step size of one) was used to identify regions containing highly co-evolving residues indirectly connected by slightly weaker coupled residues. To analyze regions with the highest co-evolutionary significance, we took the residues and further analyzed how exactly they are coupled with each other. The networks were visualized with Cytoscape (Shannon et al. 2003) and mapped on the protein structure by means of PyMOL. Sequence logos were created with WebLogo (Crooks et al. 2004).

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

This chapter has been submitted under (Schmitz et al., 2021).

8.1 Introduction

As of 2019, over 570 antibody drugs are in development with a substantial increase of late-stage antibody development over the past decade (Kaplon and Reichert 2021; Kaplon and Reichert 2019). Historically, antibody reagents were generated using cells from an animal source such as rabbit (Steinberger et al. 2000), chicken (Tsurushita, Park, et al. 2004), and more prominently murine model organisms (Gillies, Lo, and Wesolowski 1989; Bonwick et al. 1996). The downside of using antibodies with non-human origin is the elicitation of anti-drug-antibodies (ADA) in human patients (Gillies, Lo, and Wesolowski 1989; Bonwick et al. 1996; Nechansky 2010b). High titers of ADA responses usually result in reduced efficacy of the antibody drug by blocking the antigen binding site or by faster depletion of antibody drugs in the bloodstream (Holgate and M. P. Baker 2009). The reasons for the ADA response in patients are multi-factorial but often comprise sequence patterns foreign to the human system (Harding et al. 2010). This observation gave rise to humanization techniques resulting in engineered antibodies with non-human sequences interspersed among human-derived antibody segments (T. D. Jones et al. 2016; Parren, Paul J Carter, and Andreas Plückthun 2017). Here, we introduce a method based on the human-likeness (HL) assessment method IgReconstruct (Schmitz et al. 2020), and expand upon it to support the structural design of human-like antibodies. A possible application of our method is supporting the development of antibody biologics that appear human-like early in the development process. It also may be useful to simulate a possible human immune response for a particular pathogen and specific to the human donors on which the immune repertoire is based.

Essential for HL assessment are large quantities of observed human antibodies sequences, the so-called adaptive immune receptor repertoires (Schmitz et al. 2020; Wollacott et al. 2019; Seeliger 2013; Gao et al. 2013; Lazar et al. 2007). 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 (DeWitt et al. 2016; B. Briney et al. 2019; Soto, Bombardi, et al. 2019). Despite the high diversity, a small sequence overlap between individual blood donors exists (Soto, Bombardi, et al. 2019; B. Briney et al. 2019). The major mechanism of antibody diversification is comprised of 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) at its disposal. 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, with the recombination events alone producing a diversity of 10^6 sequences (Charles A Janeway et al. 2001). 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 (Teng and Papavasiliou 2007). The resulting diversity of human antibody repertoires has been estimated to exceed 10^{12} unique B-cell receptors (Charles A Janeway et al. 2001). It was shown that the human-likeness of the IgG isotype of antibodies can be modeled by assessing the single nucleotide frequency for each germline gene in the observed antibody space (Kovaltsuk et al. 2018). 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 human-likeness as was previously described (Schmitz et al. 2020), we choose the most similar V, J, and CDRH3 cluster center to assess the human-likeness of the CDRH3 region. Human-likeness 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 (Soto, Bombardi, et al. 2019). Taken together, modern sequencing methods have explored a combined sequence space of $5x10^8$ 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 from work in the Human Immunome Project, which aims to comprehensively catalog the human B- and T-cell sequence spaces (Wooden and Koff 2018). It could be shown that even though the sequence commonality between human-blood donors is greater than anticipated, the overall sequence overlap remains small (¡¡ 1% of heavy chain clonotypes) (Soto, Bombardi, et al. 2019; B. Briney et al. 2019). This low commonality is primarily a result of the high sequence diversity, and the main cause for antibody diversity is the high variability of the CDRH3 region. To accommodate for the small ratio of observed to expected antibody space, we mathematically calculate an enlarged human amino acid space from nucleotide frequencies. We hypothesize that there is additional information in nucleotide sequences that can inform the antibody space for the following reasons:

The genetic code is degenerate, which means 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. Human-likeness was previously described as independent single nucleotide

observations (Schmitz et al. 2020), suggesting that the antibody maturation process is an 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 informs 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 what 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 Rosetta (Leaver-Fay, Tyka, et al. 2011) to generate antibodies that are antigen-specific and thermodynamically stable while still maintaining human-likeness.

The computational structural design package Rosetta, allows structural sequence design of proteins. Computational design with Rosetta is mainly achieved by its scoring function, that evaluates the sequence grafted onto a protein conformation. The Rosetta scoring function comprises the weighted sum of physical, and knowledge-based potentials (Alford et al. 2017) to evaluate the conformation and sequence of proteins. 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, like alanine or site-directed mutagenesis, hydrogen–deuterium exchange mass spectrometry (HDX) or also HDX-NMR, NMR chemical shift perturbations, low-resolution cryo-EM, and chemical cross-linking data (Thornburg et al. 2013; Sivasubramanian, Chao, et al. 2006). 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) (Dunbar et al. 2014). Choosing human antibodies provides us with HL of human antibody sequences, which serves as reference for benchmark purposes. Abs designed without human restraints is expected to decrease in their HL and exhibit reduced wild-type (WT) sequence identity. Thus, Rosetta designed Ab 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 Abs 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 the sequence space down, and create antibody sequences which 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.

8.2 Results

The IgReconstruct method assesses human-likeness (HL) via single nucleotide frequency statistics from immunome repertoires (Schmitz et al. 2020), and has been compared with 10 similar approaches (Prihoda et al. n.d.). 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 can not be applied to the CDRH3 since its genes either cannot or can only be partially assigned. Instead, the immunome repertoire is clustered by V, J, and CDRH3 domains. This enables us to assign nucleotide frequencies to the CDRH3 by using the sequence of the cluster center instead of a germline gene. The following chapters 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.

8.2.1 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 24a). IgReconstruct (Schmitz et al. 2020) is a method to assess HL as nucleotide frequency profiles for each germline gene and position, and for each CDRH3 regions (length dependent). (Figure 24b). The next subchapter will describe how we expand upon this approach by creating clustered frequency profiles for genes and CDRH3 regions. The frequency profiles for V, J, and heavy chain CDR3 region (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 24c). The high diversity of the CDRH3 gives rise to the low commonality between human immunome repertoires (Soto, Bombardi, et al. 2019; B. Briney et al. 2019). 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} (Charles A Janeway et al. 2001). This study therefore suggests Bayesian statistic to model amino acid frequencies from the observed nucleotide frequencies (Equation 10). Here, it is assumed, that all positions of the antibody variable region has 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 for the human antibody space. Thus, the Bayesian statistics (Equation 10) can be simplified by making the assumption 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 11-12 take the number of different triplets (trpl) that encode for a specific amino acid into account as normalization parameter. For each amino acid probability p(aa—trpl) with a given distribution of nucleotide frequencies (trpl), a substitution score sij is calculated. The substitution score represents statistical significance of the calculated frequencies for each position and will be used as Rosetta restraint for HL antibody design. The calculation of the substitution score (Equation 13, Figure 24e) has been adapted from the description for PSI-BLAST (S. F. Altschul 1991; S. F. Altschul, Madden, et al. 1997; Stephen F. Altschul et al. 2009). 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 Section 11.10.4). The tables of amino acid frequencies and substitution scores

are then converted into a PSI-BLAST compatible ASCII file which can be parsed by Rosetta for further design (Figure 24d).

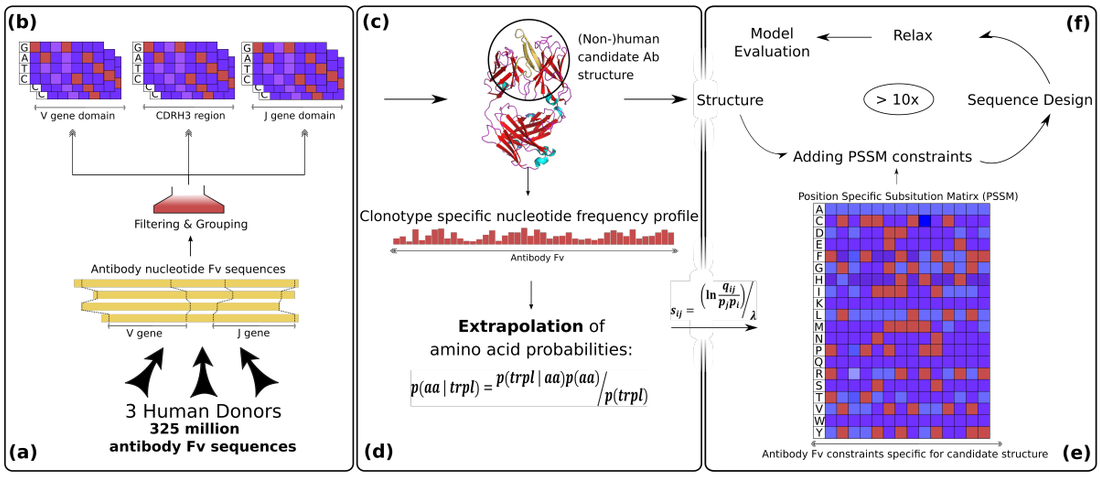

Figure 24: **From immunome repertoire processing, to statistical modeling of an amino acid sequence space, to structural human-like antibody design.** First, V, CDRH3, and J domains are extracted from the immunome repertoires (a) to generate position specific nucleotide frequency profiles for each domain (b). The structure of an antibody that is to be re-designed for increased human-likeness (c) is assigned a unique nucleotide frequency profile. A Bayesian method extrapolates possible amino acid frequencies for the antibody (d) to create amino acid substitution scores (e). The substitution scores can be used to guide the Rosetta structural design to more human-like antibodies, narrowing the human sequence space to structurally viable solutions.

8.2.2 Extending the nucleotide human-likeness metric with a clustering algorithm

Unlike the fragment 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 to assess HL via germline gene specific nucleotide frequencies. Consequently, the CDRH3 was excluded for HL calculations in our previous study (Schmitz et al. 2020). To enable CDRH3 HL assessment, we extended the positions specific substitution matrix (PSSM) generation method by implementing a basic clustering approach capable of processing large datasets quickly. 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 can be subdivided to four steps and 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 25a). 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 25b). 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 CDHR3 domain due to the distinct sequence diversity of the regions. For V, and J domains a sequence identity of 90% was used, whereas the CDRH3 clusters had a sequence identity cutoff of 30% for the following reasons: The sequence identity cutoff was determined under consideration of the sequence diversity of V, J, and CDRH3 regions, number of final clusters, and their size. The higher the sequence identity cutoff, the more and 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, to create position specific nucleotide frequencies for HL assessment. The cutoffs of 90% (V, and J domains) and 30% (CDRH3) led to 14,638 V, 390 J, and 411 CDRH3 clusters. 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.

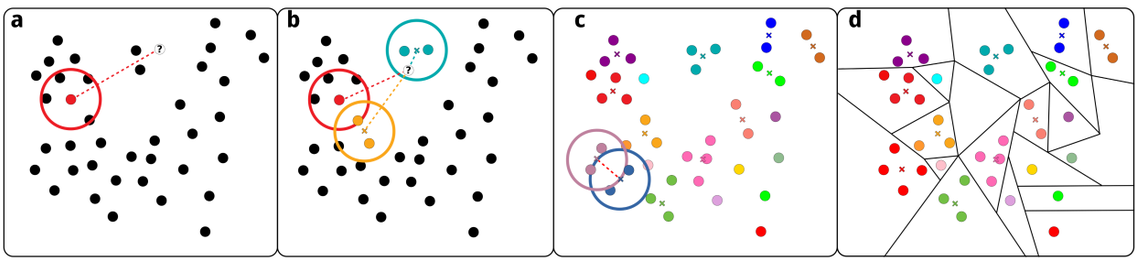

Figure 25: **Schematic of fast immunome repertoire clustering.** Each V, J, and CDRH3 sequence not assigned to a cluster in the repertoire is represented as a black dot. The arbitrary first sequence (red dot, a) is assigned to the first cluster (red circle, a). New sequences (?) are processed in random order and the sequence identity 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, clusters that cluster centers smaller than the sequence identity threshold are merged (c). Each cluster represents a unique PSSM after processing the repertoire once (d).

8.2.3 The Rosetta human-like antibody design protocol

Major change made to the IgReconstruct method (Schmitz et al. 2020) is: 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 is assigned when creating HL profiles. To create amino acid restraints that can be interpreted by RosettaLeaver-Fay, Tyka, et al. 2011, we calculate amino acid frequencies from cluster nucleotide frequencies using Bayesian statistics. In this study, we benchmark Ab designed with Rosetta that were created with the Bayesian antibody space, and without any HL restraints. From hereon out, we refer to proteins that were designed with the Rosetta suite as decoys. The antibody space is calculated

using original PSSMs, and clustered PSSMs. We then compare the 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 our 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.

A set of 27 antibody crystal structures was curated that is a) of human origin and b) high resolution (better than 2 Å), and c) available as complex bound to its antigen. For each of the antibody structures, we created a Bayesian PSSM for heavy and light chain separately. PSSM restraints were added to Rosetta in form of a PSI-Blast formatted ASCII PSSM file (Stephen F. Altschul et al. 2009). During Rosetta design, each mutation is then re-evaluated for increased human-likeness by either favoring a mutation (positive substitution score), or disfavoring 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 avoid the difference in number of mutations between control and designs to 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. From hereon out, we refer to this control group as "native". 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 74 demonstrates the close correlation of sequence identities between native, and human-like designs. Thus, for each Rosetta design, a control design can be found with comparable mutation rate.

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

To prove, that the Rosetta restraint were balanced correctly, the Rosetta energy of control and human-like decoys was compared with each other. Rosetta Energy Units (REU) are a measure for thermodynamic stability of a protein (complex) (Alford et al. 2017). The REU score can be used to compare different protein conformations, and estimate mutational changes of thermodynamic stability. The more negative the score, the higher the predicted stability. Here, we compare the REU scores of the Rosetta decoys, with the REU score of the WT crystal structure. Thus, a score smaller than 0 means an improvement compared to the WT structure. The more negative the reported results, the greater the improvement of predicted stability of the protein compared to the WT.

On average, the Rosetta energy was improved during design, compared to the relaxed wild-type

structure, by -142.8 ± 25.0 (control), or -115.6 ± 26.6 (native), or -82.7 ± 24.4 (original), or -68.7 ± 22.3 (clustered) REU. REU scores of the decoys that were restrained by original or clustered PSSMs are more positive compared to the control (e.g -142.8 ± 25.0 of the control vs. -68.7 ± 22.3 for clustered), which is expected due to the additional restraints added and is an indicator, 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 less degrees of freedom to optimize the protein. Overall, the design protocols improved the Rosetta energy compared to the WT energy in all cases (Figure 26a). The binding energy, normalized by its interface size, retained original values, suggesting a conserved specific antibody binding to its antigen (Figure 26b). We conclude, that the chosen weights (see Supplement Section 11.10.3) for HL restraints can be considered appropriate for the design task.

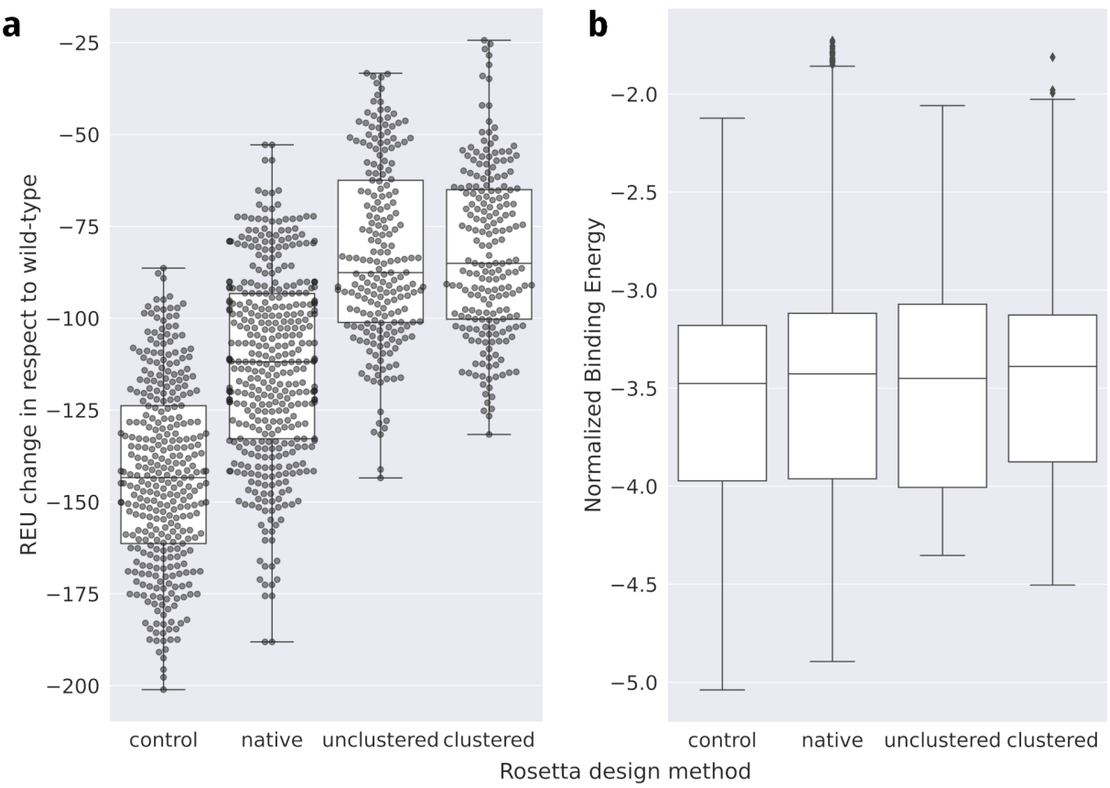

Figure 26: **Rosetta energy and binding energy of the human antibody set.** The total Rosetta score change relative to the relaxed wild type score (a) and the interface energy normalized by the interface size (b).

8.2.5 Improved human wild-type antibody sequence recovery for the V and J domain

It is hypothesized 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 more 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 region, and the CDRH3 region separately. When compared to the control group, the heavy chain sequence recovery increases from 74.5 ± 6.3% (control) to 84.8 ± 3.8%

(original), or 85.5 ± 4.6% (clustered). Similarly, the light chain sequence recovery is increased from 77.1 ± 7.2% (control) to 85.6 ± 4.3% (original), or 85.5 ± 4.6% (clustered) (Figure 27a). In contrast to the increased sequence recovery of the V and J regions, the CDRH3 sequence recovery does not change significantly from 45.6 ± 11.1% (control) to 45.0 ± 13.3% (original). With a slight decrease of sequence recovery to 40.6 ± 10.1%, the clustered human-like design approach appears to influence the average sequence recovery. 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 repertoire.

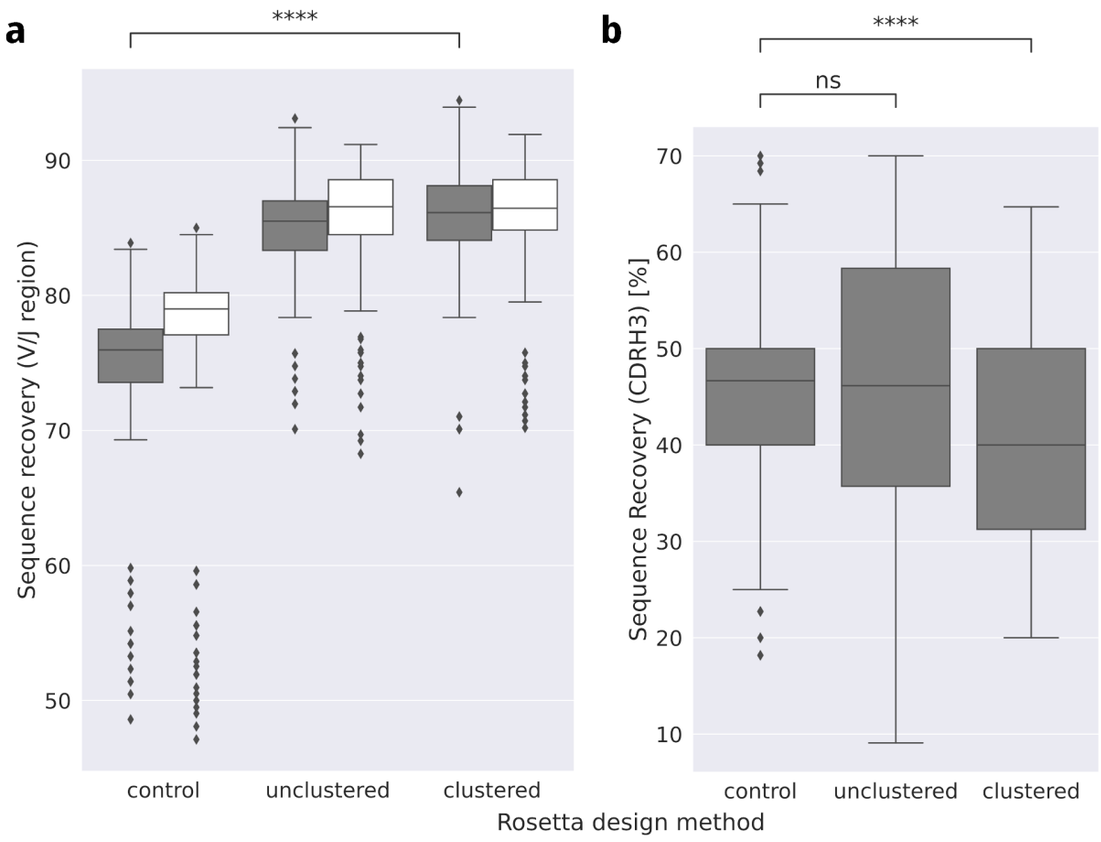

Figure 27: **Wild-type sequence recovery rates of the antibody after Rosetta design.** The sequence recovery for the V and J regions is increased for clustered and original design compared to the un-restrained control design (a). The sequence recovery of the CDRH3 region does not substantially change for the original design but is reduced when clustered design is applied (b). Heavy chains (gray), and light chains (white). Statistical annotations with the Mann-Whitney significance test (****: $p < 10e-4$; ns: not significant)

8.2.6 Increased human-likeness across the antibody framework region

Similarly, to the observed increased sequence recovery in the V and J regions of the antibody, a substantial increase of HL was observed. 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 28a 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 ± 2.8% (original) and 71.3 ± 2.7% (clustered), the human-like designs both increased substantially to 86.0 ± 2.8% (original) and 87.9 ± 2.6% (clustered). Similarly, the light chain human-likeness increased from 71.0 ± 2.8% to 86.3 ± 2.2% (original), or from 71.8 ± 2.8% to 86.8 ± 2.1% (clustered).

As a reminder, the native control group are those 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 a HL decoys. When designing an antibody in Rosetta without HL restraints, a decrease of HL is expected as the number of mutations increases. The native group does account for the different number of mutations to not artificially render the performance of the HL design protocol greater than it is. When 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 28b). In the clustered design scenario (Figure 28c) four Abs 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.

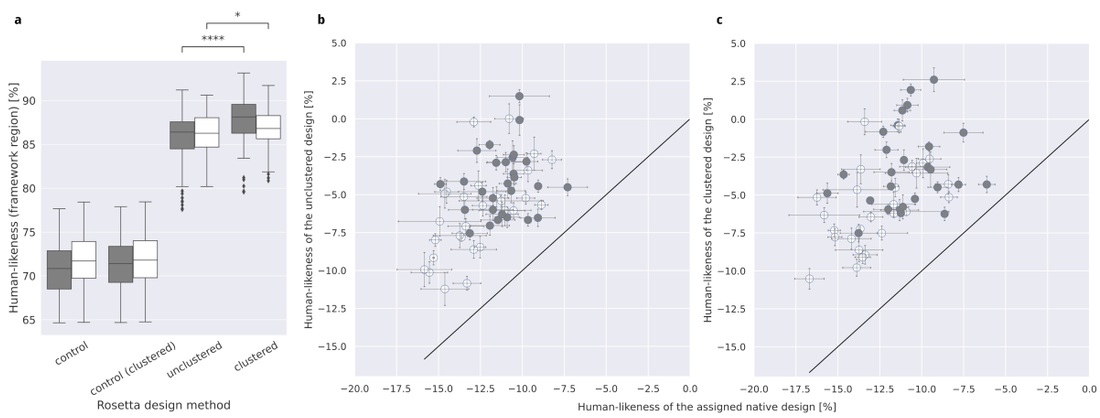

Figure 28: **Human likeness of the V and J domains after Rosetta design.** The control group was scored using original and clustered PSSMs, and is significantly lower than the human-like designs, original and clustered (a). The change of human-likeness in respect to the native designs (unrestrained Rosetta design with similar sequence identity to the wild-type used as baseline), shows an improvement of human-likeness for original (b) as well as clustered PSSM designs (c) for all antibodies in the dataset. Each data point represents a unique PDB ID. Heavy chains (gray), and light chains (white). Statistical annotations with the Mann-Whitney significance test (****: $p < 10e-4$; *: $1.00e-02 < p <= 5.00e-02$).

8.2.7 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 a HL assessment method based on single nucleotide frequencies of the observed antibody space (Schmitz et al. 2020). The untemplated and diverse character of the CDRH3 requires an alternative approach to address CDRH3 human-likeness. 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 (Soto, Bombardi, et al. 2019; B. Briney et al. 2019), and the relatively low number of observed sequences per individual (10^8 per donor versus $> 10^{12}$). To address the difficult task of defining human-likeness 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 CDHR3 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 native group (Figure 29a). In contrast, eight antibodies in our benchmark designed with clustered PSSMs exhibited a positive change (¿ 3.5%) of HL compared to their native group (Figure 29b). Structures with an increased HL compared to their natives were 1n0x (8.0 ± 0.7%), 2yc1 (3.5 ± 0.0 %), 3l5x (5.0 ± 0.7 %), 4hs6 (4.0 ± 0.6 %), 4ioi (4.3 ± 1.4%), 4j6r (3.5 ± 0.8%), 5f9o (6.1 ± 0.7%), or 5xku (6.5 ± 0.9%). Supplementary Table 8 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 used antibody repertoires were collected from healthy blood donors, it cannot be expected that the PSSMs carry the information needed to generate mature, highly specific antibody sequences in all 27 cases. Figure 29c 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 distinct sharp the frequency distribution is. Generally speaking, the more diverse a sequence set, the flatter the observed frequency distribution. Here, the HL of the CDRH3 never exceeded 40% for clustered PSSMs (Figure 29b), and less then 32.5% for original PSSMs (Figure 29a). We therefore consider the cutoff to determine an improvement in HL of 3.5% reasonable.

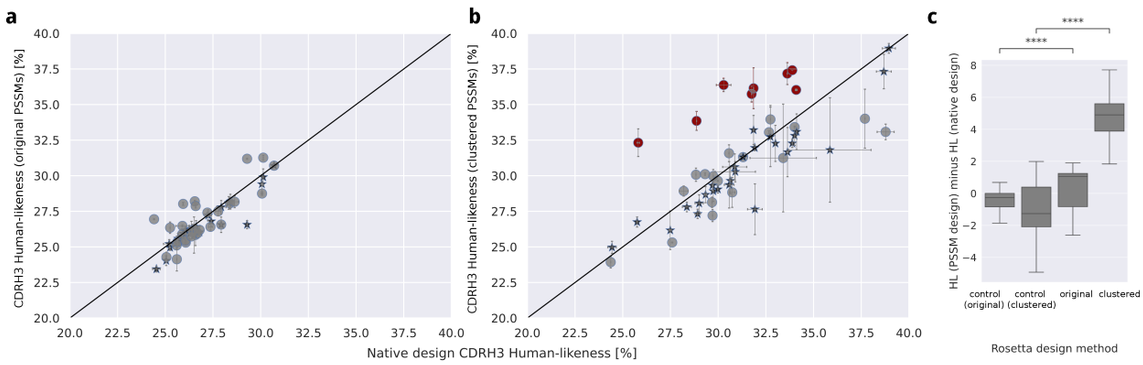

Figure 29: **Human-likeness (HL) of the CDRH3 compared to Rosetta designs with limited number of mutations (native).** HL of Rosetta designs using original (a, circles) and clustered (b, circles) PSSMs. The HL of control designs (star) was assessed with clustered and oringial PSSMs respectively. Each datapoint corresponds to a unique PDBID. In contrast to the original designs, 8 out of 27 PDBs using clustered PSSMs could improve HL when compared to its control and native designs (red). When selecting the seven antibodies with improved CDRH3 human-likeness, the significance of the change becomes visible for HL scores using clustered and original PSSMs (c). Statistical annotations with the Mann-Whitney significance test (****: $p < 10e-4$).

8.3 Discussion

Valuable research grade monoclonal antibodies are often derived from non-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 to create 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 (Soto, Bombardi, et al. 2019; B. Briney et al. 2019) of sequence repertoires is the variable CDRH3 region of the antibody, that enables specific binding to a wide variety of antigens. To model human-likeness despite these difficulties, the previously described IgReconstruct (Schmitz et al. 2020) method was improved. Amino acid frequency profiles of clustered antibody repertories were modeled from nucleotide sequences using Bayesian statistics. It is hypothesized 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 out 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 re-design existing antibodies for human-likeness. Re-design of

non-human antibodies was not conducted in the study and comes with additional challenges:

It has been shown, that human germline genes can be assigned to non-human antibody species (Schmitz et al. 2020) – which is the foundation of our method. However, the alignments naturally can be of low quality due to their low HL and low germline identity. Consequently, it can be expected that a greater number of mutations is required to humanize an antibody of non-human origin. To address humanization of non-human antibodies, decisions must be made on a case-to-case basis that include 1) which human germline gene combination should be used that optimally supports the CDRH3 conformation , 2) which areas of the variable region should be protected from mutations. 3) Evaluation of the humanness that may depend on specific project aims and include experimental evidence. Non-human antibodies are therefore ill-suited for an initial benchmark of our study as they fail to provide a HL baseline and human WT sequence for comparison. In this study, we suggest to exploit nucleotide frequencies 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 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 (Soto, Bombardi, et al. 2019; B. Briney et al. 2019). This finding implies, that antibodies specific to the same antigen can differ in its sequence significantly between different humans. 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 human-likeness on the other hand could be increased in 8 out 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 that comprise antigens from HIV, HPC, auto-antibodies, dengue virus, and more. In the cases of increased CDRH3 HL and decreased sequence identity to the wild-type, it can be assumed that the immune response of the blood donor(s) would appear differently than the antibody deposited in the PDB. A design scenario with greater practical relevance, which is less tailored for benchmarking of aggravated difficulty, e.g., starting from a non-human lead antibody, is likely to yield much more significant changes in human-likeness. Further applications may include established Rosetta design protocols that are available as RosettaScripts (Fleishman et al. 2011), and combined with our PSSMs. The RosettaScript used for this study (see Supplement Section 11.10.3) 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 (Sevy, N. C. Wu, et al. 2019; Sevy, Jacobs, et al. 2015). Another possible use case is the de-novo design with RostetaAntibodyDesign (RabD) (Adolf-Bryfogle, Kalyuzhniy, et al. 2018) 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 HL.

8.4 Methods

8.4.1 Generation of Single Nucleotide Frequency (SNF) profiles

Previous work grafted Single Nucleotide Frequency (SNF) profiles onto amino acid antibody sequences using NGS sequenced immunome repertoires (Schmitz et al. 2020). This enables the assessment of human-likeness 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 is used, but instead of pooling all sequences depending on germline gene and loop length, the immunome repertoires are clustered based on minimal sequence identity. Separation by sequence identity allows 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, Bombardi, et al. 2019. SNF profiles with an identity cutoff of 90% of V and J domains and 30% for the CDRH3 were chosen for all experiments as a compromise between number of clusters and cluster sizes.

8.4.2 Bayesian approach to model the human amino acid sequence space

We deduce amino acid substitution scores from SNF profiles. We hypothesize, that silent mutations and the degeneracy of the genetic code contain additional information which allow 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 10). We simplify Equation 10 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.

$$\Delta p(aa|trpl) = \frac{p(trpl|aa)p(aa)}{p(trpl)} \tag{10}$$

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 Opseudo, 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

Table 2: 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.

	Position1	Position2	Position 3
Triplet1	A	G	T
Triplet2	A	G	C
Triplet3	T	C	T
Triplet4	T	C	C
Triplet5	T	C	A
Triplet6	T	C	G
T_{unique}	A, T	G, C	T, C, A, G

unique nucleotides were used to infer the triplet frequency for a specific amino acid. Table 8 exemplary shows for serine. The resulting observations are independent of the varying number of triplets that encode an amino acid.

Opseudo is ultimately the sum of SNF observations of all unique nucleotides and resembles to the frequency of a specific amino acid. The probability to observe a specific amino acid resn at position resi, given the triplet observations from our SNF profile, is described in equation 11 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.

$$p(resn|trpl) = \frac{\frac{\sum_{nt}^{T_{unique}(resn)} O_{pseudo}(resi,nt)}{GCD}}{\frac{\sum_{aa}\sum_{nt}^{T_{unique}(aa)} O_{pseudo}(resi,nt)}{GCD}} \tag{11}$$

The GCD cancels out which leads us to our final Equation 12. We expect these amino acid pseudo-observations to approximate the bayesian human amino acid sequence space.

$$p(aa, resi) = \frac{\sum_{nt}^{T_{unique(resn)}} O_{pseudo}(resi, nt)}{\sum_{aa}\sum_{nt}^{T_{unqiue}(aa)} O_{pseudo}(resi, nt)} \tag{12}$$

8.4.3 Generation of a position specific substitution matrix

We use the amino acid probabilities calculated in Equation 12 to assemble position specific frequency matrices antibody variable regions. The substitution matrices are deduced from SNF profiles which are individually generated for each sequence depending on its germline gene rearrangement (Schmitz et al. 2020). We then convert these frequencies into PSI-Blast formatted position specific substitution matrices for amino acids (Stephen F. Altschul et al. 2009). The method to calculate substitution matrices from probabilities was described for Blast applications (S. F. Altschul 1991). We adopted the mathematical Equation 13 for substitution score calculation and applied it on each germline gene dependent and CDRH3 loop length dependent amino acid probability matrix p(aa, resi) (Equation 12).

$$s_{ij} = \frac{\left(ln\frac{q_{ij}}{p_i p_j}\right)}{\lambda} \tag{13}$$

The target frequency q_{ij} which describes the probability to mutate amino acid i to residue j, and background probabilities for each amino acid i, and j (p_i and p_j). The scaling parameter lambda was determined for each probability matrix individually by optimizing the spearman correlation

between substitution score and nucleotide human-likeness score PFM_{VJ} (Schmitz et al. 2020). 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 75 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.

8.4.4 Design of antibody structures with and without substitution score constraints

Crystal structures were obtained from the protein data bank (Dunbar et al. 2014), removed solvent and all non-protein and duplicate chains. For structural sequence design Rosetta was used (Leaver-Fay, Tyka, et al. 2011). First, we constraint relaxed all prepared structures succeeded by a sequence conversion to Alanine of the variable region (Fv). Using Rosetta we redesigned the Fv region with and without substitution scores, in apo and holo state if available. To add the constraints to Rosetta we used our PSI-Blast formatted PSSM in combination with the FavorSequenceProfileMover 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.

8.4.5 Human-likeness and SNF alignment generation for the dataset

Human-likeness was calculated as previously described as PFM_{VJ} is a direct measure of observed nucleotide frequencies (Schmitz et al. 2020). HL values were reported for the V and J domain as the average of nucleotide frequencies (PFMVJ) and adopted for the CDRH3 region analogously (PFM_{CDRH3}). SNF matrices were generated by from the wild-type crystal structures using IgReconstruct and the clustered version of the IgReconstruct algorithm. SNF matrices were then used to create the substitution scores/PSSMs as human-likeness 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.

8.5 Availability

The IgReconstruct webservice has been extended to output clustered and original PSSMs which can directly be used in combination with Rosetta scripts (see Supplementary Section 11.10.3). The IgReconstruct webservice is available at http://www.meilerlab.org/index.php/servers/IgReconstruct

9 Assessment and optimization of antibody expressability using Long-Short Term Memory and structural design

This chapter has been submitted to Prot. Eng. Des. Sel. (Schmitz et al., 2021).

9.1 Introduction

Although the market for monoclonal antibodies for both therapeutic use and in vaccines has increased dramatically over the past three decades (R.-M. Lu et al. 2020), there have been impediments to transitioning functional human-like antibodies to become therapeutics. The development and manufacturing of antibodies requires high concentration bioprocesses in unphysiological conditions. However, natural antibodies or antibodies derived therefrom do not experience evolutionary pressures to perform well in these altered conditions conditions of scientific experiments or industrial productions. As a consequence, the results of many antibody discovery campaigns are difficult-to-express (DTE) antibodies that yield low concentrations (Pybus, James, et al. 2014) and low product quality (Johari et al. 2015), which can compromise research efforts and increase costs when transitioned into industrial scale production. Factors that increase the likelihood of antibodies to be DTE include translation (Kallehauge et al. 2017), aggregation (Hasegawa et al. 2017), degredation (Johari et al. 2015), and folding (Jung and Alt 2004) problems. Single-point-mutations introduced during engineering efforts may improve or exacerbate expression of DTE antibodies(5), but can also be detrimental for antibody specificity (Iba et al. 1998; Winkler et al. 2000), and affinity (Wojcikiewicz and Luo 1998; Schildbach et al. 1993).

Previous studies have explored computational predictions of solubility. Tools like PROSO II (Smialowski et al. 2012), CamSol (Sormanni, Aprile, and Vendruscolo 2015), SolPro (Magnan, Randall, and Baldi 2009) are often based on a form of Machine Learning, similar to Support Vector Machines (Z. R. Yang 2004), and the manual grafting of sequence feature sets that might influence solubility. The recently described predictors DeepSol (Khurana et al. 2018) and SKADE (Raimondi et al. 2020) made use of Deep Learning and SoluProt with a Gradient Boosting Machine (Hon et al. 2021). These tools address the related challenge of protein solubility but usually are not antibody specific and are designed for the Escherichia coli expression system. The solubility definition can include non-expressing antibodies, which renders the quest for solubility and expressability predictors scientifically and technically comparable. Like DeepSol and SKADE, this study uses a Deep Learning approach to create an expressability model and investigates the possibility of conserved sequence changes that may alleviate the problem of low expressability. Unlike SKADE, this study makes use of computational re-design with Rosetta (Leaver-Fay, Tyka, et al. 2011) to identify important sequence modifications for improved expressability that are plausible from a structural point of view. The method presented here is tailored for antibodies and was developed with data from protein assays for expression in Chinese Ovarian Hamster (CHO) cells.

Therefore, this study takes a two-fold approach to predict and optimize antibody expressability. First, a Deep-Learning model was trained on a dataset of experimentally assessed concentrations. The model then was used to predict the effect of single point mutants on the expressability of par-

ticular antibody clones. Second, computational structural re-design informed by a library of single point mutant expressability was used to introduce as few mutations as possible while maximizing the expressability.

Deep mutational scanning (Fowler and Fields 2014) encompasses methods that exhaustively enlist the effect of single-point-mutants (SPM). Libraries of SPM have been used to successfully study binding tolerance (Whitehead et al. 2012) and sequence-function relationships (Fowler, Araya, et al. 2010). A recent study has shown, that SPM libraries can support the computational optimization of antibody heavy-light chain binding interfaces to improve antibody stability and even expression yields (Warszawski, Katz, et al. 2019). The concentrations of our dataset of 888 unique Flu-binding antibodies was expressed with a uniform method, and has therefore minimized experimental bias regarding expressability. Concentrations of each SPM are not available in this study, and we therefore use modern deep learning to create an exhaustive library of single point mutants.

