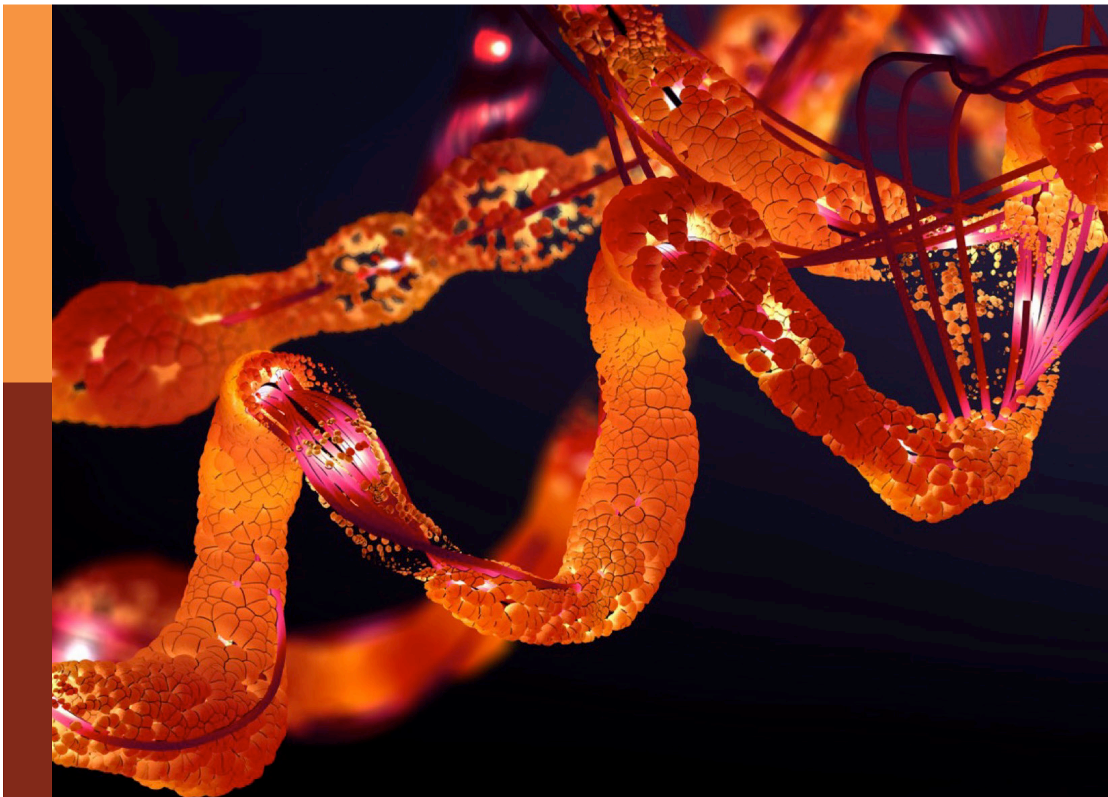
frontiers | Research Topics

Progress and challenges in computational structure-based design and development of biologic drugs

Edited by
Traian Sulea, Sandeep Kumar and Daisuke Kuroda

Published in
Frontiers in Molecular Biosciences



ISSN 1664-8714
ISBN 978-2-8325-4484-6
DOI 10.3389/978-2-8325-4484-6

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Progress and challenges in computational structure-based design and development of biologic drugs

Topic editors
Traian Sulea — Human Health Therapeutics Research Centre, National Research Council Canada (NRC), Canada
Sandeep Kumar — Boehringer Ingelheim, United States
Daisuke Kuroda — National Institute of Infectious Diseases (NIID), Japan

Citation
Sulea, T., Kumar, S., Kuroda, D., eds. (2024). *Progress and challenges in computational structure-based design and development of biologic drugs*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4484-6

Dr. Sandeep Kumar is an employee of Boehringer Ingelheim and holds patents related to this Research Topic.

Table of contents

Frontiers in Molecular Biosciences

TYPE Editorial
PUBLISHED 07 February 2024
DOI 10.3389/fmolb.2024.1360267

OPEN ACCESS

EDITED AND REVIEWED BY
Didier Devaurs,
University of Edinburgh, United Kingdom

*CORRESPONDENCE
Traian Sulea,
✉ traian.sulea@nrc-cnrc.gc.ca
Sandeep Kumar,
✉ sandeep.kumar@modernatx.com
Daisuke Kuroda,
✉ dkuroda@niid.go.jp

RECEIVED 22 December 2023
ACCEPTED 01 February 2024
PUBLISHED 07 February 2024

CITATION
Sulea T, Kumar S and Kuroda D (2024), Editorial: Progress and challenges in computational structure-based design and development of biologic drugs.
Front. Mol. Biosci. 11:1360267.
doi: 10.3389/fmolb.2024.1360267

Editorial: Progress and challenges in computational structure-based design and development of biologic drugs

Traian Sulea[1]*, Sandeep Kumar[2]* and Daisuke Kuroda[3]*

[1]Human Health Therapeutics Research Centre, National Research Council Canada, Montreal, QC, Canada, [2]Computational Protein Design and Modeling, Computational Science, Moderna Therapeutics, Cambridge, MA, United States, [3]Research Center of Drug and Vaccine Development, National Institute of Infectious Diseases, Tokyo, Japan

KEYWORDS

binding affinity, specificity, therapeutic antibody, developability, in silico method, molecular modeling, molecular simulations, machine learning

Editorial on the Research Topic
Progress and challenges in computational structure-based design and development of biologic drugs

Biotherapeutics have emerged as a major class of pharmaceuticals, encompassing monoclonal antibodies, recombinant human proteins and enzymes, fusion proteins, antibody drug conjugates, multi-specific formats, peptides, and vaccines. These modalities serve a wide range of therapeutic areas, including immune-oncology, inflammation, cardiovascular, metabolic, infectious, and rare diseases (DeFrancesco, 2019; Kang and Jung, 2020; Lu et al., 2020; Kaplon et al., 2023). Recent advancements in structure determination, structure prediction, bioanalytical characterization, and machine learning have established *in silico* approaches as a key toolbox employed in the biologic drug discovery and development pipelines (Fischman and Ofran, 2018; Norman et al., 2020; Fernandez-Quintero et al., 2023). Additionally, physics-based molecular modeling and simulation methods, along with empirical linear models, have matured to routine implementation during biotherapeutic drug candidate selection and optimization. However, the accuracy of these predictions can be improved. Further refinements will be welcomed, particularly towards binding affinity predictions and developability assessments.

At the same time, with the fast-paced infusion of artificial intelligence in various research areas that impact daily life, we are witnessing a new chapter being written in biological drug design (Kim et al., 2023). A wide range of nonlinear models, from machine learning to unsupervised deep neural networks and language models, is now emerging. These models, fueled by still modest but expanding biological and structural datasets, are complementing classical methods (Baek et al., 2021; Jumper et al., 2021; Kryshtafovych et al., 2021; Bennett et al., 2023; Lin et al., 2023; Wodak et al., 2023). There are also increasing attempts to integrate advanced computational methods with next-generation sequencing, either from synthetic or natural libraries, to enhance the efficient hit-to-lead optimization and *de novo* discovery of tight binders with favorable developability profiles (Makowski et al., 2021; Mason et al., 2021; Hie et al., 2023; Parkinson et al., 2023). While past structure-based efforts were focused on optimizing existing antibody candidates, the emerging trend is the hopeful possibility of *de novo* discovery of antibody-based and other biologic drugs via *in silico* methods. This opens prospects for extensive application of computational techniques in biologic drug discovery and development.

This Research Topic comes at an opportune time when the accuracy limits are being pushed for classical methods and foundations are being established for machine learning methods. These complementary tools are poised to enhance the entire biologics drug development pipeline, from the molecular design and property optimization to large-scale manufacturability. In this Research Topic, readers will find a breadth of computational approaches, ranging from established 3D structure and physics-based techniques to innovative explorations in sequence-based non-linear machine learning models.

An overview of the current state and opportunities for synergistic use of computation and experimentation in this field is provided by Bauer et al. The authors described their vision of Biopharmaceutical Informatics and discuss already available computational methods at each stage of the biologic drug design-discovery-optimization-development pipeline. The authors have provided useful cues on how best to apply these *in silico* methods and how to combine them with experimental approaches to maximize the odds and efficiency of arriving at biologics that are both effective and developable.

Fundamental understandings of molecular properties of the drug candidates and their targets are essential to advance both biologic discovery and development. Di Rienzo et al. focused on discerning the rules that define antibody-antigen recognition as a fundamental step in the rational design and engineering of functional antibodies with desired properties. Their novel method, which is based on the 3D Zernike polynomials to generate shape and electrostatic descriptors capturing both global and local protein surface physicochemical properties, accurately classified types of antibody-antigen interfaces solely based on paratope surface information. Fernandez-Quintero et al. took a deep dive into seemingly similar interfaces between the various Ig-folded domains that make up a monoclonal antibody structure. Using classical MD simulations and analyses, they revealed and compared contact maps that can be used to inform selection of favorable point mutations for the design of bispecific antibodies. In their case study, Paul et al. described the well-recognized yet inadequately understood trade-off between binding affinity and thermal stability, which can have significant implications during the lead candidate optimization stage. Using classical force-field methods, molecular dynamics, and amino-acid hydropathy, they observed affinity-stability correlations and patterns in key pairs of residues called hotspots.

Novel tools are also reviewed in this Research Topic. For example, Jaszczyszyn et al. assembled a timely review of recent advances of deep-learning based tools for structural modeling the variable regions of antibodies. In addition to cataloguing underlying algorithms and benchmarking their performance, the authors offered their perspective on how the emerging high accuracy of antibody paratope modeling can influence the field of biologics drug discovery. Engelberger et al. provided the energy breakdown guided protein design (ENDURE) tool for accurately assessing energetic contributions from individual and combinatorial mutations to the overall protein stability. An interesting feature is the residue depth analysis which enables tracking the energetic contributions of mutations occurring in different spatial layers of the protein structure. Spoendlin et al. introduced the second iteration of their structural profiling of antibodies to cluster by epitope (SPACE) tool, which builds upon the recent progress in machine learning antibody structure prediction and a novel clustering protocol. It improved data coverage and identified even more diverse clusters of antibodies that bind to the same epitope. These tools are expected to further advance rational design of biotherapeutics.

Proof-of-concept studies illustrating novel screening campaigns that combine computational design with experimental data are also presented. Arras et al. combined next-generation sequencing of semi-immune/semi-synthetic libraries built on a humanized VHH framework with machine learning, data processing, and model building for simultaneous optimization of affinity and developability. The proposed typical early drug discovery methodology generated diverse and potent VHH hits against NKp46 protein without requiring further humanization and developability optimization, thereby accelerating drug discovery. Gaudreault et al. focused on protein-protein docking with flexible side chains while retaining rigid protein backbone to discover novel binders against predefined target epitopes. Their approach was applied to randomized libraries of surface mutations introduced in a rigid protein scaffold called DARPin, leading to the design and experimental validation of an enriched small set of hits against a predefined epitope on the BCL-W target protein.

The literature in this field is growing rapidly. Our Research Topic does not cover all computational aspects of biologic drug discovery. Nonetheless, the articles compiled here hopefully offer timely snapshots of key components along a biologic drug's discovery, design, and development.

Author contributions

TS: Writing–original draft, Writing–review and editing. SK: Writing–original draft, Writing–review and editing. DK: Writing–original draft, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. DK acknowledges support from Japan Agency for Medical Research and Development (JP23wm0325047).

Acknowledgments

The editors would like to thank all the authors and expert reviewers who have participated in the preparation and evaluation of manuscripts presented in this Research Topic.

Conflict of interest

SK is an employee of Moderna Therapeutics.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373, 871–876. doi:10.1126/science.abj8754

Bennett, N. R., Coventry, B., Goreshnik, I., Huang, B., Allen, A., Vafeados, D., et al. (2023). Improving *de novo* protein binder design with deep learning. *Nat. Commun.* 14, 2625. doi:10.1038/s41467-023-38328-5

DeFrancesco, L. (2019). Drug pipeline 1Q19. *Nat. Biotechnol.* 37, 579–580. doi:10.1038/s41587-019-0146-7

Fernandez-Quintero, M. L., Ljungars, A., Waibl, F., Greiff, V., Andersen, J. T., Gjolberg, T. T., et al. (2023). Assessing developability early in the discovery process for novel biologics. *MAbs* 15, 2171248. doi:10.1080/19420862.2023.2171248

Fischman, S., and Ofran, Y. (2018). Computational design of antibodies. *Curr. Opin. Struct. Biol.* 51, 156–162. doi:10.1016/j.sbi.2018.04.007

Hie, B. L., Shanker, V. R., Xu, D., Bruun, T. U. J., Weidenbacher, P. A., Tang, S., et al. (2023). Efficient evolution of human antibodies from general protein language models. *Nat. Biotechnol.* 2023, 1763. doi:10.1038/s41587-023-01763-2

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2

Kang, T. H., and Jung, S. T. (2020). Reprogramming the constant region of immunoglobulin G subclasses for enhanced therapeutic potency against cancer. *Biomolecules* 10, 382. doi:10.3390/biom10030382

Kaplon, H., Crescioli, S., Chenoweth, A., Visweswaraiah, J., and Reichert, J. M. (2023). Antibodies to watch in 2023. *MAbs* 15, 2153410. doi:10.1080/19420862.2022.2153410

Kim, J., McFee, M., Fang, Q., Abdin, O., and Kim, P. M. (2023). Computational and artificial intelligence-based methods for antibody development. *Trends Pharmacol. Sci.* 44, 175–189. doi:10.1016/j.tips.2022.12.005

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2021). Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins* 89, 1607–1617. doi:10.1002/prot.26237

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379, 1123–1130. doi:10.1126/science.ade2574

Lu, R. M., Hwang, Y. C., Liu, I. J., Lee, C. C., Tsai, H. Z., Li, H. J., et al. (2020). Development of therapeutic antibodies for the treatment of diseases. *J. Biomed. Sci.* 27, 1. doi:10.1186/s12929-019-0592-z

Makowski, E. K., Wu, L., Gupta, P., and Tessier, P. M. (2021). Discovery-stage identification of drug-like antibodies using emerging experimental and computational methods. *MAbs* 13, 1895540. doi:10.1080/19420862.2021.1895540

Mason, D. M., Friedensohn, S., Weber, C. R., Jordi, C., Wagner, B., Meng, S. M., et al. (2021). Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat. Biomed. Eng.* 5, 600–612. doi:10.1038/s41551-021-00699-9

Norman, R. A., Ambrosetti, F., Bonvin, A., Colwell, L. J., Kelm, S., Kumar, S., et al. (2020). Computational approaches to therapeutic antibody design: established methods and emerging trends. *Brief. Bioinform.* 21, 1549–1567. doi:10.1093/bib/bbz095

Parkinson, J., Hard, R., and Wang, W. (2023). The RESP AI model accelerates the identification of tight-binding antibodies. *Nat. Commun.* 14, 454. doi:10.1038/s41467-023-36028-8

Wodak, S. J., Vajda, S., Lensink, M. F., Kozakov, D., and Bates, P. A. (2023). Critical assessment of methods for predicting the 3D structure of proteins and protein complexes. *Annu. Rev. Biophys.* 52, 183–206. doi:10.1146/annurev-biophys-102622-084607

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 30 September 2021
doi: 10.3389/fmolb.2021.749784

Quantitative Description of Surface Complementarity of Antibody-Antigen Interfaces

Lorenzo Di Rienzo[1], *Edoardo Milanetti*[1,2], *Giancarlo Ruocco*[1,2] *and Rosalba Lepore*[3]*

[1]Center for Life Nano and Neuro-Science, Istituto Italiano di Tecnologia, Rome, Italy, [2]Department of Physics, Sapienza University, Rome, Italy, [3]Department of Biomedicine, Basel University Hospital and University of Basel, Basel, Switzerland

Antibodies have the remarkable ability to recognise their cognate antigens with extraordinary affinity and specificity. Discerning the rules that define antibody-antigen recognition is a fundamental step in the rational design and engineering of functional antibodies with desired properties. In this study we apply the 3D Zernike formalism to the analysis of the surface properties of the antibody complementary determining regions (CDRs). Our results show that shape and electrostatic 3DZD descriptors of the surface of the CDRs are predictive of antigen specificity, with classification accuracy of 81% and area under the receiver operating characteristic curve (AUC) of 0.85. Additionally, while in terms of surface size, solvent accessibility and amino acid composition, antibody epitopes are typically not distinguishable from non-epitope, solvent-exposed regions of the antigen, the 3DZD descriptors detect significantly higher surface complementarity to the paratope, and are able to predict correct paratope-epitope interaction with an AUC = 0.75.

Keywords: surface complementarity, antibody complementarity determining regions, antibody—antigen complex, antigen recognition, zernike polynomials

OPEN ACCESS

Edited by:
Yong Wang,
Zhejiang University, China

Reviewed by:
Chun Chan,
Zhejiang University, China
Jing Huang,
Westlake University, China

****Correspondence:***
Rosalba Lepore
rosalba.lepore@unibas.ch

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *29 July 2021*
Accepted: *14 September 2021*
Published: *30 September 2021*

Citation:
Di Rienzo L, Milanetti E, Ruocco G and Lepore R (2021) Quantitative Description of Surface Complementarity of Antibody-Antigen Interfaces. Front. Mol. Biosci. 8:749784. doi: 10.3389/fmolb.2021.749784

1 INTRODUCTION

Antibodies, also known as immunoglobulins, are multimeric Y-shaped proteins that the immune system uses to recognize and neutralize foreign targets, named antigens. The antigen binding site is located on the upper tip of the molecule, and is formed by the pairing of two variable domains, the VH and the VL, each contributing three hypervariable loops or complementary determining regions (CDR). The remarkable ability of the antibodies to recognize virtually any foreign antigen stems from the sequence and length variability of the CDR, while the framework of the molecule is largely conserved (Chothia and Lesk, 1987; Chothia et al., 1989; Tramontano et al., 1990).

Early studies, based on a handful of crystallographic structures, revealed that despite the large sequence variability of CDRs, five out of the six hypervariable loops only exhibit a limited number of main-chain conformations called "canonical structures" (Chothia and Lesk, 1987; Chothia et al., 1989), where most sequence variations only modify the surface generated by the side chains on a canonical main-chain structure. Over the years, with more experimentally determined structures of antibodies becoming available, an exhaustive repertoire of canonical structures has been compiled and their relationship with the chain isotypes (Tramontano et al., 1990; Chothia et al., 1992; Foote and Winter, 1992; Tomlinson et al., 1995; Martin and Thornton, 1996; Chothia et al., 1998; Decanniere et al., 2000; Vargas-Madrazo and Paz-García, 2002; Chailyan et al., 2011; North et al., 2011; Kuroda and Gray, 2016) and packing mode of the antibody was extensively analysed (Chothia et al., 1985; De Wildt et al., 1999; Abhinandan and Martin, 2010; Jayaram et al., 2012; Dunbar et al., 2013a). This led to the development of fully automated pipelines for the prediction of immunoglobulin structures given their amino acid sequences, with predictions reaching near-native accuracy both at the global and local CDR level (Whitelegg and Rees, 2000; Marcatili et al., 2014; Messih et al., 2014; Dunbar et al., 2016; Lepore et al., 2017; Weitzner et al., 2017). In parallel, a major focus has been in understanding the structural and molecular basis of antibody function and, in particular, of antigen recognition. The identification of the portion of the antigen that is recognized by an antibody, i.e. the epitope, is indeed of central relevance for the development of vaccines and immunodiagnostics, as well as for our understanding of protective immunity (Pollard and Bijker, 2020). As a consequence, in the past years, there have been several attempts in the direction of relating the sequence and structural properties of antibody binding sites to their function, and more specifically, to the type of recognised antigen. Early work by Webster et al. in 1994 first discovered a strong correlation between the topography of the CDRs and the broad nature of the antigen, proposing that antibodies binding protein antigens are characterised by flat combining sites, while those recognising smaller antigens, like haptens and peptides, show the most concave interfaces (Webster et al., 1994). Subsequent work confirmed and extended these findings to the length and sequence composition of the CDRs based on increased availability of sequence and structural data of antibody-antigen complexes (MacCallum et al., 1996; Collis et al., 2003; Lee et al., 2006; Raghunathan et al., 2012).