Deep-Learning (DL) has demonstrated great success in mining the complex relationships hidden in biosequences (Li et al. 2019). Here, the long-short term memory (LSTM) (Hochreiter and Schmidhuber 1997) architecture was chosen, which has been applied successfully on sequences (Wainberg et al. 2018; Jurtz et al. 2017; Angermueller et al. 2016) before to predict Human-Likeness (Wollacott et al. 2019), or for de-novo structure prediction (AlQuraishi 2019). A recent study demonstrated antibody affinity maturation based on a LSTM model (Saka et al. 2021). Our study demonstrates, that an LSTM model can be used to predict the expressability of an antibody and generate SPM libraries to ultimately inform computational sequence design and optimize expressability.

The computational structural modeling suite Rosetta (Leaver-Fay, Tyka, et al. 2011) can be used for a wide variety of tasks including, but not restricted to, immunoglobulins. This includes: structure prediction, docking, antigen and antibody design (Schoeder et al. 2021). Here, we predict structures of a dataset of Flu antibodies using Rosetta homology modeling (Song et al. 2013) to successively conduct design. We restrain the sequence design to favor those mutations that are likely to improve the antibodies' overall expressability. For that, a pyrosetta (Chaudhury, Lyskov, and Gray 2010) protocol was developed that re-evaluates each mutation based on the LSTM-predicted expressability of its corresponding SPM.

9.2 Results

A set of 888 Flu antibodies were expressed in CHO cells and their expression levels were measured. Sequence diversity and chain class content are detailed in Supplementary Figures S70-S71 and Tables S6-S7. This study can be divided into two main parts. First, the antibody expressability estimation was calculated using a DL neural network trained with these CHO Flu sequences. Second, each antibody was redesigned with Rosetta in order to demonstrate the possibility of using DL-informed computational design to improve protein expressability.

The architecture of choice is a recurrent long-short term memory (LSTM) (Hochreiter and Schmidhuber 1997). The LSTM was trained and its performance evaluated via a 10-fold cross

validation. Due to the limited size of the training dataset, and to simplify the difficult computational expressability prediction task, the sequences were classified into two categories: 'expressing' or 'non-expressing'. An antibody was defined as expressing, if its concentration levels are greater than 50 µg/mL for two reasons. First, 50 µg /mL safely exceeds the minimal detectable concentration of 5µg/mL of the iQue flow cytometric detection system we used to measure immunoglobulins. Second, selecting the threshold of 50 µg/mL renders the dataset well balanced, since this cutoff results in 487 expressable (55%) and 401 non-expressing (45%) samples. This approach minimizes the risk of overfitting due to the otherwise lack of sufficient training samples in one category. Without a balanced cutoff, either the expressing or non-expressing samples are overrepresented. Unbalanced datasets would lead to biased performance numbers, since the network would learn a tendency to guess the expressibility label of the larger sample group. We do not expect any restrictions in the ability to change the expressibility cutoff for greater minimal expression yields. This would require experimental training data that either enables choosing a different cutoff, or is sizeable enough to remove excess positive or negative samples. The degree of correlation between expression levels and variable cutoffs may be explored in future studies.

9.2.1 Expressability prediction and optimization

The LSTM architecture was introduced specifically to improve the learning of long-term dependencies observed in classic recurrent neural networks (RNN) (Hochreiter and Schmidhuber 1997). Long distance relationships between residues in bio-sequences become especially relevant when residues distant from each other in the primary sequence, come spatially close to each other in the final structure.

Modern networks incorporate multiple layers of hidden states, which enable the model to learn relationships in high-dimensional data. Here, we translated the result into a probabilistic output with the help of a function called softmax (Figure 30a, Figure S63). Inputs to the network are heavy or light chain sequences that are presented as one-hot matrix, where each row corresponds to a specific amino acid identity and each column corresponds to an amino acid position assigned by the ImMunoGeneTics information system® (M.-P. Lefranc, Pommié, Kaas, et al. 2005) as unique IMGT-Number (M.-P. Lefranc, Pommié, Ruiz, et al. 2003). The IMGT numbering scheme encodes the location of framework and CDR regions, which ultimately allows us to compare different antibody sequences. We hypothesize that the alignment of the input samples to their IMGT Numbers allows the network to learn sequence features that correlate with expression characteristics in the binding interface of the paired heavy and light chain. The LSTM model was trained on either heavy (heavy), or light chain variable (light) sequences, in addition to a combination of both (paired), and the performance of all three models was assessed. The resulting input matrix for the neural net is of the shape times 21, with representing the number of unique IMGT numbers observed in the training set, and 21 features, which are the 20 canonical amino acids plus a gap symbol. When paired sequences are used, the feature dimension is doubled to 42 rows.

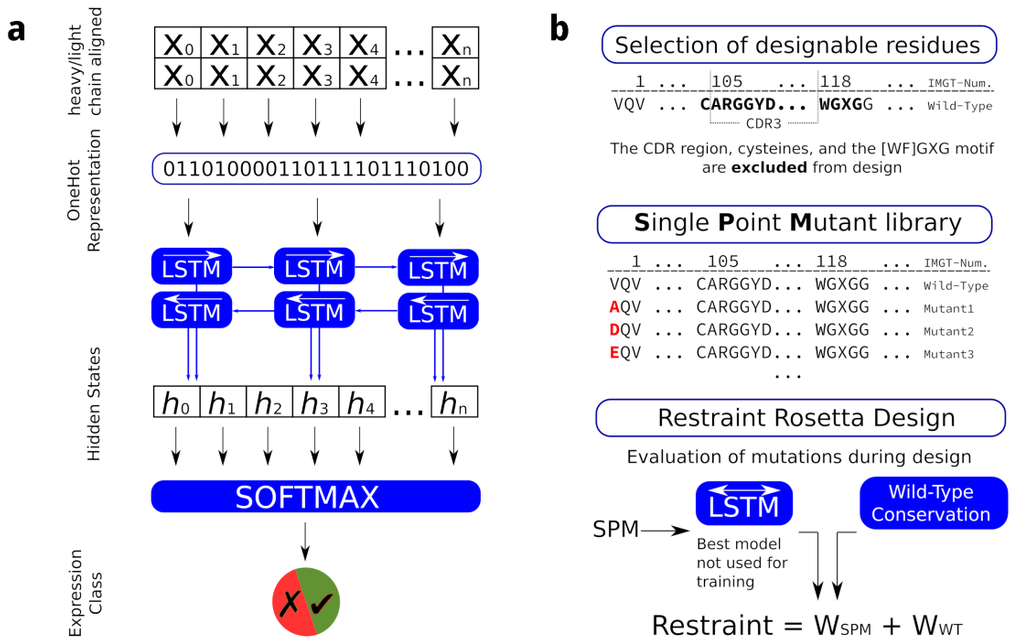

Figure 30: **LSTM architecture to binarily predict if an antibody can be expressed experimentally.** Paired heavy/light chain sequences are represented in a onehot-matrix. Each canonical amino acid plus the gap symbol is represented in one row, whereas the number of columns is determined by the longest observed training sequence. The one-hot matrix is forwarded into a bi-directional LSTM layer. A softmax function ultimately converts the LSTM prediction into the probability of belonging to either expression class (a).To improve expressability using structural re-design with Rosetta, IMGT Numbering for the input structure is assessed, to prepare the input for the LSTM and to identify framework and CDR regions. The CDR and residues with native cysteine are prohibited for mutation. For regions allowed to mutate, an exhaustive list of single point mutants is generated and its expressability predicted. To guide the re-design for increased expressability, each mutation is re-evaluated using its corresponding SPM expressability prediction (WSPM) and to a preference to retain the wild-type residue type (WWT) (b).

To increase the predicted expressability while keeping the number of changes (mutations) at a minimum, a Rosetta design protocol was developed that guides the structural design in two ways. First, each mutation is evaluated using the LSTM model. Second, a fixed term preferring the wild-type residue ("native") is calculated. Both terms interact with each other such that each possible mutation must be either structurally beneficial or especially beneficial for expressability in order to overcome the bias towards the native sequence. By adding both terms to the Rosetta scoring function, the structural re-design favors such sequences which improve expressability (Figure 30b).

9.2.2 Training performance of 10-fold cross-validation

To assess the expressability prediction performance, recall (Figure 31a) and area under the curve (AUC, Figure 31b) were recorded for each of the ten models during the ten-fold cross-validation. Control models were created by randomly shuffling the sequence labels. Random shuffling conserves the relative frequency of both expressing and non-expressing samples in the dataset as to not bias

the result.

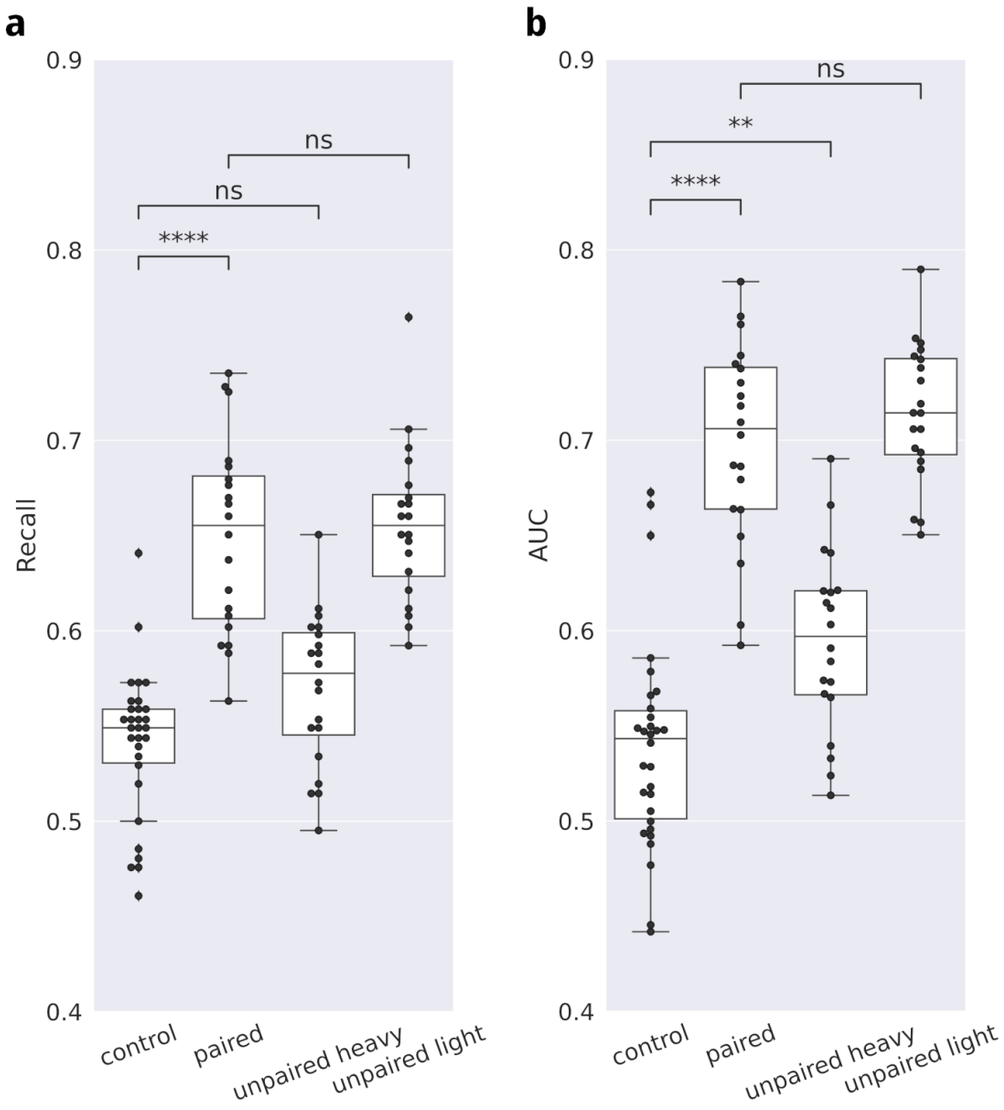

Figure 31: **Binary LSTM classification performance of (non-)expressing Flu antibodies.** Recall (a) and area under the curve (AUC, b) for two runs of ten-fold cross-validation. LSTM models were created for heavy chain, light chain, and paired heavy and light chain sequences. Each model was evaluated with a test-set of randomly relabeled sequences (Control). The paired, as well as the light chain model score, is significantly higher than the Control, without any substantial difference in performance between the two. The heavy chain model does not perform better than the Control. Mann Whitney significance tests annotated with p¡= 1e-4 (****), 1e-3¡=p¡=1e-2 (**), and p¿5e-2 (not significant; ns).

When averaged over ten models, a recall of 0.54 ± 0.04 (control), 0.65 ± 0.05 (paired), 0.57 ± 0.04 (heavy), or 0.66 ± 0.04 (light) was observed. AUC values were 0.54 ± 0.05 (control), 0.70 ± 0.05 (paired), 0.59 ± 0.05 (heavy), or 0.71 ± 0.04 (light). Precision values were 0.55 ± 0.03 (control), 0.67 ± 0.04 (paired), 0.59 ± 0.03 (heavy), or 0.67 ± 0.03 (light). Classification using light chain sequences only significantly outperformed heavy chain sequence models. In our expression experiments, a frequent over-expression of light chains was noticed. Prior to pairing, the heavy chain is chaperoned by other host proteins prior to pairing. Taking both together may indicate that the light chain drives antibody heavy and light chain pairing. We hypothesize that

the increased light chain performance of the classificator reflects the experimental conditions of our dataset.

For all further experiments, the models trained using the paired antibody sequences were used for the following reasons: First, the approach aims to remove sequence patterns that are unfavorable for expression. We hypothesize that the different chain classes can inform each other about (un-)favorable sequence patterns that may or may not be specific to chain classes or germline genes. The performance of kappa and lambda class antibodies was evaluated (Figure S72) but did not improve their performance reliably. Second, due to the limited size of the dataset a further split was avoided. We can not exclude the possibility that separate models are advantageous for different datasets.

In addition to the LSTM models, the performance of logarithmic expression was explored. To estimate the impact of the sample encoding, all models were evaluated using Kidera and Atchley factors that encode biochemical properties of amino acids, instead of the one-hot encoding of the amino acid sequence. It is hypothesized that Kidera and Atchley encoding may be advantageous, since it would allow the model to learn input feature distances based on biochemical amino acid properties. Figure S72 shows performance metrics for 13 different LSTM and regression models trained with one-hot, Kidera, and Atchley encoding using paired, light chains, kappa, or lambda sequences. To summarize, Kidera and Atchley encoding did not increase the classification performance. The performance drops for models using a lambda class antibodies, likely due to small size of the dataset (30.7% of the dataset with 273 antibodies). Overall, regression models performed comparably, but slightly worse than the LSTM models in most cases. In particular, the decreased AUC performance of regression models is likely detrimental for the re-design of antibodies for increased expressibility for Rosetta since the probability scores are converted into Rosetta restraints. We also hypothesize that the LSTM model may benefit from the continuous addition of more training data that would arise from re-training alongside expression experiments.

9.2.3 LSTM-informed structural design with Rosetta

RosettaCM (Song et al. 2013) was used to create homology models for each sequence in our Flu antibody dataset. The performance of RosettaCM homology models for antibodies have been studied in great detail in a previous study(40). The framework region is structurally conserved and RMSD vs Rosetta score plots of our homology models generally show folding tunnels, which is supportive of the idea that the models are plausible (Figure S73). The structural design does ignore the structurally much more diverse CDR3 region, which alleviates the impact of possible misfolded structures. We therefore consider the homology models reliable enough to exclude a major structural bias of misfolded structures that would impact the presented results.

To optimize the design for expressability, a library of single point mutants (SPM) was created, and the expressability of each mutant was predicted. The prediction was then converted into Rosetta energy penalties and bonuses. Each score serves as an estimate for the change each mutation has on the expressability. This change is independent of other sequence changes made during the

design. Penalties (positive scores) and bonuses (negative scores) guide the design towards sequences the LSTM would evaluate more favorably for expression.

Re-design was prevented in CDR3 regions for two main reasons. First, the CDR3 is crucial for antibody-antigen binding and mutations in this area may alter the efficacy the most. Second, the high sequence variability of the hypervariable region (Soto, Bombardi, et al. 2019) is likely insufficiently captured by the available dataset of the 888 sequences, which is a detriment to the task of making high-confidence predictions for mutations that benefit the expressability. In addition to the restriction on the CDR3 region, all cysteine residues were conserved and the introduction of new (and potentially unpaired) cysteines was prohibited.

To conserve optimal biophysical properties, and ultimately binding properties, the number of mutations was restricted by adding a Rosetta term that rewards a high identity to the wild-type ("native") sequence. This term directly opposes the newly introduced SPM expressability term and was balanced to work in harmony with the expressability term.

Our Rosetta design protocol activates two additional restraints: the LSTM expressibility restraint, and a sequence conservation restraint. Both restraints counterbalance each other, where a large expressibility weight tends to result in highly mutated antibodies, the conservation restraint reduces the number of mutations. Here a set of three weights are suggested that translates to light, moderate, and aggressive re-design by combining large and low weights for expressibility and sequence conservation. The weights were chosen that produced antibodies with expressibilities and number of mutations that appeared reasonable to us, but can be freely modified by the user of the protocol (Table 3). As a result, the "low intensity" weight combination introduces few mutations with the lowest increase in expressibility, "medium intensity" results in a moderate increase in both expressibility and number of mutations, and "strong intensity" which results in the largest number of mutations and the greatest increase in expressibility.

9.2.4 Predicted expressability before and after re-design

Figure 32 visualizes the three design intensities in red (strong), yellow (medium), or blue (low). A completely unrestraint control design resulted in large number of mutations without a significant gain in expressability. A restraint control (control cst., gray) was therefore introduced, where expressability was not optimized during re-design and the number of mutations restricted in the same way as in the low intensity design. See Table 3 for exact weighting schemes.

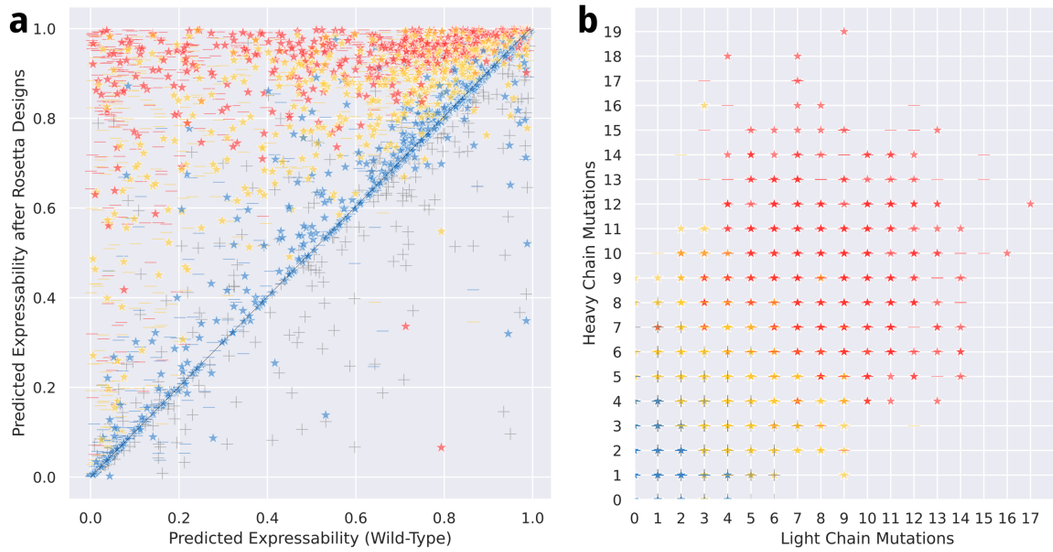

Figure 32: **Design performance of 888 Flu antibodies that did (star), or did not express (-) before the Rosetta re-design.** The LSTM prediction of the antibody expressability of the wild-type sequence versus the best scoring Rosetta design (a) and the number of heavy and light chain mutations (b). All antibodies were design with weak (blue), medium (yellow), or strong (red) expressability sequence restraints. Out of 888 antibodies considered, 352 were labeled as non-expressers before design (-). For 277 (weak), 32 (medium), or 16 (strong) of out 888 antibodies, the design resulted in a decrease of predicted expressability. Control designs with limited number mutations and disabled expressability term (gray) did not result in increased expressability on average.

In comparing the predicted expressability before or after re-design (Figure 32a), our method improves the expressability in 611 (low), 856 (medium), 872 (strong), or 447 (control cst) out of 888 cases, and the expressability changed on average by 3.6 ± 12% (low), 25.4 ± 24% (medium), 37.1 ± 29% (strong), and 0.0 ± 14% (control cst). The large standard deviation is direct result of the variable starting expressability of the wild-type sequences. Even though the gain in predicted low expressability for the low intensity design is moderate, the number of designs with a predicted decrease in expressability decreases compared to the control (Figure 32a, blue and gray marks below the diagonal).

Low intensity design can be considered the least conservative, and strong intensity design the most aggressive. Similar to the increased predicted expressability, the number of mutations also increases allowing a greater degree of freedom to increase expressability (Figure 32b). The median of introduced light chain mutations was 1 (low), 5 (medium), 9 (strong), or 1 (control cst.). The median of introduced heavy chain mutations was 1 (low), 5 (medium), 9 (strong), or 2 (control cst.).

To summarize, the number of mutations increases with increasing design intensity, whereas the predicted expressability substantially improves alongside. With about 4 to 5 mutations per chain on medium settings, excluding the CDR region, it appears plausible that the binding activity of the antibody can be conserved while increasing the chance for successful expression in CHO cells.

Design with the strong setting allows the reliable design of antibodies with high expressibility with the expense of up to 19 mutations. It is recommended to choose or adapt the weighting schema on a case-by-case basis to explore the optimal ratio of expressibility gained to the number of mutations. This also may involve creating new fine-tuned weighting schemes for the desired outcome.

9.2.5 No evidence for reduced structural stability after re-design

Rosetta protein design can come with the risk of impaired structural stability, especially when additional restraints are added. Here, these terms favor the native sequence and expressability. To evaluate Rosetta designs, Rosetta Energy Units (REU) serve as a metric to compare the wild-type structure with the designs. Additional restraint terms inevitably decrease the REU and potentially aggravate the effect of structures with impaired stability.

To assess a potential negative effect of the re-design protocol on the protein stability, the change of REU compared to the wild-type was assessed for each design intensity (Figure 33a). The largest effect on the REU was 18.2 ± 10 (low), followed by 3.0 ± 13 (medium), -3.4 ± 13 (strong), -72.6 ± 18 (control), or 12.3 ± 12 (control cst.). Here, in addition to the previously used control that enabled the native sequence restraint (control cst.), the results also were compared to a completely unrestrained design (control). To generalize, the more mutations that were introduced, the more degrees of freedom were available to compensate for mutations that were less favorable without restraint, but were beneficial for expressability. This effect is exaggerated in the unrestrained control design, which exhibits the greatest REU improvement to the expense of a large number of mutations. By reducing the number of allowed mutations (control cst.) the REU change avareges on a similar like the low intensity design, showing that our method does exhibit energy changes within expectations. Among the restrained designs, the greatest impact was observed for low intensity, which is comparable to its control (control cst.). Strong intensity designs exhibit on average a statistically significantly lower reduction in REU than the restraint control design (control cst.), and remains close to the wild-type energy. Thus, it can be concluded that our design approach does not have the tendency to reduce the structural integrity, and remains in ranges that can be considered normal for the specific structures and the chosen design.

The predicted expressability after re-design significantly improved with low intensity (56.1 ± 31%), medium (78.0 ± 23%), and strong intensity (90.0 ± 15%), compared to the control (57.4 ± 32%) and restraint control (52.0 ± 32%) (Figure 33b).

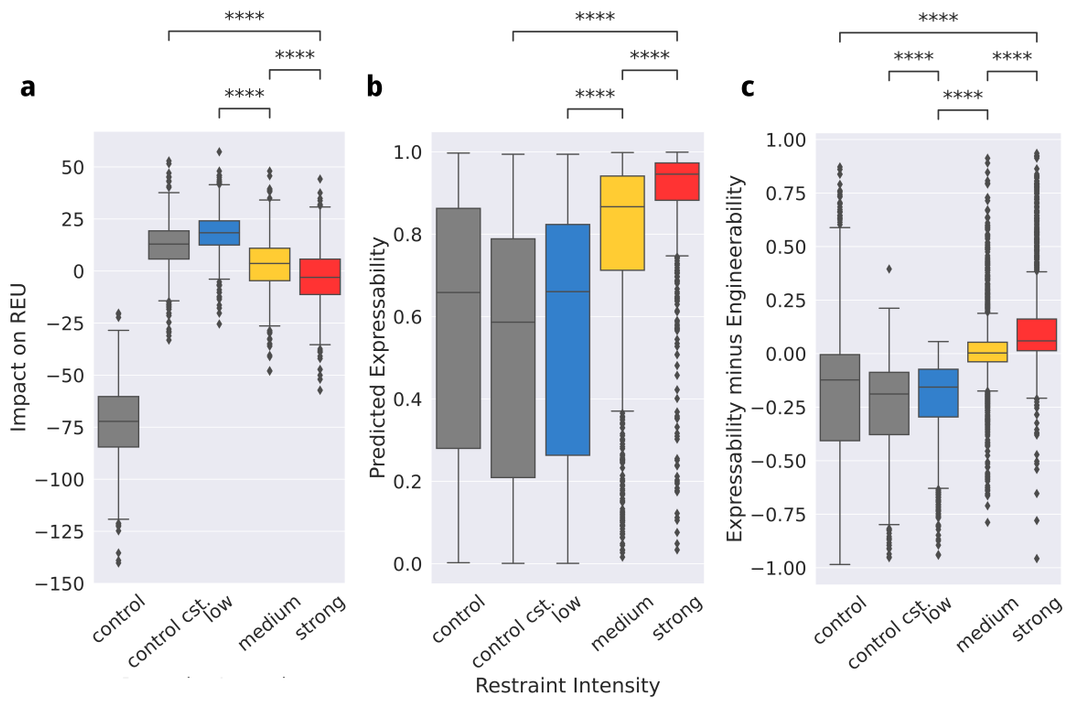

Figure 33: **Effect on predicted expressability, engineerability, and Rosetta energy of re-engineered antibodies.** The increase (worsening) of Rosetta energy relative to the wild-type is highest with low intensities (low number of mutations) and comparable for medium or strong restraint intensities. The energy impact on the two control groups is lowest without restraints (control, maximum number of mutations). The control with limited mutations (control cst.) has similar effect on the REU as the design with low intensity (a). The predicted expressability significantly increases when using low versus medium versus strong restraint intensitites. Low intensities and controls without restraints (control) and limited amount of mutations (control cst.) do not improve the predicted expressability clearly (b). The engineerability term, which serves as an estimate on how well an re-engineered anitbody can improve expressability is least predictive for the low (¡0), on average accurate for medium (on average 0), and in most cases represents a minimal expressability improvement for strong intensity (¿0) (c). Mann Whitney significance tests annotated with p¡= 1e-4 (****), 1e-3¡=p¡=1e-2 (**), and p¡5e-2 (*).

To estimate the degree to which an antibody's expressability can be improved in advance of re-design, we introduce the term engineerability. Here, engineerability is equal to the SPM with the highest predicted expressability. To demonstrate the predictive potential of the engineerability term, the expressability of the best re-design was compared with its engineerability. Figure 33c shows the difference between engineerability and expressability for each antibody. For both control groups, the expressability stays below zero and rarely reaches the engineerability value (-19.6 ± 33% control and -25.1 ± 21% control cst.). Similarly, the low intensity design (-21.0 ± 18%) remains below the engineerability value, whereas medium intensity designs, on average, come close to their engineerability (0.0 ± 17%), and strong intensity designs, on average, surpass the engineerability (12.6 ± 21%). Engineerability therefore can be considered as a predictive tool, when keeping the used restraint intensities in mind.

9.2.6 Re-engineered antibodies show a preference for certain residues

In this study, we demonstrated that a LSTM informed re-design is able to improve the predicted antibody expressability. To assess how the improved expressability was achieved, we assessed mutational preferences of the most successful Rosetta designs. The most successful designs were chosen as follows: First, the predicted expressability must increase by at least 80% compared to the wild-type expressability. Second, the final expressability prediction after design is at least 90%. Out of 888 cases, 142 unique antibodies remained matching the criteria and were analyzed further for potential sequence patterns.

Figure 34 visualizes the mutation rate for heavy (a, b) or light chains (c, d). The overall mutation frequency can be seen, which indicates how often an antibody has been mutated at a specific residue (panel a for heavy, and c for light chains). Column X is the cumulative frequency for each residue, indicating how often a position was mutated, disregarding its specific amino acid type. The specific distribution of amino acids for each residue is presented in panel b (heavy chain), or d (light chain).

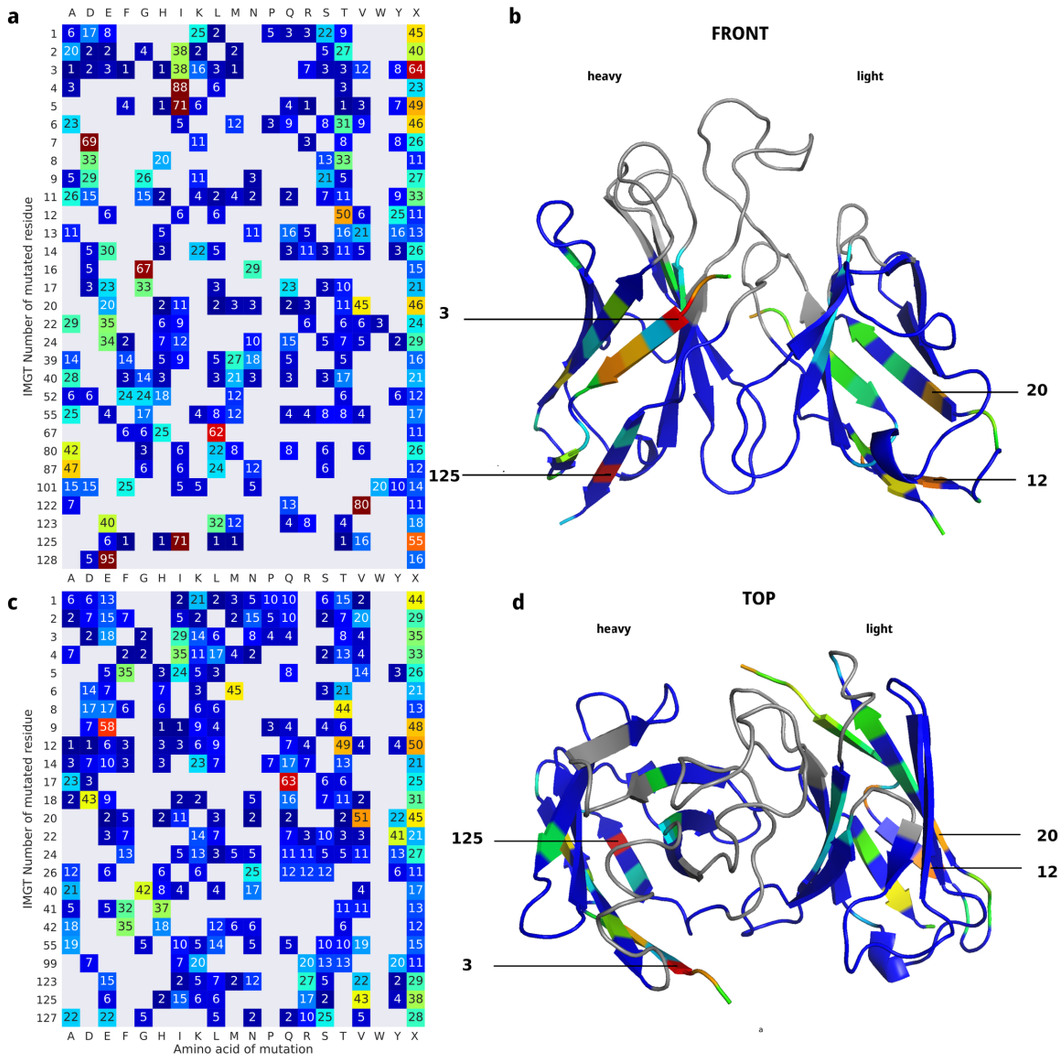

Figure 34: **Frequency of heavy chain mutations of the re-engineered designs with strong intensity.** A subset of antibodies is represented (best Rosetta score, expressibility greater than 90%, and improvement by re-design at least 80%). The numbers are percent rounded down to integers. The frequency of mutations in the heavy chain observed for each IMGT number and their total frequency of all designs (X). Mutation frequency is plotted separately for heavy chain (a) and light chain (c). For example: position 125 was mutated in 55% of all designs, most frequently substituted with Isoleucin (I, 71%). The mutations were mapped onto a typical immunoglobulin structure that represents the used dataset well (gray: CDR; colors match the heatmap). Most frequent mutations (3, 5, 12, 125) are located in either the N or C termini (b) which are both located at the surface of the structure. Mutations between the CDR1 and CDR2 (39-55) located in between the heavy and light chain interface exhibit only low mutation frequencies (d).

In heavy chains, the most frequently mutated residues was IMGT residue 3 with 64%, followed by 125 (56%), 5 (51%), 20 (49%), 6 (46%), 1 (42%), 2 (39%), 11 (35%), 9 (29%), or 24 (28%) each (Figure 34a, column X). Positions 2, 3, 5, and 125 were frequently replaced with isoleucine (41%, 39%, 71%, or 73%, Figure 34b). For most other most positions with high mutation frequencies, the amino acid preferences were slightly more ambiguous, ranging from a hydrophobic valine at position 20 (46%), alanine at position 11 (29%), charged amino acids aspartatic acid or glutamatic acid at positions 9 (29%), 14 (29%), or 24 (33%). Overall, it can be summarized, that for very

successful designs with high positive change in expressability, there is a preference towards N and C-terminal mutations with primarily hydrophobic (isoleucine, valine, arginine), and also less frequently charged amino acids (aspartic or glutamatic acid). This overall trend barely changes, when "successful design" is defined more generously, with an expressability improvement of at least 50% (284 unique antibodies applicable) instead of 80%, indicating that for this dataset, the observed mutational preference is robust (Figure S68).

Sequence logos for the complete Flu antibody dataset, as well sequence logos for Flu sequences labeled as expressing or non-expressing (Figures S64-S66) show, that the mutational preferences after Rosetta design (Figure S67) do not match the sequence conservation observed in the dataset. It can be concluded, that the LSTM guided re-design does not simply recapitulate amino acid identities that are already present in the set of already expressing sequences. The extent to which the specific mutations can be reproduced with different antibody lineages, and experimental assays requires further studies.

9.3 Methods

9.3.1 Plasmablasts isolation and paired heavy and light chain variable regions sequencing

Generation a panel of monoclonal antibodies that were isolated from plasmablasts was described previously (Zost et al. 2021). Briefly, PBMCs were isolated during natural influenza A H3N2 virus infection on day 7 from symptom onset and stained with the following phenotyping antibodies; anti-CD19, -CD27 and -CD38 (BD Biosciences) to identify plasmblasts. Plasmablasts were single-cell sorted in bulk using an Aria III flow cytometer (BD Biosciences) and carried through single-cell RNA sequencing using the 10X Genomics Chromium platform with enrichment using the VDJ amplification kit (10X Genomics) according to the manufacturer's instructions. Amplicons were sequenced on an Illumina Novaseq 6000, and data were processed using the CellRanger software v3.1.0 (10X Genomics).

9.3.2 Antibody production, purification, and quantification

cDNA encoding heavy or light chains of interest were synthesized and cloned into IgG1 or IgK/IgL expression vectors, respectively (Twist Bioscience). Heavy and light chain plasmids were transfected into 96-well cultures of ExpiCHO cells (ThermoFisher Scientific,) for microscale expression and then purified using previously described methods (Gilchuk et al. 2020). High-throughput quantification of microscale-purified mAbs was performed using the Cy-Clone Plus Kit and an iQue Plus Screener flow cytometer (IntelliCyt) according to the vendor's protocol. Only a small fraction of antibodies (¡ 10%) did aggregated visibly; low antibody concentration measurement values therefore represent either, low expression or aggregation after purification.

9.3.3 Training of LSTM models

TensorFlow 2.1 (Abadi et al. 2016) short-term memory written in python3.6. The two layer bi-directional LSTM (Figure S63) was trained by splitting the dataset into train, dev, test datasets from 888 antibodies in a ratio of 80%, 10%, 10%. As a pre-processing step, sequences were assigned IMGT numbering using IgReconstruct (Schmitz et al. 2020) and presented as an aligned one-hot matrix, where each column represents a unique IMGT number. The rows of each sample correspond to one of the 20 canonical amino acids plus a gap symbol. Models were created for either heavy or light chain individually, or for both chains paired together. In the case of paired heavy/light chain models, matrices for heavy and light chains were concatenated. The LSTM results were translated into a binary classifier to estimate the expressability using the Dense network of the tensorflow/keras framework. To evaluate the training performance ten fold cross validation was chosen and recall and AUC metrics reported. Each model was trained for up to 100 epochs, while logging the weights and loss after each epoch. To minimize the risk of over-fitting, the model weights with the smallest loss was chosen. As a consequence, some models may have trained for fewer than 100 epochs. Labels were represented as a onehot vector where 1 represents $>$ 50 µg/mL and 0 represents $\leqslant$ 50 µg/mL measured titers. A threshold of 50µg/mL rendered the dataset available for this study balanced with 487 sequences labeled as expressing and 401 sequences labeled as non-expressing.

9.3.4 Expressability prediction

For expressability prediction, only those IMGT numbers (columns) observed during training can be considered. Non-observed IMGT numbers in each are equivalent to gaps. 10 models were generated with random bootstrapping of the training set using either light chains, heavy chains, or aligned heavy and light chains. After evaluating the classification performance of the models, the 10 models generated with paired heavy and light chains were chosen to predict expressability before and after design. Each sequence was scored with up to 10 models while the only models that were considered were those not previously part of the training set. The result with the highest confidence in expressability was chosen as the result.

9.3.5 Structural antibody homology modeling with Rosetta

500 homology models for each antibody in our dataset were created using a multi-template protocol (Kodali et al. 2021) based on RosettaCM (Song et al. 2013). Each structure was refined with RosettaRelax (Nivón, Moretti, and David Baker 2013) five times and the best model was chosen purely by its best Rosetta energy.