The study of molecular interactions in proteins, and antibodies in particular, poses well known challenges. Existing experimental methods, such as Xray crystallography, mass spectrometry, phage display and mutagenesis analysis are intrinsically expensive, laborious, and time consuming (Sela-Culang et al., 2013). Hence, computational methods have established themselves as a valuable complement to experimental biology efforts for the analysis and characterization of the vast repertoire of molecular interactions at the atomic level. Early studies by Lee and Richards (1971) proposed the first description of protein solvent-accessible surface, which was later refined by Connolly (1983), allowing to distinguish surface atoms from buried atoms and opening the way to efficient graphical representation and comparison of molecular surface properties. Subsequent methods relied on the application of spherical harmonics descriptors (Leicester et al., 1988; Max and Getzoff, 1988) and Fourier correlation theory to shape complementarity and electrostatic interaction analysis (Gabb et al., 1997). Additionally, approaches based on tessellation (Walls and Sternberg, 1992; Li et al., 2007), void volume (Jones and Thornton, 1996) and surface density (Norel et al., 1995) provided an efficient way for representation and matching of protein surfaces, including protein-protein interaction sites, ligand binding sites and functional sites (Via et al., 2000; Mitra and Pal, 2010).

In this study we rely on a surface representation of antibodies and their cognate antigens based on the 3D Zernike Descriptors (3DZD). The Zernike polynomials were first described by Fritz Zernike in 1934 (Zernike and Stratton, 1934) as a framework for the analysis of aberrations in optical systems and subsequently generalized to three-dimensions (Ming-Kuei Hu, 1962; Canterakis, 1999; Novotni and Klein, 2004). One of the convenient features of Zernike polynomials is that their rotational symmetry allows the polynomials to be expressed as products of radial terms and functions of angle, where the coordinate system can be rotated without changing the form of the polynomial. Hence, they allow a concise, roto-translationally invariant characterization of 3D objects, comparing favourably to other moment-based descriptors in terms of shape retrieval and robustness to topological and geometrical artifacts (Novotni and Klein, 2004). When applied to molecular surfaces, the 3DZD have been shown to capture both global and local protein surface properties and to adequately represent their physico-chemical properties (Venkatraman et al., 2009a; Venkatraman et al., 2009b; Kihara et al., 2011; Di Rienzo et al., 2017; Daberdaku and Ferrari, 2018; Daberdaku and Ferrari, 2019; Alba et al., 2020; Di Rienzo et al., 2020). Here we apply the 3DZD to provide a quantitative description of the shape and electrostatic properties of Ab–Ag interfaces, leading to an accurate classification of the antibodies according to the type of their cognate antigens solely based on the information of the CDR surface, with overall AUC = 0.85 and accuracy of 81%.

Additionally, we show that while in terms of surface size, solvent accessibility and amino acid composition, antibody epitopes are not distinguishable from non-epitope, solvent-exposed regions of the antigen, they display significantly higher surface complementarity to the antibody paratope, both in terms of shape and electrostatic 3DZD, leading to a prediction performance in terms of ROC AUC of 0.75 and 0.61 respectively.

2 MATERIALS AND METHODS

2.1 Dataset

We selected 326 antibodies with redundancy lower than 90% and resolution <3.0 Å using the SabDab database (Dunbar et al., 2013b). 229 antibodies were solved in complex with protein antigens, 71 with haptens, 19 with carbohydrates and 7 with nucleic acids. The sequence of each antibody was renumbered according to the Chothia numbering scheme (Chothia and Lesk, 1987; Chothia et al., 1989) using an in-house python script.

2.2 Solvent Accessible Surface and Electrostatics Surface

For each antibody and protein antigen 3D structure, atomic partial charges and radii were assigned using PDB2PQR with default parameters (Dolinsky et al., 2004). Solvent Accessible Surface (SAS) was computed using GROMACS (Abraham et al., 2015). Electrostatic surface (ES) potential was computed using the Bluues software (options -srf and -srfpot) (Fogolari et al., 2012). Each molecular surface point was assigned to the electrostatic potential of the corresponding residue. The “geometry” (Habel et al., 2019) and “Bio3D” (Grant et al., 2006) packages available in R were used for PDB structure processing and analysis.

2.3 Voxelization Procedure

The set of selected molecular surface points was scaled to the unit sphere and placed into a 3D grid of dimension 128^3. To avoid boundary effects, the size of the bounding box of the point cloud was set so as to be contained within 80% of the unit sphere (Grandison et al., 2009). Voxelization was performed separately for SAS and ES. In SAS voxelization, each voxel was assigned a value of 1 if the center of the voxel was closer than 1.7 to any SAS point, 0 otherwise. In ES voxelization, each voxel was assigned the mean ES value of the enclosed points, 0 otherwise.

Since the Zernike formalism does not differentiate positive and negative values (Chikhi et al., 2010; Daberdaku and Ferrari, 2018), but only patterns of non-zero values in the 3D space, voxels were initialized for positive and negative patterns separately using a similar approach as done in (Chikhi et al., 2010), as follows:

$$f^{+}_{elec} = 0 \quad if \quad f_{elec} < 0 \qquad f^{+}_{elec} = f_{elec} \quad if \quad f_{elec} > 0 \quad (1)$$

$$f^{-}_{elec} = f_{elec} \quad if \quad f_{elec} < 0 \qquad f^{-}_{elec} = 0 \quad if \quad f_{elec} > 0 \quad (2)$$

In summary, voxels with positive electrostatics values were initialized to 1 and all other voxels with negative electrostatics values were set to zero, and vice versa. The resulting voxels, one for SAS values, and two for positive and negative ES values, respectively, were considered as three different 3D functions, f(x), each expanded into the 3DZD as described in the next section.

2.4 3D Zernike Descriptors

For the quantitative description of the binding sites, we rely on a representation based on the Zernike polynomials and their corresponding moments. Moment-based representations are a class of mathematical descriptors of shape, originally developed for pattern recognition and subsequently generalized to three-dimensions (Ming-Kuei Hu, 1962; Canterakis, 1999; Novotni and Klein, 2004).

A surface described by a function $f(r, \theta, \phi)$ in polar coordinates can be represented by a series expansion in an orthonormal sequence of polynomials (Canterakis, 1999):

$$f(r, \theta, \phi) = \sum_{n=0}^{\infty} \sum_{l=0}^{n} \sum_{m=-l}^{l} C_{nlm} Z^{m}_{nl}(r, \theta, \phi) \quad (3)$$

where the indices n, m and l are the order, degree and repetition, respectively.

The Zernike polynomials can be written as:

$$Z^{m}_{nl}(r, \theta, \phi) = R_{nl}(r) Y^{m}_{l}(\theta, \phi) \quad (4)$$

where the Y functions are complex spherical harmonics depending on both θ and ϕ while R only depends on the radius r, which is given by

$$R_{nl}(r) = \sum_{k=0}^{\frac{(n-l)}{2}} N_{nlk} r^{n-2k} \quad (5)$$

where N is a normalization factor.

The 3D Zernike moments of a surface described by a function $f(r, \theta, \phi)$ are defined as the coefficients of the expansion of f(r) in the Zernike polynomial basis, i.e.:

$$C_{nlm} = \int_{|r| \le 1} f(\mathbf{r}) \overline{Z^{m}_{nl}(r, \theta, \phi)} d\mathbf{r} \quad (6)$$

where $\bar{Z}$ is the polynomial complex conjugate.

Their rotation invariant norms, i.e. the 3DZD, are defined as:

$$D_{nl} = \|C_{nlm}\| = \sqrt{\sum_{m=-l}^{l} (C_{nlm})^2}. \quad (7)$$

The Zernike formalism can be as detailed as desired by modulating the order of the expansion n. In our implementation, the function f represents the geometric or the (positive or negative) electrostatic potential of the molecular surface, and the maximum order of expansion was set to 20, giving a total of 121 invariants.

2.5 Generation of Native Epitopes and Surface Decoys

Given the dataset of Antibody-Antigen complexes containing protein antigens, the native geometric epitope was defined as the set of residues of the antigen having a distance lower than 6 Å to any residue of the antibody. The pivot residue was defined as the residue with the lowest mean distance to any residue of the native geometric epitope. The native electrostatic epitope was defined as the set of residues of the antigen having a distance shorter than 15 Å to any residue of the antibody. For the set of native geometric epitope residues, the Solvent Accessible Surface Area (SASA) was computed using GROMACS. The mean and standard deviation values of the computed global and residue-based SASA were used to generate an alternative set of surface patches, i.e. decoy epitopes. The algorithm first selects a decoy pivot residue, i.e. by randomly selecting any solvent exposed residue having a value of SASA within half standard deviation of the mean SASA value measured over all pivot residues of the native epitopes, i.e. $SASA = 0.48 \pm 0.33\,\mathrm{nm}^2$ (**Supplementary Figure S1**). The algorithm proceeds by adding neighboring solvent accessible residues, i.e. having relative SASA >0.2 (Tien et al., 2013), until the decoy geometric epitope reaches a similar global SASA to that of the native epitope. To ensures continuous coverage of the antigen protein surface (**Supplementary Figure S2**) and diversity of the generated patches, a maximum 50% surface patch overlap was allowed between native and decoy epitopes. Electrostatic decoy epitopes were defined by calculating the electrostatic potential over the region defined by a geometric decoy epitope considering all the charged residues within 15 $\mathring{A}$ to the pivot residue.

2.6 Comparison of the 3DZD Descriptors

Given a pair of ordered set of 3DZD, x and y, their cosine distance is measured as:

$$D(x, y) = 1 - S_c(x, y) = 1 - \frac{xy}{\|x\| \; \|y\|} \quad (8)$$

where $S_c(x, y)$ is the cosine similarity as measured by the "proxy" R package (Meyer and Buchta, 2019).