9.3.6 Rosetta design with and without expressability restraints

The Rosetta design protocol was implemented in pyrosetta (Chaudhury, Lyskov, and Gray 2010) and implemented in the RosettaScripts (**fleisman˙rosettascripts˙2011**) framework for the purpose of combining our protocol with existing or future Rosetta protocols that may, for example,

reduce immunogenic effects by adding additional design constraints. An exhaustive list of all single point mutations was created and scored for expressability. The Rosetta protocol implemented in pyrosetta uses the predicted expressability score of the SNPs to assign a Rosetta score penality and bonuses during Rosetta sequence design respectively. The expressability term E(resi) for any given mutation resi is therefore expressed as the difference between the expressability of its corresponding SPM (ESPM), and the WT expressability (EWT). The expressability difference between WT and SPM may be small and thus the prediction has a lower confidence than more distinguishable scores. To account for close-to-wildtype scores, the SPM – WT difference was scaled according to equation 14. The effect of the scaling is visualized in Figure S69.

$$S(resi) = \begin{cases} E(resi) \times (1 - E_{WT}) & if \quad E(resi) > 0 \\ E(resi) \times E_{WT} & if \quad E(resi) < 0; \quad with \quad E(resi) = E_{SPM} - E_{WT} \\ 0 & if \quad E(resi) = 0 \end{cases} \tag{14}$$

The scores S(resi) is clamped into a range of -W to W to create Rosetta scores. This approach guarantees that the Rosetta restraint term has values close to the user specified weight W and allows the protocol to behave in a predictable way. However, to account for outliers in the set of S(resi) scores, the final constraint bonuses and penalties CR(resi) are first normalized by its 90th percentile (P90) and then clamped into the range [-W; W] (equation 15). The percentile is calculated for each chain individually.

$$C(resi) = \frac{S(resi)}{P_{90}(S)} \times W; \quad with \quad C(resi) \in [-W; W] \tag{15}$$

The influenza dataset used in this study comprises antibody variable regions of paired heavy and light chains, with only residues within the variable region subjected to design. In addition, Rosetta design was restricted to not introduce cysteines and to conserve already existing cysteines. No mutations were allowed in the CDR region comprising all 6 CDR loops of the heavy and light chains. The location of the CDR loops 1-3 was inferred from IMGT numbers, that is IMGT numbers 27-38 for CDR1, 56-65 for CDR2, and 105-117 for CDR3 (M.-P. Lefranc, Pommié, Ruiz, et al. 2003). Equally to the CDR, the four residues after the CDR3 (the [WF]GXG motif) also were prohibited for re-design. The [WF]GXG motif is a highly conserved antibody sequence pattern indicating the end of the CDRH3. The remaining Fv region was free to mutate using 19 canonical amino-acids (cysteine excluded).

To minimize the number of mutations, the WT sequence was favored. Three expressability design intensities were chosen by combining the strength to favor expressability as well as a low number of mutations. Two controls, one completely unrestrained, the other restrained exclusively to favor the WT sequence (Table 3). A custom energy constraint was implemented in pyrosetta for the Expressability weight, whereas RosettaScript's FavorNativeResidue was used to limit the number of mutations.

Table 3: Rosetta weights used to increase expressability and keep the number of mutations at a minimum

Intensity	Expressability weight (W)	WT sequence weight
low	4	3
medium	3	2
strong	4	2
control cst.	0	2
control	0	0

9.4 Availability

Pyrosetta scripts using tensorflow 2.1 for training and predicting expressability are at the Rosetta Commons github repository under https://github.com/RosettaCommons/AbExpress. Ten pre-trained models using our Flu dataset are included, but it is recommended to re-train the model for individual use. IgReconstruct (Schmitz et al. 2020) is available to provide IMGT numbered sequence alignments at http://www.meilerlab.org/index.php/servers/IgReconstruct.

9.5 Discussion

This study presents a Deep Learning based method to predict antibody expressability in CHO cells. It could be shown that the predictor can be used to inform computational structural design that can alter the sequence to improve the expressability of antibodies without disrupting the protein structures. The predicted expressability could be significantly increased. To increase the chance of keeping biophysical properties intact, the number of mutations can be controlled and minimized. The method developed here is comparable to scientifically related solubility predictors. The presented method is antibody specific and makes use of computational structural design to optimize expressability. Sequence analysis of optimized antibodies exhibited mutational hot-spots at the N and C-termini of the variable region with predominantly hydrophobic, and to a much lesser extent, charged amino-acids. Even though attention based neural networks previously have attributed high importance to N-and C-termini for solubility (Raimondi et al. 2020), the underlying driving forces may be different. First, N and C-terminal regions tend to be more conserved due to their germline gene templating. The artificial neural network may conceal potential mutational preferences with a higher degree of uncertainty for more variable regions, that have a lower training sample coverage. Generalization of these observations may be aggravated by the low degree of shared sequence space in antibody heavy chains (Soto, Bombardi, et al. 2019). Thus, the optimal use case scenario may involve training custom expressability LSTM models for each antibody lineage alongside ongoing experimental characterization.

Whether these indications for mutational preferences, and whether the expressability optimization protocol can be observed with other antibody lineages, must ultimately be decided by future efforts.

9.5.1 Acknowledgement

We thank Pavlo Gilchuk, Rachel Nargi and Robert Carnahan at Vanderbilt for sharing the antibody sequences and their experimental results.

10 Conclusion and Future Directions

10.1 Human-likeness from large sequence datasets

In the first part of this work (Chapter 6), the technological foundation (IgReconstruct) is laid to statistically assess complete immunome repertoires of more than 300 million unique sequences. IgReconstruct as an approach to links the nucleotide sequence space with resources of Abs where primarily amino acid information is available, like de-novo computational models or structural databases (Berman et al. 2000; Dunbar et al. 2014). by creating individual nucleotide frequency alignments with each antibody.

Nucleotide statistics of the variable region derived from large human immunome repertoires are capable of estimating the similarity of an antibody the observed human antibody sequence space. by predicting nucleotide sequences from amino acid sequences. In the process, our statistical model performed *on-par* with a Deep-Learning approach to estimate antibody human-likeness (Wollacott et al. 2019) and in addition is capable of suggesting nucleotide-sequences. Future studies may explore how these sequences compare with codon-optimized sequences routinely generated by biotech companies.

The rise of modern Deep-Learning techniques has recently found widespread applications in structural biology, like AlphaFold (Senior et al. 2020; Jumper et al. 2021), RoseTTAFold (Baek et al. 2021), or Long short-term memory (LSTM) based antibody affinity maturation (Saka et al. 2021). Deep-Learning approaches can be implemented and evaluated quickly, and do not require an elaborate hypothesis due to the automatic feature-extraction from data that is inherent to the success of this approach. At the same time, Deep-Learning suffers from a lack of explainability. While extraordinary success could be demonstrated in the area of *de-novo* protein-folding, it has not improved the understanding of protein-folding mechanics.

By choosing a statistical model for IgReconstruct, it can be concluded that the human-likeness is primarily a function of the underlying germline gene rearrangement and its individual nucleotide frequencies. This indirectly supports the hypothesis that the B-Cell maturation process is a primarily an undirected stochastic process succeeded by - as opposed to guided by - a selection process for affinity and productivity.

The greatest challenge and limitation of this method is the high variability of CDRH3 region, aggravated by a lack of germline genes with a high-confidence alignment. Germline gene rearrangements are fundamental to assessing human-likeness with IgReconstruct. Thus, to solely estimate human-likeness, or more generally speaking, the similarity to a set of sequences, Deep-Learning approaches certainly would outperform our CDRH3 human-likeness prediction as well as back-translation. In this study, we build upon this technique to address this shortcoming (Chapter 8).

10.2 Co-evolving residues characterize protein function and flexibility

Co-evolving residues span a network across the protein, with the majority in physical contact with each other ($> 91\%$), but others account for protein dynamics and function (e.g. interac-

tion partners). It was shown that incorporating co-evolutionary information into protein design helps Rosetta to better design stable and functional proteins while conserving functionally relevant residues (Chapter 7).

Traditionally protein design with Rosetta occurs on one or few conformations with no or very little information about the protein environment. Thus, the primary hypothesis is, that evolutionary information comprising structural, flexibility and functional information can be leveraged to inform the protein design. This is especially useful for proteins with little available additional structural information, i.e. properties of the protein that are 'hidden' in a single structure/sequence pair. It could be shown that the designed proteins have more natural sequences that are more similar to homologous proteins, and tend to conserve in many cases the amino acid identity of functionally relevant residues. In future studies, the extent to which the designed proteins can also conserve the protein's conformation and flexibility can be investigated in greater detail. We further could show that filtering of the network of co-evolving residues can highlight areas of interest that overlap with regions known to be relevant for function.

Implications towards antibody design in conjunction with large immunome repertoires are as follows: Future studies may develop methods to assess antibody datasets based on their function. Fingerprinting antibodies according to their coupled residues to ultimately support antibody discovery by directly scanning for and grouping by patterns of covariant residue frequencies. The design with antibodies sharing the same fingerprint may support antibody sequence design and potentially inform the epitope-focused immunogen design to elicit a more potent immune response with Rosetta.

The applicability to antibodies was limited by the lack of antibody sequence data and requires a large amount of annotated antibody sequences of the same antibody lineage that can support the development and validation of the method. Thus, the development of the antibody design approach was continued by using a clustering approach on large immunome repertoires without information about its specificity.

10.3 Modeling the antibody sequence space and human-like antibody design

For many years, the development of structure-based computational methods has been proceeding, resulting in a multitude of protocols in Rosetta. Various structural design protocols are specific for antibody design. One of the most comprehensive Rosetta Antibody design protocols is RabD (Rosetta Antibody Design) (Adolf-Bryfogle, Kalyuzhniy, et al. 2018) and allows the addition of sequence profiles from structurally clustered antibodies (Adolf-Bryfogle, Q. Xu, et al. 2015). This approach is significantly limited by the comparatively small amount of available antibody structures. At the time of writing, 1,000 - 2,000 human antibodies were available at the time of writing at the curated antibody database SAbDab (Dunbar et al. 2014). It seems obvious that a small amount of sequences is insufficient to recapitulate the (human) antibody sequence space, which is conservatively estimated to be in the range of 10^{13} unique sequences. Even though substantial progress was made to experimentally increase the observed antibody sequence space (DeWitt et al.

2016; B. Briney et al. 2019; Soto, Bombardi, et al. 2019), the cumulative body of sequences amounts to few billion antibodies from highly diverse sources (Corrie et al. 2018).

In this study, both challenges were addressed, the limited observed antibody sequence space from diverse sources and their insufficient integration into protein design with Rosetta, and mitigation of the sequenced fraction of immune repertoires via a mathematical model (Chapter 8).

The previously discussed human-likeness evaluation of antibodies using nucleotide frequency statistics has been discussed earlier (Chapter 6). The limitation of making conclusions about the CDRH3 region has been pointed out. The primary challenge in the CDHR3 region is the lack of high confidence diversity (D) germline genes and the untemplated character of CDRH3 junction regions in general. This enables the high specificity of antibodies to its antigen, but at the same time, this diminishes the meaning of suggested back-translations and human-likeness predictions. Templated regions on the other hand allow the estimation of the diversification of sequences using the germline genes as reference.

To improve the value of the CDRH3 predictions, a clustering approach was employed to generate nucleotide statistics for similar sequences. A set of similar sequences may convey sequence patterns that support the binding and conformation of a given structural design. Thus, clustered human-likeness or human-likeness on a specific antibody lineage may reveal patterns that are characteristic for binding and function.

To address the limited number of single-source observed sequences, the degeneration of the genetic code was leveraged to model amino acid probabilities that arise from the nucleotide type at the first, second, and third position of a triplet which is shared between certain amino acids. In this work we have demonstrated increased sequence similarity of designed antibodies to their respective wild-type crystal structures and increased human-likeness. For the CDRH3 region, the human-likeness increased only for a fraction of the benchmarked antibodies. This was expected due to the high variability and low commonality of antibodies, rendering the used repertoire in some cases more, and in other cases less prone to develop a response.

Future studies may use this technology to improve upon the limitations in the nucleotide frequency statistics. For one, the heavy and light chains were unpaired - a limitation of the dataset - and during clustering, the clonotype was ignored. One improvement may thus be to include the usage of a sequence dataset of paired sequences and with V, J, and CDRH3 clusters that depends on the V and J gene combination (Clonotype). Co-evolutionary analysis on top of the clustering may reveal distinct patterns that indicate functional activity, facilitating the grouping of repertoires by function.

In combination with Rosetta, an enhanced method may allow for simulation of the immune response to a specific antigen to ultimately optimize the epitope focused vaccine design in order to elicit the most effective immune response with the lowest probability of escape mutants. The combination of multi-state design using different human-likeness restraints at the same time that have previously been determined to be functionally relevant with Rosetta may support the design of bi-specific antibodies that appear human-like. As it is common for Research and Development

projects, meaningful outcomes can only be achieved in a closed cycle of computational prediction and experimental validation to establish a positive feedback loop. Thus, the greatest limitation of this approach is the requirement of close collaboration with experimental antibody discovery.

10.4 Prediction of antibody expressability

The complex biophysical cascade that involves the expression of a protein from transcription to secretion has been described in Chapter 5.7. To address the challenge of predicting if an antibody can be expressed is a multivariate challenge and the public data is sparse that fulfill the requirements to develop expressability pipelines. In the optimal case, antibodies are required to be screened for bottlenecks in translation, folding, and secretion. The minimal requirements to develop an expressability predictor is a dataset with experimental expression yields determined by one single experimental approach.

In this study, a Flu dataset was examined with expression yields. Due to the lack of any further sequence annotation a Deep Learning approach was employed to predict expressability and to design antibodies with increased expressability. Even though, the performance of the predictor was mediocre with an AUC of 0.70, it could be shown that in combination with Rosetta, the predicted expressability could be increased in all cases. Thus, a use-case of this approach is to improve the expressability while keeping the number of mutations low to avoid changes in the binding mode.

To further improve performance of expressability predictors, the single most important approach is the integration with an experimental workflow - usually well established in antibody discovery laboratories - to attain a continuous feedback loop with fresh training data. Furthermore, predictors or experimental data regarding post-translational modification (PTM) or structural stability may to some degree improve the classification. However, due to the vast number of unknowns, it would be advised to re-train the predictor individually for specific antibody lineages and experimental setups. In this work it was observed that the light chains of the studied Flu antibodies had a dis-proportionally high effect on the classificator performance compared to the heavy chains. This phenomenon remains unexplained and may or may not be reproducible with other datasets. Ultimately, the project may facilitate the study binding characteristics of antibodies that have been elusive to experiements due to low yields.

11 Appendices

11.1 Antibody human-likeness via back-translation

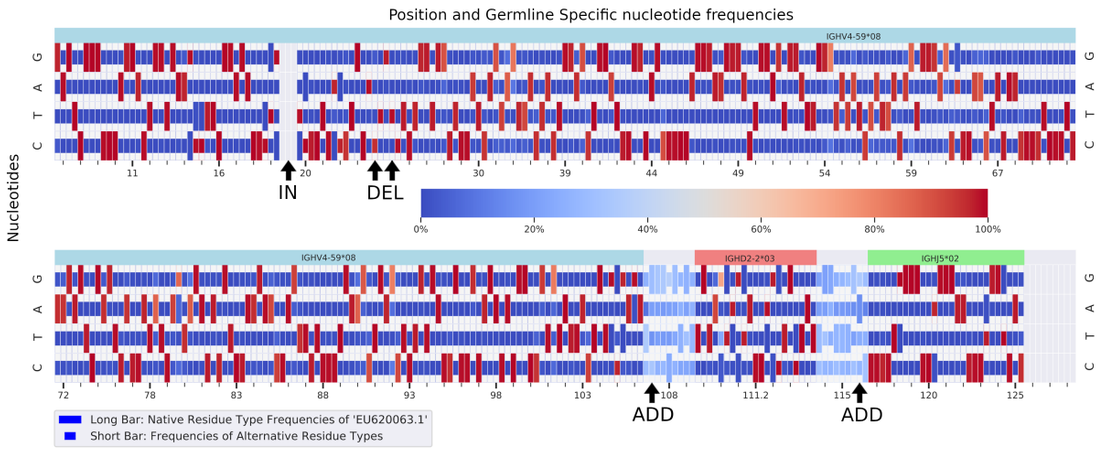

Figure 35: **Heatmap of single nucleotide frequences for the heavy chain sequence with GenBank ID EU6200063.1**

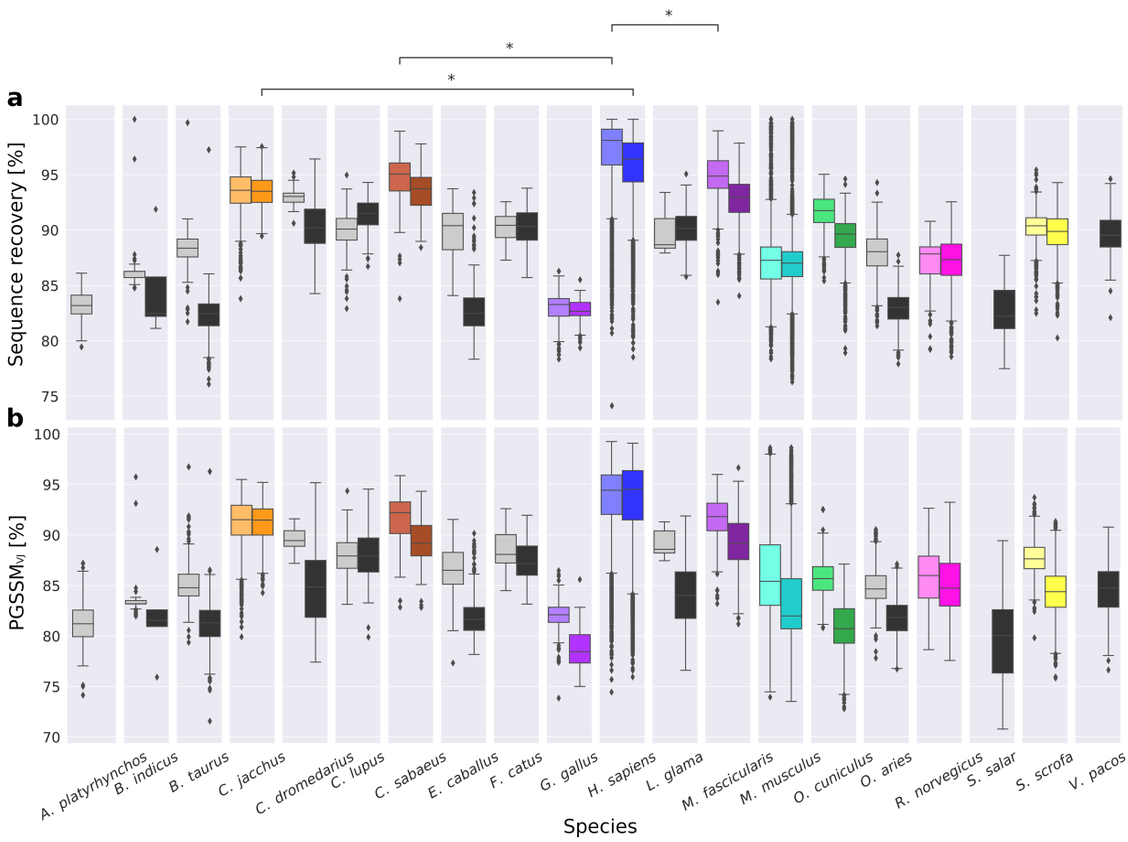

Figure 36: **Sequence recovery and human-likeness scores for all 20 species.**

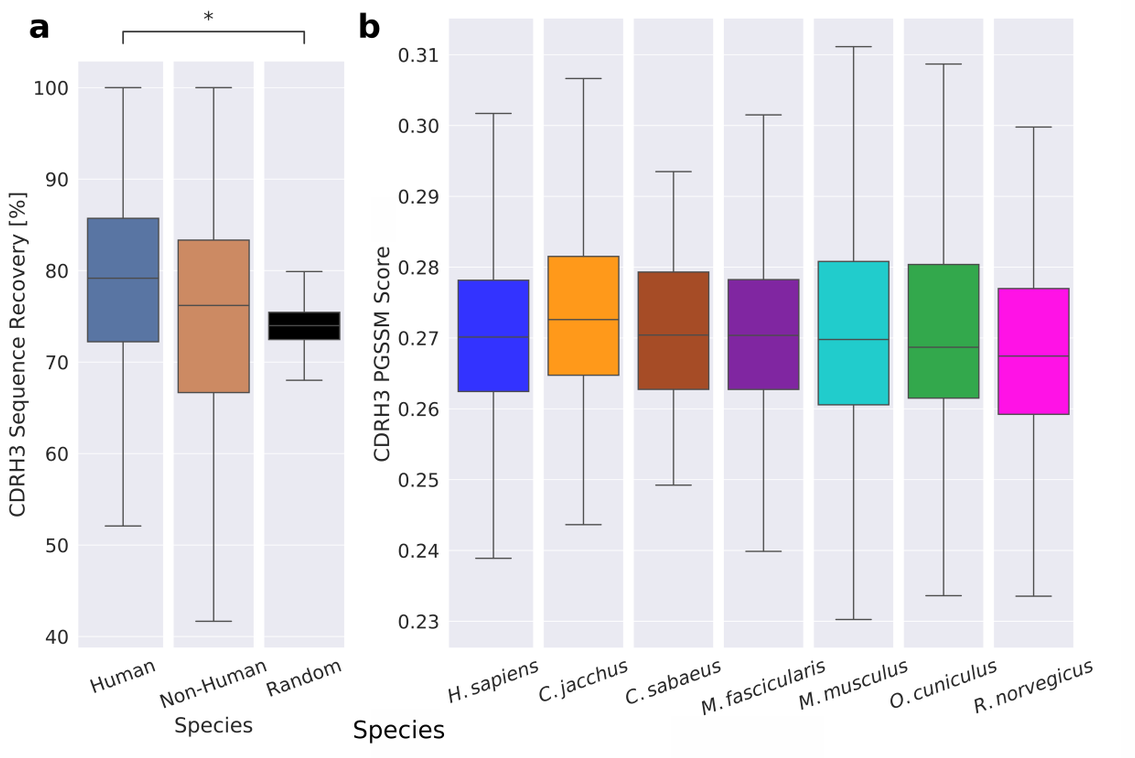

Figure 37: **Nucleotide sequence recovery for the CDRH3 loop for human and non-human sequences.**

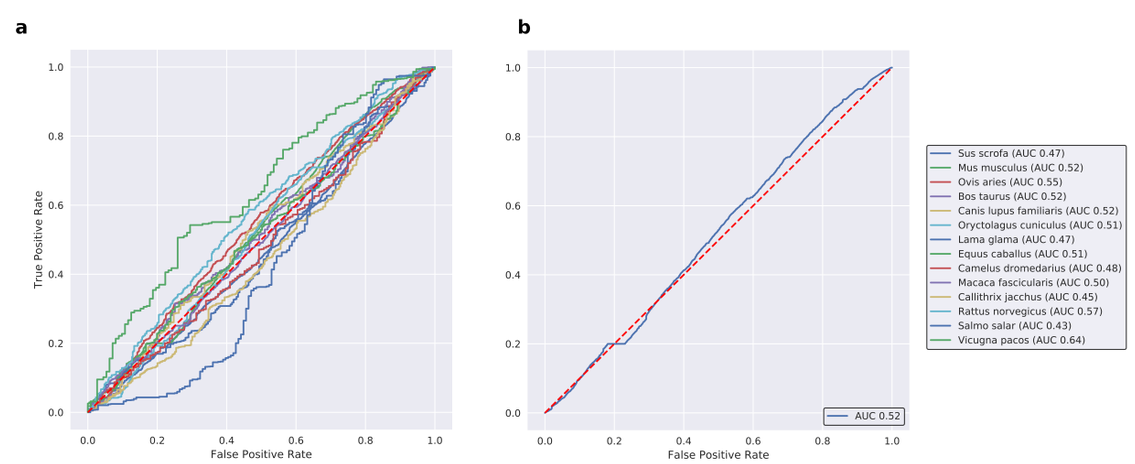

Figure 38: **CDRH3 classification performance using the CDRH3 human-likeness score.**

11.1.1 Random back-translation results in nucleotide sequence identity of roughly 74%

To roughly estimate the chance of guessing a nucleotide sequence correctly given a fixed amino acid sequence the following considerations were made: To simplify the calculation, the degree of degeneration of the amino acid translation table was quantified. Six distinct groups of codons, which differ in how many different nucleotides are possible at all three positions, were identified. Group one comprises two amino acids with only one triplet (W, M). Group two comprises nine amio acids (H, F, Y, E, N, K, D, E, C) with two triplets, which differ at the third position. Group three of one amino acid (I) with three triplets differing at the third nucleotide and so on. All groups are listed in detail in Supplementary Table 1. For each group we calculated the chance, T, to guess the wild type nucleotide triplet given the wild type amino acid, which is the average of the probability to guess each individual nucleotide (Equation 16).

$$T = \frac{\frac{1}{A} + \frac{1}{P} + \frac{1}{E}}{3} \tag{16}$$

T := Chance to guess the correct triplet for a given amino acid

A := Number of possible nucleotides at the first triplet position

P:= Number of possible nucleotides at the second triplet position

E:= Number of possible nucleotides at the third triplet position

Amino acids R, L, and S are encoded by two groups. The total chance to guess correctly in such cases is the average of both groups (Supplementary Table 4, column 5). The calculation was simplified with the assumption that all amino acids are observed equally, often on average. To obtain the probability to guess group one to six correctly, the total chance was multiplied with the number of amino acids in each group. The final chance to guess the nucleotide sequence of a fixed amino acid sequence correctly is then the sum of all probabilities divided by 20 (number of canonical amino acids). This simplified calculation results in an average probability to guess the correct nucleotide sequence of 73.68%. To validate our calculation, 206,165 GenBank sequences were back-translated from our benchmark dataset three times. The sequence set contained 130,768 heavy chain sequences of the species Homo sapiens (92,787), Callithrix jacchus (547), Chlorocebus sabaeus (123), Macaca fascicularis (4780), Mus musculus (31,070), Oryctolagus cuniculus (608), and Rattus norvegicus (865). It further contained 75,384 light chain sequences of species Homo sapiens (57,427), Callithrix jacchus (828), Chlorocebus sabaeus (129), Macaca fascicularis (735), Mus musculus (13,619), Oryctolagus cuniculus (2,141), and Rattus norvegicus (505). Supplementary Figure 5 shows the average sequence identity of three random back-translations for each of the sequences. We used uniformly distributed probability for each triplet encoding the given amino acid at each position. On average, the sequence identity over all species was 73.46 $\pm$ 2.60 % (compare: 73.68% calculated).

Table 4: Expected nucleotide sequence recovery for random back-translation. The rightmost column summed up results in a probability of 0.7368

Amino acids	Codons of **bold** amino acid	Number of different nucleotides to choose from position one to three	Chance to choose the correct nucleotide	Total chance	Total chance multiplied by number of amino acids
W, M	TGG	1, 1, 1	$\frac{1+1+1}{3} = 1$	1	1
H, F, Y, E, N, K, D, E, C	CAT CAC	1, 1, 2	$\frac{1+1+1/2}{3} = \frac{5}{6}$	$\frac{5}{6}$	$\frac{15}{2}$
I	ATT ATC ATA	1, 1, 3	$\frac{1+1+1/3}{3} = \frac{7}{9}$	$\frac{7}{9}$	$\frac{7}{9}$
P, V, A, T	CCT CCC CCA CCG	1, 1, 4	$\frac{1+1+1/4}{3} = \frac{3}{4}$	$\frac{3}{4}$	$\frac{15}{4}$
R, L	CGT CGC CGA CGG	2, 1, 4	$\frac{1/2+1+1/4}{3} = \frac{7}{12}$	$\frac{7/12+2/3}{2}$ $\frac{5}{8}$	$\frac{5}{4}$
R, L	AGA AGG	2, 1, 2	$\frac{1/2+1+1/2}{3} = \frac{2}{3}$	$\frac{7/12+2/3}{2}$ $\frac{5}{8}$	$\frac{5}{4}$
S	TCT TCC TCA TCG	2, 2, 4	$\frac{1/2+1/2+1/4}{3} = \frac{5}{12}$	$\frac{5/12+1/2}{2}$ $\frac{11}{24}$	$\frac{11}{24}$
S	AGT AGC	2, 2, 2	$\frac{1/2+1/2+1/2}{3} = \frac{1}{2}$	$\frac{5/12+1/2}{2}$ $\frac{11}{24}$	$\frac{11}{24}$

11.2 Rosetta design with co-evolutionary restraints and benchmark description

11.2.1 Benchmak Protein Description

Here, additional data on the other benchmark proteins can be found, highlighting the broad use of ResCue. For a detailed description of the protein design methods used, see Section 2.

11.2.2 Overview of all ten benchmark proteins

In all cases ResCue designs showed a lower energy increase compared to SeqProf and RECON, while having a large increase in crs values (Fig S1). All ResCue designs occupied a more favorable area of this energy landscapes. In all of our benchmark proteins there is a clear separation of the design methods visible.

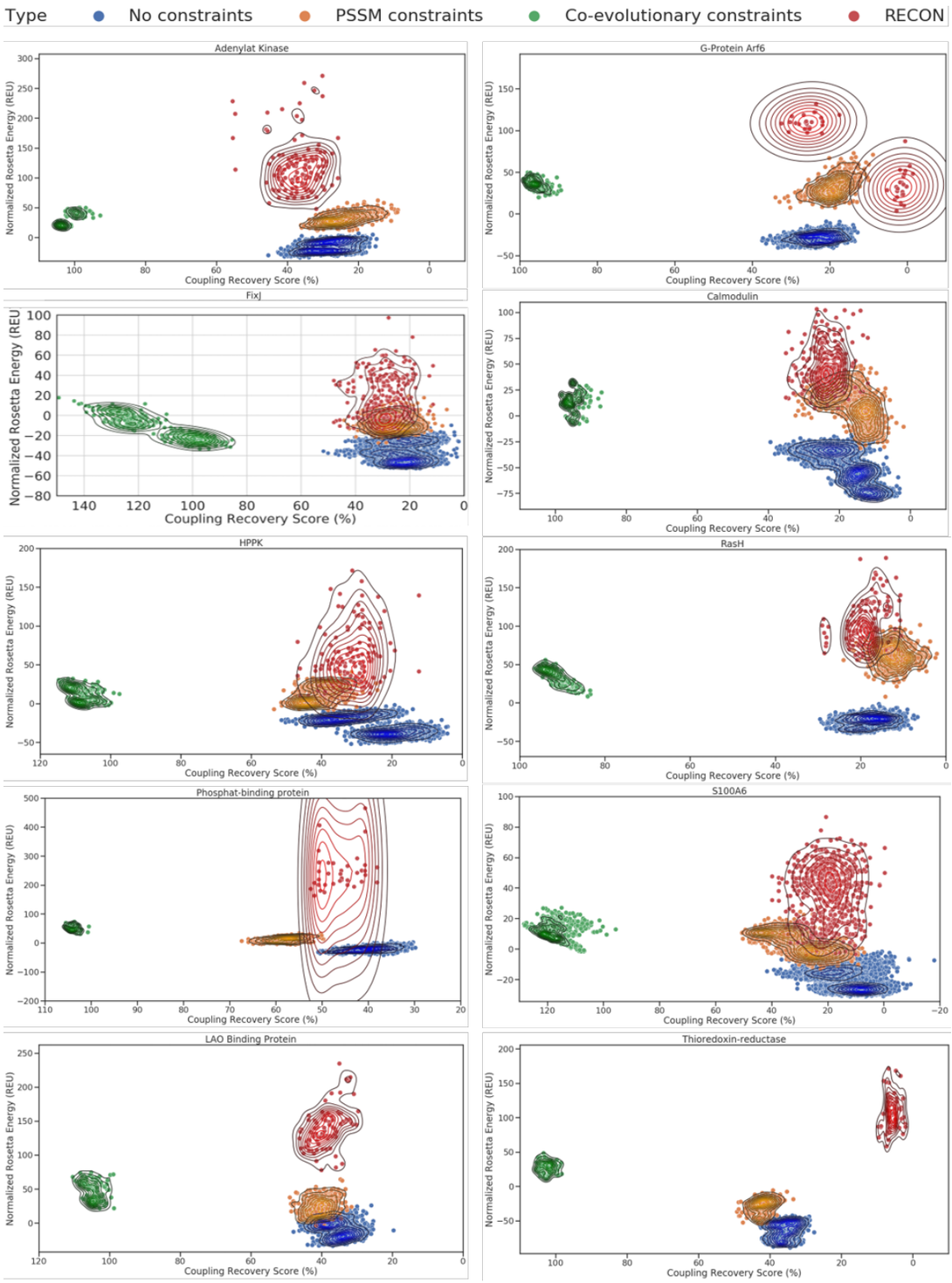

Figure 39: **Energy landscapes for designed sequences.** The normalized Rosetta Energy in REU is plotted against the Coupling Recovery Score *crs* (%). Each dot represents a designed sequence and is colored by the used design protocol. Each plot is one of the ten proteins used as benchmark. (Note that in typical Rosetta fashion the x-axis is inverted, to highlight the energy funnel. Additionally the *crs* is represented in % to enable comparison between the proteins.)

11.2.3 A network of coupled residues is involved in the binding of ATP in the HPPK

The 6-Hydroxymethyl-7,8-dihydropterin pyrophosphokinase (HPPK) belongs to a class of enzymes catalyzing the pyrophosphoryl transfer from ATP to 6-hydroxymethyl-7,8-dihydropterin (HP), which is the first reaction in the folate pathway (Xiao et al. 1999; Switzer and Gibson 1978; Blaszczyk et al. 2000). Here, the residue interaction network formed by the 20% highest coupled

residues $res_{cc}^{20}(HPPK)$ is mostly involved in the binding of ATP (Q74, R88, W89, H115, Y116, R121) (Fig S2, red network). In contrast, the other found small networks (purple, green, yellow) are not annotated in the literature. Other residues critical for the substrate/cofactor binding besides ATP were not part of our found networks, but some of the residues are conserved. This conservation excludes co-evolutionary mutations with other residues. Comparing the sequence logos of designed sequences reveals that there is almost no difference in sampling between RoSSD and RECON MSD (Fig S3). This lack of improvement is not surprising since the RMSD between the two conformational states is only 0.5Å, and optimizing over two almost identical state has no additional benefit. In the case of the two residues H115 and R121, sequences designed with SeqProf constraints sample the native amino acid more often than sequences designed with ResCue. Analyzing the MSA revealed that these two positions are conserved residues, preventing a co-evolutionary signal.

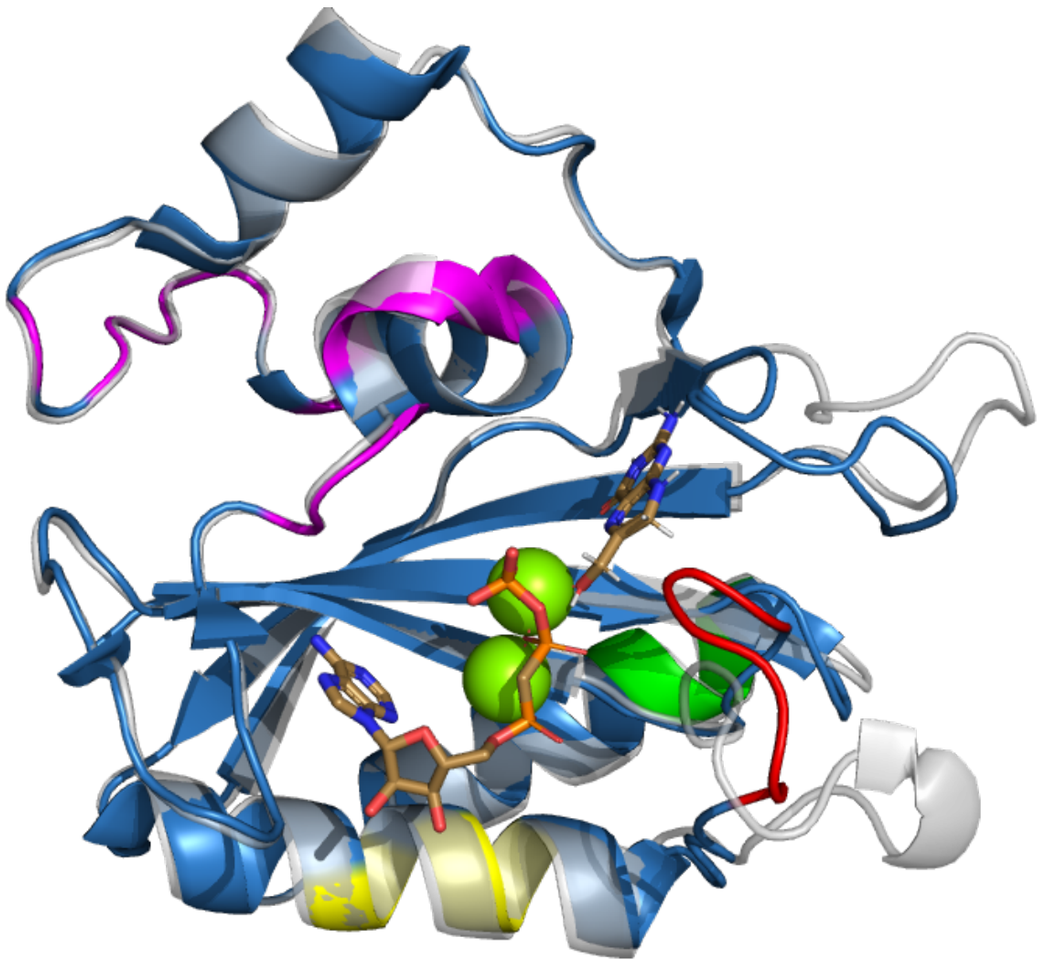

Figure 40: **Localization of highly coupled residues in HPPK.** Network of highly coupled residues (red, yellow, green, purple) displayed on the structure of HPPK (PDB ID: Unbound 1HKA, Bound 1Q0N). Alignment of the bound state (blue) and the unbound state (grey). The substrate ATP and a HP analog are shown as sticks. Bound magnesium is depicted as green spheres.

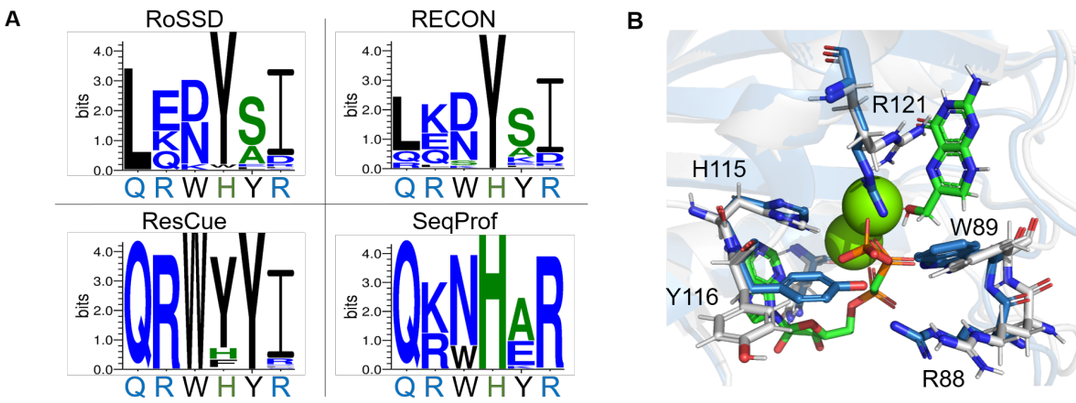

Figure 41: **Sequence logos resulting from four design protocols for HPPK.** The native sequences are listed below the logos. (**A**) Shown are sequence logos for six binding site residues for each design approach. (**B**) Graphical view of the binding site residues and the substrates ATP and a HP analog (green) shown in stick representation. Native structure in closed state is depicted in blue, while a protein designed with ResCue is shown in grey. For the protein designs, the ligands were not part of the starting structure. (Bound state PDB ID: 1Q0N)

11.2.4 The calcium sensor mechanism of S100A6 relies on the coupled residues at the two binding sites

S100A6 is a member of the S100 family of calcium-binding proteins and undergoes a conformational shift after binding (Otterbein et al. 2002). The protein functions as a calcium sensor through a helix-loop-helix motif called EF-hand, similar to calmodulin and troponin C (Lewit-Bentley and Réty 2000). Here, the residue interaction network formed by the 20% highest coupled residues $res_{cc}^{20}(S100A6)$ is separated into two distinct networks at the two calcium-binding sites (Fig S4). Interestingly, while the red network at the second binding site captures all crucial binding residues, the yellow network at the first binding site is small and only in close spacial proximity. This result is perhaps explained by the fact that the coordination of the calcium in the first site primarily involves main chain carbonyls compared to the involvement of side chains in the second binding site.