Given two patches A and B, the similarity between their 3DZD is computed as:

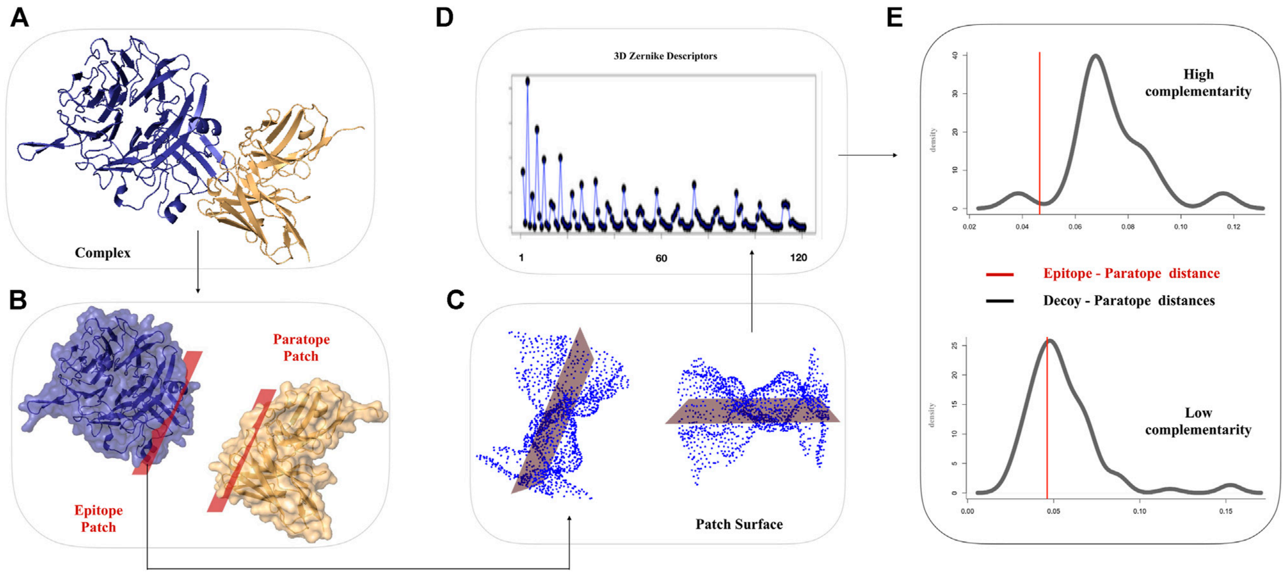


FIGURE 1 | Schematic workflow for the comparison of Ab-Ag interfaces based on 3DZD. **(A)** Molecular representation of a given Ab-Ag complex. Antibody and antigen are shown in gold and blue, respectively. **(B)** The interacting surfaces are selected according to inter-molecular atomic distance threshold. **(C)** Solvent accessible and electrostatic surfaces are computed on the selected regions **(D)** 3DZD Zernike descriptors are computed for each molecular surface. **(E)** Distribution of 3DZD surface complementary complementarity between paratope and non-epitope surface decoys. The red line denotes 3DZD surface complementarity between the antibody paratope and their cognate epitope.

$$[A-B]_{shape} = D(X^{A}_{shape}, X^{B}_{shape}) \tag{9}$$

$$[A-B]_{elec} = \frac{(D(X^{+,A}_{elec}, X^{+,B}_{elec}) + D(X^{-,A}_{elec}, X^{-,B}_{elec}))}{2} \tag{10}$$

where X_{shape}, X^{+}_{elec} and X^{-}_{elec} are, respectively, the shape, the electrostatic positive potential, and the electrostatic negative potential 3DZD.

The surface complementarity between A and B is defined as follows:

$$[A-B]_{shape} = D(X^{A}_{shape}, X^{B}_{shape}) \tag{11}$$

$$[A-B]_{elec} = \frac{(D(X^{+,A}_{elec}, X^{-,B}_{elec}) + D(X^{-,A}_{elec}, X^{+,B}_{elec}))}{2} \tag{12}$$

3 RESULTS

In this work we aim at providing a quantitative description of the geometric and electrostatic properties of antibody-antigen interaction through a mathematical representation of the interacting surfaces. To this aim, we rely on a dataset of experimentally determined 3D structures of antibody-antigen complexes and a moment-based representation of the interacting surface using the 3D Zernike descriptors (3DZD) (Novotni and Klein, 2004; Venkatraman et al., 2009b; Daberdaku and Ferrari, 2018).

The 3DZD descriptors provide a compact, roto-translationally invariant representation of 3D objects, thus enabling effective comparison of both global and local properties of molecular surfaces by standard pairwise similarity metrics. The order n of the series expansion determines the resolution of the descriptor. In this study, 3DZD were computed at different levels of truncation of the expansion, with n ranging from 10 to 20, which correspond to vectors of 36 and 121 invariants, respectively. The overall scheme of the procedure used in this work is shown in **Figure 1**.

3.1 Antibody Classification Based on Surface Shape and Electrostatic 3DZD Descriptors of CDRs

We have previously shown that a 3DZD-based description of the surface of the antibody CDRs provides an effective metric for antibody classification according to their specificity towards protein and non-protein antigens (Di Rienzo et al., 2017). Here we extend this approach to the analysis of both the shape and electrostatic properties of the CDRs and analyze the classification performance of both descriptors at different orders of the Zernike expansion. For each CDR we generated two sets of 121-dimensional vectors, representing the 3DZD of the shape and the electrostatic surface, similar to what done in (Chikhi et al., 2010; Di Rienzo et al., 2020). The similarity between each set of descriptors is then computed to perform an all-against-all comparison of CDRs, according to **Eq. 9**, **10** in Methods section. For each CDR, we then selected the nearest neighbors set as the 5% most similar CDRs in terms of shape and electrostatic surface and analyse the number of protein binding antibodies (N_{pb}) in the neighbours set. As it is shown in **Figures 2A,B**, protein-binding antibodies (green curve) are typically characterized by an higher number of N_{pb} ($mean(N^{shape}_{pb}) = 13.37 \pm 2.61$, $mean(N^{elec}_{pb}) = 13.54 \pm 3.24$) in

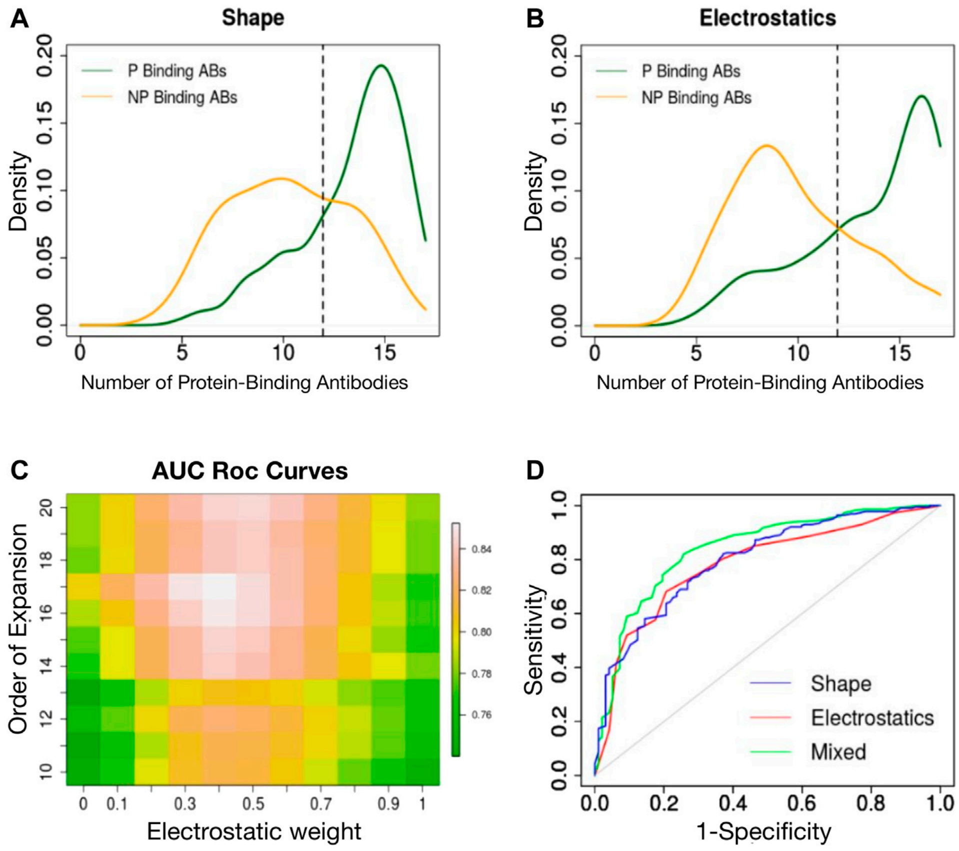


FIGURE 2 | (A) Density distribution of protein binding antibodies (N_{pb}) in the neighbours set of protein binding (green curve) and non-protein binding antibodies (orange curve) based on surface shape similarity. **(B)** Density distribution of protein binding antibodies (N_{pb}) in the neighbours set of protein binding (green curve) and non-protein binding antibodies (orange curve) based on electrostatic surface similarity. **(C)** Classification performance (ROC AUC) is reported as a function of the order n of the Zernike expansion and weight of the average. **(D)** ROC curve of the best classifier based on shape 3DZD (blue curve), electrostatic 3DZD (red curve) and weighted average N_{pb} (green curve).

the neighbors set as compared to non protein-binding antibodies (orange curve) ($mean(N_{pb}^{shape}) = 10.31 \pm 2.99$, $mean(N_{pb}^{elec}) = 9.93 \pm 3.13$) and to random expectation (i.e., $Ex\,[N_{pb}] = N_{Prot}/N_{tot}$, where $Ex\,[N_{pb}]$ is the expected number of protein-binding antibodies if they were distributed uniformly, N_{prot} represents the number of protein-binding in the dataset and N_{tot} is the total number of antibodies in the dataset.).

We next analyzed the performance of each descriptor in classifying the CDRs as a function of the antigen type, using a leave-one-out approach. In summary, for each CDR, if the N_{pb} was greater than Ex (N_{pb}) the CDR was labeled as protein-binding, non protein-binding otherwise. The obtained classification accuracy for the shape and electrostatic descriptors at order n = 20 is 75 and 73%, respectively. Using a Receiver Operating Curve (ROC) analysis, both descriptors achieved an Area Under the Curve (AUC) of 0.78. We next analyzed the classification performance when assigning the class label based on the weighted contribution of shape and electrostatics, as follows:

$$\overline{N_{pb}} = AN_{pb}^{elec} + (1 - A)N_{pb}^{shape} \quad A \in [0, 1] \tag{13}$$

where N_{pb}^{shape} and N_{pb}^{elec} correspond to the N_{pb} computed based on shape and electrostatic descriptors, respectively, and A is the weight ranging from 0 to 1. The results are shown in **Figure 2C**, where the ROC AUC is reported as a function of the weight A and the order n of the Zernike expansion. As it can be noticed, overall performance increases with increasing values of n. Higher AUC values are achieved when both descriptors contribute with similar weight in the classification. Top classification performance indeed is obtained with A = 0.4 and n = 17, leading to an AUC = 0.85 and accuracy of 81%. A very similar performance is obtained with n = 20 and A = 0.4 (AUC = 0.83).

3.2 CDRs vs. Antibody Paratope

The sequence and structure analysis of antibodies, as well as antibody engineering experiments, crucially rely on the precise identification of the CDRs from the antibody sequence (Chothia and Lesk, 1987; Chothia et al., 1989; Kabat et al., 1992; MacCallum et al., 1996; Lefranc, 2011). On the other hand, it is well known that the CDRs only provide a proxy of the actual antigen-binding site, i.e. the antibody paratope (Kunik et al., 2012; Olimpieri et al., 2013). Indeed, early studies showed that only 20–30% of residues within the CDRs

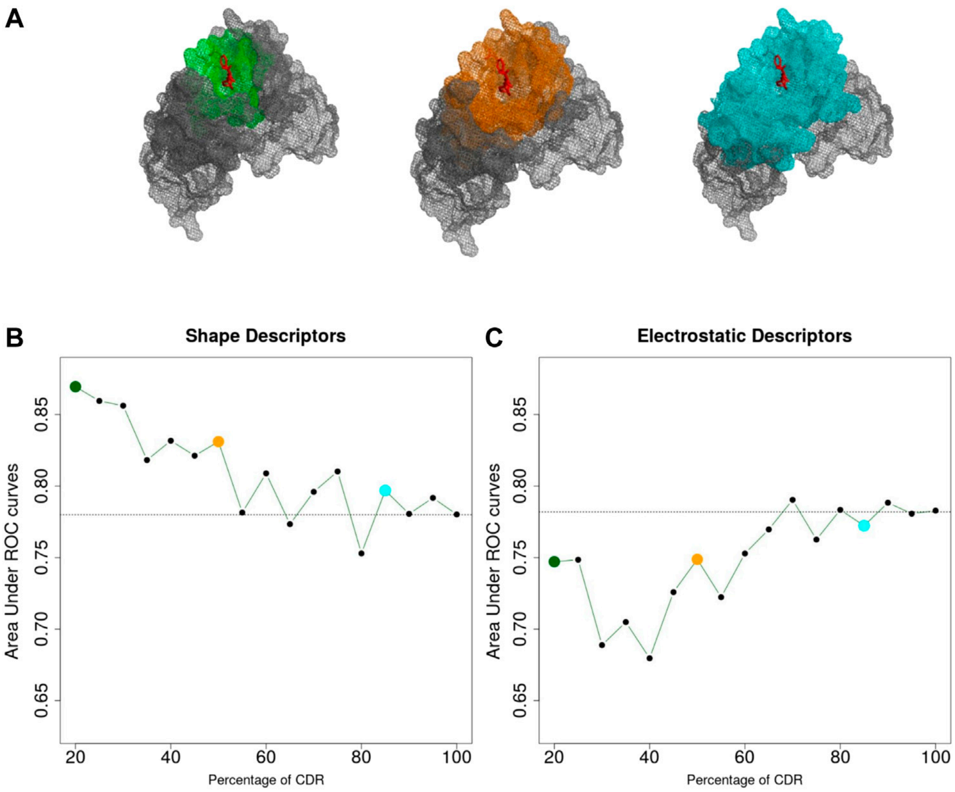


FIGURE 3 | (A) Portion of the CDR surface used for classification. **(B,C)** Area Under the ROC Curve achieved considering different portions of the CDR, based on shape **(B)** and electrostatics **(C)** 3DZD descriptors. Dashed lines indicate the performances obtained considering the entire CDR surface (AUC = 0.78 for both descriptors).

are directly involved in the interaction with the antigen (Padlan, 1994; Sela-Culang et al., 2013). To quantify to what extent this approximation affects our predictions, we analyzed the classification performance as a function of distance from the center of the antibody-antigen interface. For each Antibody-Antigen complex, we defined a *centerpoint*, *b*, as the centroid of the 10 interface atoms of the antibody closer to the antigen and computed the 3DZD for increasing concentric shells around *b*.

We then applied the same classification procedure as described previously, by fixing the order $n = 20$ for both shape and electrostatic 3DZD. The results are shown in **Figure 3** where the ROC AUC of the individual classifiers are reported as a function of the percentage of the CDR surface included in the analysis.

As it can be noticed in **Figure 3B**, the performance of the shape-based classifier shows a maximum when the selected surface region around b extends up to including 20% of the CDRs (ROC AUC = 0.88) followed by a linear decrease when larger surfaces are considered. These results are consistent with the previous notion that shape recognition of the antigen is largely mediated by smaller interacting surfaces contained within the CDR, i.e. the antibody paratope. In summary, while the overall CDR surface can inform about the function of the antibody, this analysis highlights that the information of the paratope can significantly increase our ability to predict antibody specificity. On the other hand, in **Figure 3C**, the classification performance based on the electrostatic descriptor shows a different trend. Indeed, while the classifier shows an overall lower performance compared to the shape-based classifier, performance increases when larger CDR surfaces are considered, reaching a maximum when almost the entire CDR surface is included in the analysis.

3.3 Geometric and Electrostatic Complementarity of Antibody-Antigen Interfaces

A key feature of the 3DZD description is that it is invariant under rotation and translation of the represented surface. This implies that two interacting protein regions with perfect surface complementarity yield identical sets of 3DZD descriptors (Venkatraman et al., 2009a). In line with this principle, here we focus on the application of 3DZD to the analysis of surface complementarity between antibody CDRs and their cognate protein antigens (Details in Methods). The results are shown in **Figure 4**, where the average surface shape and electrostatic complementarity computed on 229 antibody-antigen complexes are reported as a function of the interaction cutoff distance

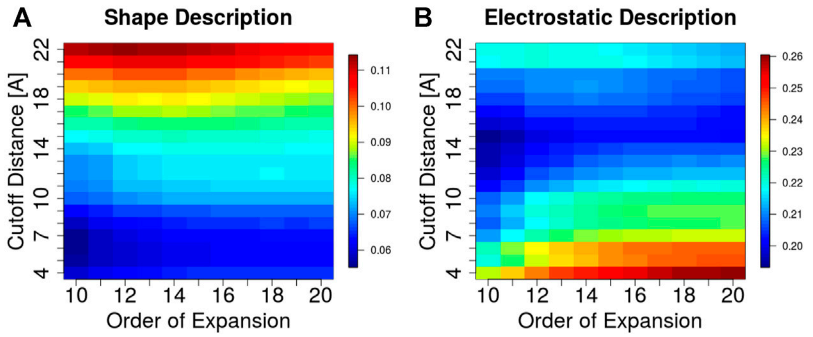


FIGURE 4 | Surface complementarity of antibody-antigen interacting surfaces based on shape **(A)** and electrostatic **(B)** 3DZD descriptors as a function of the interaction cutoff distance (y-axis) and order n of the series expansion (x-axis).

between the antibody and the antigen, and the order n of the series expansion. As expected, shape complementarity decreases at higher values of the cutoff distance, i.e. as regions of the antibody/antigen that are distant from the interaction interface are progressively included in the analysis. On the other hand, electrostatic complementarity increases at higher distances, reaching a maximum when the distance cutoff is 15Å. Notably, in both cases, results are consistent at different orders n of the series expansion. These results indicate that the two descriptors are competent in capturing both short- and long-range effects occurring during antibody-antigen recognition. As further validation of our approach, we measured the surface complementarity at the paratope-epitope interface and compared it with that measured between the paratope and a set of non-epitope, solvent-exposed regions of the antigen, i.e. surface decoys. The results are reported in **Figure 5**, where both shape and electrostatic complementarity are reported for each paratope as normalized Z-score distances to native epitopes and surface decoys, respectively. Notably, while in terms of amino acid composition, surface size, and solvent accessibility the antibody epitopes are essentially not distinguishable from the decoys (**Supplementary Figure S3**), they display significantly higher surface shape and electrostatics complementarity to the paratope. In summary, the metric is able to distinguish the correct paratope-epitope pair among the set of decoys with a classification performance of AUC = 0.75 based on the shape descriptor, and AUC = 0.61 based on the electrostatic 3DZD. Additionally, we compared the 3DZD complementarity observed between specific paratope-epitope pairs and that between the antibody paratopes and non-native epitopes. The results (**Supplementary Figure S4**) show that only a relatively low number, i.e. 68% (72%) of the antibodies in our dataset show a higher shape (electrostatic) complementarity to their cognate epitope compared to non-native epitopes, highlighting the limitation of this metric in the very elusive task of predicting which antibody recognises specifically a given antigen.

4 DISCUSSIONS

In this work we describe a computational protocol based on the 3D Zernike descriptors formalism, which allows a fast, superposition-free comparison of molecular surfaces, and has been applied here to the study of the interacting regions of the antibodies and their cognate antigens. The method represents a significant upgrade compared to our previous implementation (Di Rienzo et al., 2017) as it includes two relevant modifications found to improve its performance, namely, the selection of the molecular patch of interest and the description of its electrostatic properties. Using this new version of the method we are able to classify the antibodies according to the nature of their recognized antigens with a classification performance of 81%. Notably, the method only takes as input the information of the antibody CDR surface. However, when the analysis is restricted to the CDR surface that is in direct contact with the antigen, i.e. the antibody paratope, the classifier based on the shape 3DZD descriptor alone reaches a maximum performance of AUC = 0.88.