Comparing the sequence logos of residues crucial for calcium-binding at both sites of different design approaches further demonstrates the difference between the two binding sites (Fig S5). Here, we expected RECON do perform better at sampling the native amino acids at the second binding site than the first. This expectation is based on the fact that a large conformational shift occurs at the second binding site while the first site shows almost no change. Indeed, RECON samples the native amino acids at the second binding site more often than at the first. While the residues at the first binding site are already well sampled by RoSSD design, the residues at the second site have only in ResCue designs the native amino acids.

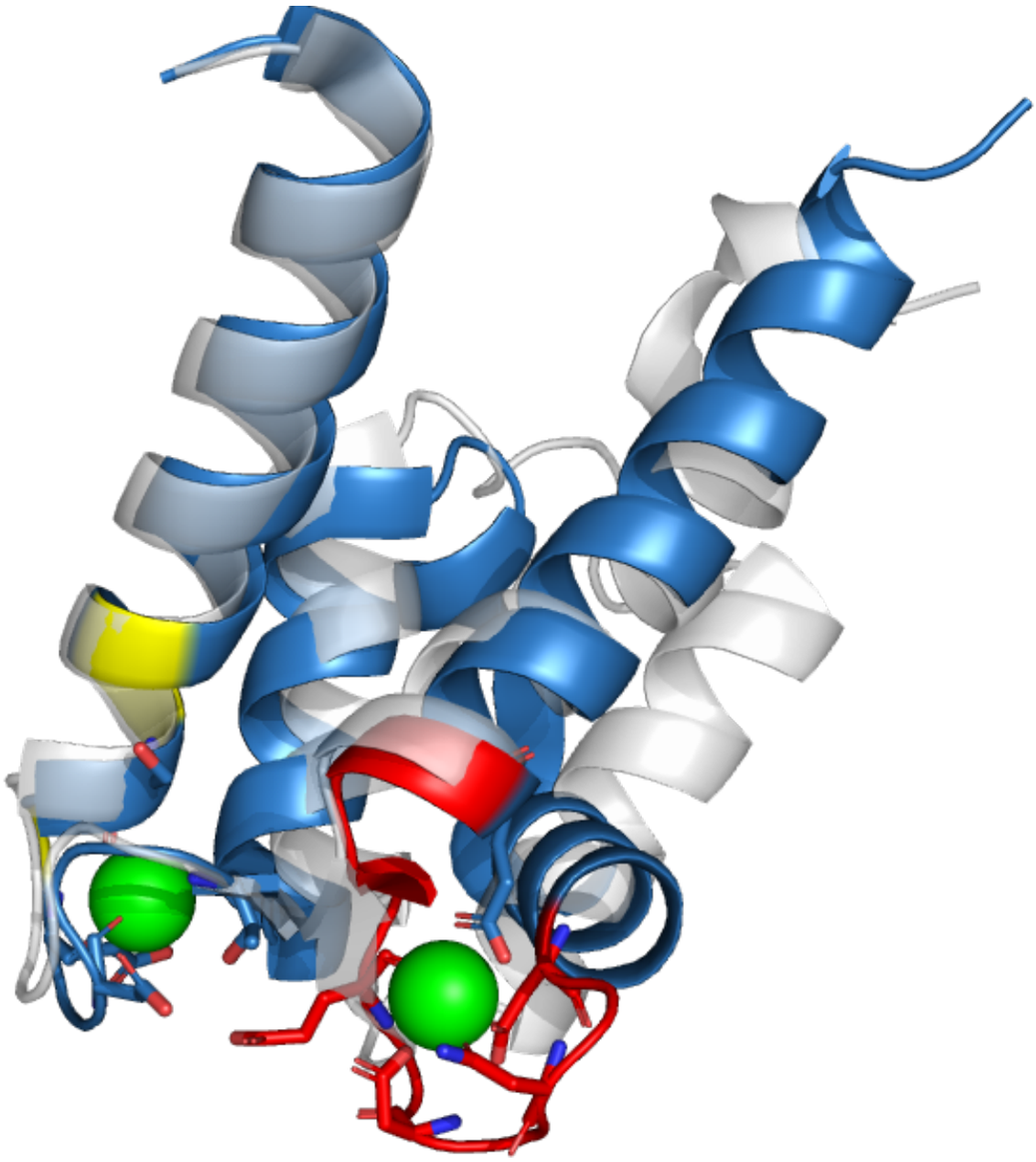

Figure 42: **Localization of highly coupled residues in S100A6.** Network of highly coupled residues (red, yellow) displayed on the structure of S100A6 (PDB ID: Calcium bound 1K9K, Calcium free 1K9P). Alignment of the calcium bound state (blue) and the calcim free state (grey). Bound magnesium is depicted as green spheres. Residues critical for calcium-binding are shown as sticks.

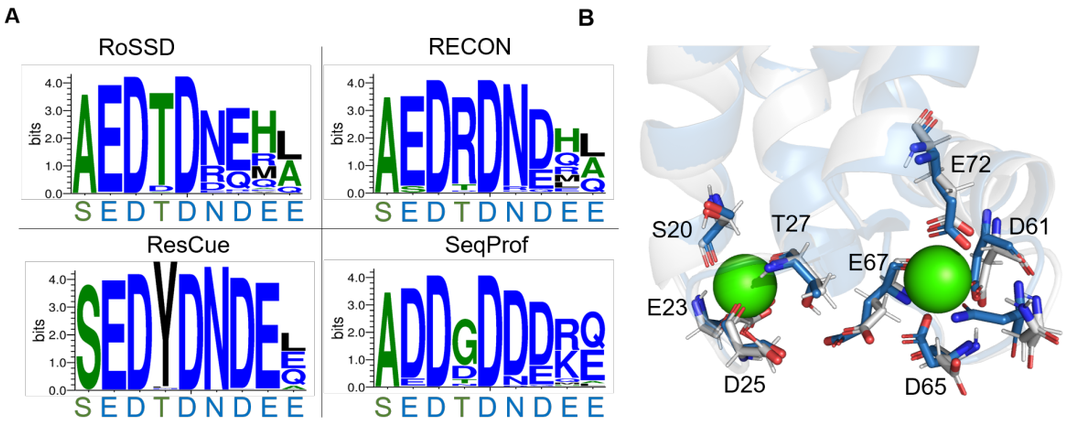

Figure 43: **Sequence logos resulting from four design protocols for S100A6.** The native sequences are listed below the logos. (**A**) Shown are sequence logos for nine binding site residues for each design approach. (**B**) Native structure in calcium bound state is depicted in blue, while a protein designed with ResCue is shown in grey. For the protein designs, the ligands were not part of the starting structure. (Bound state PDB ID: 1K9K)

11.2.5 A network of coupled residues contributes to the conformational shift after phosphate binding in the Phosphate-Binding Protein

The phosphate-binding protein (PBP) is responsible for the active transport of phosphate in bacterial cells and is highly specific for phosphate. The binding of phosphate is stabilized by twelve hydrogen bonds, as well as one salt link (Yao et al. 1996). We found that the residue interaction network formed by the 20% highest coupled residues $res_{cc}^{20}(PBP)$ connects the positions forming hydrogen bonds with the phosphate with residues further away from the binding site that undergo conformational changes upon binding (Fig S6, Residues T10, F11, A13, Y33, S38, D56, N137,R135, S139, G140, T141, S142, G176, N177, E195, Y198, T256, F257). Analyzing the designed sequences of our different approaches with sequence logos highlighted how only our ResCue approach samples the residues necessary to create hydrogen bonds with the ligand (Fig S7). The only residue that is not sampled in all different methods is R135, which forms a hydrogen bond with a water molecule. Again, water is not commonly included in protein design and thus remains a challenge for Rosetta.

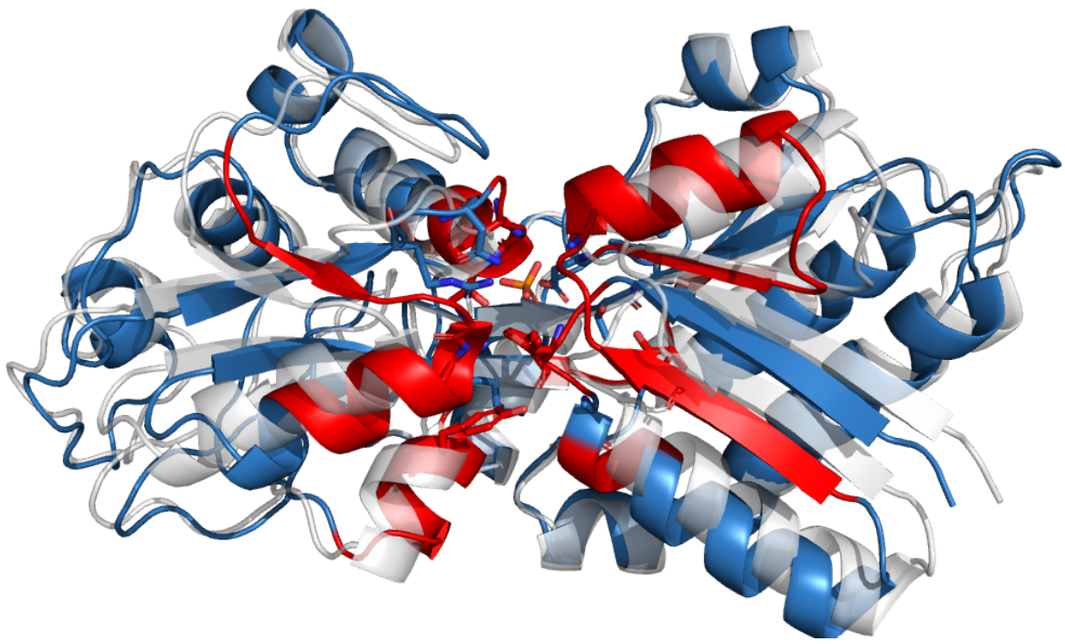

Figure 44: **Localization of highly coupled residues in PBP.** Network of highly coupled residues (red) displayed on the structure of PBP (PDB ID: Unbound 1OIB, Bound 1QUK). Alignment of the bound state (blue) and the unbound state (grey). The substrate phosphate is shown as sticks. Residues known to be crucial for substrate binding are shown as sticks.

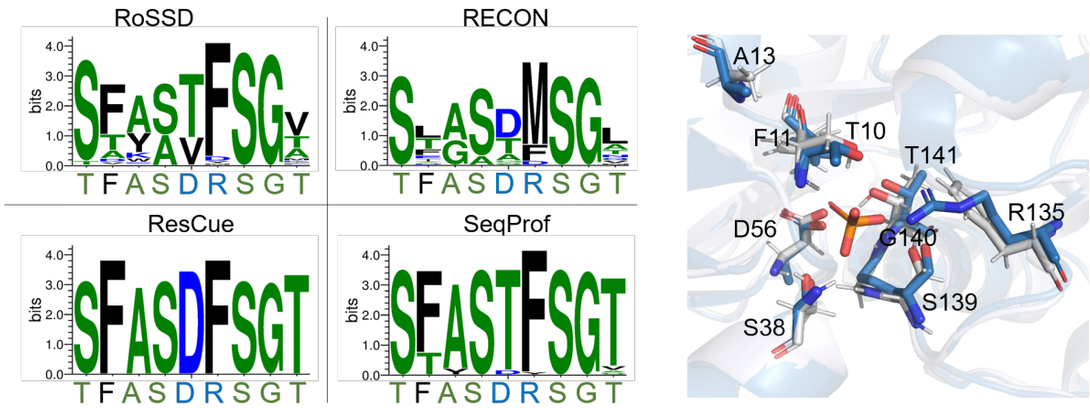

Figure 45: **Sequence logos resulting from four design protocols for PBP.** The native sequences are listed below the logos. (**A**) Shown are sequence logos for nine binding site residues for each design approach. (**B**)Graphical view of the binding site residues and the ligand phosphate (orange) shown in stick representation. Native structure in closed state is depicted in blue, while a protein designed with ResCue is shown in grey. For the protein designs, the ligand was not part of the starting structure. (Bound state PDB ID: 1QUK)

11.2.6 A network of coupled residues is involved in the binding of GTP in the small G protein Arf6-GDP

Arf6 localizes at the periphery of the cell and plays an essential role in endocytotic pathways (Ménétrey et al. 2000; Pasqualato et al. 2001). Here, the residue interaction network formed by the 20% highest coupled residues $res_{cc}^{20}(Arf6)$ is involved in the binding of GTP (T41, I42, D63,

V64, G65, G66) (Fig S9, red network). Comparing the sequence logos of designed sequences reveals that RoSSD often samples the native amino acids at positions 41, 63, 64 and 65. RECON does a slightly better job at sampling the native sequence, for example in the case of G66 (Fig S8). In the case of the residues I42 and D63 and G66, only ResCue designs show a clear bias towards the native sequence.

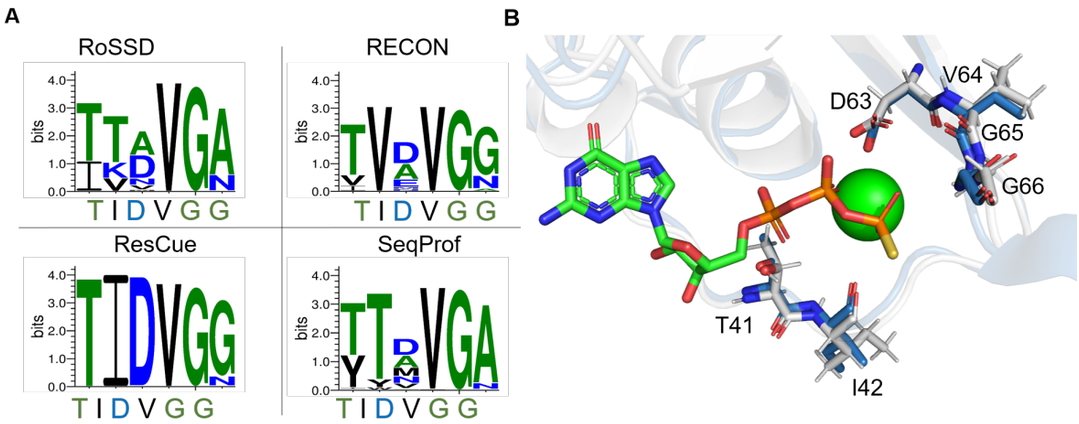

Figure 46: **Sequence logos resulting from four design protocols for Arf6.** The native sequences are listed below the logos. (**A**) Shown are sequence logos for six binding site residues for each design approach. (**B**) Graphical view of the binding site residues and the ligand shown in stick representation. Native structure in GTP bound state is depicted in blue, while a protein designed with ResCue is shown in grey. For the protein designs, the ligand was not part of the starting structure. (PDB ID: With GDP 1E0S, With GTP 2J5X)

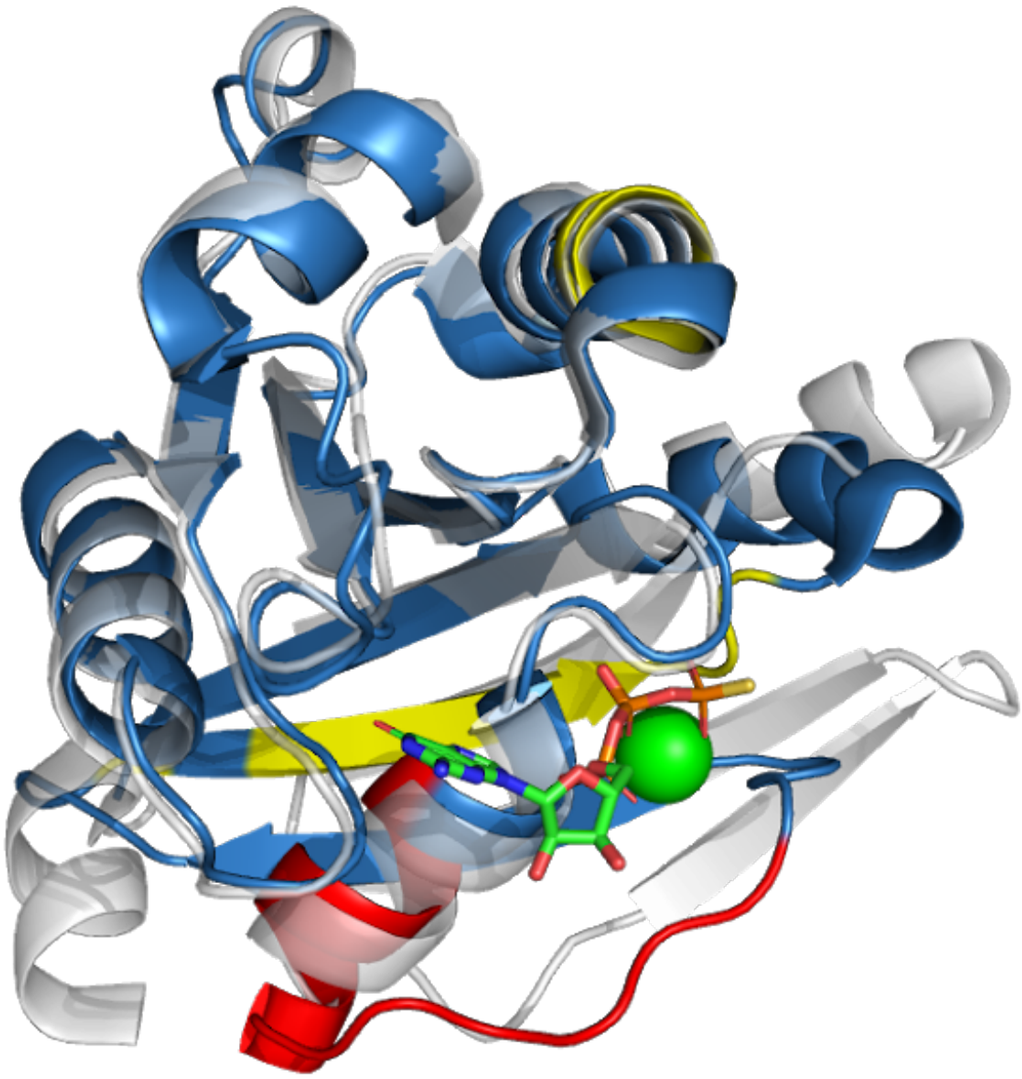

Figure 47: **Localization of highly coupled residues in Arf6.** Network of highly coupled residues (red) displayed on the structure of Arf6 (PDB ID: With GDP 1E0S, With GTP 2J5X). Alignment of the GTP bound state (blue) and the GDP bound state (grey). The substrate is shown as sticks.

11.2.7 A network of coupled residues is involved in the binding of AMP in the Adenylate Kinase

Adenylate kinases are nucleoside monophosphate (NMP) kinases and consist of a large CORE domain, a small NMP-binding domain and a LID domain (Müller, Schlauderer, et al. 1996; Müller and Schulz 1992). Here, the residue interaction network formed by the 20% highest coupled residues $res_{cc}^{20}(AdenylatKinase)$ is involved in the binding of AMP (Network residues: 29-40, 57, 58, 60, 61, 81-93, Binding residues: T31, R36, K57, L58, G85, F86, P87, R88) (Fig S10, red network). Comparing the sequence logos of designed sequences reveals that all methods sample the native amino acids for G85, F86 and P87 (Fig S11). However, in the case of the residues T31, R36 and K57, only ResCue designs show a clear bias towards the native sequence.

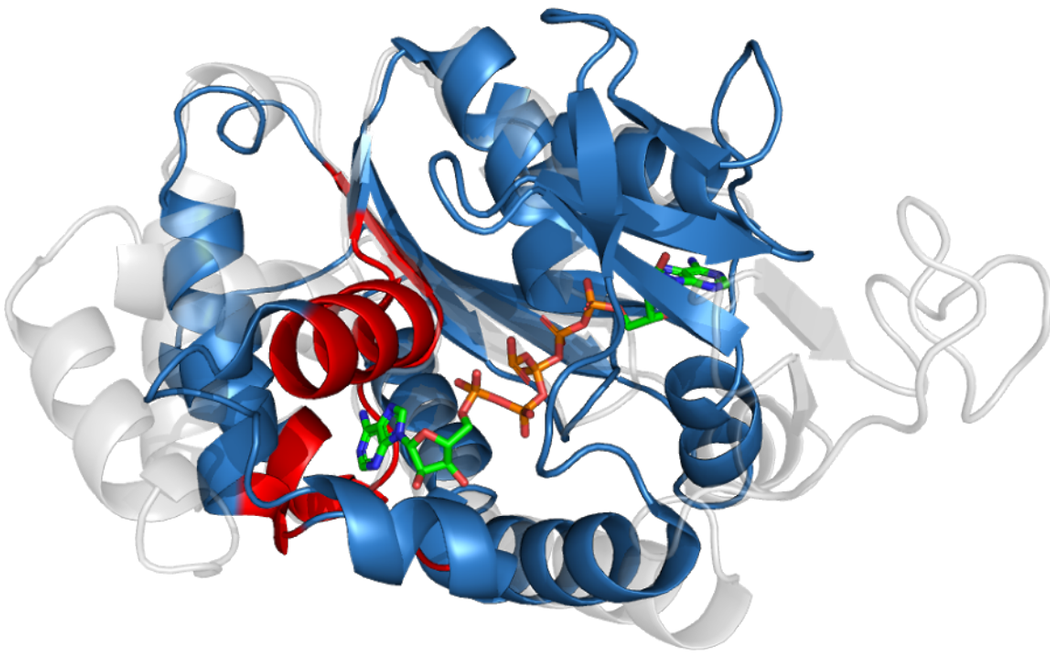

Figure 48: **Localization of highly coupled residues in the Adenylate kinase.** Network of highly coupled residues (red) displayed on the structure of the Adenylate kinase (PDB ID: Ap5A bound 1AKE, unbound 4AKE). Alignment of the Ap5A bound state (blue) and the unbound state (grey). The substrate is shown as sticks.

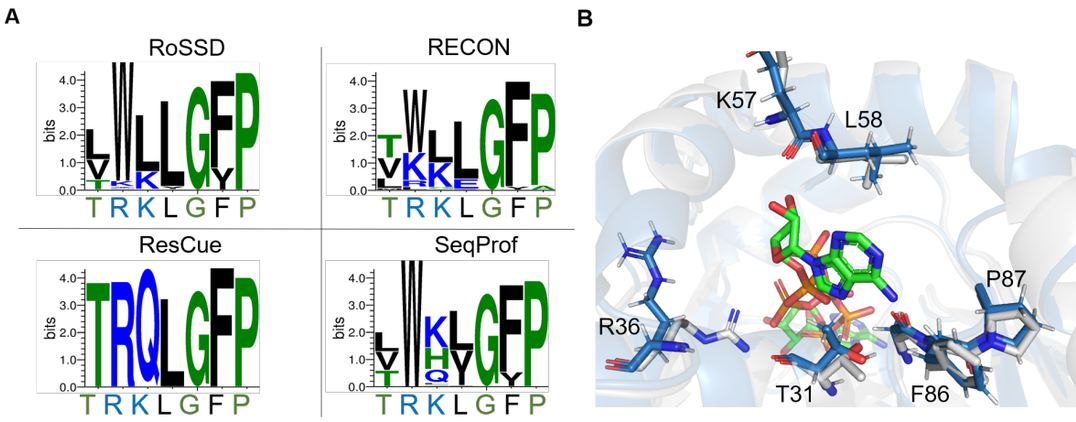

Figure 49: **Sequence logos resulting from four design protocols for the Adenylate kinase.** The native sequences are listed below the logos. (**A**) Shown are sequence logos for seven binding site residues for each design approach. (**B**) Graphical view of the binding site residues and the ligand shown in stick representation. Native structure in Ap5A bound state is depicted in blue, while a protein designed with ResCue is shown in grey. For the protein designs, the ligand was not part of the starting structure. (PDB ID: Ap5A bound 1AKE, unbound 4AKE)

11.2.8 A network of coupled residues is involved in the binding of FAD in the Thioredoxin reductase

In the thioredoxin reductase, cycles of reduction and reoxidation of FAD depend on rate-limiting rearrangements of the FAD and NADPH domains (Lennon, Williams, and Ludwig 2000; Waksman et al. 1994). Here, the residue interaction network formed by the 20% highest coupled residues

$res_{cc}^{20}(Thioredoxinreductase)$ is involved in the intearction with FAD (Network residues: 44-53, 132-144, 159-168, 172, 173, 179-182, 184, 291-309, Binding residues: N51, R181, R293, Q294, A295) (Fig S12, red network). Comparing the sequence logos of designed sequences reveals that RoSSD samples the native amino acids for N51 and Q294(Fig S13). However, in the case of the residues R181 and R293, only ResCue designs show a clear bias towards the native sequence.

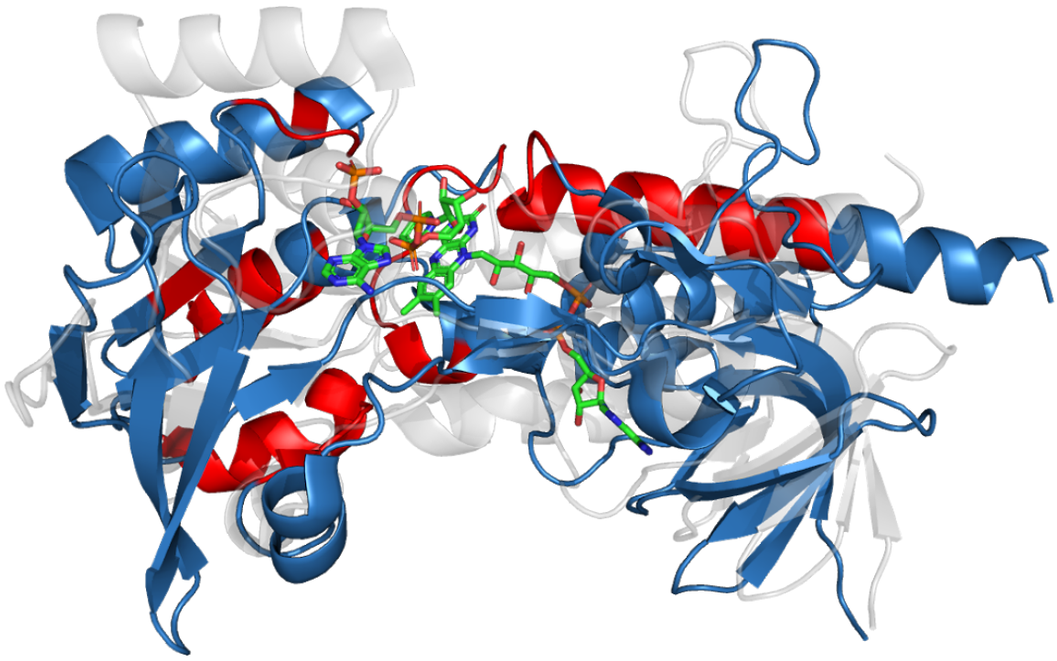

Figure 50: **Localization of highly coupled residues in the Thioredoxin reductase.** Network of highly coupled residues (red) displayed on the structure of the Thioredoxin reductase (PDB ID: AADP+ bound 1F6M, unbound 1E0S). Alignment of the AADP+ bound state (blue) and the unbound state (grey). The substrate is shown as sticks.

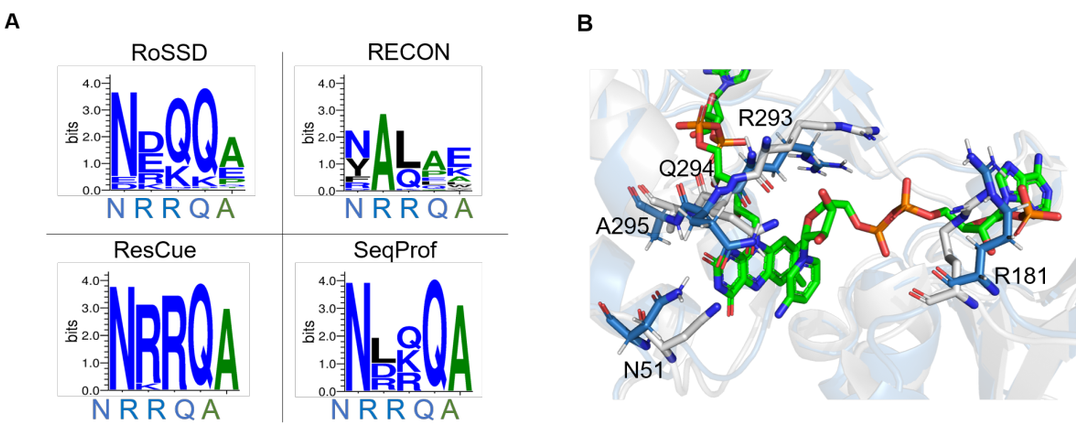

Figure 51: **Sequence logos resulting from four design protocols for the Thioredoxin reductase.** The native sequences are listed below the logos. (**A**) Shown are sequence logos for five binding site residues for each design approach. (**B**) Graphical view of the binding site residues and the ligand shown in stick representation. Native structure in AADP+ bound state is depicted in blue, while a protein designed with ResCue is shown in grey. For the protein designs, the ligand was not part of the starting structure. (PDB ID: AADP+ bound 1F6M, unbound 1E0S)

11.2.9 Rosetta Design Protocols

11.2.10 Clean and relax

All proteins were cleaned and relaxed before design. An ensemble of five relaxed structures was used as a starting point for all designs.

Listing 1: PDB cleaning commands

```
./Rosetta/tools/protein_tools/scripts/clean_pdb.py $PDBid $CHAIN
```

Listing 2: Rosetta relax commands

```
./Rosetta/main/source/bin/relax.default.linuxgccrelease -s $PDB -use_input_sc -
   ↪ nstruct 5 -relax:constrain_relax_to_start_coords -scorefile relax.fasc -out
   ↪ :suffix _relax
```

Listing 3: RosettaScripts design command file for unconstrained designs.

```
./Rosetta/main/source/bin/rosetta_scripts.default.linuxgccrelease @design.options
-parser:protocol /design.xml -out:suffix_design -scorefile design.fasc -s $pdb
```

Listing 4: RosettaScripts XML file for unconstrained designs.

```
<ROSETTASCRIPTS>
        <SCOREFXNS>
        </SCOREFXNS>
        <TASKOPERATIONS>
                <InitializeFromCommandline name="ifcl"/>
        <ReadResfile name="rrf" filename="design.resfile"/>
        </TASKOPERATIONS>
        <MOVERS>
        <PackRotamersMover name="design" scorefxn="REF2015" task_operations="ifcl,
            ↪ rrf" />
        </MOVERS>
        <FILTERS>
        </FILTERS>
        <APPLY_TO_POSE>
        </APPLY_TO_POSE>
        <PROTOCOLS>
        <Add mover="design" />
        </PROTOCOLS>
        <OUTPUT scorefxn="REF2015" />
</ROSETTASCRIPTS>
```

11.2.11 Unconstraint Rosetta Single State Design (RoSSD)

The following options were used to design proteins in the benchmark without any additional constraints (control group).

Listing 5: Rosetta design options and residue file for unconstrained designs.

```
-linmem_ig 5 -ex1 -ex2 -nstruct 10000
design.resfile:
ALLAAxc
start
```

11.2.12 Design with co-evolutionary constraints (ResCue)

The following options and commands were used for the new ResCue protocol (with the same options and resfile as above).

Listing 6: RosettaScripts command for co-evolutionary constraint designs.

```
./Rosetta/main/source/bin/rosetta_scripts.default.linuxgccrelease @design.options
-parser:protocol design.xml -out:suffix _design -scorefile design.fasc -s $pdb
```

Listing 7: RosettaScripts XML file for co-evolutionary constraint designs.

```
<ROSETTASCRIPTS>
        <SCOREFXNS>
                <ScoreFunction name="scorefxn_cst" weights="ref2015.wts">
                        <Reweight scoretype="res_type_linking_constraint" weight
                            ↪ ="1.0"/>
                </ScoreFunction>
                <ScoreFunction name="scorefxn" weights="ref2015.wts"/>
        </SCOREFXNS>
        <TASKOPERATIONS>
                <InitializeFromCommandline name="ifcl"/>
                <ReadResfile name="rrf" filename="design.resfile"/>
        </TASKOPERATIONS>
        <MOVERS>
                <AddResidueCouplingConstraint name="favor" tensor_file=".
                    ↪ tensorBinary.bin" index_file="indexList" strength="1.0"
                    ↪ alphabet="ARNDCQEGHILKMFPSTWYV-"/>
                <PackRotamersMover name="design" scorefxn="scorefxn_cst"
                    ↪ task_operations="ifcl,rrf" />
        </MOVERS>
        <FILTERS>
        </FILTERS>
```

```
        <PROTOCOLS>
                <Add mover="favor" />
                <Add mover="design" />
        </PROTOCOLS>
        <OUTPUT scorefxn="scorefxn" />
</ROSETTASCRIPTS>
```

11.2.13 RECON Multistate Designs (MSD)

Following commands and options were used for RECON multistate design (with the same Options and resfile as above):

Listing 8: RosettaScripts design command for RECON MSD design.

```
./Rosetta/main/source/bin/recon.default.linuxgccrelease @design.options -parser:
   ↪ protocol design.xml -out:suffix _multiDesign -scorefile design.fasc -s $pdb
   ↪  $pdb2
```

Listing 9: RosettaScripts XML file for RECON design.

```
<ROSETTASCRIPTS>


        <TASKOPERATIONS>
                <InitializeFromCommandline name="ifcl"/>
        </TASKOPERATIONS>
        <MOVERS>


                <PackRotamersMover name="design" scorefxn="REF2015" task_operations
                    ↪ ="ifcl" />


                <MSDMover name="msd1" design_mover="design" constraint_weight="0.5"
                    ↪  resfiles="design.resfile, design.resfile" />
                <MSDMover name="msd2" design_mover="design" constraint_weight="1"
                    ↪ resfiles="design.resfile, design.resfile"/>
                <MSDMover name="msd3" design_mover="design" constraint_weight="1.5"
                    ↪  resfiles="design.resfile, design.resfile" />
                <MSDMover name="msd4" design_mover="design" constraint_weight="2"
                    ↪ resfiles="design.resfile, design.resfile" />


                <FindConsensusSequence name="finish" scorefxn="REF2015" resfiles="
                    ↪ design.resfile, design.resfile" />
        </MOVERS>
        <FILTERS>
```

```
        </FILTERS>
        <PROTOCOLS>
                <Add mover="msd1" />
                <Add mover="msd2" />
                <Add mover="msd3" />
                <Add mover="msd4" />

                <Add mover="finish" />
        </PROTOCOLS>
        <OUTPUT scorefxn="REF2015" />
</ROSETTASCRIPTS>
```

11.2.14 Design with a position specific scoring matrix (PSSM)

Following commands were used to design with a PSSM.

Listing 10: RosettaScripts design command and XML file for design constraint with PSSM.

```
./Rosetta/main/source/bin/rosetta_scripts.default.linuxgccrelease
@design.options -parser:protocol design.xml -out:suffix _design -scorefile design
    ↪ .fasc -s $pdb
```

Listing 11: RosettaScripts XML for design constraint with PSSM.

```
<ROSETTASCRIPTS>
        <SCOREFXNS>
        <ScoreFunction name="scorefxn" weights="ref2015.wts">
                <Reweight scoretype="res_type_constraint" weight="0.0"/>
        </ScoreFunction>
        <ScoreFunction name="scorefxn_cst" weights="ref2015.wts">
                <Reweight scoretype="res_type_constraint" weight="1.0"/>
        </ScoreFunction>
        </SCOREFXNS>
        <TASKOPERATIONS>
                <InitializeFromCommandline name="ifcl"/>
                <ReadResfile name="rrf" filename="design.resfile"/>
        </TASKOPERATIONS>
        <MOVERS>
                <FavorSequenceProfile name="favorSequence" scaling="global" weight
                    ↪ ="5" pssm="pssm.txt" scorefxns="scorefxn_cst" />
                <PackRotamersMover name="design" scorefxn="scorefxn_cst"
                    ↪ task_operations="ifcl,rrf" />
        </MOVERS>
        <FILTERS>
        </FILTERS>
        <PROTOCOLS>
                <Add mover="favorSequence"/>
                <Add mover="design" />
        </PROTOCOLS>
        <OUTPUT scorefxn="scorefxn" />
</ROSETTASCRIPTS>
```

11.2.15 Design favoring the wild-type sequence

Following commands were used to design with a limited amount of mutations.

Listing 12: RosettaScripts design command and XML file for design constraint to the native sequence.

```
./Rosetta/main/source/bin/rosetta_scripts.default.linuxgccrelease
@design.options -parser:protocol design.xml -out:suffix _design -scorefile design
    ↪ .fasc -s $pdb -"parser:script_vars weight=WEIGHT"
```

Listing 13: RosettaScripts XML for design constraint to the native sequence.

```
<ROSETTASCRIPTS>
    <SCOREFXNS>
        <ScoreFunction name="scorefxn" weights="ref2015.wts">
            <Reweight scoretype="res_type_constraint" weight="0.0"/>
        </ScoreFunction>
        <ScoreFunction name="scorefxn_cst" weights="ref2015.wts">
            <Reweight scoretype="res_type_constraint" weight="%%weight%%"/>
        </ScoreFunction>
    </SCOREFXNS>
    <TASKOPERATIONS>
        <InitializeFromCommandline name="ifcl"/>
            <ReadResfile name="rrf" filename="/home/erteltm/moritz/Protein_Designs
                ↪ /Constrained/Full/design.resfile"/>
    </TASKOPERATIONS>
    <MOVERS>
        <FavorSequenceProfile name="favorSequence" weight="1.3" use_current="true"
            ↪  matrix="IDENTITY" scorefxns="scorefxn_cst" />
        <PackRotamersMover name="design" scorefxn="scorefxn_cst" task_operations="
            ↪ ifcl,rrf" />
    </MOVERS>
    <FILTERS>
    </FILTERS>
    <APPLY_TO_POSE>
    </APPLY_TO_POSE>
    <PROTOCOLS>
        <Add mover="favorSequence"/>
        <Add mover="design" />
    </PROTOCOLS>
    <OUTPUT scorefxn="scorefxn" />
</ROSETTASCRIPTS>
```

For each benchmark protein, the weight of the score function term *res_type_constraint* was optimized to roughly reflect the average native sequence recovery of the ResCue protocol. This allows to compare the coupling recovery across the protocols. Table 1 lists the used weights for the FavorNative protocol.

Table 5: Weights used for the FavorNative protocol for each benchmark protein.

PDB1	PDB2	Weight	Protein Description
1CKK	1CFD	1.8	Calmodulin
1EOS	2J5X	1.4	G-protein Arf6
6Q21	4Q21	1.25	RasH
1TDE	1F6M	1.0	Thioredoxin reductase
1QUK	1OIB	1.0	Phosphate-binding protein
2LAO	1LAF	0.8	LAO Binding protein
1K9P	1K9K	0.6	S100A6
1AKE	4AKE	0.7	Adenylate kinase
1HKA	1Q0N	0.7	HPPK
1D5W	1DBW	0.4	FixJ

11.2.16 ResCue full length sequence logos

Here, full length weblogos are provided for the ResCue design on the benchmark dataset (Figures S1-S10). The weblogos visualize high wild-type sequence recovery over the full protein length.