As 3DZD descriptors are roto-translation invariant, they are also adept at capturing and quantifying surface complementarity at protein-protein interfaces (Venkatraman et al., 2009a). Here we exploit this property to study the surface shape and electrostatic complementarity between antibody CDRs and their bound protein antigens. Our results indicate that maximum surface shape complementarity is achieved at the docking interface, i.e. at 4 to 8 Angstrom distance cutoff between antibody and antigen residues, and decreases when larger distance cutoffs are considered. In contrast, electrostatic complementarity increases at larger distance cutoffs, reaching a maximum between 14 and 17 Å. For both descriptors, results are consistent at different orders n of the series expansion. Hence, we tested the ability of the surface complementarity metric in recognising antigenic surface epitopes among a set of non-epitope, solvent

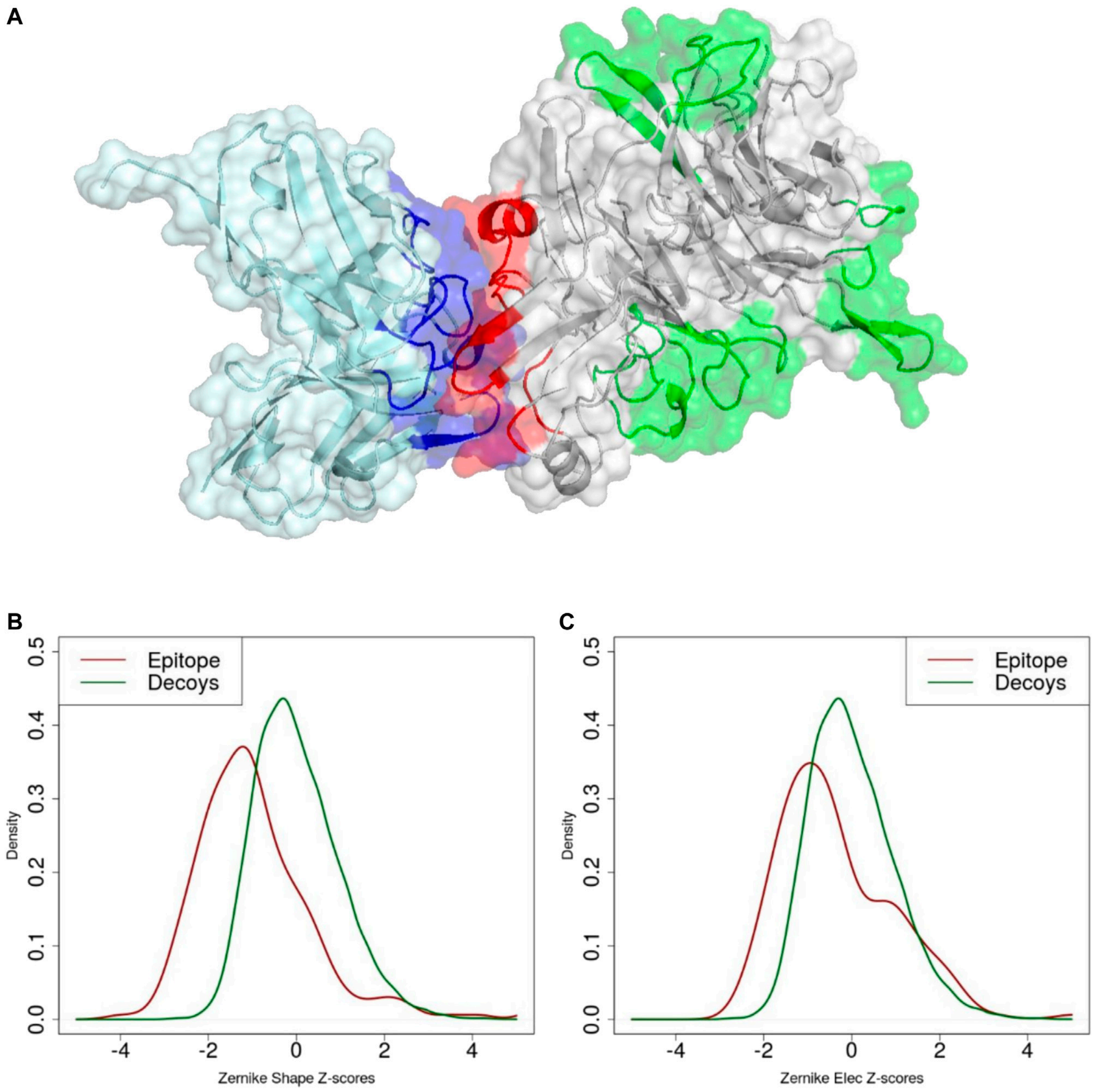


FIGURE 5 | (A) Molecular representation of experimental paratope (blue), experimental epitope (red) and decoys (green). Z-score distribution of **(B)** shape and **(C)** electrostatic surface complementarity based on the 3DZD descriptors between paratope-epitope (red) and paratope-decoy surfaces (green).

exposed regions of the antigen, i.e. surface decoys. Notably, while in terms of surface size, solvent accessibility and amino acid composition the selected surface decoys are not distinguishable from true epitopes, they display significantly lower surface complementarity to the paratope. Indeed, when the 3DZD-based complementarity metric is used to select the correct paratope-epitope pair among a set of surface decoys, we show that shape complementarity alone can lead to a prediction performance of ROC AUC = 0.75. These results show that 3DZD provide a valid quantitative metric for the analysis of surface complementarity at the antibody-antigen interface, which is expected to find applications in many areas, including the identification and design of optimal antibody-antigen interfaces.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

RL contributed to conception and design of the study. LDR collected the datasets, wrote the software and performed the analyses. EM and GR contributed to analysis and interpretation of the results. RL and LDR wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.749784/full#supplementary-material

REFERENCES

Abhinandan, K. R., and Martin, A. C. R. (2010). Analysis and Prediction of VH/VL Packing in Antibodies. *Protein Eng. Des. Selection* 23 (9), 689–697. doi:10.1093/protein/gzq043

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., et al. (2015). Gromacs: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* 1-2, 19–25. doi:10.1016/j.softx.2015.06.001

Alba, J., Di Rienzo, L., Milanetti, E., Acuto, O., and D'Abramo, M. (2020). Molecular Dynamics Simulations Reveal Canonical Conformations in Different Pmhc/tcr Interactions. *Cells* 9 (4), 942. doi:10.3390/cells9040942

Canterakis, N. (1999).3d Zernike Moments and Zernike Affine Invariants for 3d Image Analysis and Recognition. In Proceedings of the 11th Scandinavian Conference on Image Analysis, Kangerlusssuaq, Greenland, June 7–11, 1999. Pattern Recognition Society of Denmark.

Chailyan, A., Marcatili, P., Cirillo, D., and Tramontano, A. (2011). Structural Repertoire of Immunoglobulin λ Light Chains. *Proteins* 79 (5), 1513–1524. doi:10.1002/prot.22979

Chikhi, R., Sael, L., and Kihara, D. (2010). Real-time Ligand Binding Pocket Database Search Using Local Surface Descriptors. *Proteins* 78 (9), 2007–2028. doi:10.1002/prot.22715

Chothia, C., Gelfand, I., and Kister, A. (1998). Structural Determinants in the Sequences of Immunoglobulin Variable Domain 1 1Edited by A. R. Fersht. *J. Mol. Biol.* 278 (2), 457–479. doi:10.1006/jmbi.1998.1653

Chothia, C., and Lesk, A. M. (1987). Canonical Structures for the Hypervariable Regions of Immunoglobulins. *J. Mol. Biol.* 196 (4), 901–917. doi:10.1016/0022-2836(87)90412-8

Chothia, C., Lesk, A. M., Gherardi, E., Tomlinson, I. M., Walter, G., Marks, J. D., et al. (1992). Structural Repertoire of the Human Vh Segments. *J. Mol. Biol.* 227 (3), 799–817. doi:10.1016/0022-2836(92)90224-8

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., et al. (1989). Conformations of Immunoglobulin Hypervariable Regions. *Nature* 342 (6252), 877–883. doi:10.1038/342877a0

Chothia, C., Novotný, J., Bruccoleri, R., and Karplus, M. (1985). Domain Association in Immunoglobulin Molecules. *J. Mol. Biol.* 186 (3), 651–663. doi:10.1016/0022-2836(85)90137-8

Collis, A. V. J., Brouwer, A. P., and Martin, A. C. R. (2003). Analysis of the Antigen Combining Site: Correlations between Length and Sequence Composition of the Hypervariable Loops and the Nature of the Antigen. *J. Mol. Biol.* 325 (2), 337–354. doi:10.1016/s0022-2836(02)01222-6

Connolly, M. L. (1983). Analytical Molecular Surface Calculation. *J. Appl. Cryst.* 16 (5), 548–558. doi:10.1107/s0021889883010985

Daberdaku, S., and Ferrari, C. (2019). Antibody Interface Prediction with 3d Zernike Descriptors and Svm. *Bioinformatics* 35 (11), 1870–1876. doi:10.1093/bioinformatics/bty918

Daberdaku, S., and Ferrari, C. (2018). Exploring the Potential of 3D Zernike Descriptors and SVM for Protein-Protein Interface Prediction. *BMC bioinformatics* 19 (1), 35. doi:10.1186/s12859-018-2043-3

De Wildt, R. M. T., Hoet, R. M. A., van Venrooij, W. J., Tomlinson, I. M., and Winter, G. (1999). Analysis of Heavy and Light Chain Pairings Indicates that Receptor Editing Shapes the Human Antibody Repertoire. *J. Mol. Biol.* 285 (3), 895–901. doi:10.1006/jmbi.1998.2396

Decanniere, K., Muyldermans, S., and Wyns, L. (2000). Canonical Antigen-Binding Loop Structures in Immunoglobulins: More Structures, More Canonical Classes. *J. Mol. Biol.* 300 (1), 83–91. doi:10.1006/jmbi.2000.3839

Di Rienzo, L., Milanetti, E., Alba, J., and D'Abramo, M. (2020). Quantitative Characterization of Binding Pockets and Binding Complementarity by Means of Zernike Descriptors. *J. Chem. Inf. Model.* 60 (3), 1390–1398. doi:10.1021/acs.jcim.9b01066

Di Rienzo, L., Milanetti, E., Lepore, R., Olimpieri, P. P., and Tramontano, A. (2017). Superposition-free Comparison and Clustering of Antibody Binding Sites: Implications for the Prediction of the Nature of Their Antigen. *Sci. Rep.* 7, 45053. doi:10.1038/srep45053

Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., and Baker, N. A. (2004). PDB2PQR: an Automated Pipeline for the Setup of Poisson-Boltzmann Electrostatics Calculations. *Nucleic Acids Res.* 32 (Suppl. l_2), W665–W667. doi:10.1093/nar/gkh381

Dunbar, J., Fuchs, A., Shi, J., and Deane, C. M. (2013). ABangle: Characterising the VH-VL Orientation in Antibodies. *Protein Eng. Des. Selection* 26 (10), 611–620. doi:10.1093/protein/gzt020

Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., et al. (2013). Sabdab: the Structural Antibody Database. *Nucl. Acids Res.* 42 (D1), D1140–D1146. doi:10.1093/nar/gkt1043

Dunbar, J., Krawczyk, K., Leem, J., Marks, C., Nowak, J., Regep, C., et al. (2016). Sabpred: a Structure-Based Antibody Prediction Server. *Nucleic Acids Res.* 44 (W1), W474–W478. doi:10.1093/nar/gkw361

Fogolari, F., Corazza, A., Yarra, V., Jalaru, A., Viglino, P., and Esposito, G. (2012). Bluues: a Program for the Analysis of the Electrostatic Properties of Proteins Based on Generalized Born Radii. *BMC bioinformatics* 13 Suppl. 4 (4), S18. doi:10.1186/1471-2105-13-S4-S18

Foote, J., and Winter, G. (1992). Antibody Framework Residues Affecting the Conformation of the Hypervariable Loops. *J. Mol. Biol.* 224 (2), 487–499. doi:10.1016/0022-2836(92)91010-m

Gabb, H. A., Jackson, R. M., and Sternberg, M. J. E. (1997). Modelling Protein Docking Using Shape Complementarity, Electrostatics and Biochemical Information 1 1Edited by J. Thornton. *J. Mol. Biol.* 272 (1), 106–120. doi:10.1006/jmbi.1997.1203

Grandison, S., Roberts, C., and Morris, R. J. (2009). The Application of 3D Zernike Moments for the Description of "Model-free" Molecular Structure, Functional Motion, and Structural Reliability. *J. Comput. Biol.* 16 (3), 487–500. doi:10.1089/cmb.2008.0083

Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A., and Caves, L. S. D. (2006). Bio3d: an R Package for the Comparative Analysis of Protein Structures. *Bioinformatics* 22 (21), 2695–2696. doi:10.1093/bioinformatics/btl461

Habel, K., Grasman, R., Gramacy, R. B., Mozharovskyi, P., and Sterratt, D. C. (2019). Geometry: Mesh Generation and Surface Tessellation, R Package. Available at: https://CRAN.R-project.org/package=geometry.

Jayaram, N., Bhowmick, P., and Martin, A. C. R. (2012). Germline VH/VL Pairing in Antibodies. *Protein Eng. Des. Selection* 25 (10), 523–530. doi:10.1093/protein/gzs043

Jones, S., and Thornton, J. M. (1996). Principles of Protein-Protein Interactions. *Proc. Natl. Acad. Sci.* 93 (1), 13–20. doi:10.1073/pnas.93.1.13

Kabat, E. A., Te Wu, T., Perry, H. M., Foeller, C., and Gottesman, K. S. (1992). *Sequences of Proteins of Immunological Interest*. Darby, PA: DIANE publishing Co.

Kihara, D., Sael, L., Chikhi, R., and Esquivel-Rodriguez, J. (2011). Molecular Surface Representation Using 3d Zernike Descriptors for Protein Shape Comparison and Docking. *Cpps* 12 (6), 520–530. doi:10.2174/138920311796957612

Kunik, V., Ashkenazi, S., and Ofran, Y. (2012). Paratome: an Online Tool for Systematic Identification of Antigen-Binding Regions in Antibodies Based on Sequence or Structure. *Nucleic Acids Res.* 40 (W1), W521–W524. doi:10.1093/nar/gks480

Kuroda, D., and Gray, J. J. (2016). Shape Complementarity and Hydrogen Bond Preferences in Protein-Protein Interfaces: Implications for Antibody Modeling and Protein-Protein Docking. *Bioinformatics* 32 (16), 2451–2456. doi:10.1093/bioinformatics/btw197

Lee, B., and Richards, F. M. (1971). The Interpretation of Protein Structures: Estimation of Static Accessibility. *J. Mol. Biol.* 55 (3), 379–IN4. doi:10.1016/0022-2836(71)90324-x

Lee, M., Lloyd, P., Zhang, X., Schallhorn, J. M., Sugimoto, K., Leach, A. G., et al. (2006). Shapes of Antibody Binding Sites: Qualitative and Quantitative Analyses Based on a Geomorphic Classification Scheme. *J. Org. Chem.* 71 (14), 5082–5092. doi:10.1021/jo052659z

Lefranc, M.-P. (2011).Antibody Nomenclature, *MAbs*, 3, 1–2. doi:10.4161/mabs.3.1.14151

Leicester, S. E., Finney, J. L., and Bywater, R. P. (1988). Description of Molecular Surface Shape Using Fourier Descriptors. *J. Mol. Graphics* 6 (2), 104–108. doi:10.1016/0263-7855(88)85008-2

Lepore, R., Olimpieri, P. P., Messih, M. A., and Tramontano, A. (2017). Pigspro: Prediction of Immunoglobulin Structures V2. *Nucleic Acids Res.* 45 (W1), W17–W23. doi:10.1093/nar/gkx334

Li, N., Sun, Z., and Jiang, F. (2007). SOFTDOCK Application to Protein-Protein Interaction Benchmark and CAPRI. *Proteins* 69 (4), 801–808. doi:10.1002/prot.21728

MacCallum, R. M., Martin, A. C. R., and Thornton, J. M. (1996). Antibody-antigen Interactions: Contact Analysis and Binding Site Topography. *J. Mol. Biol.* 262 (5), 732–745. doi:10.1006/jmbi.1996.0548

Marcatili, P., Olimpieri, P. P., Chailyan, A., and Tramontano, A. (2014). Antibody Modeling Using the Prediction of ImmunoGlobulin Structure (PIGS) Web Server. *Nat. Protoc.* 9 (12), 2771–2783. doi:10.1038/nprot.2014.189

Martin, A. C. R., and Thornton, J. M. (1996). Structural Families in Loops of Homologous Proteins: Automatic Classification, Modelling and Application to Antibodies. *J. Mol. Biol.* 263 (5), 800–815. doi:10.1006/jmbi.1996.0617

Max, N. L., and Getzoff, E. D. (1988). Spherical Harmonic Molecular Surfaces. *IEEE Comput. Grap. Appl.* 8 (4), 42–50. doi:10.1109/38.7748

Messih, M. A., Lepore, R., Marcatili, P., and Tramontano, A. (2014). Improving the Accuracy of the Structure Prediction of the Third Hypervariable Loop of the Heavy Chains of Antibodies. *Bioinformatics* 30 (19), 2733–2740. doi:10.1093/bioinformatics/btu194

Meyer, D., and Buchta, C. (2019). *Proxy: Distance and Similarity Measures, R Package Version 0*, 4–23. Available at: https://CRAN.R-project.org/package=proxy.

Ming-Kuei Hu, M.-K. (1962). Visual Pattern Recognition by Moment Invariants. *IEEE Trans. Inform. Theor.* 8 (2), 179–187. doi:10.1109/tit.1962.1057692

Mitra, P., and Pal, D. (2010). New Measures for Estimating Surface Complementarity and Packing at Protein-Protein Interfaces. *FEBS Lett.* 584 (6), 1163–1168. doi:10.1016/j.febslet.2010.02.021

Norel, R., Lin, S. L., Wolfson, H. J., and Nussinov, R. (1995). Molecular Surface Complementarity at Protein-Protein Interfaces: the Critical Role Played by Surface Normals at Well Placed, Sparse, Points in Docking. *J. Mol. Biol.* 252 (2), 263–273. doi:10.1006/jmbi.1995.0493

North, B., Lehmann, A., and Dunbrack, R. L., Jr (2011). A New Clustering of Antibody Cdr Loop Conformations. *J. Mol. Biol.* 406 (2), 228–256. doi:10.1016/j.jmb.2010.10.030

Novotni, M., and Klein, R. (2004). Shape Retrieval Using 3d Zernike Descriptors. *Computer-Aided Des.* 36 (11), 1047–1062. doi:10.1016/j.cad.2004.01.005

Olimpieri, P. P., Chailyan, A., Tramontano, A., and Marcatili, P. (2013). Prediction of Site-specific Interactions in Antibody-Antigen Complexes: the Proabc Method and Server. *Bioinformatics* 29 (18), 2285–2291. doi:10.1093/bioinformatics/btt369

Padlan, E. A. (1994). Anatomy of the Antibody Molecule. *Mol. Immunol.* 31 (3), 169–217. doi:10.1016/0161-5890(94)90001-9

Pollard, A. J., and Bijker, E. M. (2020). A Guide to Vaccinology: from Basic Principles to New Developments. *Nat. Rev. Immunol.*, 1–18. doi:10.1038/s41577-020-00479-7

Raghunathan, G., Smart, J., Williams, J., and Almagro, J. C. (2012). Antigen-binding Site Anatomy and Somatic Mutations in Antibodies that Recognize Different Types of Antigens. *J. Mol. Recognit.* 25 (3), 103–113. doi:10.1002/jmr.2158

Sela-Culang, I., Kunik, V., and Ofran, Y. (2013). The Structural Basis of Antibody-Antigen Recognition. *Front. Immunol.* 4, 302. doi:10.3389/fimmu.2013.00302

Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J., and Wilke, C. O. (2013). Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PloS one* 8 (11), e80635. doi:10.1371/journal.pone.0080635

Tomlinson, I. M., Cox, J. P., Gherardi, E., Lesk, A. M., and Chothia, C. (1995). The Structural Repertoire of the Human V Kappa Domain. *EMBO J.* 14 (18), 4628–4638. doi:10.1002/j.1460-2075.1995.tb00142.x

Tramontano, A., Chothia, C., and Lesk, A. M. (1990). Framework Residue 71 Is a Major Determinant of the Position and Conformation of the Second Hypervariable Region in the Vh Domains of Immunoglobulins. *J. Mol. Biol.* 215 (1), 175–182. doi:10.1016/s0022-2836(05)80102-0

Vargas-Madrazo, E., and Paz-García, E. (2002). Modifications to Canonical Structure Sequence Patterns: Analysis for L1 and L3. *Proteins* 47 (2), 250–254. doi:10.1002/prot.10187

Venkatraman, V., Sael, L., and Kihara, D. (2009). Potential for Protein Surface Shape Analysis Using Spherical Harmonics and 3d Zernike Descriptors. *Cell Biochem Biophys* 54 (1-3), 23–32. doi:10.1007/s12013-009-9051-x

Venkatraman, V., Yang, Y. D., Sael, L., and Kihara, D. (2009). Protein-protein Docking Using Region-Based 3d Zernike Descriptors. *BMC bioinformatics* 10 (1), 407. doi:10.1186/1471-2105-10-407

Via, A., Ferrè, F., Brannetti, B., and Helmer-Citterich*, M. (2000). Protein Surface Similarities: a Survey of Methods to Describe and Compare Protein Surfaces. *Cmls, Cel. Mol. Life Sci.* 57 (13), 1970–1977. doi:10.1007/pl00000677

Walls, P. H., and Sternberg, M. J. E. (1992). New Algorithm to Model Protein-Protein Recognition Based on Surface Complementarity. *J. Mol. Biol.* 228 (1), 277–297. doi:10.1016/0022-2836(92)90506-f