Figure 52: **Full sequence weblogo for the ResCue design on LAO**

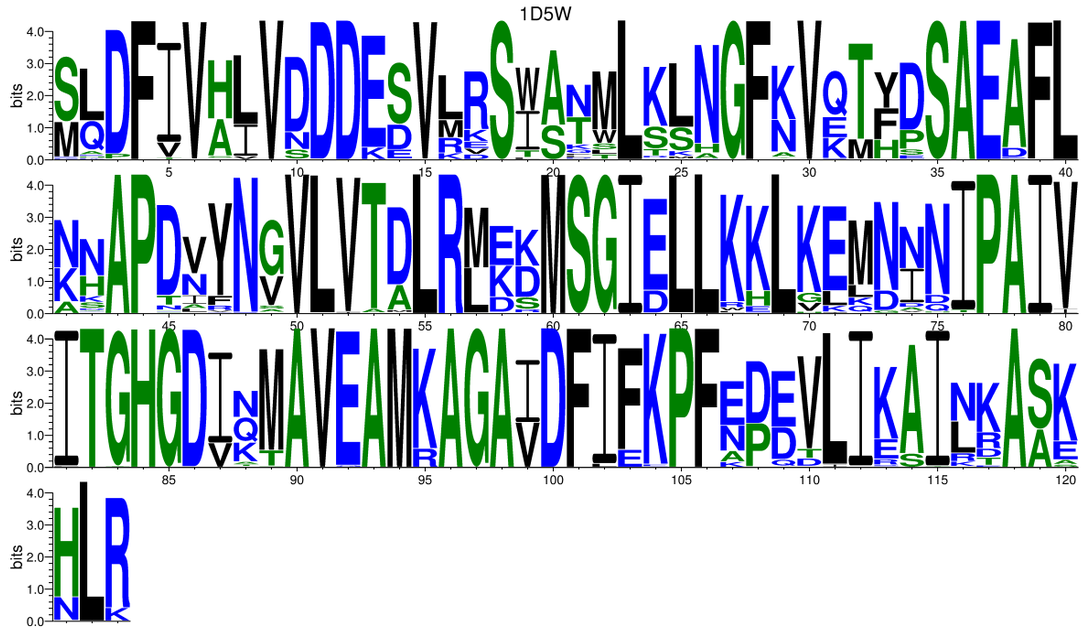

Figure 53: **Full sequence weblogo for the ResCue design on FixJ**

Figure 54: **Full sequence weblogo for the ResCue design on RasH**

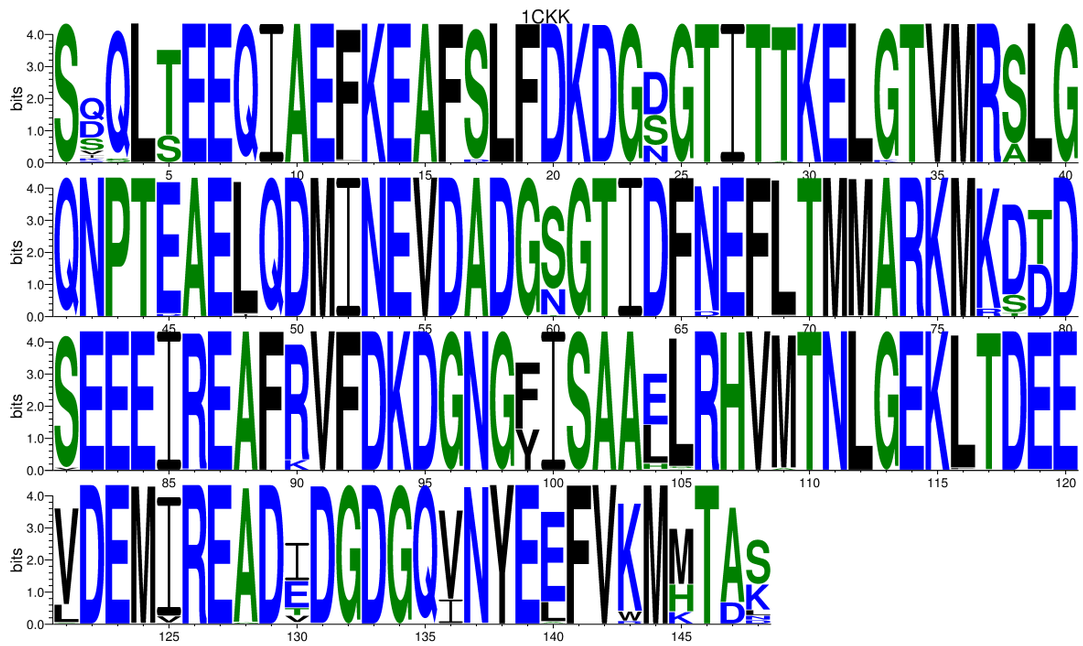

Figure 55: **Full sequence weblogo for the ResCue design on Calmodulin**

Figure 56: **Full sequence weblogo for the ResCue design on HPPK**

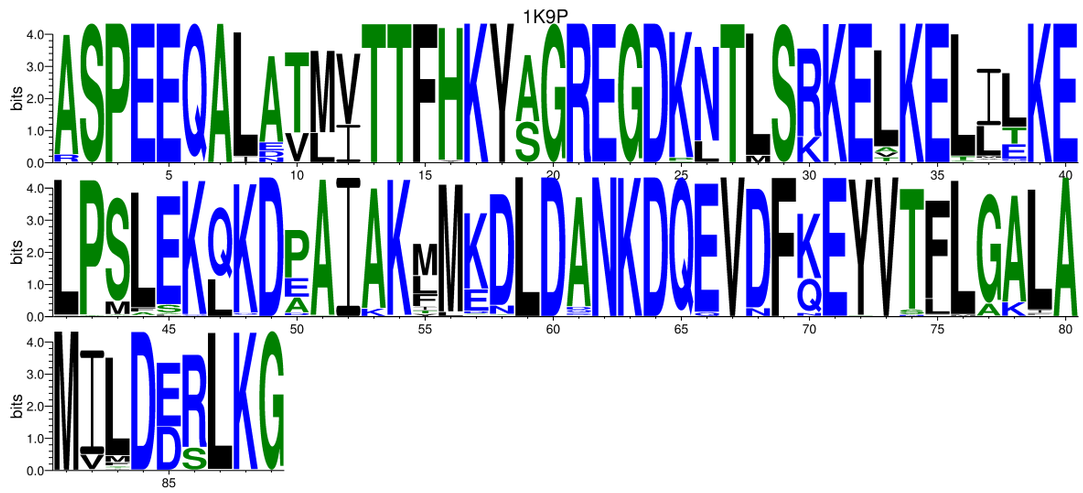

Figure 57: **Full sequence weblogo for the ResCue design on S100A6**

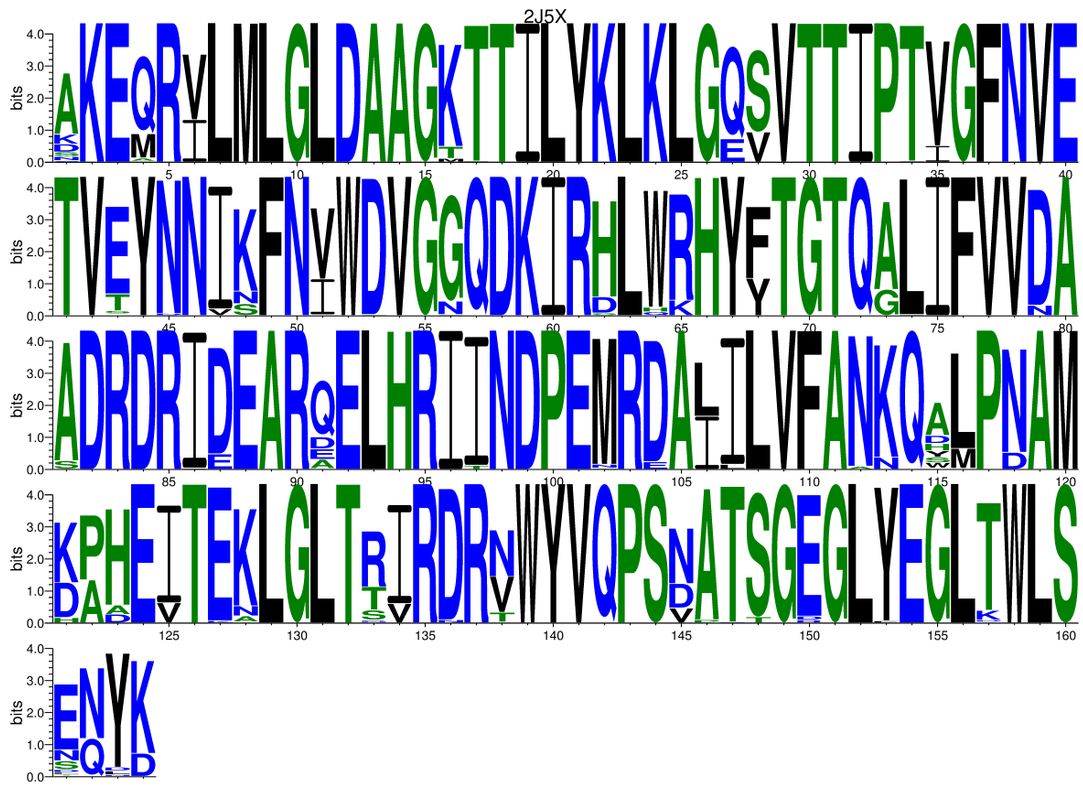

Figure 58: **Full sequence weblogo for the ResCue design on Arf 6**

Figure 59: **Full sequence weblogo for the ResCue design on thioredoxin reductase**

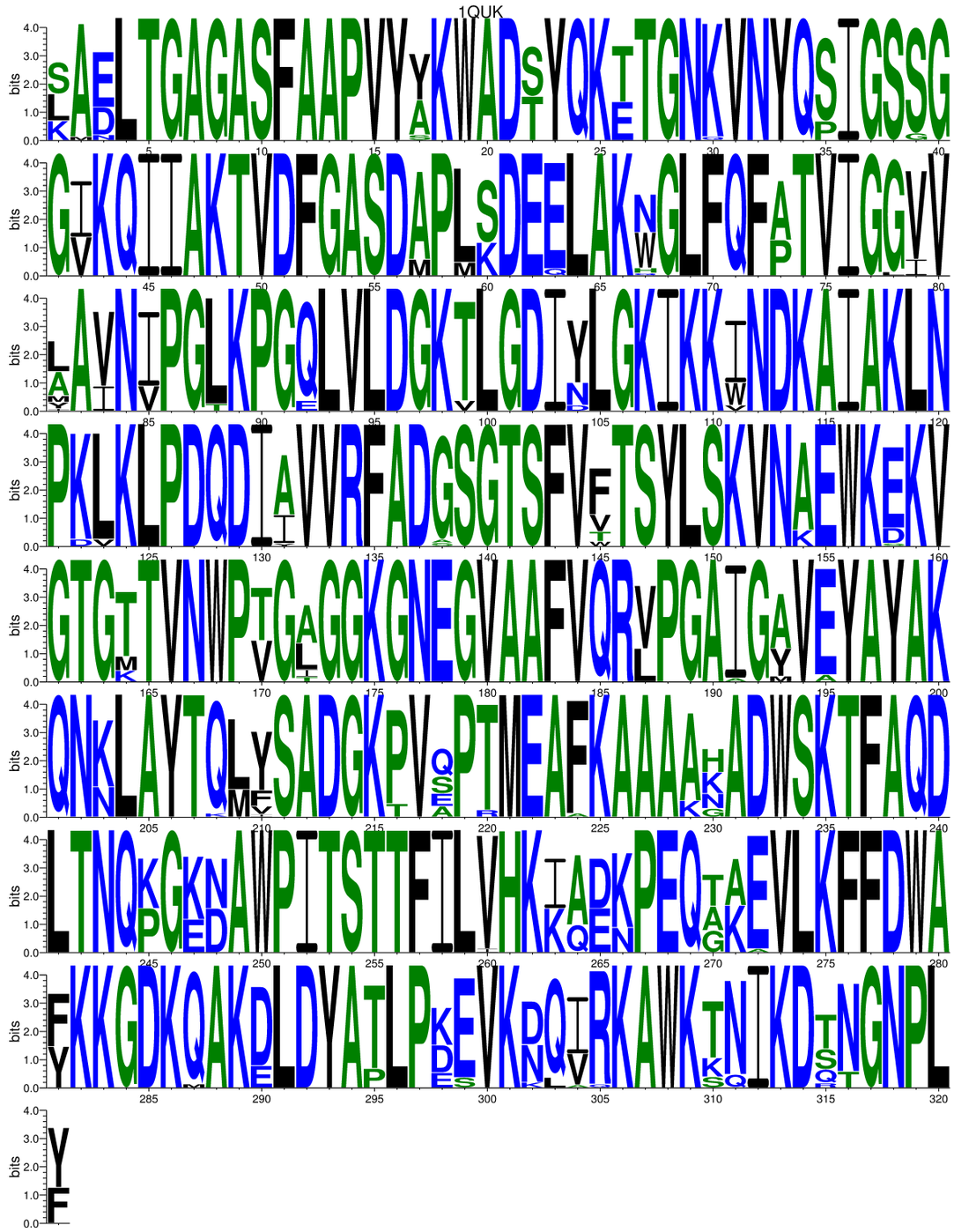

Figure 60: **Full sequence weblogo for the ResCue design on Phosphate binding protein**

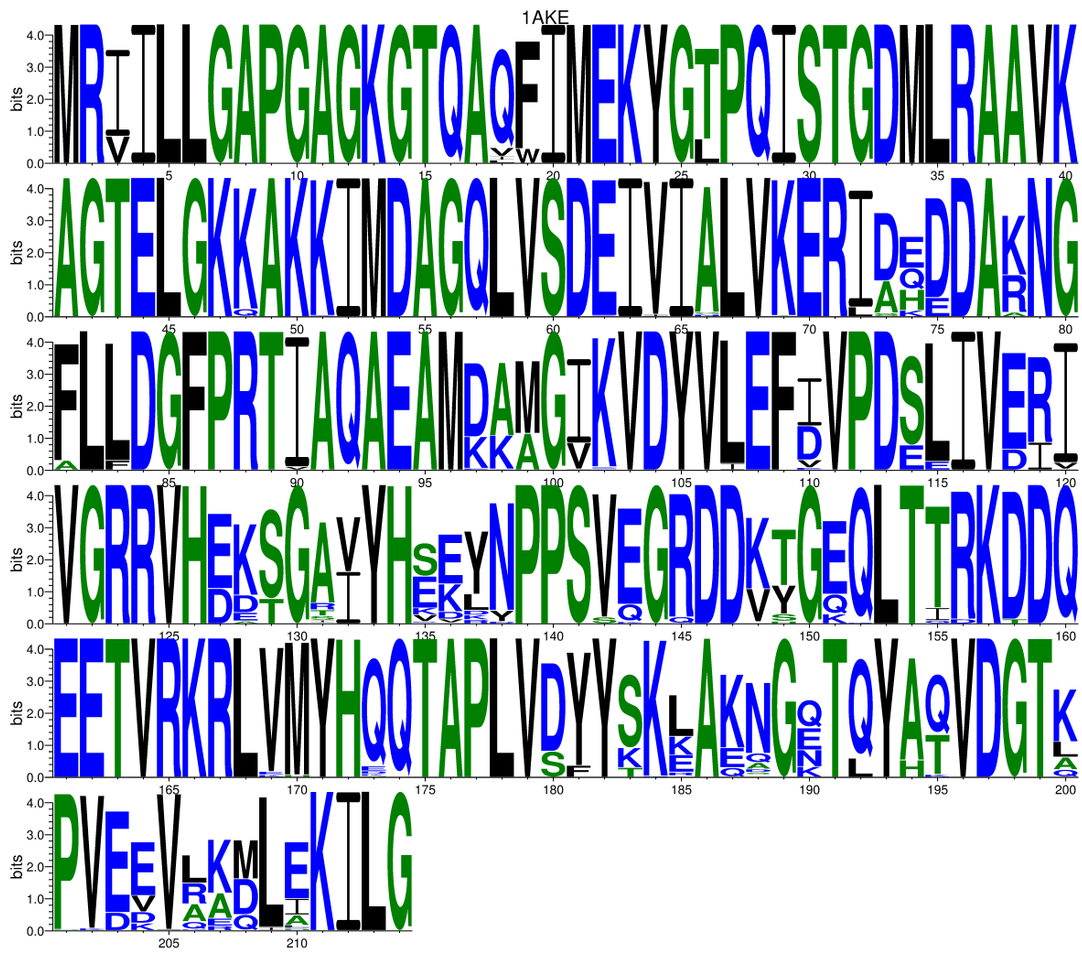

Figure 61: **Full sequence weblogo for the ResCue design on adenylate kinase**

11.2.17 Coupling strength of functionally relevant residues

Figure S1 shows the coupling strength cs(seq) of functional relevant residues in the benchmark proteins. Residues were chosen according when mentioned in literature as functional (see Methods of manuscript). The coupling strength cs(seq) of the chosen set of residues for each protein was calculated and visualized as a bar plot. Compared to the native sequence, the coupling strength of ResCue was on average 103±50%, followed by FavorNative with 35±55%, SeqProf with 23±58%, RECON with 20±63% and RoSSD with 4±56% (Fig8). The improvement of ResCue was statistically significant compared to all other design methods (MW p ¡ 5.0e-04) (Fig 8). The increased sequence recovery of the ResCue protocol can therefore be attributed to the collective interaction of couplings.

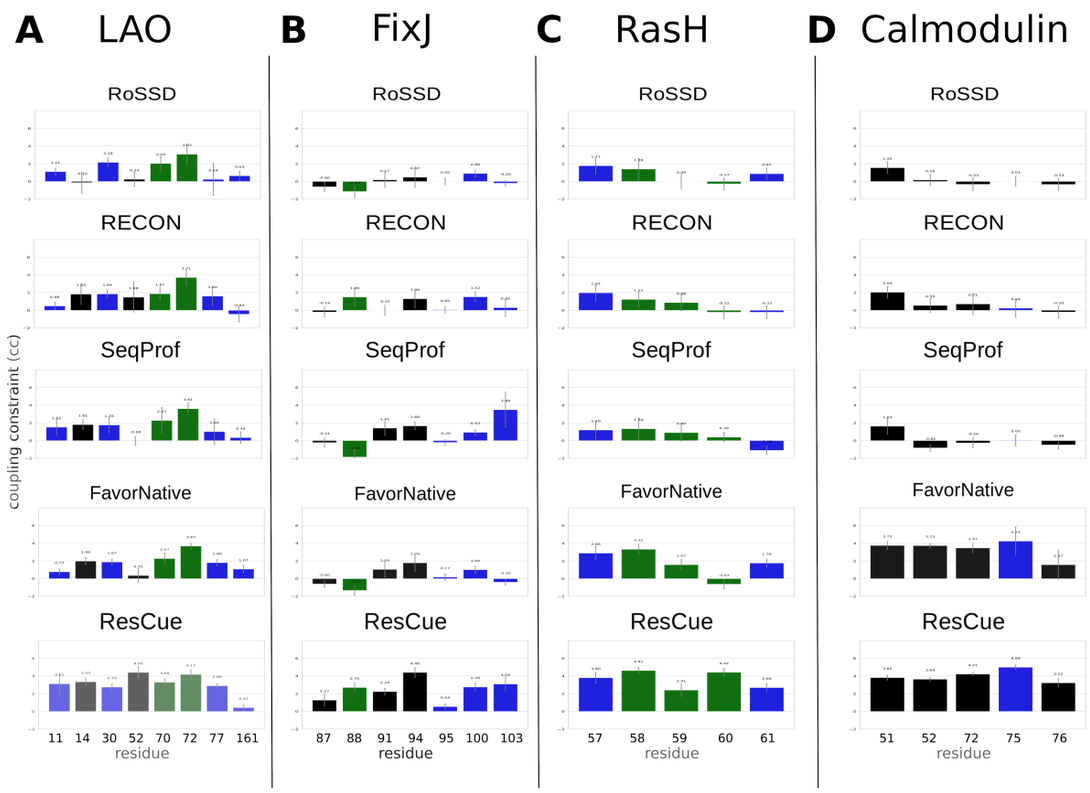

Figure 62: **Coupling strengths for residues relevant to function.** The methods are sorted by the average coupling strength, and increases in the order: RoSSD, RECON, SeqProf, FavorNative and ResCue (top to down). The observation was made for all proteins, (A) LAO binding site, eight residues. (B) FixJ dimer interface, seven residues (C) RasH binding site, five residues. (D) calmodulin-binding site, five residues.

11.3 Antibody expressability prediction and engineering using LSTM and Rosetta

11.4 Tensorflow model chart

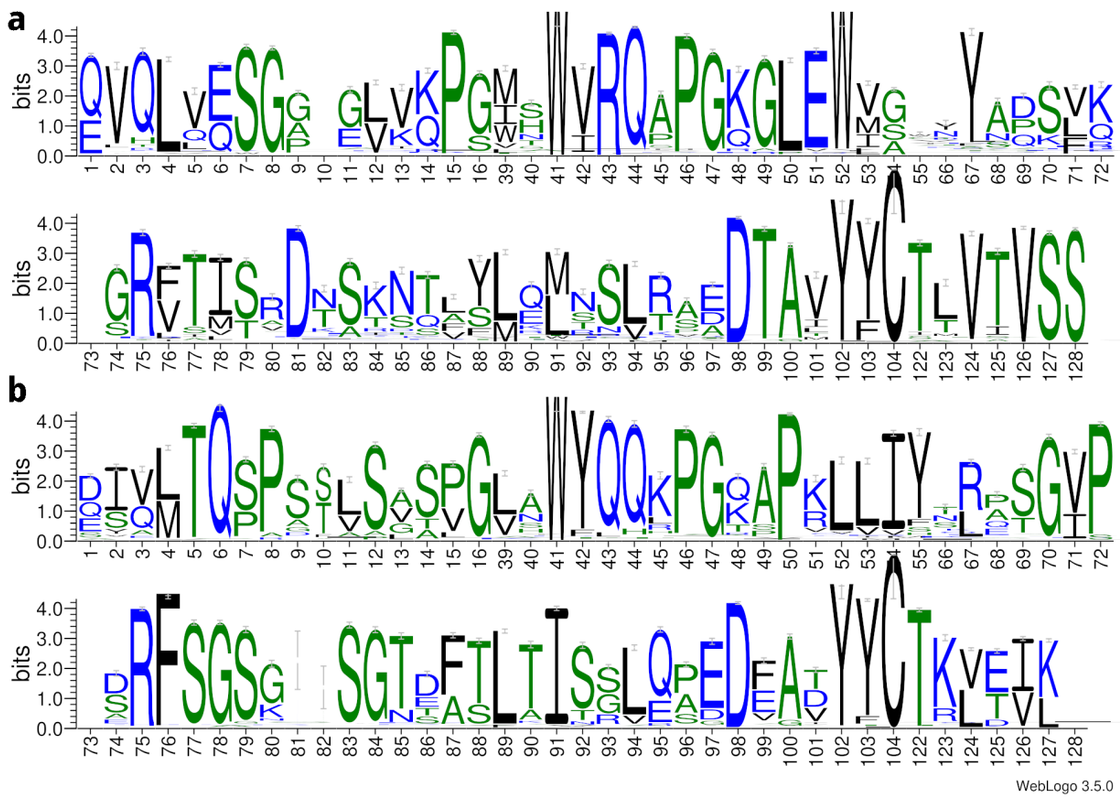

Figure 63: **The detailed architecture implemented in tensorflow consists of primarily two bi-directional GRU/LSTM layers.** The input layer of the visualized model was created for 140 unique IMGT Numbers (timesteps), and 42 features (the number time-steps may vary depending on the training data). The 42 features of paired heavy/light sequences comprise 20 amino acids and a gap symbols for each sequence. The GRU layer consists of 128 units and returns the full sequence, whereas the LSTM layer consists of 32 units and returns its last output. Dropout layers with rates of 0.1 avoid over-fitting. The last layer is a dense network with a softmax function to classify the samples as either expressing or non- expressing

11.5 WebLogos of the Flu dataset

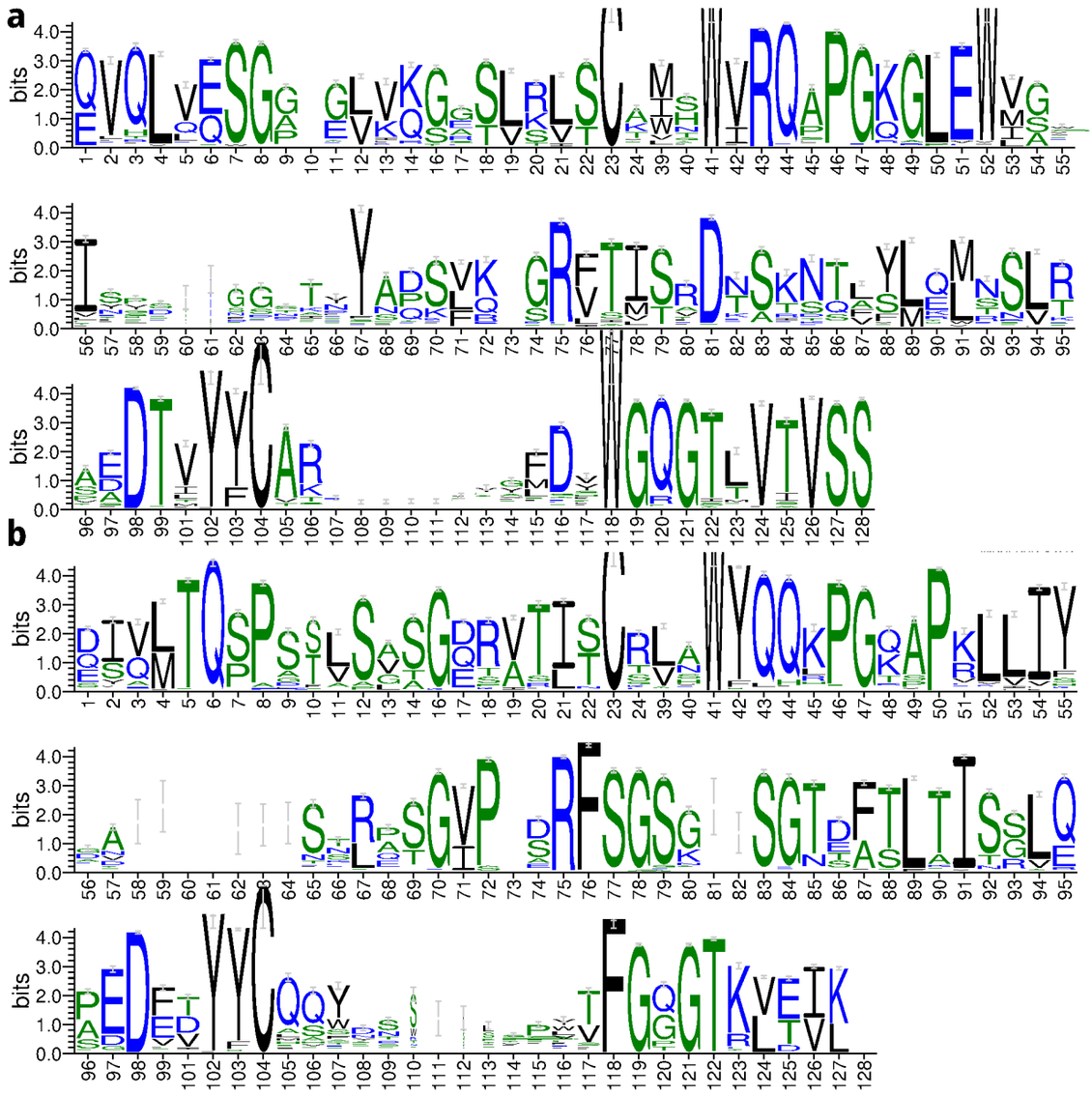

Figure 64: **Weblogo of all Flu antibodies** classified as expressing (titer > 50 µg/mL) and non-expressing (titer $\leqslant 50$µg/mL) for heavy chains (a) and light chains (b). Residues are numbered according to the IMGT numbering schema.

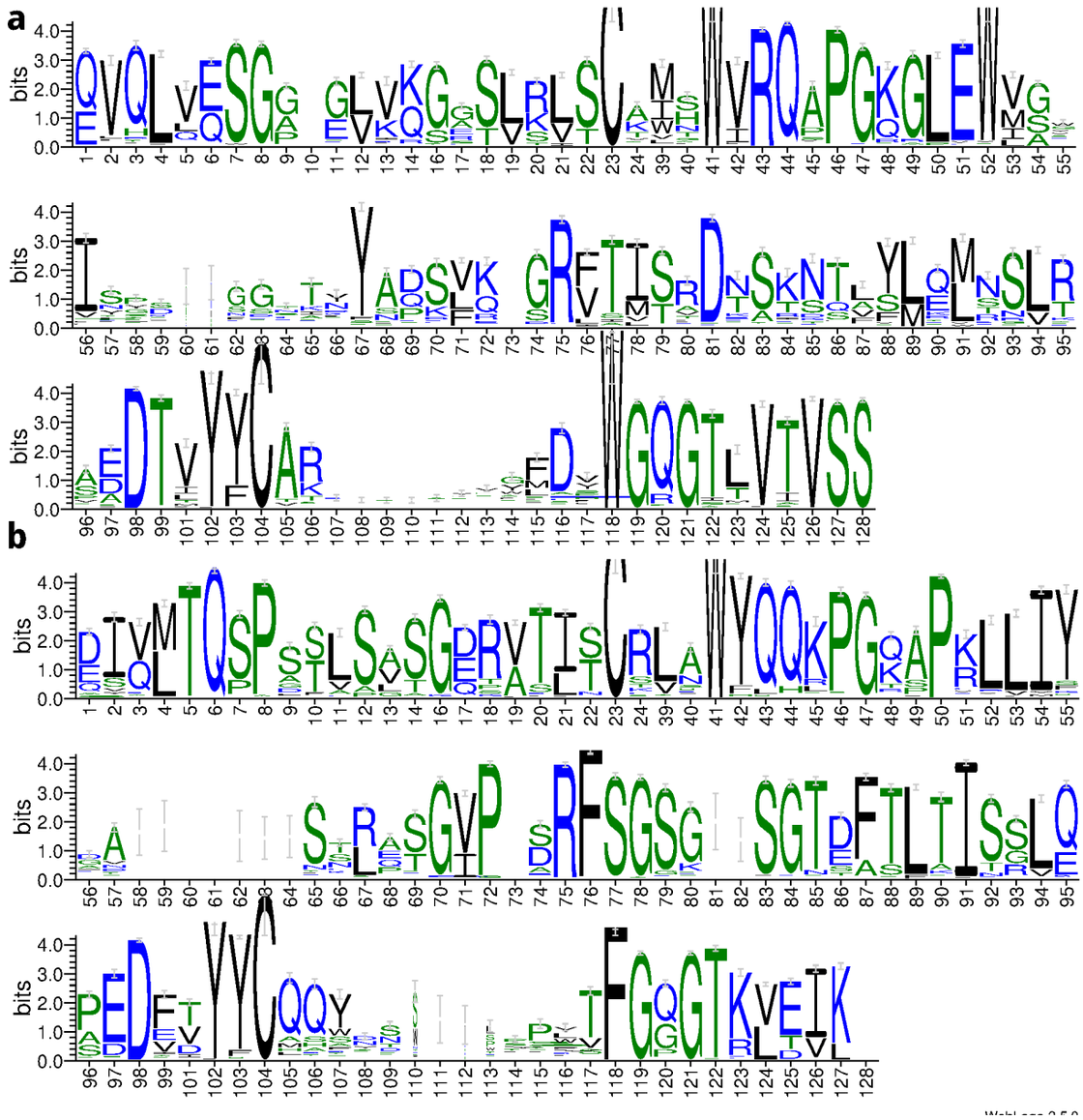

Figure 65: **Weblogo of all Flu antibodies classified as non-expressing** ($\leqslant$ 50µg/mL) for heavy chains (a) and light chains (b). Residues are numbered according to the IMGT numbering scheman.

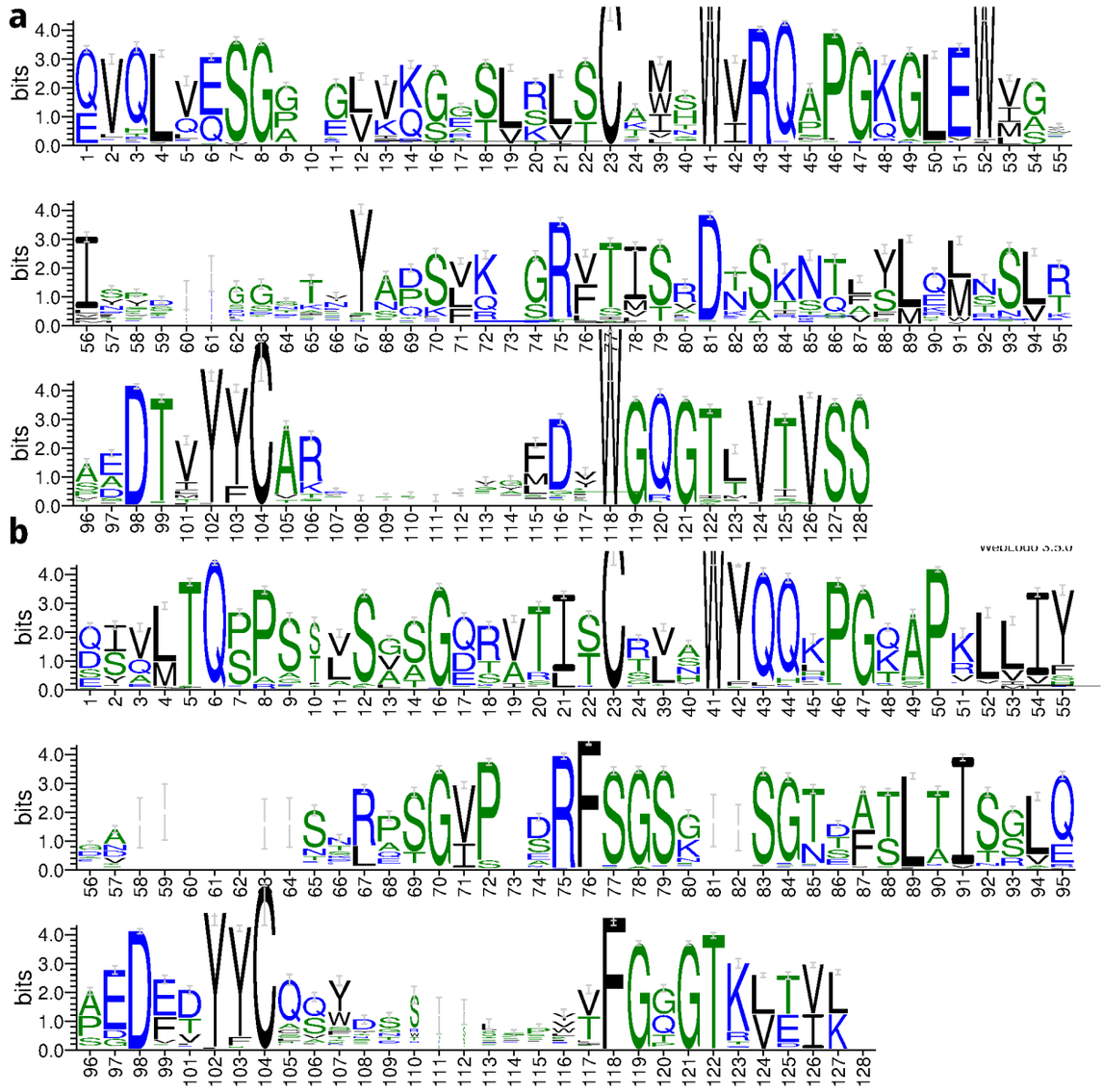

Figure 66: **Weblogo of all Flu antibodies classified as non-expressing** ($\leqslant$ 50µg/mL) for heavy chains (a) and light chains (b). Residues are numbered according to the IMGT numbering schema. Framework regions 1-4 are displayed that were used for Rosetta re-design. No significant differences in the sequence logo of expressing antibodies can be identified.

11.6 WebLogos of designed Flu antibodies with Rosetta

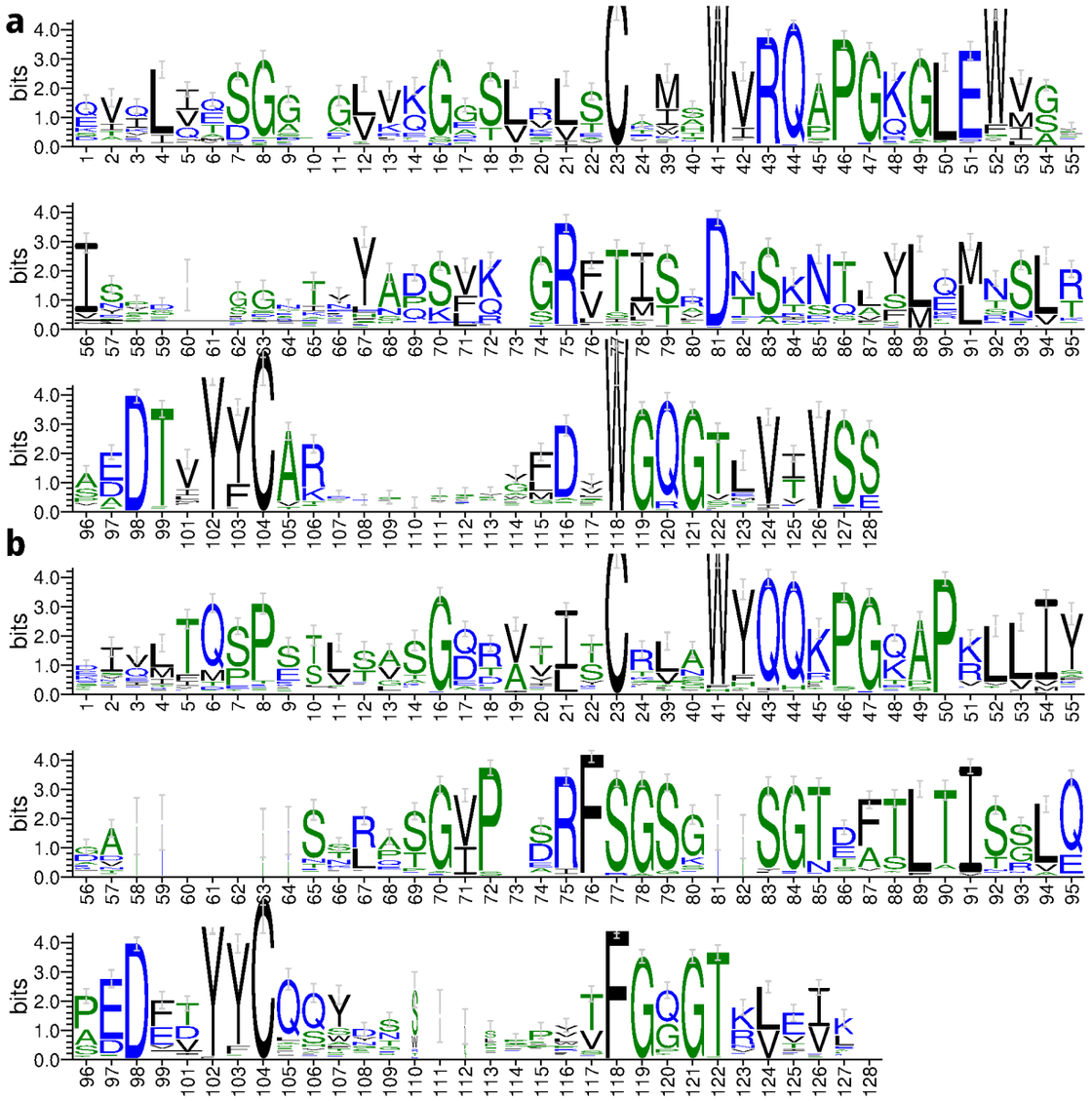

Figure 67: **WebLogo of re-engineered Flu antibodies with Rosetta.** Heavy chains (a) and light chains (b). The amino acid distributions appear not to recapitulate either the complete Flu dataset, or the subset of (non)-expressing sequences (compare: heavy chain residues 3, 20, and 125 in sequence logos of Figures S2-S4).

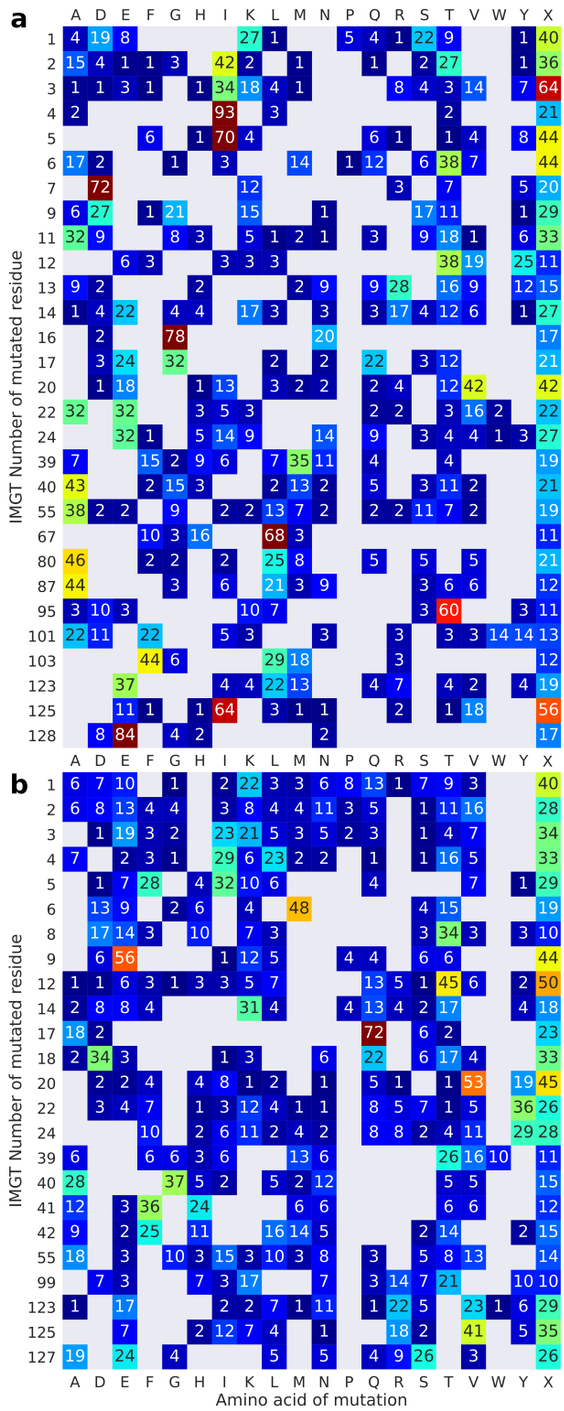

Figure 68: **Frequency of heavy chain muations of the re-engineered designs with strong intensity.** A subset of antibodies is represented (best Rosetta score, expressability greater than 90%, and improvement by re-design at least 50%). Compared to main Figure 5, which shows antibodies with an expressability improvement of at least 80%, the mutational preferences do not visibly change.

11.7 Rosetta score term scaling using single point mutant expressability predictions

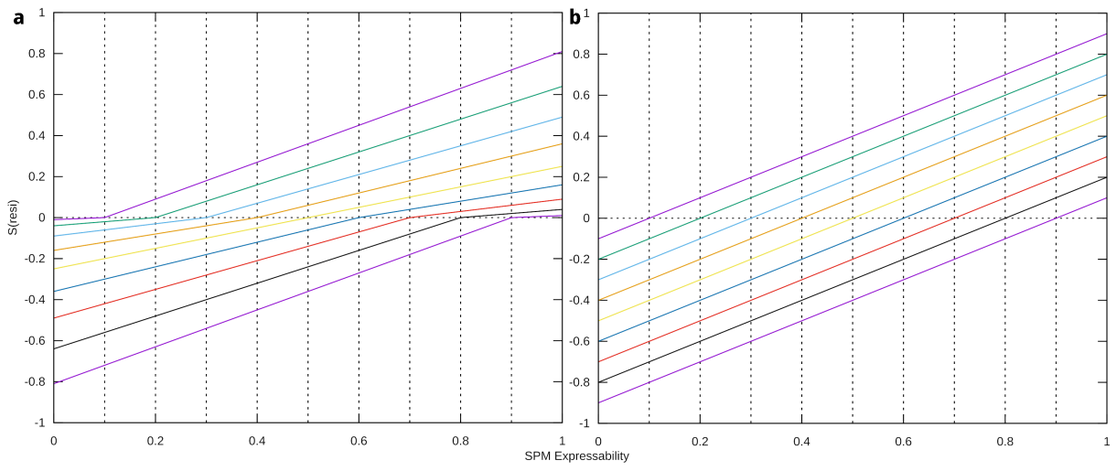

Figure 69: **Visualization of the re-scaling used for converting predicted expressability values into a Rosetta scoring term.** The scaled score (a) compared to the unscaled score (b) mediates the contribution of single point mutants (SPM) with a very low or very high probability to express. This allows a greater degree of freedom for mutations when the mutational score already approaches the positive and negative limit of S(resi). By reducing the weight of mutations with a small relative gain (when already close to optimal), we hypothesize that it is more easy to escape from local minima. S(resi) is negated for the Rosetta scoring term to penalize low and favor high expressability.

11.7.1 Antibody sequence dataset description

The Flu dataset used for creating LSTM models consists of 888 unique antibody sequences. The dataset was aligned by its IMGT numbers, and the pairwise sequence identities were calculated. Heavy chains have two populations of frequent sequence identities at 48% and 72%, kappa chains at 67% and 81%, lambda chains at 58% (Figure S70). For the sequence identity calculation, the CDRH3 region is excluded. The standardized antibody germline gene nomenclature divides the germline genes into group (IG, TR), gene type (V, variable; D, diversity; J, joining; C, constant), and the subgroup category groups the genes that contain a nucleotide sequence identity of at least 75%1. Thus, the measured sequence identities align with the prevalent selection of gene subgroups in the dataset, with 3 heavy chain germline gene subgroups resembling the majority (¿= 90%) of the dataset (Table S6) and 3 to 5 light chain germline gene subgroups (Table 7). The iQue expresability threshold of 50µg/mL was used to distinguish between antibodies that express and antibodies that do not express. With this threshold, 71 of 273 are lambda antibodies and are labeled as expressing (26.0%), while 416 of 615 kappa antibodies are labeled as expressing (67.6%). When splitting the dataset by its chain class, 615 (69.3%) belong to the kappa and 273 (30.7%) lambda class (Figure S71b) likely aggravating the challenge to predict lambda antibody expressability.

Table 6: V germline gene subgroups of the used Flu antibody dataset sorted by their highest frequency. The majority (⩾ 90%) of sequences was annotated with germline genes belonging to one of the top three germline gene subgroups (bold)

	Heavy	Nr. of sequences	Kappa	Nr. of sequences	Lambda	Nr. of sequences
1	**IGHV3**	**378**	**IGKV1**	**325**	**IGLV3**	**90**
2	**IGHV1**	**243**	**IGKV3**	**195**	**IGLV1**	**86**
3	**IGHV4**	**190**	**IGKV2**	**39**	**IGLV2**	**76**
4	IGHV5	37	IGKV4	35	IGLV8	7
5	IGHV2	30	IGKV3D	16	IGLV4	5
6	IGHV6	6	IGKV1D	5	IGLV6	3
7	IGHV4/OR15	4			IGLV5	3
8					IGLV7	2
9					IGLV9	1

Table 7: J germline gene subgroups of the used Flu antibody dataset sorted by their highest frequency. The majority (⩾ 90%) of sequences was annotated with germline genes belonging to one of the top three to five germline gene subgroups (bold)

	Heavy	Nr. of sequences	Kappa	Nr. of sequences	Lambda	Nr. of sequences
1	**IGHJ4**	**230**	**IGKJ2**	**182**	**IGLJ2**	**114**
2	**IGHJ5**	**224**	**IGKJ1**	**155**	**IGLJ1**	**72**
3	**IGHJ6**	**223**	**IGKJ4**	**116**	**IGLJ3**	**65**
4	**IGHJ3**	**151**	**IGKJ3**	**82**	IGLJ7	20
5	IGHJ1	41	**IGKJ5**	**80**	IGLJ6	2
6	IGHJ2	19				

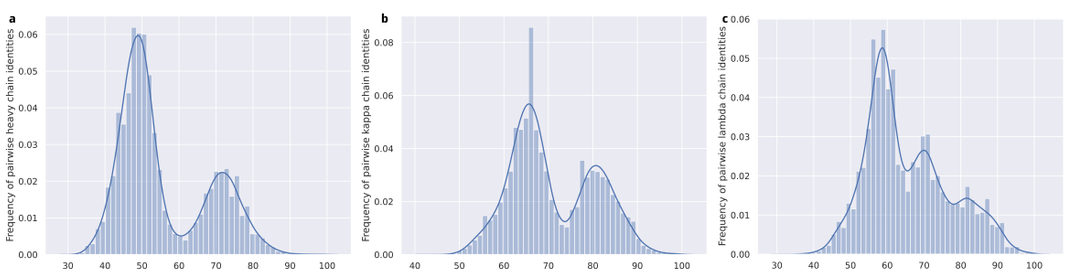

Figure 70: **Histograms of pairwise sequence identities of the Flu dataset.** Heavy chain (a), kappa light chain (b), and lambda light chain (c) pairwise sequence identity frequencies. The sequence identity calculation incudes IMGT residues 1-104 and 119-127 and does not include the CDRH3 region.

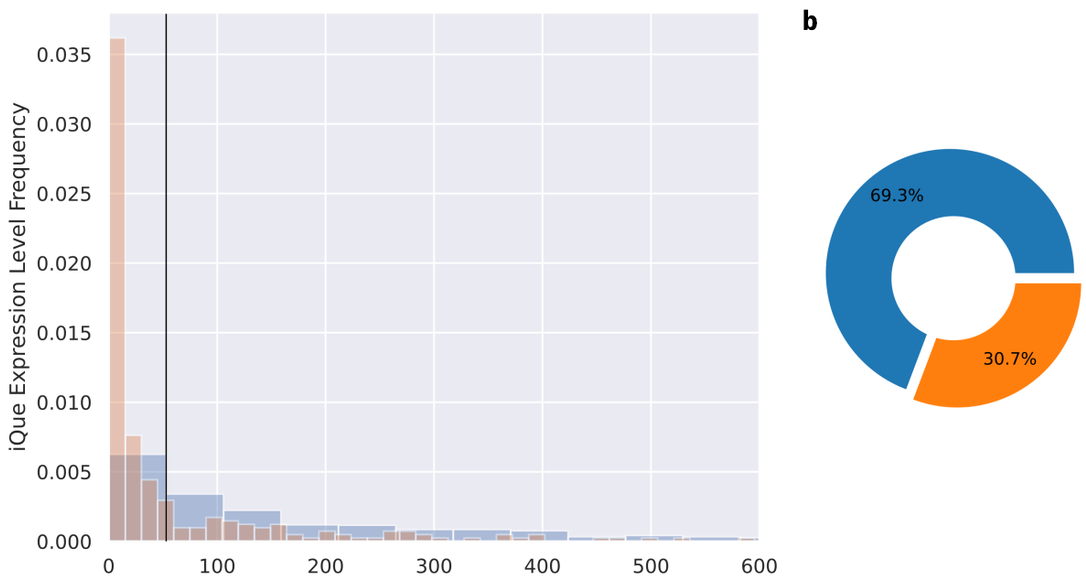

Figure 71: **Expression levels and chain class content of the Flu dataset.** The iQue expression levels are distinguishable between kappa (blue) and lambda (orange) chains. The threshold used to distinguish between antbodies that express and non-expressing antibodies at 50µg/mL (vertical line, a). The majority of antibodies in the dataset are kappa (blue) antibodies, only 30.7% of the antibody chains are lambda class (b).

11.8 Performance metrics of LSTM and Regression models

As an alternative to the LSTM models, logarithmic regression models were evaluated on the same influenza dataset. The generalization capability of the models was estimated via 10fold cross-validation and visualized. Figure S10 shows AUC, Accuracy, Precision, and Recall for all 13 predictor types. For both logarithmic regression and LSTM models, the input was encoded as one-hot matrix, Kidera, or Atchley factors. Models were generated for "paired" antibodies (meaning heavy and light chains as one sample), lambda class light chains, kappa class light chains, heavy chains, or paired lambda, or paired kappa antibodies. In addition to one-hot encoding, each column that represents a specific antibody residue (IMGT number), was replaced by Kidera, or Atchley factors. Kidera and Atchley factors describe the biochemical properties of amino acids and could have the potential to act as a biochemical similarity measure between the amino acids. This could potentially improve the classificators' performance by including a distance metric to the embedding space, which is equidistant in the case of the one-hot encoding. Overall, the performance of the regression and LSTM models are for the most part comparable, especially when compared to the paired one-hot LSTM model used in the study. The regression models' performance breaks down when the dataset is split into kappa and lambda antibodies – likely due to the reduced size of the dataset. The LSTM models' performance appears to be more robust in these cases. For 6 of the 13 model types, the LSTM's AUC scores are superior, which may be advantageous for the Rosetta structural design protocol in this study. Superior ranking capabilities of the model are preferred since the scores are converted to sequence design restraints.

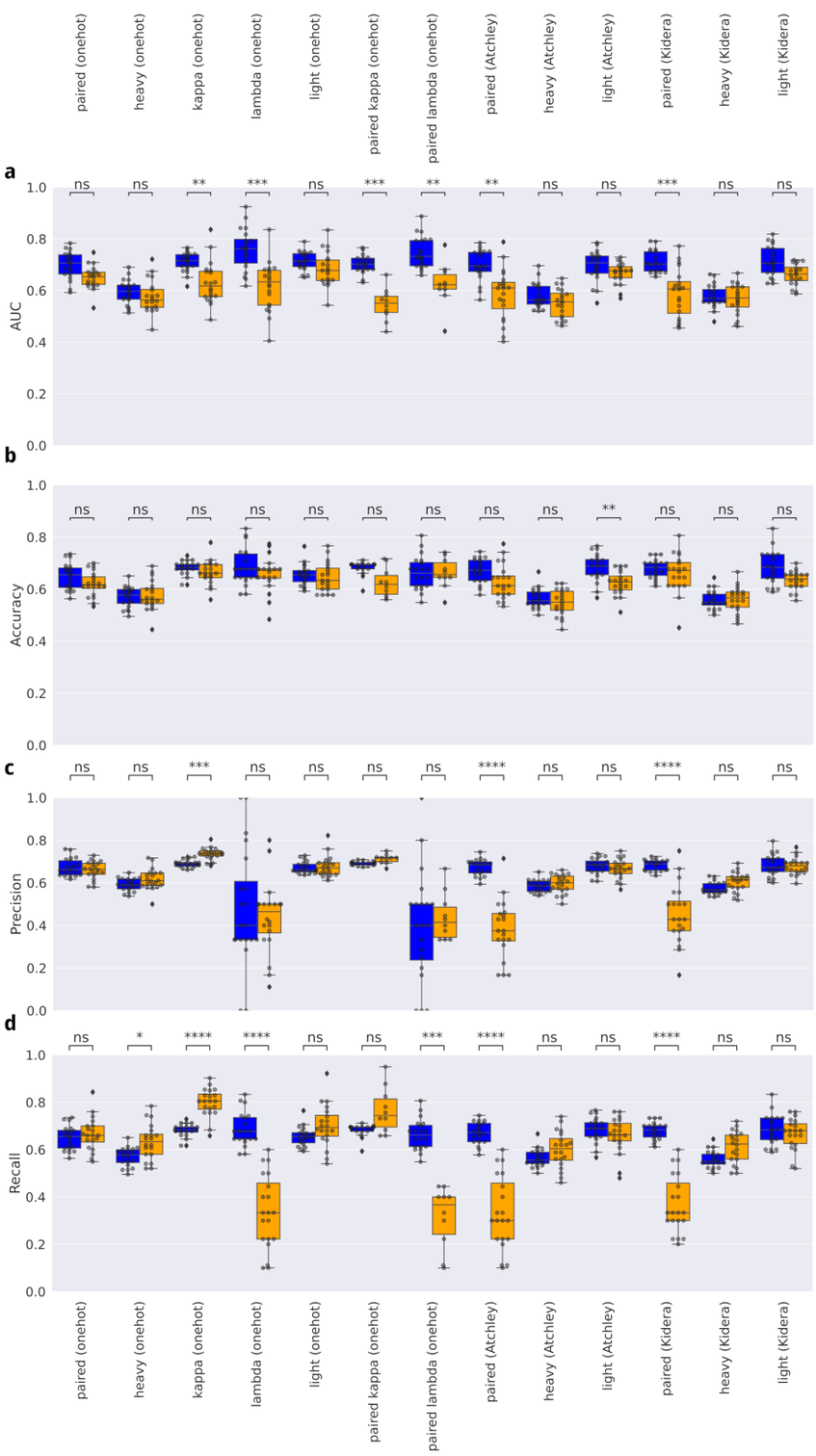

Figure 72: **Performance metrics of logarithmic regression and LSTM expressability classifiers.** Area under the curve (AUC, a), Accuracy (b), Precision (c), and Recall (d) are reported for LSTM (blue) and logarithmic regression models (orange) for 20 models each. Models were trained either with heavy, light (kappa and lambda), or kappa, or lambda chains only, or IMGT alignments of paired heavy and light chains (paired). Models were either trained with a onehot matrix representing the amino acid sequence, or by using Kidera, or Atchley amino acid descriptors. Mann-Whitney-Wilcoxon test with p-value annotation ns: $5.00e{-}02 < p \leqslant 1.00e{+}00 * : 1.00e{-}02 < p \leqslant 5.00e{-}02 ** : 1.00e{-}03 < p \leqslant 1.00e{-}02 *** : 1.00e{-}04 < p \leqslant 1.00e{-}03 **** : p \leqslant 1.00e{-}04$

11.9 RosettaCM structure predictions of the antibody dataset

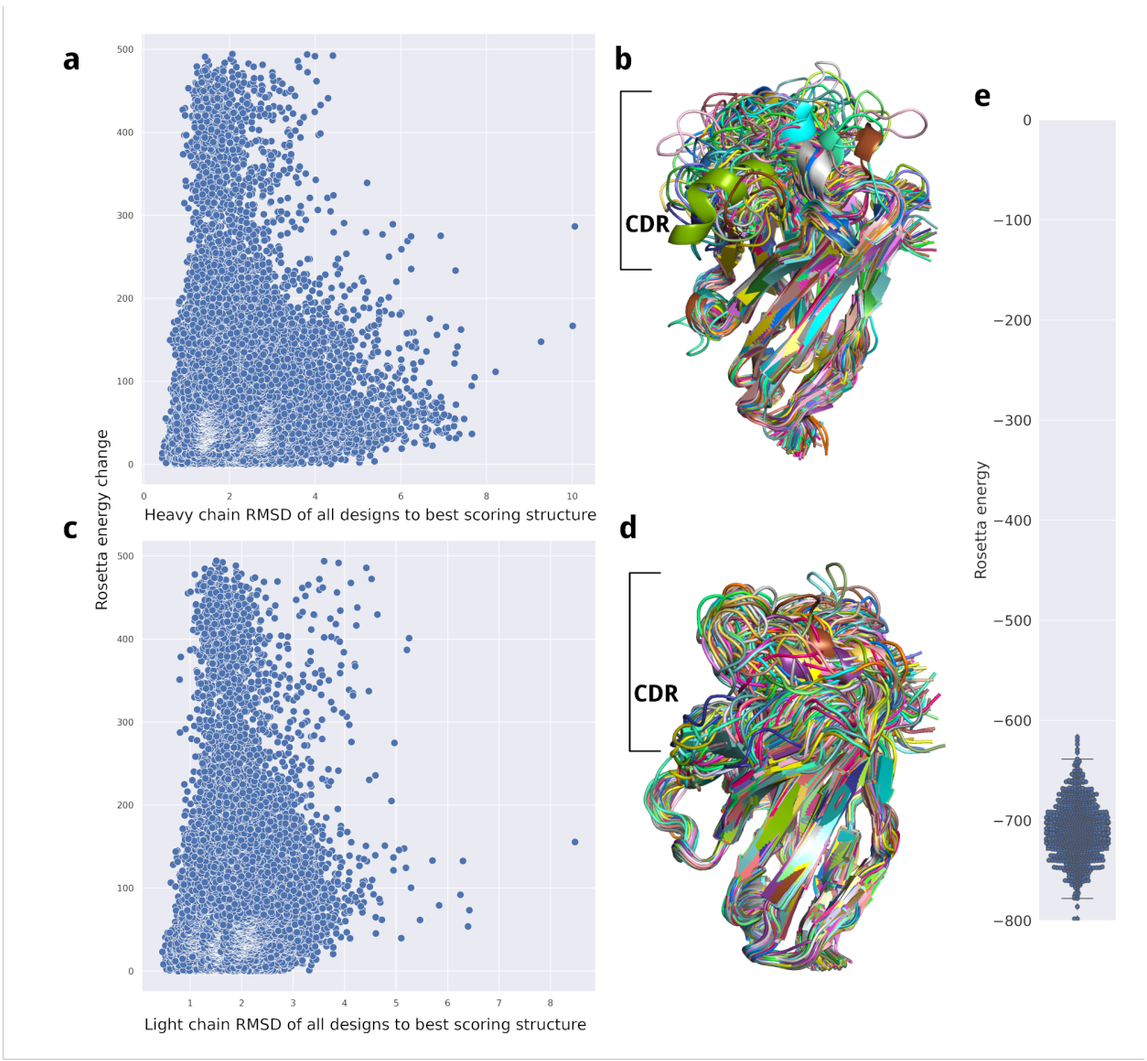

Figure 73: **Assessment of Rosetta homology models do not suggest any folding defects.** Folding tunnels for heavy chain (a), and light chains (c) of the homology models were created by subtracting the RMSD and Rosetta Energy of each of the 888 influenza antibodies from each of its alternative designs. RMSD vs Energy plots for all 888 homology models indicate folding tunnels suggesting confidence of Rosetta's scoring function in the results. All heavy chains (b) and light chains (d) superimposed appear as expected, with constant low RMSD framework regions and highly variable CDR loops. The overall Rosetta energy is negative (the more negative the better) in the range of -800 to -615 REU (e). The variability in REU is expected due to the diverse length of the antibodies and does not indicate misfolded outliers.

11.10 Rosetta design with human-like sequence restraints

11.10.1 Dataset of 27 co-crystalized human antibodies

The benchmark dataset of 27 crystal-structures was chosen by scanning SabDab for antibodies annotated as human, available as co-crystal structure, and with a resolution of < 2 Å.

Table 1 PDB ID number, binding partner, CDRH3 length, and change of the CDRH3 human-likeness (ΔHL) compared to the native designs, and antibodies for which the CDRH3 human-likeness could be improved (bold).

Table 8: PDB ID number, binding partner, CDRH3 length, and change of the CDRH3 human-likeness (ΔHL) compared to the native designs, and antibodies for which the CDRH3 human-likeness could be improved (bold).

#	PDB ID	Specificity	CDRH3 length (aa)	Δ HL
1	1n0x	HIV-1	20	**+8.0 ± 0.7%**
2	2vxq	Pollen Allergen Phl P2	10	−3.7 ± 2.1%
3	2xwt	TSHR Autoantibody	12	−2.5 ± 0.4%
4	2yc1	Cn2 toxin (Centruroides noxius Hoffman)	10	**3.5 ± 0.0%**
5	3fn0	HIV-1	19	0.4 ± 0.10%
6	3l5x	IL-13	13	**5.0 ± 0.7%**
7	3uji	Anti-HIV-1	16	0.7 ± 0.3%
8	4al8	Dengue virus DIII	10	0.0 ± 0.5%
9	4dgy	HPC glycoprotein E2	16	1.2 ± 0.5%
10	4h8w	HIV-1	12	−2.3 ± 0.3%
11	4hpo	HIV-1	19	−0.44 ± 0.4%
12	4hs6	HPC glycoprotein E2	12	**4.0 ± 0.6%**
13	4ioi	Cancer (trastuzumab)	13	**4.3 ± 1.4%**
14	4j6r	HIV-1	14	**3.5 ± 0.8%**
15	4lkx	IgE	10	−5.7 ± 0.6%
16	4m1d	HIV-1	22	−1.9 ± 1.0%
17	4m62	HIV-1	20	−0.3 ± 0.1%
18	4nzr	HIV-1	20	0.8 ± 0.1%
19	4xc1	HIV-1	20	1.0 ± 0.6%
20	4xmp	HIV-1	25	0.2 ± 0.5%
21	5cin	HIV-1	18	−0.6 ± 0.0%
22	5f9o	HIV-1	15	**6.1 ± 0.7%**
23	5ig7	Gluten peptides (B-Cell epitope)	16	−1.5 ± 0.4%
24	5l6y	IL-13 (tralokinumab)	15	1.2 ± 1.0%
25	5ob5	gro-beta	12	−2.2 ± 3.0%
26	5uek	Histone chaperone ASF1	10	1.9 ± 2.0%
27	5xku	57N9	17	**6.5 ± 0.9%**

11.10.2 The sequence identity of the unrestraint "native" Rosetta designs is comparable to that of HL designs

To benchmark the HL design approach, a control group (called "native") with a limited number of mutation was designed (Rosetta's FavorNativeResidue mover). To each human-likness design using the amino acid substitution scores, one of the native designs with the closest sequence identity to the wild-type was assigned. Native designs were assigned separately for the V/J region, and the CDRH3 region. The native group was used as baseline to estimate the amino acid change of an already human antibody after Rosetta design.

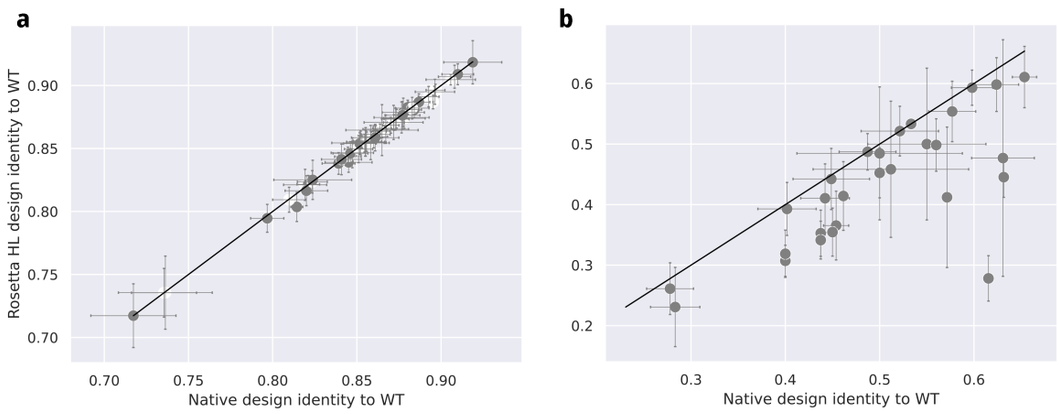

Figure 74: **The sequence identity between human-like Rosetta designs and wild-type**, and between its assigned native design and the wild-type for the VJ region (a) and the CDRH3 region (b). For the VJ region, native designs with almost perfect sequence identity could be found, whereas some of the best fitting native designs for the CDRH3 loop had a slightly higher sequence identity to the wild-type.

11.10.3 Rosetta design methods with and without human-likeness restraints

Each antibody in the dataset was designed 13 times. First, the unrestraint control group was designed without any additional restraints. Second, the native designs were generated using the FavorNativeResidue mover using 8 different weights (0.5 to 4.0 with an interval of 0.5). Human-likeness designs were achieved by adding the PSSM via the FavorSequenceProfile mover. Designs took place with a clustered and original PSSM separately using the weights 3 and 4.

For each antibody, resfiles were generated allowing only design within the Fv region, allowing all amino acids but Cysteins. Design was not allowed for residues that are Cysteins in the wild-type structure.

Unrestraint control designs

```
<ROSETTASCRIPTS>
 <SCOREFXNS>
   <ScoreFunction name="scorefxn" weights="\%\%scorefxn\%\%.wts">
     <Reweight scoretype="res_type_constraint" weight="0.0"/>
   </ScoreFunction>
   <ScoreFunction name="scorefxn_cst" weights="\%\%scorefxn\%\%.wts">
```

```
    <Reweight scoretype="res_type_constraint" weight="1.0"/>
  </ScoreFunction>
 </SCOREFXNS>
 <RESIDUE_SELECTORS>
 </RESIDUE_SELECTORS>
 <TASKOPERATIONS>
  <InitializeFromCommandline name="ifcl"/>
  <ReadResfile name="rrf" filename="\%\%resfile\%\%"/>
 </TASKOPERATIONS>
 <MOVE_MAP_FACTORIES>
 </MOVE_MAP_FACTORIES>
 <FILTERS>
 </FILTERS>
 <MOVERS>
  <PackRotamersMover name="design" scorefxn="scorefxn_cst" task_operations="rrf,
      ↪ ifcl"/>
 </MOVERS>
 <PROTOCOLS>
  <Add mover="design"/>
 </PROTOCOLS>
 <OUTPUT scorefxn="scorefxn"/>
</ROSETTASCRIPTS>
```

Native designs

```
<ROSETTASCRIPTS>
 <SCOREFXNS>
  <ScoreFunction name="scorefxn" weights="\%\%scorefxn\%\%.wts">
    <Reweight scoretype="res_type_constraint" weight="0.0"/>
  </ScoreFunction>
  <ScoreFunction name="scorefxn_cst" weights="\%\%scorefxn\%\%.wts">
    <Reweight scoretype="res_type_constraint" weight="1.0"/>
  </ScoreFunction>
 </SCOREFXNS>
 <RESIDUE_SELECTORS>
 </RESIDUE_SELECTORS>
 <TASKOPERATIONS>
  <InitializeFromCommandline name="ifcl"/>
  <ReadResfile name="rrf" filename="\%\%resfile\%\%"/>
 </TASKOPERATIONS>
 <MOVE_MAP_FACTORIES>
```

```
 </MOVE_MAP_FACTORIES>
 <FILTERS>
 </FILTERS>
 <MOVERS>
   <FavorNativeResidue name="native" bonus="\%\%weight\%\%"/>
   <PackRotamersMover name="design" scorefxn="scorefxn_cst" task_operations="rrf,
       ↪ ifcl"/>
 </MOVERS>
 <PROTOCOLS>
   <Add mover="native"/>
   <Add mover="design"/>
 </PROTOCOLS>
 <OUTPUT scorefxn="scorefxn"/>
</ROSETTASCRIPTS>
```

Design with human-likeness restraints

```
<ROSETTASCRIPTS>
 <SCOREFXNS>
   <ScoreFunction name="scorefxn" weights="\%\%scorefxn\%\%.wts">
     <Reweight scoretype="res_type_constraint" weight="0.0"/>
   </ScoreFunction>
   <ScoreFunction name="scorefxn_cst" weights="\%\%scorefxn\%\%.wts">
     <Reweight scoretype="res_type_constraint" weight="1.0"/>
   </ScoreFunction>
 </SCOREFXNS>
 <RESIDUE_SELECTORS>
 </RESIDUE_SELECTORS>
 <TASKOPERATIONS>
   <InitializeFromCommandline name="ifcl"/>
   <ReadResfile name="rrf" filename="\%\%resfile\%\%"/>
 </TASKOPERATIONS>
 <MOVE_MAP_FACTORIES>
 </MOVE_MAP_FACTORIES>
 <FILTERS>
 </FILTERS>
 <MOVERS>
   <FavorSequenceProfile name="profile" weight="\%\%weight\%\%" scaling="\%\%
       ↪ scaling\%\%" pssm="\%\%pssm\%\%" chain="\%\%chainnum\%\%" scorefxns="
       ↪ scorefxn_cst"/>
   <PackRotamersMover name="design" scorefxn="scorefxn_cst" task_operations="rrf,
```

```
      ↪ ifcl"/>
 </MOVERS>
 <PROTOCOLS>
   <Add mover="profile"/>
   <Add mover="design"/>
 </PROTOCOLS>
 <OUTPUT scorefxn="scorefxn"/>
</ROSETTASCRIPTS>
```

All designs were succeeded by a constraint relax run using the Rosetta relax application the option

```
 -relax:constrain_relax_to_start_coords
```

11.10.4 The effect of Powell optimization of lambda on the substitution scores

Powell optimization of the lambda parameter was introduced to increase the correlation of the amino acid substitution scores with the human-likeness to maximize the effect of the HL design protocol. Each CDRH3 profile has an individual distribution and therefore is generated for each PSSM individually.

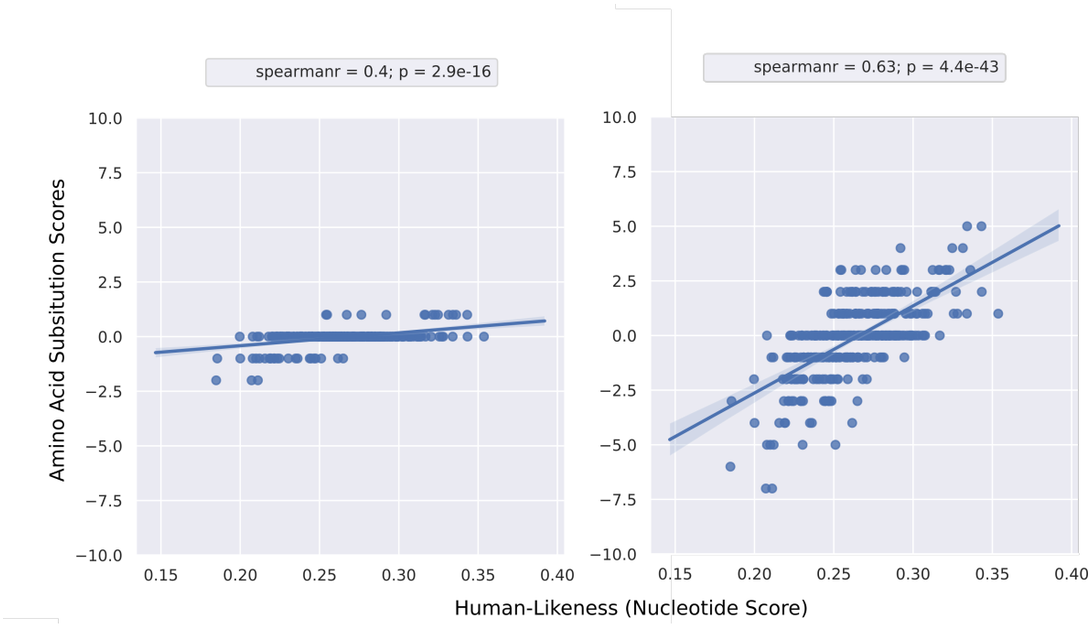

Figure 75: **Correlation between substitution scores and human-likeness before and after Powell optimization.** Optimization of the lambda parameter maximizes the effect of the substitution score for the design. The distribution of the substiution scores for a CDRH3 of the length 19 is flat before (left) and is likely to have little effect on the design favors and disfavors certain subsitutions more clearly after optimization (right). Before optimization, a default lambda for 0.09 is chosen and was determined to be 0.027 after optimization in this case.

References

Abadi, Martín et al. (May 2016). "TensorFlow: A system for large-scale machine learning". In: *arXiv:1605.08695 [cs]*. arXiv: 1605.08695. URL: http://arxiv.org/abs/1605.08695 (visited on 12/09/2020).

Abhinandan, K. R. and Andrew C. R. Martin (June 2007). "Analyzing the "degree of humanness" of antibody sequences". eng. In: *J Mol Biol* 369.3, pp. 852–862. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2007.02.100.

Adolf-Bryfogle, Jared, Oleks Kalyuzhniy, et al. (Apr. 2018). "RosettaAntibodyDesign (RAbD): A general framework for computational antibody design". eng. In: *PLoS Comput Biol* 14.4, e1006112. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006112.

Adolf-Bryfogle, Jared, Qifang Xu, et al. (Jan. 2015). "PyIgClassify: a database of antibody CDR structural classifications". eng. In: *Nucleic Acids Res* 43.Database issue, pp. D432–438. ISSN: 1362-4962. DOI: 10.1093/nar/gku1106.

Afif, Waqqas et al. (May 2010). "Clinical utility of measuring infliximab and human anti-chimeric antibody concentrations in patients with inflammatory bowel disease". eng. In: *Am J Gastroenterol* 105.5, pp. 1133–1139. ISSN: 1572-0241. DOI: 10.1038/ajg.2010.9.

Agron, Peter G. and Donald R. Helinski (1995). "Symbiotic Expression of Rhizobium meliloti Nitrogen Fixation Genes Is Regulated by Oxygen". en. In: *Two-Component Signal Transduction*. Section: 17 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1128/9781555818319.ch17. John Wiley & Sons, Ltd, pp. 275–287. ISBN: 978-1-68367-271-5. DOI: 10.1128/9781555818319.ch17. URL: https://onlinelibrary.wiley.com/doi/abs/10.1128/9781555818319.ch17 (visited on 09/13/2021).

Aguiar, Rodrigo Barbosa de et al. (Feb. 2016). "Blocking FGF2 with a new specific monoclonal antibody impairs angiogenesis and experimental metastatic melanoma, suggesting a potential role in adjuvant settings". en. In: *Cancer Letters* 371.2, pp. 151–160. ISSN: 0304-3835. DOI: 10.1016/j.canlet.2015.11.030. URL: https://www.sciencedirect.com/science/article/pii/S030438351500720X (visited on 09/30/2021).

Alberts, Bruce et al. (2017). *Molecular biology of the cell*. WW Norton & Company. ISBN: 1-315-73536-9.

Alford, Rebecca F. et al. (June 2017). "The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design". eng. In: *J Chem Theory Comput* 13.6, pp. 3031–3048. ISSN: 1549-9626. DOI: 10.1021/acs.jctc.7b00125.

Almagro, Juan C., Mary Pat Beavers, et al. (Nov. 2011). "Antibody modeling assessment". eng. In: *Proteins* 79.11, pp. 3050–3066. ISSN: 1097-0134. DOI: 10.1002/prot.23130.

Almagro, Juan C., Alexey Teplyakov, et al. (Aug. 2014). "Second antibody modeling assessment (AMA-II)". eng. In: *Proteins* 82.8, pp. 1553–1562. ISSN: 1097-0134. DOI: 10.1002/prot.24567.

AlQuraishi, Mohammed (Apr. 2019). "End-to-End Differentiable Learning of Protein Structure". en. In: *Cell Systems* 8.4, 292–301.e3. ISSN: 2405-4712. DOI: 10.1016/j.cels.2019.03.006.

URL: https://www.sciencedirect.com/science/article/pii/S2405471219300766 (visited on 03/06/2021).

Altschul, S. F. (June 1991). "Amino acid substitution matrices from an information theoretic perspective". eng. In: *J Mol Biol* 219.3, pp. 555–565. ISSN: 0022-2836. DOI: 10.1016/0022-2836(91)90193-a.

Altschul, S. F. and E. V. Koonin (Nov. 1998). "Iterated profile searches with PSI-BLAST–a tool for discovery in protein databases". eng. In: *Trends Biochem Sci* 23.11, pp. 444–447. ISSN: 0968-0004. DOI: 10.1016/s0968-0004(98)01298-5.

Altschul, S. F., T. L. Madden, et al. (Sept. 1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". eng. In: *Nucleic Acids Res* 25.17, pp. 3389–3402. ISSN: 0305-1048. DOI: 10.1093/nar/25.17.3389.

Altschul, Stephen F. et al. (Feb. 2009). "PSI-BLAST pseudocounts and the minimum description length principle". eng. In: *Nucleic Acids Res* 37.3, pp. 815–824. ISSN: 1362-4962. DOI: 10.1093/nar/gkn981.

Amann, Thomas et al. (2019). "Genetic engineering approaches to improve posttranslational modification of biopharmaceuticals in different production platforms". en. In: *Biotechnology and Bioengineering* 116.10. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bit.27101, pp. 2778–2796. ISSN: 1097-0290. DOI: https://doi.org/10.1002/bit.27101. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bit.27101 (visited on 01/14/2021).

Angermueller, Christof et al. (July 2016). "Deep learning for computational biology". In: *Molecular Systems Biology* 12.7. Publisher: John Wiley & Sons, Ltd, p. 878. ISSN: 1744-4292. DOI: 10.15252/msb.20156651. URL: https://www.embopress.org/doi/full/10.15252/msb.20156651 (visited on 03/06/2021).

Anishchenko, Ivan et al. (Aug. 2017). "Origins of coevolution between residues distant in protein 3D structures". In: *Proc Natl Acad Sci U S A* 114.34, pp. 9122–9127. ISSN: 0027-8424. DOI: 10.1073/pnas.1702664114. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5576787/ (visited on 09/02/2021).

Appenzeller-Herzog, Christian (Mar. 2011). "Glutathione- and non-glutathione-based oxidant control in the endoplasmic reticulum". eng. In: *J Cell Sci* 124.Pt 6, pp. 847–855. ISSN: 1477-9137. DOI: 10.1242/jcs.080895.

Babor, Mariana and Tanja Kortemme (June 2009). "Multi-constraint computational design suggests that native sequences of germline antibody H3 loops are nearly optimal for conformational flexibility". eng. In: *Proteins* 75.4, pp. 846–858. ISSN: 1097-0134. DOI: 10.1002/prot.22293.

Baek, Minkyung et al. (Aug. 2021). "Accurate prediction of protein structures and interactions using a three-track neural network". In: *Science* 373.6557. Publisher: American Association for the Advancement of Science, pp. 871–876. DOI: 10.1126/science.abj8754. URL: https://www.science.org/doi/abs/10.1126/science.abj8754 (visited on 10/08/2021).

Balakrishnan, Sivaraman et al. (Apr. 2011). "Learning generative models for protein fold families". eng. In: *Proteins* 79.4, pp. 1061–1078. ISSN: 1097-0134. DOI: 10.1002/prot.22934.

Baran, Dror et al. (Oct. 2017). “Principles for computational design of binding antibodies”. eng. In: *Proc Natl Acad Sci U S A* 114.41, pp. 10900–10905. ISSN: 1091-6490. DOI: 10.1073/pnas.1707171114.

Bender, Brian J. et al. (Aug. 2016). “Protocols for Molecular Modeling with Rosetta3 and RosettaScripts”. eng. In: *Biochemistry* 55.34, pp. 4748–4763. ISSN: 1520-4995. DOI: 10.1021/acs.biochem.6b00444.

Bender, Brian Joseph et al. (Mar. 2019). “Structural Model of Ghrelin Bound to its G Protein-Coupled Receptor”. eng. In: *Structure* 27.3, 537–544.e4. ISSN: 1878-4186. DOI: 10.1016/j.str.2018.12.004.

Benson, Dennis A. et al. (Jan. 2013). “GenBank”. eng. In: *Nucleic Acids Res* 41.Database issue, pp. D36–42. ISSN: 1362-4962. DOI: 10.1093/nar/gks1195.

Berman, H. M. et al. (Jan. 2000). “The Protein Data Bank”. eng. In: *Nucleic Acids Res* 28.1, pp. 235–242. ISSN: 0305-1048. DOI: 10.1093/nar/28.1.235.

Bhatti, Adnan Bashir, Muhammad Usman, and Venkataramana Kandi (Mar. 2016). “Current Scenario of HIV/AIDS, Treatment Options, and Major Challenges with Compliance to Antiretroviral Therapy”. eng. In: *Cureus* 8.3, e515. ISSN: 2168-8184. DOI: 10.7759/cureus.515.

Birck, C. et al. (Dec. 1999). “Conformational changes induced by phosphorylation of the FixJ receiver domain”. eng. In: *Structure* 7.12, pp. 1505–1515. ISSN: 0969-2126. DOI: 10.1016/s0969-2126(00)88341-0.

Bitbol, Anne-Florence et al. (Oct. 2016). “Inferring interaction partners from protein sequences”. eng. In: *Proc Natl Acad Sci U S A* 113.43, pp. 12180–12185. ISSN: 1091-6490. DOI: 10.1073/pnas.1606762113.

Blaszczyk, J. et al. (Oct. 2000). “Catalytic center assembly of HPPK as revealed by the crystal structure of a ternary complex at 1.25 A resolution”. eng. In: *Structure* 8.10, pp. 1049–1058. ISSN: 0969-2126. DOI: 10.1016/s0969-2126(00)00502-5.

Bolotin, Dmitriy A. et al. (May 2015). “MiXCR: software for comprehensive adaptive immunity profiling”. en. In: *Nat Methods* 12.5. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 5 Primary_atype: Correspondence Publisher: Nature Publishing Group Subject_term: Adaptive immunity;Genetics research;Software Subject_term_id: adaptive-immunity;genetics-research;software, pp. 380–381. ISSN: 1548-7105. DOI: 10.1038/nmeth.3364. URL: https://www.nature.com/articles/nmeth.3364 (visited on 09/04/2021).

Bolt, S. et al. (Feb. 1993). “The generation of a humanized, non-mitogenic CD3 monoclonal antibody which retains in vitro immunosuppressive properties”. eng. In: *Eur J Immunol* 23.2, pp. 403–411. ISSN: 0014-2980. DOI: 10.1002/eji.1830230216.

Bonetti, Daniela et al. (May 2016). “Identification and Structural Characterization of an Intermediate in the Folding of the Measles Virus X Domain”. eng. In: *J Biol Chem* 291.20, pp. 10886–10892. ISSN: 1083-351X. DOI: 10.1074/jbc.M116.721126.

Bonwick, G. A. et al. (Sept. 1996). "Production of murine monoclonal antibodies against sulcofuron and flucofuron by in vitro immunisation". eng. In: *J Immunol Methods* 196.2, pp. 163–173. ISSN: 0022-1759. DOI: 10.1016/0022-1759(96)00098-1.

Braakman, Ineke and Daniel N. Hebert (May 2013). "Protein Folding in the Endoplasmic Reticulum". In: *Cold Spring Harb Perspect Biol* 5.5. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a013201. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3632058/ (visited on 05/08/2019).

Briney, Bryan et al. (Feb. 2019). "Commonality despite exceptional diversity in the baseline human antibody repertoire". eng. In: *Nature* 566.7744, pp. 393–397. ISSN: 1476-4687. DOI: 10.1038/s41586-019-0879-y.

Briney, Bryan S. et al. (Sept. 2012). "Frequency and genetic characterization of V(DD)J recombinants in the human peripheral blood antibody repertoire". eng. In: *Immunology* 137.1, pp. 56–64. ISSN: 1365-2567. DOI: 10.1111/j.1365-2567.2012.03605.x.

Brochet, Xavier, Marie-Paule Lefranc, and Véronique Giudicelli (July 2008). "IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis". eng. In: *Nucleic Acids Res* 36.Web Server issue, W503–508. ISSN: 1362-4962. DOI: 10.1093/nar/gkn316.

Brovkina, Olga I. et al. (2018). "The Ethnic-Specific Spectrum of Germline Nucleotide Variants in DNA Damage Response and Repair Genes in Hereditary Breast and Ovarian Cancer Patients of Tatar Descent". eng. In: *Front Oncol* 8, p. 421. ISSN: 2234-943X. DOI: 10.3389/fonc.2018.00421.

Burger, Lukas and Erik van Nimwegen (2008). "Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method". eng. In: *Mol Syst Biol* 4, p. 165. ISSN: 1744-4292. DOI: 10.1038/msb4100203.

Campeotto, Ivan et al. (Jan. 2017). "One-step design of a stable variant of the malaria invasion protein RH5 for use as a vaccine immunogen". eng. In: *Proc Natl Acad Sci U S A* 114.5, pp. 998–1002. ISSN: 1091-6490. DOI: 10.1073/pnas.1616903114.

Canutescu, Adrian A. and Roland L. Dunbrack (May 2003). "Cyclic coordinate descent: A robotics algorithm for protein loop closure". eng. In: *Protein Sci* 12.5, pp. 963–972. ISSN: 0961-8368. DOI: 10.1110/ps.0242703.

Charles A Janeway, Jr et al. (2001). "The generation of diversity in immunoglobulins". en. In: *Immunobiology: The Immune System in Health and Disease. 5th edition.* Publisher: Garland Science. URL: https://www.ncbi.nlm.nih.gov/books/NBK27140/ (visited on 11/25/2021).

Chaudhury, Sidhartha, Monica Berrondo, et al. (2011). "Benchmarking and analysis of protein docking performance in Rosetta v3.2". eng. In: *PLoS One* 6.8, e22477. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0022477.

Chaudhury, Sidhartha and Jeffrey J. Gray (Sept. 2008). "Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles". eng. In: *J Mol Biol* 381.4, pp. 1068–1087. ISSN: 1089-8638. DOI: 10.1016/j.jmb.2008.05.042.

Chaudhury, Sidhartha, Sergey Lyskov, and Jeffrey J. Gray (Mar. 2010). "PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta". eng. In: *Bioinformatics* 26.5, pp. 689–691. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btq007.

Chen, Chia-Lin et al. (May 2017). "Crystal Structure of a Homogeneous IgG-Fc Glycoform with the N-Glycan Designed to Maximize the Antibody Dependent Cellular Cytotoxicity". eng. In: *ACS Chem Biol* 12.5, pp. 1335–1345. ISSN: 1554-8937. DOI: 10.1021/acschembio.7b00140.

Choi, Yoonjoo et al. (Nov. 2015). "Antibody humanization by structure-based computational protein design". In: *mAbs* 7.6. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19420862.2015.1076600, pp. 1045–1057. ISSN: 1942-0862. DOI: 10.1080/19420862.2015.1076600. URL: https://doi.org/10.1080/19420862.2015.1076600 (visited on 09/10/2021).

Chothia, C. et al. (Dec. 1989). "Conformations of immunoglobulin hypervariable regions". eng. In: *Nature* 342.6252, pp. 877–883. ISSN: 0028-0836. DOI: 10.1038/342877a0.

Cisneros, Alberto et al. (Aug. 2019). "Role of antibody heavy and light chain interface residues in affinity maturation of binding to HIV envelope glycoprotein". en. In: *Mol. Syst. Des. Eng.* 4.4. Publisher: The Royal Society of Chemistry, pp. 737–746. ISSN: 2058-9689. DOI: 10.1039/C8ME00080H. URL: https://pubs.rsc.org/en/content/articlelanding/2019/me/c8me00080h (visited on 09/07/2021).

Cock, Peter J. A. et al. (June 2009). "Biopython: freely available Python tools for computational molecular biology and bioinformatics". eng. In: *Bioinformatics* 25.11, pp. 1422–1423. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp163.

Corrie, Brian D. et al. (July 2018). "iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories". eng. In: *Immunol Rev* 284.1, pp. 24–41. ISSN: 1600-065X. DOI: 10.1111/imr.12666.

Crooks, Gavin E. et al. (June 2004). "WebLogo: a sequence logo generator". eng. In: *Genome Res* 14.6, pp. 1188–1190. ISSN: 1088-9051. DOI: 10.1101/gr.849004.

Da Re, S., S. Bertagnoli, et al. (May 1994). "Intramolecular signal transduction within the FixJ transcriptional activator: in vitro evidence for the inhibitory effect of the phosphorylatable regulatory domain". eng. In: *Nucleic Acids Res* 22.9, pp. 1555–1561. ISSN: 0305-1048. DOI: 10.1093/nar/22.9.1555.

Da Re, S., J. Schumacher, et al. (Nov. 1999). "Phosphorylation-induced dimerization of the FixJ receiver domain". eng. In: *Mol Microbiol* 34.3, pp. 504–511. ISSN: 0950-382X. DOI: 10.1046/j.1365-2958.1999.01614.x.

Dago, Angel E. et al. (June 2012). "Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis". eng. In: *Proc Natl Acad Sci U S A* 109.26, E1733–1742. ISSN: 1091-6490. DOI: 10.1073/pnas.1201301109.

David, M. et al. (Aug. 1988). "Cascade regulation of nif gene expression in Rhizobium meliloti". eng. In: *Cell* 54.5, pp. 671–683. ISSN: 0092-8674. DOI: 10.1016/s0092-8674(88)80012-6.

Del Alamo, Diego et al. (Jan. 2020). "Rapid Simulation of Unprocessed DEER Decay Data for Protein Fold Prediction". eng. In: *Biophys J* 118.2, pp. 366–375. ISSN: 1542-0086. DOI: 10.1016/j.bpj.2019.12.011.

DeWitt, William S. et al. (Aug. 2016). "A Public Database of Memory and Naive B-Cell Receptor Sequences". en. In: *PLOS ONE* 11.8. Publisher: Public Library of Science, e0160853. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0160853. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0160853 (visited on 01/11/2021).

Dias Neto, E. et al. (Mar. 2000). "Shotgun sequencing of the human transcriptome with ORF expressed sequence tags". eng. In: *Proc Natl Acad Sci U S A* 97.7, pp. 3491–3496. ISSN: 0027-8424. DOI: 10.1073/pnas.97.7.3491.

Dondelinger, Mathieu et al. (Oct. 2018). "Understanding the Significance and Implications of Antibody Numbering and Antigen-Binding Surface/Residue Definition". In: *Front Immunol* 9, p. 2278. ISSN: 1664-3224. DOI: 10.3389/fimmu.2018.02278. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6198058/ (visited on 09/09/2021).

Doria-Rose, Nicole A. et al. (May 2014). "Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies". en. In: *Nature* 509.7498. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7498 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Bioinformatics;HIV infections;Immunology;Virology Subject_term_id: bioinformatics;hiv-infections;immunology;virology, pp. 55–62. ISSN: 1476-4687. DOI: 10.1038/nature13036. URL: https://www.nature.com/articles/nature13036 (visited on 09/04/2021).

Ducancel, Frédéric and Bruno H. Muller (Aug. 2012). "Molecular engineering of antibodies for therapeutic and diagnostic purposes". eng. In: *MAbs* 4.4, pp. 445–457. ISSN: 1942-0870. DOI: 10.4161/mabs.20776.

Dunbar, James et al. (Jan. 2014). "SAbDab: the structural antibody database". eng. In: *Nucleic Acids Res* 42.Database issue, pp. D1140–1146. ISSN: 1362-4962. DOI: 10.1093/nar/gkt1043.

Ekeberg, Magnus et al. (Jan. 2013). "Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models". eng. In: *Phys Rev E Stat Nonlin Soft Matter Phys* 87.1, p. 012707. ISSN: 1550-2376. DOI: 10.1103/PhysRevE.87.012707.

Ekiert, Damian C. et al. (Sept. 2012). "Cross-neutralization of influenza A viruses mediated by a single antibody loop". eng. In: *Nature* 489.7417, pp. 526–532. ISSN: 1476-4687. DOI: 10.1038/nature11414.

Ellgaard, Lars and Ari Helenius (Mar. 2003). "Quality control in the endoplasmic reticulum". eng. In: *Nat Rev Mol Cell Biol* 4.3, pp. 181–191. ISSN: 1471-0072. DOI: 10.1038/nrm1052.

Finn, Jessica A. et al. (2016). "Improving Loop Modeling of the Antibody Complementarity-Determining Region 3 Using Knowledge-Based Restraints". eng. In: *PLoS One* 11.5, e0154811. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0154811.

Fleishman, Sarel J. et al. (2011). "RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite". eng. In: *PLoS One* 6.6, e20161. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0020161.

Fowler, Douglas M., Carlos L. Araya, et al. (Sept. 2010). "High-resolution mapping of protein sequence-function relationships". eng. In: *Nat Methods* 7.9, pp. 741–746. ISSN: 1548-7105. DOI: 10.1038/nmeth.1492.

Fowler, Douglas M. and Stanley Fields (Aug. 2014). "Deep mutational scanning: a new style of protein science". eng. In: *Nat Methods* 11.8, pp. 801–807. ISSN: 1548-7105. DOI: 10.1038/nmeth.3027.

Fraussen, Judith et al. (Aug. 2013). "Autoantigen induced clonal expansion in immortalized B cells from the peripheral blood of multiple sclerosis patients". eng. In: *J Neuroimmunol* 261.1-2, pp. 98–107. ISSN: 1872-8421. DOI: 10.1016/j.jneuroim.2013.05.002.

Gaëta, Bruno A. et al. (July 2007). "iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences". eng. In: *Bioinformatics* 23.13, pp. 1580–1587. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btm147.

Gao, Sean H. et al. (July 2013). "Monoclonal antibody humanness score and its applications". eng. In: *BMC Biotechnol* 13, p. 55. ISSN: 1472-6750. DOI: 10.1186/1472-6750-13-55.

Gilchuk, Pavlo et al. (Nov. 2020). "Integrated pipeline for the accelerated discovery of antiviral antibody therapeutics". en. In: *Nature Biomedical Engineering* 4.11. Number: 11 Publisher: Nature Publishing Group, pp. 1030–1043. ISSN: 2157-846X. DOI: 10.1038/s41551-020-0594-x. URL: https://www.nature.com/articles/s41551-020-0594-x (visited on 04/26/2021).

Gillies, S. D., K. M. Lo, and J. Wesolowski (Dec. 1989). "High-level expression of chimeric antibodies using adapted cDNA variable region cassettes". eng. In: *J Immunol Methods* 125.1-2, pp. 191–202. ISSN: 0022-1759. DOI: 10.1016/0022-1759(89)90093-8.

Giudicelli, Véronique, Denys Chaume, and Marie-Paule Lefranc (Jan. 2005). "IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes". In: *Nucleic Acids Res* 33.Database Issue, pp. D256–D261. ISSN: 0305-1048. DOI: 10.1093/nar/gki010. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC539964/ (visited on 01/12/2021).

Glanville, Jacob et al. (Dec. 2009). "Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire". eng. In: *Proc Natl Acad Sci U S A* 106.48, pp. 20216–20221. ISSN: 1091-6490. DOI: 10.1073/pnas.0909775106.

Goldenzweig, Adi and Sarel J. Fleishman (June 2018). "Principles of Protein Stability and Their Application in Computational Design". eng. In: *Annu Rev Biochem* 87, pp. 105–129. ISSN: 1545-4509. DOI: 10.1146/annurev-biochem-062917-012102.

Goldenzweig, Adi, Moshe Goldsmith, et al. (July 2016). "Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability". eng. In: *Mol Cell* 63.2, pp. 337–346. ISSN: 1097-4164. DOI: 10.1016/j.molcel.2016.06.012.

Gouet, P. et al. (Dec. 1999). "Structural transitions in the FixJ receiver domain". eng. In: *Structure* 7.12, pp. 1517–1526. ISSN: 0969-2126. DOI: 10.1016/s0969-2126(00)88342-2.

Gray, Jeffrey J. et al. (Aug. 2003). "Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations". eng. In: *J Mol Biol* 331.1, pp. 281–299. ISSN: 0022-2836. DOI: 10.1016/s0022-2836(03)00670-3.

Guindon, Stéphane et al. (May 2010). "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0". eng. In: *Syst Biol* 59.3, pp. 307–321. ISSN: 1076-836X. DOI: 10.1093/sysbio/syq010.

Harding, Fiona A. et al. (June 2010). "The immunogenicity of humanized and fully human antibodies: residual immunogenicity resides in the CDR regions". eng. In: *MAbs* 2.3, pp. 256–265. ISSN: 1942-0870. DOI: 10.4161/mabs.2.3.11641.

Hasegawa, Haruki et al. (July 2017). "Single amino acid substitution in LC-CDR1 induces Russell body phenotype that attenuates cellular protein synthesis through eIF2 phosphorylation and thereby downregulates IgG secretion despite operational secretory pathway traffic". In: *mAbs* 9.5. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19420862.2017.1314875, pp. 854–873. ISSN: 1942-0862. DOI: 10.1080/19420862.2017.1314875. URL: https://doi.org/10.1080/19420862.2017.1314875 (visited on 04/12/2021).

Henderson, Kylie A. et al. (Nov. 2007). "Structure of an IgNAR-AMA1 complex: targeting a conserved hydrophobic cleft broadens malarial strain recognition". eng. In: *Structure* 15.11, pp. 1452–1466. ISSN: 0969-2126. DOI: 10.1016/j.str.2007.09.011.

Henikoff, S. and J. G. Henikoff (Nov. 1992). "Amino acid substitution matrices from protein blocks". eng. In: *Proc Natl Acad Sci U S A* 89.22, pp. 10915–10919. ISSN: 0027-8424. DOI: 10.1073/pnas.89.22.10915.

Hochreiter, Sepp and Jürgen Schmidhuber (Dec. 1997). "Long Short-term Memory". In: *Neural computation* 9, pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

Holgate, Robert G. E. and Matthew P. Baker (Apr. 2009). "Circumventing immunogenicity in the development of therapeutic antibodies". eng. In: *IDrugs* 12.4, pp. 233–237. ISSN: 2040-3410.

Holinski, Alexandra et al. (Feb. 2017). "Combining ancestral sequence reconstruction with protein design to identify an interface hotspot in a key metabolic enzyme complex". eng. In: *Proteins* 85.2, pp. 312–321. ISSN: 1097-0134. DOI: 10.1002/prot.25225.

Hon, Jiri et al. (Jan. 2021). "SoluProt: prediction of soluble protein expression in Escherichia coli". In: *Bioinformatics* 37.1, pp. 23–28. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa1102. URL: https://doi.org/10.1093/bioinformatics/btaa1102 (visited on 04/21/2021).

Honegger, A. and A. Plückthun (June 2001). "Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool". eng. In: *J Mol Biol* 309.3, pp. 657–670. ISSN: 0022-2836. DOI: 10.1006/jmbi.2001.4662.

Hopf, Thomas A. et al. (Sept. 2014). "Sequence co-evolution gives 3D contacts and structures of protein complexes". eng. In: *Elife* 3. ISSN: 2050-084X. DOI: 10.7554/eLife.03430.

Hu, Xiangqian, David N. Beratan, and Weitao Yang (Oct. 2009). "A gradient-directed Monte Carlo method for global optimization in a discrete space: application to protein sequence design and folding". eng. In: *J Chem Phys* 131.15, p. 154117. ISSN: 1089-7690. DOI: 10.1063/1.3236834.

Hu, Xihao et al. (Nov. 2018). "Evaluation of immune repertoire inference methods from RNA-seq data". eng. In: *Nat Biotechnol* 36.11, p. 1034. ISSN: 1546-1696. DOI: 10.1038/nbt.4294.

Hu, Zengjian et al. (June 2007). "Ligand binding and circular permutation modify residue interaction network in DHFR". eng. In: *PLoS Comput Biol* 3.6, e117. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.0030117.

Hwang, William Ying Khee and Jefferson Foote (May 2005). "Immunogenicity of engineered antibodies". eng. In: *Methods* 36.1, pp. 3–10. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2005.01.001.

Iba, Y. et al. (May 1998). "Changes in the specificity of antibodies against steroid antigens by introduction of mutations into complementarity-determining regions of the V(H) domain". eng. In: *Protein Eng* 11.5, pp. 361–370. ISSN: 0269-2139. DOI: 10.1093/protein/11.5.361.

Jacob, Etai, Ron Unger, and Amnon Horovitz (Sept. 2015). "Codon-level information improves predictions of inter-residue contacts in proteins by correlated mutation analysis". eng. In: *Elife* 4, e08932. ISSN: 2050-084X. DOI: 10.7554/eLife.08932.

Jain, Tushar, David S. Cerutti, and J. Andrew McCammon (2006). "Configurational-bias sampling technique for predicting side-chain conformations in proteins". en. In: *Protein Science* 15.9. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1110/ps.062165906, pp. 2029–2039. ISSN: 1469-896X. DOI: 10.1110/ps.062165906. URL: https://onlinelibrary.wiley.com/doi/abs/10.1110/ps.062165906 (visited on 09/02/2021).

Jain, Tushar, Tingwan Sun, et al. (Jan. 2017). "Biophysical properties of the clinical-stage antibody landscape". eng. In: *Proc Natl Acad Sci U S A* 114.5, pp. 944–949. ISSN: 1091-6490. DOI: 10.1073/pnas.1616408114.

Jarasch, Alexander et al. (June 2015). "Developability assessment during the selection of novel therapeutic antibodies". eng. In: *J Pharm Sci* 104.6, pp. 1885–1898. ISSN: 1520-6017. DOI: 10.1002/jps.24430.

Jayapal, Karthik P. et al. (Oct. 2007). "Recombinant protein therapeutics from CHO Cells - 20 years and counting". In: *Chemical Engineering Progress* 103.10, pp. 40–47. ISSN: 0360-7275. URL: http://www.scopus.com/inward/record.url?scp=41849140828&partnerID=8YFLogxK (visited on 10/01/2021).

Jeon, Jouhyun et al. (Sept. 2011). "Molecular evolution of protein conformational changes revealed by a network of evolutionarily coupled residues". eng. In: *Mol Biol Evol* 28.9, pp. 2675–2685. ISSN: 1537-1719. DOI: 10.1093/molbev/msr094.

Jiang, Ning et al. (Feb. 2013). "Lineage Structure of the Human Antibody Repertoire in Response to Influenza Vaccination". EN. In: *Science Translational Medicine*. Publisher: American Association for the Advancement of Science. URL: https://www.science.org/doi/abs/10.1126/scitranslmed.3004794 (visited on 09/04/2021).

Johari, Yusuf B. et al. (Dec. 2015). "Integrated cell and process engineering for improved transient production of a "difficult-to-express" fusion protein by CHO cells". eng. In: *Biotechnol Bioeng* 112.12, pp. 2527–2542. ISSN: 1097-0290. DOI: 10.1002/bit.25687.

Jones, David T. et al. (Jan. 2012). "PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments". eng. In: *Bioinformatics* 28.2, pp. 184–190. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btr638.

Jones, Tim D. et al. (2016). "The INNs and outs of antibody nonproprietary names". eng. In: *MAbs* 8.1, pp. 1–9. ISSN: 1942-0870. DOI: 10.1080/19420862.2015.1114320.

Jost, Christian and Andreas Plückthun (Aug. 2014). "Engineered proteins with desired specificity: DARPins, other alternative scaffolds and bispecific IgGs". eng. In: *Curr Opin Struct Biol* 27, pp. 102–112. ISSN: 1879-033X. DOI: 10.1016/j.sbi.2014.05.011.

Joyce, M. Gordon et al. (July 2016). "Vaccine-Induced Antibodies that Neutralize Group 1 and Group 2 Influenza A Viruses". English. In: *Cell* 166.3. Publisher: Elsevier, pp. 609–623. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2016.06.043. URL: https://www.cell.com/cell/abstract/S0092-8674(16)30851-0 (visited on 09/04/2021).

Jumper, John et al. (Aug. 2021). "Highly accurate protein structure prediction with AlphaFold". en. In: *Nature* 596.7873. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 7873 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational biophysics;Machine learning;Protein structure predictions;Structural biology Subject_term_id: computational-biophysics;machine-learning;protein-structure-predictions;structural-biology, pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2. URL: https://www.nature.com/articles/s41586-021-03819-2 (visited on 10/08/2021).

Jung, David and Frederick W Alt (Jan. 2004). "Unraveling V(D)J Recombination: Insights into Gene Regulation". en. In: *Cell* 116.2, pp. 299–311. ISSN: 0092-8674. DOI: 10.1016/S0092-8674(04)00039-X. URL: https://www.sciencedirect.com/science/article/pii/S009286740400039X (visited on 09/10/2021).

Jung, David, Cosmas Giallourakis, et al. (Jan. 2006). "Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus". eng. In: *Annu Rev Immunol* 24, pp. 541–570. ISSN: 1545-3278. DOI: 10.1146/annurev.immunol.23.021704.115830. URL: https://doi.org/10.1146/annurev.immunol.23.021704.115830 (visited on 09/10/2021).

Jurtz, Vanessa Isabell et al. (Nov. 2017). "An introduction to deep learning on biological sequence data: examples and solutions". In: *Bioinformatics* 33.22, pp. 3685–3690. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx531. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5870575/ (visited on 03/06/2021).

Kahn, D. and G. Ditta (Apr. 1991). "Modular structure of FixJ: homology of the transcriptional activator domain with the -35 binding domain of sigma factors". eng. In: *Mol Microbiol* 5.4, pp. 987–997. ISSN: 0950-382X. DOI: 10.1111/j.1365-2958.1991.tb00774.x.

Kallehauge, Thomas Beuchert et al. (Jan. 2017). "Ribosome profiling-guided depletion of an mRNA increases cell growth rate and protein secretion". en. In: *Scientific Reports* 7.1. Number: 1 Publisher: Nature Publishing Group, p. 40388. ISSN: 2045-2322. DOI: 10.1038/srep40388. URL: https://www.nature.com/articles/srep40388 (visited on 04/12/2021).

Kamisetty, Hetunandan, Sergey Ovchinnikov, and David Baker (Sept. 2013). "Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era". eng. In: *Proc Natl Acad Sci U S A* 110.39, pp. 15674–15679. ISSN: 1091-6490. DOI: 10.1073/pnas.1314045110.

Kaplon, Hélène and Janice M. Reichert (Mar. 2019). "Antibodies to watch in 2019". eng. In: *MAbs* 11.2, pp. 219–238. ISSN: 1942-0870. DOI: 10.1080/19420862.2018.1556465.

— (Dec. 2021). "Antibodies to watch in 2021". eng. In: *MAbs* 13.1, p. 1860476. ISSN: 1942-0870. DOI: 10.1080/19420862.2020.1860476.

Khurana, Sameer et al. (Aug. 2018). "DeepSol: a deep learning framework for sequence-based protein solubility prediction". In: *Bioinformatics* 34.15, pp. 2605–2613. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty166. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6355112/ (visited on 04/21/2021).

Koboldt, Daniel C. et al. (Sept. 2013). "The next-generation sequencing revolution and its impact on genomics". eng. In: *Cell* 155.1, pp. 27–38. ISSN: 1097-4172. DOI: 10.1016/j.cell.2013.09.006.

Kodali, Pranav et al. (2021). "RosettaCM for antibodies with very long HCDR3s and low template availability". en. In: *Proteins: Structure, Function, and Bioinformatics* n/a.n/a (). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.26166. ISSN: 1097-0134. DOI: 10.1002/prot.26166. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26166 (visited on 06/28/2021).

Köhler, G. and C. Milstein (Aug. 1975). "Continuous cultures of fused cells secreting antibody of predefined specificity". eng. In: *Nature* 256.5517, pp. 495–497. ISSN: 0028-0836. DOI: 10.1038/256495a0.

Kolar, Grant R. et al. (Nov. 2004). "Human fetal, cord blood, and adult lymphocyte progenitors have similar potential for generating B cells with a diverse immunoglobulin repertoire". eng. In: *Blood* 104.9, pp. 2981–2987. ISSN: 0006-4971. DOI: 10.1182/blood-2003-11-3961.

Kortemme, Tanja, Alexandre Morozov, and David Baker (Mar. 2003). "An Orientation-dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein–Protein Complexes". In: *Journal of molecular biology* 326, pp. 1239–59. DOI: 10.1016/S0022-2836(03)00021-4.

Kovaltsuk, Aleksandr et al. (Oct. 2018). "Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires". eng. In: *J Immunol* 201.8, pp. 2502–2509. ISSN: 1550-6606. DOI: 10.4049/jimmunol.1800708.

Krause, Jens C. et al. (Oct. 2011). "A Broadly Neutralizing Human Monoclonal Antibody That Recognizes a Conserved, Novel Epitope on the Globular Head of the Influenza H1N1 Virus Hemagglutinin". In: *Journal of Virology* 85.20. Publisher: American Society for Microbiology, pp. 10905–10908. DOI: 10.1128/JVI.00700-11. URL: https://journals.asm.org/doi/10.1128/JVI.00700-11 (visited on 09/04/2021).

Krawczyk, Konrad et al. (Oct. 2019). "Looking for therapeutic antibodies in next-generation sequencing repositories". In: *mAbs* 11.7. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19420862.20

pp. 1197–1205. ISSN: 1942-0862. DOI: 10.1080/19420862.2019.1633884. URL: https://doi.org/10.1080/19420862.2019.1633884 (visited on 09/10/2021).

Kuboniwa, H. et al. (Sept. 1995). "Solution structure of calcium-free calmodulin". eng. In: *Nat Struct Biol* 2.9, pp. 768–776. ISSN: 1072-8368. DOI: 10.1038/nsb0995-768.

Kuenze, Georg et al. (Nov. 2019). "Integrative Protein Modeling in RosettaNMR from Sparse Paramagnetic Restraints". eng. In: *Structure* 27.11, 1721–1734.e5. ISSN: 1878-4186. DOI: 10.1016/j.str.2019.08.012.

Kufareva, Irina and Ruben Abagyan (2012). "Methods of protein structure comparison". In: *Methods Mol Biol* 857, pp. 231–257. ISSN: 1064-3745. DOI: 10.1007/978-1-61779-588-6_10. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4321859/ (visited on 09/01/2021).

Kuhlman, B. and D. Baker (Sept. 2000). "Native protein sequences are close to optimal for their structures". eng. In: *Proc Natl Acad Sci U S A* 97.19, pp. 10383–10388. ISSN: 0027-8424. DOI: 10.1073/pnas.97.19.10383.

Kuhlman, Brian et al. (Nov. 2003). "Design of a novel globular protein fold with atomic-level accuracy". eng. In: *Science* 302.5649, pp. 1364–1368. ISSN: 1095-9203. DOI: 10.1126/science.1089427.

Kunert, Renate and David Reinhart (2016). "Advances in recombinant antibody manufacturing". In: *Appl Microbiol Biotechnol* 100, pp. 3451–3461. ISSN: 0175-7598. DOI: 10.1007/s00253-016-7388-9. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4803805/ (visited on 09/30/2021).

Lapidoth, Gideon D. et al. (Aug. 2015). "AbDesign: An algorithm for combinatorial backbone design guided by natural conformations and sequences". eng. In: *Proteins* 83.8, pp. 1385–1406. ISSN: 1097-0134. DOI: 10.1002/prot.24779.

Lawrence, M. C. and P. M. Colman (Dec. 1993). "Shape complementarity at protein/protein interfaces". eng. In: *J Mol Biol* 234.4, pp. 946–950. ISSN: 0022-2836. DOI: 10.1006/jmbi.1993.1648.

Lazar, Greg A. et al. (Mar. 2007). "A molecular immunology approach to antibody humanization and functional optimization". eng. In: *Mol Immunol* 44.8, pp. 1986–1998. ISSN: 0161-5890. DOI: 10.1016/j.molimm.2006.09.029.

Al-Lazikani, B., A. M. Lesk, and C. Chothia (Nov. 1997). "Standard conformations for the canonical structures of immunoglobulins". eng. In: *J Mol Biol* 273.4, pp. 927–948. ISSN: 0022-2836. DOI: 10.1006/jmbi.1997.1354.

Leaver-Fay, Andrew, Ron Jacak, et al. (July 2011). "A Generic Program for Multistate Protein Design". en. In: *PLOS ONE* 6.7. Publisher: Public Library of Science, e20937. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0020937. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0020937 (visited on 09/06/2021).

Leaver-Fay, Andrew, Michael Tyka, et al. (2011). "ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules". eng. In: *Methods Enzymol* 487, pp. 545–574. ISSN: 1557-7988. DOI: 10.1016/B978-0-12-381270-4.00019-6.

Leavy, Olive (Dec. 2016). "The birth of monoclonal antibodies". en. In: *Nat Immunol* 17.1. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group, S13–S13. ISSN: 1529-2916. DOI: 10.1038/ni.3608. URL: https://www.nature.com/articles/ni.3608 (visited on 09/30/2021).

Lefranc, M. P. (Nov. 1997). "Unique database numbering system for immunogenetic analysis". eng. In: *Immunol Today* 18.11, p. 509. ISSN: 0167-5699. DOI: 10.1016/s0167-5699(97)01163-8.

— (1998). "IMGT (ImMunoGeneTics) locus on focus. A new section of Experimental and Clinical Immunogenetics". eng. In: *Exp Clin Immunogenet* 15.1, pp. 1–7. ISSN: 0254-9670. DOI: 10.1159/000019049.

Lefranc, Marie-Paule, Véronique Giudicelli, et al. (Jan. 2015). "IMGT®, the international ImMunoGeneTics information system® 25 years on". eng. In: *Nucleic Acids Res* 43.Database issue, pp. D413–422. ISSN: 1362-4962. DOI: 10.1093/nar/gku1056.

Lefranc, Marie-Paule, Christelle Pommié, Quentin Kaas, et al. (2005). "IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains". eng. In: *Dev Comp Immunol* 29.3, pp. 185–203. ISSN: 0145-305X. DOI: 10.1016/j.dci.2004.07.003.

Lefranc, Marie-Paule, Christelle Pommié, Manuel Ruiz, et al. (Jan. 2003). "IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains". eng. In: *Dev Comp Immunol* 27.1, pp. 55–77. ISSN: 0145-305X. DOI: 10.1016/s0145-305x(02)00039-3.

Lehmann, Andreas et al. (2015). "Stability engineering of anti-EGFR scFv antibodies by rational design of a lambda-to-kappa swap of the VL framework using a structure-guided approach". eng. In: *MAbs* 7.6, pp. 1058–1071. ISSN: 1942-0870. DOI: 10.1080/19420862.2015.1088618.

Lennon, Brett W., Charles H. Williams, and Martha L. Ludwig (Aug. 2000). "Twists in Catalysis: Alternating Conformations of Escherichia coli Thioredoxin Reductase". In: *Science* 289.5482. Publisher: American Association for the Advancement of Science, pp. 1190–1194. DOI: 10.1126/science.289.5482.1190. URL: https://www.science.org/doi/abs/10.1126/science.289.5482.1190 (visited on 10/15/2021).

Lewit-Bentley, Anita and Stéphane Réty (Dec. 2000). "EF-hand calcium-binding proteins". In: *Current Opinion in Structural Biology* 10.6, pp. 637–643. ISSN: 0959-440X. DOI: 10.1016/S0959-440X(00)00142-1. URL: http://www.sciencedirect.com/science/article/pii/S0959440X00001421 (visited on 05/10/2019).

Li, Yu et al. (Aug. 2019). "Deep learning in bioinformatics: Introduction, application, and perspective in the big data era". eng. In: *Methods* 166, pp. 4–21. ISSN: 1095-9130. DOI: 10.1016/j.ymeth.2019.04.008.

Liao, Hua-Xin et al. (Apr. 2013). "Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus". en. In: *Nature* 496.7446. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7446 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Antibodies;HIV infections Subject_term_id: antibodies;hiv-infections, pp. 469–476. ISSN: 1476-4687.

DOI: 10.1038/nature12053. URL: https://www.nature.com/articles/nature12053 (visited on 09/04/2021).

Löffler, Patrick et al. (June 2017). "Rosetta:MSF: a modular framework for multi-state computational protein design". en. In: *PLOS Computational Biology* 13.6. Publisher: Public Library of Science, e1005600. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005600. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005600 (visited on 07/22/2021).

Longo, Nancy S. et al. (May 2017). "Mechanisms That Shape Human Antibody Repertoire Development in Mice Transgenic for Human Ig H and L Chain Loci". eng. In: *J Immunol* 198.10, pp. 3963–3977. ISSN: 1550-6606. DOI: 10.4049/jimmunol.1700133.

Lu, Ruei-Min et al. (Jan. 2020). "Development of therapeutic antibodies for the treatment of diseases". In: *Journal of Biomedical Science* 27.1, p. 1. ISSN: 1423-0127. DOI: 10.1186/s12929-019-0592-z. URL: https://doi.org/10.1186/s12929-019-0592-z (visited on 12/09/2020).

Lu, Xiaojun et al. (Jan. 2019). "Deamidation and isomerization liability analysis of 131 clinical-stage antibodies". eng. In: *MAbs* 11.1, pp. 45–57. ISSN: 1942-0870. DOI: 10.1080/19420862.2018.1548233.

Magnan, Christophe N., Arlo Randall, and Pierre Baldi (Sept. 2009). "SOLpro: accurate sequence-based prediction of protein solubility". In: *Bioinformatics* 25.17, pp. 2200–2207. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp386. URL: https://doi.org/10.1093/bioinformatics/btp386 (visited on 04/21/2021).

Maloney, D. G. et al. (Oct. 1997). "IDEC-C2B8: results of a phase I multiple-dose trial in patients with relapsed non-Hodgkin's lymphoma". eng. In: *J Clin Oncol* 15.10, pp. 3266–3274. ISSN: 0732-183X. DOI: 10.1200/JCO.1997.15.10.3266.

Marino Buslje, Cristina et al. (Nov. 2010). "Networks of high mutual information define the structural proximity of catalytic sites: implications for catalytic residue identification". eng. In: *PLoS Comput Biol* 6.11, e1000978. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000978.

Marks, Debora S., Lucy J. Colwell, et al. (2011). "Protein 3D structure computed from evolutionary sequence variation". eng. In: *PLoS One* 6.12, e28766. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0028766.

Marks, Debora S., Thomas A. Hopf, and Chris Sander (Nov. 2012). "Protein structure prediction from sequence variation". eng. In: *Nat Biotechnol* 30.11, pp. 1072–1080. ISSN: 1546-1696. DOI: 10.1038/nbt.2419.

Martin, Andrew C. R. and Anthony R. Rees (Oct. 2016). "Extracting human antibody sequences from public databases for antibody humanization: high frequency of species assignment errors". eng. In: *Protein Eng Des Sel* 29.10, pp. 403–408. ISSN: 1741-0134. DOI: 10.1093/protein/gzw018.

Marze, Nicholas A., Sergey Lyskov, and Jeffrey J. Gray (Oct. 2016). "Improved prediction of antibody VL-VH orientation". eng. In: *Protein Eng Des Sel* 29.10, pp. 409–418. ISSN: 1741-0134. DOI: 10.1093/protein/gzw013.

Mathias, Sven et al. (2020). “Unraveling what makes a monoclonal antibody difficult-to-express: From intracellular accumulation to incomplete folding and degradation via ERAD”. en. In: *Biotechnology and Bioengineering* 117.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bit.27196, pp. 5–16. ISSN: 1097-0290. DOI: https://doi.org/10.1002/bit.27196. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bit.27196 (visited on 01/13/2021).

Mayrhofer, Patrick and Renate Kunert (2021). “Nomenclature of humanized mAbs: Early concepts, current challenges and future perspectives”. In: *Hum Antibodies* 27.1 (), pp. 37–51. ISSN: 1093-2607. DOI: 10.3233/HAB-180347. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6294595/ (visited on 01/18/2021).

McLellan, Jason S. et al. (Nov. 2011). “Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9”. eng. In: *Nature* 480.7377, pp. 336–343. ISSN: 1476-4687. DOI: 10.1038/nature10696.

Ménétrey, Julie et al. (June 2000). “Structure of Arf6–GDP suggests a basis for guanine nucleotide exchange factors specificity”. en. In: *Nat Struct Mol Biol* 7.6. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6 Primary_atype: Research Publisher: Nature Publishing Group, pp. 466–469. ISSN: 1545-9985. DOI: 10.1038/75863. URL: https://www.nature.com/articles/nsb0600_466 (visited on 10/15/2021).

Metzker, Michael L. (Jan. 2010). “Sequencing technologies — the next generation”. In: *Nature Reviews Genetics* 11.1, pp. 31–46. ISSN: 1471-0064. DOI: 10.1038/nrg2626. URL: https://doi.org/10.1038/nrg2626.

Milburn, M. V. et al. (Feb. 1990). “Molecular switch for signal transduction: structural differences between active and inactive forms of protooncogenic ras proteins”. eng. In: *Science* 247.4945, pp. 939–945. ISSN: 0036-8075. DOI: 10.1126/science.2406906.

Mimura, Yusuke et al. (Jan. 2018). “Glycosylation engineering of therapeutic IgG antibodies: challenges for the safety, functionality and efficacy”. eng. In: *Protein Cell* 9.1, pp. 47–62. ISSN: 1674-8018. DOI: 10.1007/s13238-017-0433-3.

Morcos, Faruck et al. (Dec. 2011). “Direct-coupling analysis of residue coevolution captures native contacts across many protein families”. eng. In: *Proc Natl Acad Sci U S A* 108.49, E1293–1301. ISSN: 1091-6490. DOI: 10.1073/pnas.1111471108.

Morea, V. et al. (Jan. 1998). “Conformations of the third hypervariable region in the VH domain of immunoglobulins”. eng. In: *J Mol Biol* 275.2, pp. 269–294. ISSN: 0022-2836. DOI: 10.1006/jmbi.1997.1442.

Müller, C. W., G. J. Schlauderer, et al. (Feb. 1996). “Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding”. eng. In: *Structure* 4.2, pp. 147–156. ISSN: 0969-2126. DOI: 10.1016/s0969-2126(96)00018-4.

Müller, C. W. and G. E. Schulz (Mar. 1992). “Structure of the complex between adenylate kinase from Escherichia coli and the inhibitor Ap5A refined at 1.9 A resolution. A model for a catalytic transition state”. eng. In: *J Mol Biol* 224.1, pp. 159–177. ISSN: 0022-2836. DOI: 10.1016/0022-2836(92)90582-5.

Murphy, Kenneth, Casey Weaver, and Charles Janeway (2017). *Janeway's immunobiology*. English. OCLC: 933586700. ISBN: 978-0-8153-4505-3 978-0-8153-4551-0 978-0-8153-4445-2 978-0-8153-4550-3.

Nechansky, Andreas (Jan. 2010a). "HAHA – nothing to laugh about. Measuring the immunogenicity (human anti-human antibody response) induced by humanized monoclonal antibodies applying ELISA and SPR technology". en. In: *Journal of Pharmaceutical and Biomedical Analysis* 51.1, pp. 252–254. ISSN: 0731-7085. DOI: 10.1016/j.jpba.2009.07.013. URL: http://www.sciencedirect.com/science/article/pii/S0731708509004725 (visited on 01/18/2021).

— (Jan. 2010b). "HAHA–nothing to laugh about. Measuring the immunogenicity (human anti-human antibody response) induced by humanized monoclonal antibodies applying ELISA and SPR technology". eng. In: *J Pharm Biomed Anal* 51.1, pp. 252–254. ISSN: 1873-264X. DOI: 10.1016/j.jpba.2009.07.013.

Neria, Eyal, Stefan Fischer, and Martin Karplus (Aug. 1996). "Simulation of activation free energies in molecular systems". In: *J. Chem. Phys.* 105.5. Publisher: American Institute of Physics, pp. 1902–1921. ISSN: 0021-9606. DOI: 10.1063/1.472061. URL: https://aip.scitation.org/doi/10.1063/1.472061 (visited on 09/01/2021).

Nguyen, Elizabeth Dong et al. (July 2013). "Assessment and Challenges of Ligand Docking into Comparative Models of G-Protein Coupled Receptors". en. In: *PLOS ONE* 8.7. Publisher: Public Library of Science, e67302. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0067302. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0067302 (visited on 09/06/2021).

Nivón, Lucas Gregorio, Rocco Moretti, and David Baker (Apr. 2013). "A Pareto-Optimal Refinement Method for Protein Design Scaffolds". In: *PLoS One* 8.4. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0059004. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3614904/ (visited on 04/16/2021).

Norn, Christoffer H., Gideon Lapidoth, and Sarel J. Fleishman (Jan. 2017). "High-accuracy modeling of antibody structures by a search for minimum-energy recombination of backbone fragments". eng. In: *Proteins* 85.1, pp. 30–38. ISSN: 1097-0134. DOI: 10.1002/prot.25185.

North, Benjamin, Andreas Lehmann, and Roland L. Dunbrack (Feb. 2011). "A new clustering of antibody CDR loop conformations". eng. In: *J Mol Biol* 406.2, pp. 228–256. ISSN: 1089-8638. DOI: 10.1016/j.jmb.2010.10.030.

Oh, B. H. et al. (May 1993). "Three-dimensional structures of the periplasmic lysine/arginine/ornithine-binding protein with and without a ligand". eng. In: *J Biol Chem* 268.15, pp. 11348–11355. ISSN: 0021-9258.

Olimpieri, Pier Paolo, Paolo Marcatili, and Anna Tramontano (Feb. 2015). "Tabhu: tools for antibody humanization". eng. In: *Bioinformatics* 31.3, pp. 434–435. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btu667.

Osawa, M. et al. (Sept. 1999). "A novel target recognition revealed by calmodulin in complex with Ca2+-calmodulin-dependent kinase kinase". eng. In: *Nat Struct Biol* 6.9, pp. 819–824. ISSN: 1072-8368. DOI: 10.1038/12271.

Otterbein, Ludovic R. et al. (Apr. 2002). "Crystal structures of S100A6 in the Ca(2+)-free and Ca(2+)-bound states: the calcium sensor mechanism of S100 proteins revealed at atomic resolution". eng. In: *Structure* 10.4, pp. 557–567. ISSN: 0969-2126. DOI: 10.1016/s0969-2126(02)00740-2.

Ovchinnikov, Sergey, Hetunandan Kamisetty, and David Baker (May 2014). "Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information". In: *eLife* 3. Ed. by Benoit Roux. Publisher: eLife Sciences Publications, Ltd, e02030. ISSN: 2050-084X. DOI: 10.7554/eLife.02030. URL: https://doi.org/10.7554/eLife.02030 (visited on 09/02/2021).

Ovchinnikov, Sergey, David E. Kim, et al. (2016). "Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta". en. In: *Proteins: Structure, Function, and Bioinformatics* 84.S1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.24974, pp. 67–75. ISSN: 1097-0134. DOI: 10.1002/prot.24974. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.24974 (visited on 09/02/2021).

Pancera, Marie et al. (Aug. 2010). "Crystal structure of PG16 and chimeric dissection with somatically related PG9: structure-function analysis of two quaternary-specific antibodies that effectively neutralize HIV-1". eng. In: *J Virol* 84.16, pp. 8098–8110. ISSN: 1098-5514. DOI: 10.1128/JVI.00966-10.

Parray, Hilal Ahmed et al. (Aug. 2020). "Hybridoma technology a versatile method for isolation of monoclonal antibodies, its applicability across species, limitations, advancement and future perspectives". en. In: *International Immunopharmacology* 85, p. 106639. ISSN: 1567-5769. DOI: 10.1016/j.intimp.2020.106639. URL: https://www.sciencedirect.com/science/article/pii/S156757692031105X (visited on 09/30/2021).

Parren, Paul W H I, Paul J Carter, and Andreas Plückthun (Aug. 2017). "Changes to International Nonproprietary Names for antibody therapeutics 2017 and beyond: of mice, men and more". eng. In: *MAbs* 9.6, pp. 898–906. ISSN: 1942-0870. DOI: 10.1080/19420862.2017.1341029. URL: https://europepmc.org/articles/PMC5590622 (visited on 08/04/2021).

Pasqualato, Sebastiano et al. (Mar. 2001). "The structural GDP/GTP cycle of human Arf6". In: *EMBO Rep* 2.3, pp. 234–238. ISSN: 1469-221X. DOI: 10.1093/embo-reports/kve043. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1083839/ (visited on 10/15/2021).

Pierce, Niles A. and Erik Winfree (Oct. 2002). "Protein design is NP-hard". eng. In: *Protein Eng* 15.10, pp. 779–782. ISSN: 0269-2139. DOI: 10.1093/protein/15.10.779.

Poiron, C (2021). *IMGT/mAb-DB: the IMGT® database for therapeutic monoclonal antibodies*. URL: https://scholar.googleusercontent.com/scholar?q=cache:n6cNQHNblWEJ:scholar.google.com/&hl=en&as_sdt=0,43 (visited on 09/10/2021).

Ponder, J. W. and F. M. Richards (Feb. 1987). "Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes". eng. In: *J Mol Biol* 193.4, pp. 775–791. ISSN: 0022-2836. DOI: 10.1016/0022-2836(87)90358-5.

Prihoda, David et al. (n.d.). "BioPhi: A platform for antibody design, humanization and humanness evaluation based on natural antibody repertoires and deep learning". en. In: (), p. 19.

Protopapadakis, Evdokia et al. (June 2005). "Isolation and characterization of human anti-acetylcholine receptor monoclonal antibodies from transgenic mice expressing human immunoglobulin loci". eng. In: *Eur J Immunol* 35.6, pp. 1960–1968. ISSN: 0014-2980. DOI: 10.1002/eji.200526173.

Pybus, Leon P., Greg Dean, et al. (Feb. 2014). "Model-directed engineering of "difficult-to-express" monoclonal antibody production by Chinese hamster ovary cells". eng. In: *Biotechnol Bioeng* 111.2, pp. 372–385. ISSN: 1097-0290. DOI: 10.1002/bit.25116.

Pybus, Leon P., David C. James, et al. (Feb. 2014). "Predicting the expression of recombinant monoclonal antibodies in Chinese hamster ovary cells based on sequence features of the CDR3 domain". eng. In: *Biotechnol Prog* 30.1, pp. 188–197. ISSN: 1520-6033. DOI: 10.1002/btpr.1839.

Queen, C. et al. (Dec. 1989). "A humanized antibody that binds to the interleukin 2 receptor". eng. In: *Proc Natl Acad Sci U S A* 86.24, pp. 10029–10033. ISSN: 0027-8424. DOI: 10.1073/pnas.86.24.10029.

Raimondi, Daniele et al. (Apr. 2020). "Insight into the protein solubility driving forces with neural attention". en. In: *PLOS Computational Biology* 16.4. Publisher: Public Library of Science, e1007722. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007722. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007722 (visited on 04/21/2021).

Raveh, Barak et al. (Apr. 2011). "Rosetta FlexPepDock ab-initio: simultaneous folding, docking and refinement of peptides onto their receptors". eng. In: *PLoS One* 6.4, e18934. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0018934.

Remmert, Michael et al. (Dec. 2011). "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment". eng. In: *Nat Methods* 9.2, pp. 173–175. ISSN: 1548-7105. DOI: 10.1038/nmeth.1818.

Renfrew, P. Douglas, Glenn L. Butterfoss, and Brian Kuhlman (2008). "Using quantum mechanics to improve estimates of amino acid side chain rotamer energies". en. In: *Proteins: Structure, Function, and Bioinformatics* 71.4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.21845, pp. 1637–1646. ISSN: 1097-0134. DOI: 10.1002/prot.21845. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.21845 (visited on 09/01/2021).

Richter, Florian et al. (2011). "De novo enzyme design using Rosetta3". eng. In: *PLoS One* 6.5, e19230. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0019230.

Rohl, Carol A. et al. (2004). "Protein structure prediction using Rosetta". eng. In: *Methods Enzymol* 383, pp. 66–93. ISSN: 0076-6879. DOI: 10.1016/S0076-6879(04)83004-0.

Rubelt, Florian et al. (Dec. 2017). "Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data". en. In: *Nature Immunology* 18.12.

Number: 12 Publisher: Nature Publishing Group, pp. 1274–1278. ISSN: 1529-2916. DOI: 10.1038/ni.3873. URL: https://www.nature.com/articles/ni.3873 (visited on 01/12/2021).

Russ, Daniel E., Kwan-Yuet Ho, and Nancy S. Longo (May 2015). "HTJoinSolver: Human immunoglobulin VDJ partitioning using approximate dynamic programming constrained by conserved motifs". eng. In: *BMC Bioinformatics* 16, p. 170. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0589-x.

Saada, Ravit et al. (June 2007). "Models for antigen receptor gene rearrangement: CDR3 length". eng. In: *Immunol Cell Biol* 85.4, pp. 323–332. ISSN: 0818-9641. DOI: 10.1038/sj.icb.7100055.

Saka, Koichiro et al. (Mar. 2021). "Antibody design using LSTM based deep generative model from phage display library for affinity maturation". en. In: *Scientific Reports* 11.1. Number: 1 Publisher: Nature Publishing Group, p. 5852. ISSN: 2045-2322. DOI: 10.1038/s41598-021-85274-7. URL: https://www.nature.com/articles/s41598-021-85274-7 (visited on 04/15/2021).

Sauer, Marion F. et al. (Feb. 2020). "Multi-state design of flexible proteins predicts sequences optimal for conformational change". eng. In: *PLoS Comput Biol* 16.2, e1007339. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007339.

Schildbach, J. F. et al. (Oct. 1993). "Heavy chain position 50 is a determinant of affinity and specificity for the anti-digoxin antibody 26-10". eng. In: *J Biol Chem* 268.29, pp. 21739–21747. ISSN: 0021-9258.

Schmitz, Samuel et al. (Jan. 2020). "Human-likeness of antibody biologics determined by back-translation and comparison with large antibody variable gene repertoires". In: *mAbs* 12.1. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19420862.2020.1758291, p. 1758291. ISSN: 1942-0862. DOI: 10.1080/19420862.2020.1758291. URL: https://doi.org/10.1080/19420862.2020.1758291 (visited on 03/03/2021).

Schoeder, Clara T. et al. (Mar. 2021). "Modeling Immunity with Rosetta: Methods for Antibody and Antigen Design". In: *Biochemistry* 60.11. Publisher: American Chemical Society, pp. 825–846. ISSN: 0006-2960. DOI: 10.1021/acs.biochem.0c00912. URL: https://doi.org/10.1021/acs.biochem.0c00912 (visited on 04/15/2021).

Schroff, R. W. et al. (Feb. 1985). "Human anti-murine immunoglobulin responses in patients receiving monoclonal antibody therapy". eng. In: *Cancer Res* 45.2, pp. 879–885. ISSN: 0008-5472.

Schug, Alexander et al. (Dec. 2009). "High-resolution protein complexes from integrating genomic information with molecular simulation". eng. In: *Proc Natl Acad Sci U S A* 106.52, pp. 22124–22129. ISSN: 1091-6490. DOI: 10.1073/pnas.0912100106.

Seeliger, Daniel (2013). "Development of scoring functions for antibody sequence assessment and optimization". eng. In: *PLoS One* 8.10, e76909. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0076909.

Seifert, Marc et al. (Feb. 2009). "A model for the development of human IgD-only B cells: Genotypic analyses suggest their generation in superantigen driven immune responses". eng. In: *Mol Immunol* 46.4, pp. 630–639. ISSN: 0161-5890. DOI: 10.1016/j.molimm.2008.07.032.

Senior, Andrew W. et al. (Jan. 2020). "Improved protein structure prediction using potentials from deep learning". en. In: *Nature* 577.7792. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7792 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Machine learning;Protein structure predictions Subject_term_id: machine-learning;protein-structure-predictions, pp. 706–710. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1923-7. URL: https://www.nature.com/articles/s41586-019-1923-7 (visited on 10/08/2021).

Sevy, Alexander M., Tim M. Jacobs, et al. (July 2015). "Design of Protein Multi-specificity Using an Independent Sequence Search Reduces the Barrier to Low Energy Sequences". eng. In: *PLoS Comput Biol* 11.7, e1004300. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004300.

Sevy, Alexander M., Swetasudha Panda, et al. (Feb. 2018). "Integrating linear optimization with structural modeling to increase HIV neutralization breadth". eng. In: *PLoS Comput Biol* 14.2, e1005999. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005999.

Sevy, Alexander M., Cinque Soto, et al. (Dec. 2019). "Immune repertoire fingerprinting by principal component analysis reveals shared features in subject groups with common exposures". en. In: *BMC Bioinformatics* 20.1, p. 629. ISSN: 1471-2105. DOI: 10.1186/s12859-019-3281-8. URL: https://doi.org/10.1186/s12859-019-3281-8 (visited on 09/04/2021).

Sevy, Alexander M., Nicholas C. Wu, et al. (Jan. 2019). "Multistate design of influenza antibodies improves affinity and breadth against seasonal viruses". eng. In: *Proc Natl Acad Sci U S A* 116.5, pp. 1597–1602. ISSN: 1091-6490. DOI: 10.1073/pnas.1806004116.

Shannon, Paul et al. (Nov. 2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks". eng. In: *Genome Res* 13.11, pp. 2498–2504. ISSN: 1088-9051. DOI: 10.1101/gr.1239303.

Sheffler, Will and David Baker (Jan. 2009). "RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation". eng. In: *Protein Sci* 18.1, pp. 229–239. ISSN: 1469-896X. DOI: 10.1002/pro.8.

Sheng, Zizhang et al. (2017). "Gene-Specific Substitution Profiles Describe the Types and Frequencies of Amino Acid Changes during Antibody Somatic Hypermutation". eng. In: *Front Immunol* 8, p. 537. ISSN: 1664-3224. DOI: 10.3389/fimmu.2017.00537.

Shirai, H., A. Kidera, and H. Nakamura (Dec. 1996). "Structural classification of CDR-H3 in antibodies". eng. In: *FEBS Lett* 399.1-2, pp. 1–8. ISSN: 0014-5793. DOI: 10.1016/s0014-5793(96)01252-5.

Simons, Kim T., Charles Kooperberg, et al. (Apr. 1997). "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions11Edited by F. E. Cohen". en. In: *Journal of Molecular Biology* 268.1, pp. 209–225. ISSN: 0022-2836. DOI: 10.1006/jmbi.1997.0959. URL: https://www.sciencedirect.com/science/article/pii/S0022283697909591 (visited on 09/01/2021).

Simons, Kim T., Ingo Ruczinski, et al. (1999). "Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins". fr. In: *Proteins: Structure, Function, and Bioinformatics* 34.1. _eprint: https://onlinelibrary.wiley.com/do

0134%2819990101%2934%3A1%3C82%3A%3AAID-PROT7%3E3.0.CO%3B2-A, pp. 82–95. ISSN: 1097-0134. DOI: 10.1002/(SICI)1097-0134(19990101)34:1<82::AID-PROT7>3.0.CO;2-A. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0134%2819990101%2934%3A1%3C82%3A%3AAID-PROT7%3E3.0.CO%3B2-A (visited on 09/01/2021).

Sircar, Aroop and Jeffrey J. Gray (Jan. 2010). "SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models". eng. In: *PLoS Comput Biol* 6.1, e1000644. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000644.

Sircar, Aroop, Eric T. Kim, and Jeffrey J. Gray (July 2009). "RosettaAntibody: antibody variable region homology modeling server". eng. In: *Nucleic Acids Res* 37.Web Server issue, W474–479. ISSN: 1362-4962. DOI: 10.1093/nar/gkp387.

Sivasubramanian, Arvind, Ginger Chao, et al. (Mar. 2006). "Structural model of the mAb 806-EGFR complex using computational docking followed by computational and experimental mutagenesis". eng. In: *Structure* 14.3, pp. 401–414. ISSN: 0969-2126. DOI: 10.1016/j.str.2005.11.022.

Sivasubramanian, Arvind, Aroop Sircar, et al. (Feb. 2009). "Toward high-resolution homology modeling of antibody Fv regions and application to antibody-antigen docking". eng. In: *Proteins* 74.2, pp. 497–514. ISSN: 1097-0134. DOI: 10.1002/prot.22309.

Smialowski, Pawel et al. (June 2012). "PROSO II–a new method for protein solubility prediction". eng. In: *FEBS J* 279.12, pp. 2192–2200. ISSN: 1742-4658. DOI: 10.1111/j.1742-4658.2012.08603.x.

Smith, Colin A. and Tanja Kortemme (July 2008). "Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction". eng. In: *J Mol Biol* 380.4, pp. 742–756. ISSN: 1089-8638. DOI: 10.1016/j.jmb.2008.05.023.

Socolich, Michael et al. (Sept. 2005). "Evolutionary information for specifying a protein fold". eng. In: *Nature* 437.7058, pp. 512–518. ISSN: 1476-4687. DOI: 10.1038/nature03991.

Sok, Devin et al. (Sept. 2016). "Priming HIV-1 broadly neutralizing antibody precursors in human Ig loci transgenic mice". eng. In: *Science* 353.6307, pp. 1557–1560. ISSN: 1095-9203. DOI: 10.1126/science.aah3945.

Song, Yifan et al. (Oct. 2013). "High-resolution comparative modeling with RosettaCM". eng. In: *Structure* 21.10, pp. 1735–1742. ISSN: 1878-4186. DOI: 10.1016/j.str.2013.08.005.

Sormanni, Pietro, Francesco A. Aprile, and Michele Vendruscolo (Jan. 2015). "The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility". en. In: *Journal of Molecular Biology* 427.2, pp. 478–490. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2014.09.026. URL: https://www.sciencedirect.com/science/article/pii/S0022283614005312 (visited on 04/21/2021).

Soto, Cinque, Robin G. Bombardi, et al. (Feb. 2019). "High frequency of shared clonotypes in human B cell receptor repertoires". eng. In: *Nature* 566.7744, pp. 398–402. ISSN: 1476-4687. DOI: 10.1038/s41586-019-0934-8.

Soto, Cinque, Jessica A. Finn, et al. (July 2020). "PyIR: a scalable wrapper for processing billions of immunoglobulin and T cell receptor sequences using IgBLAST". en. In: *BMC Bioinformatics* 21.1, p. 314. ISSN: 1471-2105. DOI: 10.1186/s12859-020-03649-5. URL: https://doi.org/10.1186/s12859-020-03649-5 (visited on 07/22/2021).

Spiess, Christoph, Qianting Zhai, and Paul J. Carter (Oct. 2015). "Alternative molecular formats and therapeutic applications for bispecific antibodies". en. In: *Molecular Immunology*. Therapeutic Antibodies: Discovery, Design and Deployment 67.2, Part A, pp. 95–106. ISSN: 0161-5890. DOI: 10.1016/j.molimm.2015.01.003. URL: http://www.sciencedirect.com/science/article/pii/S016158901500005X (visited on 01/13/2021).

Stein, Amelie and Tanja Kortemme (2013). "Improvements to robotics-inspired conformational sampling in rosetta". eng. In: *PLoS One* 8.5, e63090. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0063090.

Steinberger, P. et al. (Nov. 2000). "Generation and characterization of a recombinant human CCR5-specific antibody. A phage display approach for rabbit antibody humanization". eng. In: *J Biol Chem* 275.46, pp. 36073–36078. ISSN: 0021-9258. DOI: 10.1074/jbc.M002765200.

Suárez, Eduardo et al. (Apr. 2006). "Rearrangement of only one human IGHV gene is sufficient to generate a wide repertoire of antigen specific antibody responses in transgenic mice". eng. In: *Mol Immunol* 43.11, pp. 1827–1835. ISSN: 0161-5890. DOI: 10.1016/j.molimm.2005.10.015.

Süel, Gürol M. et al. (Jan. 2003). "Evolutionarily conserved networks of residues mediate allosteric communication in proteins". eng. In: *Nat Struct Biol* 10.1, pp. 59–69. ISSN: 1072-8368. DOI: 10.1038/nsb881.

Switzer, R. L. and K. J. Gibson (1978). "Phosphoribosylpyrophosphate synthetase (ribose-5-phosphate pyrophosphokinase) from Salmonella typhimurium". eng. In: *Methods Enzymol* 51, pp. 3–11. ISSN: 0076-6879. DOI: 10.1016/s0076-6879(78)51003-3.

Teng, Grace and F. Nina Papavasiliou (Dec. 2007). "Immunoglobulin Somatic Hypermutation". In: *Annu. Rev. Genet.* 41.1. Publisher: Annual Reviews, pp. 107–120. ISSN: 0066-4197. DOI: 10.1146/annurev.genet.41.110306.130340. URL: https://www.annualreviews.org/doi/10.1146/annurev.genet.41.110306.130340 (visited on 08/04/2021).

Thornburg, Natalie J. et al. (Oct. 2013). "Human antibodies that neutralize respiratory droplet transmissible H5N1 influenza viruses". eng. In: *J Clin Invest* 123.10, pp. 4405–4409. ISSN: 1558-8238. DOI: 10.1172/JCI69377.

Thullier, Philippe et al. (Mar. 2010). "The humanness of macaque antibody sequences". eng. In: *J Mol Biol* 396.5, pp. 1439–1450. ISSN: 1089-8638. DOI: 10.1016/j.jmb.2009.12.041.

Tian, Ming et al. (Sept. 2016). "Induction of HIV Neutralizing Antibody Lineages in Mice with Diverse Precursor Repertoires". eng. In: *Cell* 166.6, 1471–1484.e18. ISSN: 1097-4172. DOI: 10.1016/j.cell.2016.07.029.

Tian, Pengfei and Robert B. Best (Oct. 2017). "How Many Protein Sequences Fold to a Given Structure? A Coevolutionary Analysis". eng. In: *Biophys J* 113.8, pp. 1719–1730. ISSN: 1542-0086. DOI: 10.1016/j.bpj.2017.08.039.

Tian, Pengfei, John M. Louis, et al. (May 2018). "Co-Evolutionary Fitness Landscapes for Sequence Design". eng. In: *Angew Chem Int Ed Engl* 57.20, pp. 5674–5678. ISSN: 1521-3773. DOI: 10.1002/anie.201713220.

Tsurushita, Naoya, Paul R. Hinton, and Shankar Kumar (May 2005). "Design of humanized antibodies: from anti-Tac to Zenapax". eng. In: *Methods* 36.1, pp. 69–83. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2005.01.007.

Tsurushita, Naoya, Minha Park, et al. (Dec. 2004). "Humanization of a chicken anti-IL-12 monoclonal antibody". en. In: *Journal of Immunological Methods* 295.1, pp. 9–19. ISSN: 0022-1759. DOI: 10.1016/j.jim.2004.08.018. URL: https://www.sciencedirect.com/science/article/pii/S0022175904003035 (visited on 08/04/2021).

UniProt Consortium (Jan. 2019). "UniProt: a worldwide hub of protein knowledge". eng. In: *Nucleic Acids Res* 47.D1, pp. D506–D515. ISSN: 1362-4962. DOI: 10.1093/nar/gky1049.

Vergara, Renan et al. (Feb. 2020). "The interplay of protein-ligand and water-mediated interactions shape affinity and selectivity in the LAO binding protein". eng. In: *FEBS J* 287.4, pp. 763–782. ISSN: 1742-4658. DOI: 10.1111/febs.15019.

Wainberg, Michael et al. (Oct. 2018). "Deep learning in biomedicine". en. In: *Nat Biotechnol* 36.9, pp. 829–838. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.4233. URL: http://www.nature.com/articles/nbt.4233 (visited on 03/06/2021).

Waksman, G. et al. (Feb. 1994). "Crystal structure of Escherichia coli thioredoxin reductase refined at 2 A resolution. Implications for a large conformational change during catalysis". eng. In: *J Mol Biol* 236.3, pp. 800–816. ISSN: 0022-2836.

Waldmann, Herman (2019). "Human Monoclonal Antibodies: The Benefits of Humanization". eng. In: *Methods Mol Biol* 1904, pp. 1–10. ISSN: 1940-6029. DOI: 10.1007/978-1-4939-8958-4_1.

Wang, Chu, Philip Bradley, and David Baker (Oct. 2007). "Protein-protein docking with backbone flexibility". eng. In: *J Mol Biol* 373.2, pp. 503–519. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2007.07.050.

Warren, René L. et al. (May 2011). "Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes". eng. In: *Genome Res* 21.5, pp. 790–797. ISSN: 1549-5469. DOI: 10.1101/gr.115428.110.

Warszawski, Shira, Aliza Borenstein Katz, et al. (Aug. 2019). "Optimizing antibody affinity and stability by the automated design of the variable light-heavy chain interfaces". en. In: *PLOS Computational Biology* 15.8. Publisher: Public Library of Science, e1007207. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007207. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007207 (visited on 12/09/2020).

Warszawski, Shira, Ravit Netzer, et al. (Dec. 2014). "A "fuzzy"-logic language for encoding multiple physical traits in biomolecules". eng. In: *J Mol Biol* 426.24, pp. 4125–4138. ISSN: 1089-8638. DOI: 10.1016/j.jmb.2014.10.002.

Weitzner, Brian D., Roland L. Dunbrack, and Jeffrey J. Gray (Feb. 2015). "The origin of CDR H3 structural diversity". eng. In: *Structure* 23.2, pp. 302–311. ISSN: 1878-4186. DOI: 10.1016/j.str.2014.11.010.

Weitzner, Brian D. and Jeffrey J. Gray (Jan. 2017). "Accurate Structure Prediction of CDR H3 Loops Enabled by a Novel Structure-Based C-Terminal Constraint". eng. In: *J Immunol* 198.1, pp. 505–515. ISSN: 1550-6606. DOI: 10.4049/jimmunol.1601137.

Weitzner, Brian D., Jeliazko R. Jeliazkov, et al. (Feb. 2017). "Modeling and docking of antibody structures with Rosetta". eng. In: *Nat Protoc* 12.2, pp. 401–416. ISSN: 1750-2799. DOI: 10.1038/nprot.2016.180.

Weitzner, Brian D., Daisuke Kuroda, et al. (Aug. 2014). "Blind prediction performance of RosettaAntibody 3.0: grafting, relaxation, kinematic loop modeling, and full CDR optimization". eng. In: *Proteins* 82.8, pp. 1611–1623. ISSN: 1097-0134. DOI: 10.1002/prot.24534.

Whitehead, Timothy A. et al. (May 2012). "Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing". eng. In: *Nat Biotechnol* 30.6, pp. 543–548. ISSN: 1546-1696. DOI: 10.1038/nbt.2214.

Willis, Jordan R., Bryan S. Briney, et al. (Apr. 2013). "Human germline antibody gene segments encode polyspecific antibodies". eng. In: *PLoS Comput Biol* 9.4, e1003045. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003045.

Willis, Jordan R., Gopal Sapparapu, et al. (June 2015). "Redesigned HIV antibodies exhibit enhanced neutralizing potency and breadth". eng. In: *J Clin Invest* 125.6, pp. 2523–2531. ISSN: 1558-8238. DOI: 10.1172/JCI80693.

Winkler, K. et al. (Oct. 2000). "Changing the antigen binding specificity by single point mutations of an anti-p24 (HIV-1) antibody". eng. In: *J Immunol* 165.8, pp. 4505–4514. ISSN: 0022-1767. DOI: 10.4049/jimmunol.165.8.4505.

Wojcikiewicz, Richard J. H. and Su Ge Luo (Apr. 1998). "Differences Among Type I, II, and III Inositol-1,4,5-Trisphosphate Receptors in Ligand-Binding Affinity Influence the Sensitivity of Calcium Stores to Inositol-1,4,5-Trisphosphate". en. In: *Molecular Pharmacology* 53.4, pp. 656–662. ISSN: 0026-895X, 1521-0111. DOI: 10.1124/mol.53.4.656. URL: http://molpharm.aspetjournals.org/lookup/doi/10.1124/mol.53.4.656 (visited on 02/12/2019).

Wollacott, Andrew M et al. (Dec. 2019). "Quantifying the nativeness of antibody sequences using long short-term memory networks". In: *Protein Engineering, Design and Selection* 32.7, pp. 347–354. ISSN: 1741-0126. DOI: 10.1093/protein/gzz031. URL: https://doi.org/10.1093/protein/gzz031 (visited on 03/06/2021).

Wollenberg, K. R. and W. R. Atchley (Mar. 2000). "Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap". eng. In: *Proc Natl Acad Sci U S A* 97.7, pp. 3288–3291. ISSN: 0027-8424. DOI: 10.1073/pnas.070154797.

Wooden, Stacey L. and Wayne C. Koff (Sept. 2018). "The Human Vaccines Project: Towards a comprehensive understanding of the human immune response to immunization". In: *Human Vaccines & Immunotherapeutics* 14.9. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/21645515.2018.1

pp. 2214–2216. ISSN: 2164-5515. DOI: 10.1080/21645515.2018.1476813. URL: https://doi.org/10.1080/21645515.2018.1476813 (visited on 06/18/2021).

Wu, Xudong and Tom A. Rapoport (Aug. 2018). "Mechanistic insights into ER-associated protein degradation". eng. In: *Curr Opin Cell Biol* 53, pp. 22–28. ISSN: 1879-0410. DOI: 10.1016/j.ceb.2018.04.004.

Wu, Xueling, Zhenhai Zhang, et al. (Apr. 2015). "Maturation and Diversity of the VRC01-Antibody Lineage over 15 Years of Chronic HIV-1 Infection". eng. In: *Cell* 161.3, pp. 470–485. ISSN: 1097-4172. DOI: 10.1016/j.cell.2015.03.004.

Wu, Xueling, Tongqing Zhou, et al. (Sept. 2011). "Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing". eng. In: *Science* 333.6049, pp. 1593–1602. ISSN: 1095-9203. DOI: 10.1126/science.1207532.

Xiao, B. et al. (May 1999). "Crystal structure of 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase, a potential target for the development of novel antimicrobial agents". eng. In: *Structure* 7.5, pp. 489–496. ISSN: 0969-2126. DOI: 10.1016/s0969-2126(99)80065-3.

Xu, Yingda et al. (Oct. 2013). "Addressing polyspecificity of antibodies selected from an in vitro yeast presentation system: a FACS-based, high-throughput selection and analytical tool". In: *Protein Engineering, Design and Selection* 26.10, pp. 663–670. ISSN: 1741-0126. DOI: 10.1093/protein/gzt047. URL: https://doi.org/10.1093/protein/gzt047 (visited on 01/14/2021).

Yang, Zheng Rong (Dec. 2004). "Biological applications of support vector machines". eng. In: *Brief Bioinform* 5.4, pp. 328–338. ISSN: 1467-5463. DOI: 10.1093/bib/5.4.328.

Yao, N. et al. (Feb. 1996). "Modulation of a salt link does not affect binding of phosphate to its specific active transport receptor". eng. In: *Biochemistry* 35.7, pp. 2079–2085. ISSN: 0006-2960. DOI: 10.1021/bi952686r.

Ye, Jian et al. (July 2013). "IgBLAST: an immunoglobulin variable domain sequence analysis tool". eng. In: *Nucleic Acids Res* 41.Web Server issue, W34–40. ISSN: 1362-4962. DOI: 10.1093/nar/gkt382.

Zhou, Zhipeng et al. (Oct. 2016). "Codon usage is an important determinant of gene expression levels largely through its effects on transcription". In: *Proc Natl Acad Sci U S A* 113.41, E6117–E6125. ISSN: 0027-8424. DOI: 10.1073/pnas.1606724113. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5068308/ (visited on 10/01/2021).

Zimmermann, Lukas et al. (July 2018). "A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core". eng. In: *J Mol Biol* 430.15, pp. 2237–2243. ISSN: 1089-8638. DOI: 10.1016/j.jmb.2017.12.007.

Zost, Seth J. et al. (Aug. 2021). "Canonical features of human antibodies recognizing the influenza hemagglutinin trimer interface". eng. In: *J Clin Invest* 131.15, p. 146791. ISSN: 1558-8238. DOI: 10.1172/JCI146791.