Webster, D. M., Henry, A. H., and Rees, A. R. (1994). Antibody-antigen Interactions. *Curr. Opin. Struct. Biol.* 4 (1), 123–129. doi:10.1016/s0959-440x(94)90070-1

Weitzner, B. D., Jeliazkov, J. R., Lyskov, S., Marze, N., Kuroda, D., Frick, R., et al. (2017). Modeling and Docking of Antibody Structures with Rosetta. *Nat. Protoc.* 12 (2), 401–416. doi:10.1038/nprot.2016.180

Whitelegg, N. R. J., and Rees, A. R. (2000). Wam: an Improved Algorithm for Modelling Antibodies on the Web. *Protein Eng.* 13 (12), 819–824. doi:10.1093/protein/13.12.819

Zernike, F., and Stratton, F. J. M. (1934). Diffraction Theory of the Knife-Edge Test and its Improved Form, the Phase-Contrast Method. *Monthly Notices R. Astronomical Soc.* 94, 377–384. doi:10.1093/mnras/94.5.377

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Molecular Biosciences

ORIGINAL RESEARCH
published: 26 January 2022
doi: 10.3389/fmolb.2022.812750

Comparing Antibody Interfaces to Inform Rational Design of New Antibody Formats

*Monica L. Fernández-Quintero[1†], Patrick K. Quoika[1†], Florian S. Wedl[1], Clarissa A. Seidler[1], Katharina B. Kroell[1], Johannes R. Loeffler[1], Nancy D. Pomarici[1], Valentin J. Hoerschinger[1], Alexander Bujotzek[2], Guy Georges[2], Hubert Kettenberger[2] and Klaus R. Liedl[1]**

[1]Department of General, Inorganic and Theoretical Chemistry, Center for Molecular Biosciences Innsbruck (CMBI), University of Innsbruck, Innsbruck, Austria, [2]Roche Pharma Research and Early Development, Large Molecule Research, Roche Innovation Center Munich, Penzberg, Germany

OPEN ACCESS

Edited by:
*Maya Topf,
Leibniz Institute of Experimental
Biology and UKE, Germany*

Reviewed by:
*Didier Devaurs,
University of Edinburgh,
United Kingdom
Stéphane Téletchéa,
Université de Nantes, France*

***Correspondence:**
*Klaus R. Liedl
Klaus.Liedl@uibk.ac.at*

†These authors have contributed equally to this work

Specialty section:
*This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences*

***Received:** 10 November 2021*
***Accepted:** 10 January 2022*
***Published:** 26 January 2022*

Citation:
Fernández-Quintero ML, Quoika PK, Wedl FS, Seidler CA, Kroell KB, Loeffler JR, Pomarici ND, Hoerschinger VJ, Bujotzek A, Georges G, Kettenberger H and Liedl KR (2022) Comparing Antibody Interfaces to Inform Rational Design of New Antibody Formats. Front. Mol. Biosci. 9:812750. doi: 10.3389/fmolb.2022.812750

As the current biotherapeutic market is dominated by antibodies, the design of different antibody formats, like bispecific antibodies and other new formats, represent a key component in advancing antibody therapy. When designing new formats, a targeted modulation of pairing preferences is key. Several existing approaches are successful, but expanding the repertoire of design possibilities would be desirable. Cognate immunoglobulin G antibodies depend on homodimerization of the fragment crystallizable regions of two identical heavy chains. By modifying the dimeric interface of the third constant domain (C_H3-C_H3), with different mutations on each domain, the engineered Fc fragments form rather heterodimers than homodimers. The first constant domain (C_H1-C_L) shares a very similar fold and interdomain orientation with the C_H3-C_H3 dimer. Thus, numerous well-established design efforts for C_H3-C_H3 interfaces, have also been applied to C_H1-C_L dimers to reduce the number of mispairings in the Fabs. Given the high structural similarity of the C_H3-C_H3 and C_H1-C_L domains we want to identify additional opportunities in comparing the differences and overlapping interaction profiles. Our vision is to facilitate a toolkit that allows for the interchangeable usage of different design tools from crosslinking the knowledge between these two interface types. As a starting point, here, we use classical molecular dynamics simulations to identify differences of the C_H3-C_H3 and C_H1-C_L interfaces and already find unexpected features of these interfaces shedding new light on possible design variations. Apart from identifying clear differences between the similar C_H3-C_H3 and C_H1-C_L dimers, we structurally characterize the effects of point-mutations in the C_H3-C_H3 interface on the respective dynamics and interface interaction patterns. Thus, this study has broad implications in the field of antibody engineering as it provides a structural and mechanistical understanding of antibody interfaces and thereby presents a crucial aspect for the design of bispecific antibodies.

Keywords: antibodies, structure, interface characterization, interface dynamics, antibody design, bispecific antibody formats

INTRODUCTION

Antibodies play a central role in the adaptive immune system, as they can recognize and neutralize foreign antigens (Chiu et al., 2019). In the last years, antibodies emerged as a new class of pharmaceuticals (Kaplon et al., 2020; Kaplon and Reichert, 2021), with over one hundred antibody-based drugs being marketed or pending approval.

Structurally, antibodies consist of two heavy and two light chains and have a unique modular anatomy facilitating their engineering and design (Davies and Chacko, 1993). The immunoglobulin heavy and light chains are composed of various discrete protein domains. Especially interesting is that these domains all have a similar folded structure, which is known as the immunoglobulin fold (Chiu et al., 2019). However, even though they share a similar fold, there are distinct structural differences between these domains (**Figure 1**). In general, antibodies can be divided into a crystallizable fragment (Fc) and two identical antigen-binding fragments (Fabs). The Fab can further be subdivided into constant (C_H1-C_L) and variable (V_H-V_L) domains (Davies and Chacko, 1993; Röthlisberger et al., 2005). The variable domains of the heavy and the light chain (V_H and V_L) shape the antigen binding site and are responsible for antigen binding and recognition (Colman and Dixon, 1988; Addis et al., 2014; Fernández-Quintero et al., 2020c). The variable and the constant domains in the Fab are linked via a so-called switch region (Stanfield et al., 2006). The C_H1-C_L heterodimer plays an essential role for antibody assembly and secretion in the cell (Adachi et al., 2003). Comparison of the V_H-V_L and the C_H1-C_L heterodimers revealed that the C_H1-C_L heterodimer is more stable than the V_H–V_L heterodimer (Röthlisberger et al., 2005). The individual C_H1 domain is not stable in folded form and requires interactions with either the chaperone BiP or the C_L domain for folded state stability (Vanhove et al., 2001; Feige et al., 2014). The crystallizable fragment is composed of a C_H2-C_H2 and a C_H3-C_H3 homodimer (Teplyakov et al., 2013). The C_H2-C_H2 domain has no direct protein interactions in the interface as the interface is formed by glycans (Teplyakov et al., 2013). Thus, the C_H2-C_H2 domain differs from all other domains and consequently will not be discussed in this manuscript. The C_H3 domains bind tightly with each other by hydrophobic interactions at the center, surrounded by salt bridges and thereby forming the foundation for the heavy chain dimer association (Teplyakov et al., 2013). Mutations in the C_H3-C_H3 interface have been shown to strongly influence the stability and the association of the two domains (Rose et al., 2013).

The concept of having an antibody with two different antigen binding sites was established more than 50 years ago by Nisonoff and co-workers and evolved alongside numerous advances and technical innovations in the field of antibody engineering, leading to more than 100 bispecific antibody (bsAb) formats known up to now (Nisonoff and Rivers, 1961; Fudenberg et al., 1964). BsAb formats expand the functionality of traditional antibodies by their ability to target effector cells to kill tumor cells, to enhance tissue specificity or to combine the antigen binding of two antibodies in a single molecule to simultaneously target two signaling pathways (Brinkmann and Kontermann, 2017; Sedykh et al., 2018). BsAbs can be assembled from different heavy and light chains. To suppress random assembly of different chains, resulting in various non-desired molecules, engineering efforts are required (Bönisch et al., 2017). A major breakthrough in the development of bsAb formats was the invention of the knobs-into-holes (KiH) technology for C_H3-C_H3 interfaces (Ridgway et al., 1996; Elliott et al., 2014). Precisely, advances like the KiH technology for C_H3-C_H3 interfaces represented a novel and effective design strategy for engineering heavy chain homodimers towards heterodimers, to reduce the risk of random assembly of different chains (Ridgway et al., 1996; Elliott et al., 2014; Kuglstatter et al., 2017). Thus, the idea of modifying the interfaces has motivated numerous studies to find variations of this approach by following a number of different strategies, such as alterations of the charge polarity in the interfaces compared to the homodimer, e.g., inverted charge interactions (DE-KK and DD-KK variants) (Ha et al., 2016; Moore et al., 2019). More recently, also KiH mutations in combination with charge inversions have been introduced into both Fab interfaces, C_H1-C_L and V_H-V_L, enforcing the correct pairings of light chains with the corresponding heavy chains (Bönisch et al., 2017; Dillon et al., 2017; Regula et al., 2018).

In this study, we use classical molecular dynamics simulations to provide a systematic and extensive comparison of different antibody interfaces, which are in the spotlight of antibody engineering as they offer numerous design opportunities for bispecific antibody formats (Brinkmann and Kontermann, 2017; Sedykh et al., 2018). As C_H1-C_L dimers are inherently heterodimers, we compare them with the homo-and-heterodimeric C_H3-C_H3 domains. We aim to identify different and overlapping interaction profiles of either the C_H3-C_H3 or C_H1-C_L interfaces with the intention to crosslink the knowledge covering the two interfaces (Jost Lopez et al., 2020). Apart from that, we compare the interface flexibilities of C_H3-C_H3 or C_H1-C_L domains and provide key determinants that contribute to the stability and their tendency to heterodimerize.

The investigated C_H3-C_H3 and C_H1-C_L dimers and their respective PDB accession codes are summarized in **Supplementary Table S1**, covering a variety of different design strategies to enforce the formation of heterodimers. The Fabs to study the C_H1-C_L dimers were chosen based on their availability of experimentally determined structure and stability data and their light chain isotypes. We also included in our dataset three antibody Fabs with mutations in the C_H1-C_L, which facilitate selective Fab assembly in combination with previously described KiH mutations for preferential heavy chain heterodimerization.

RESULTS

Structural Architecture of the Investigated Antibody Interfaces

First, we introduce and structurally characterize different antibody interfaces and their respective architectures

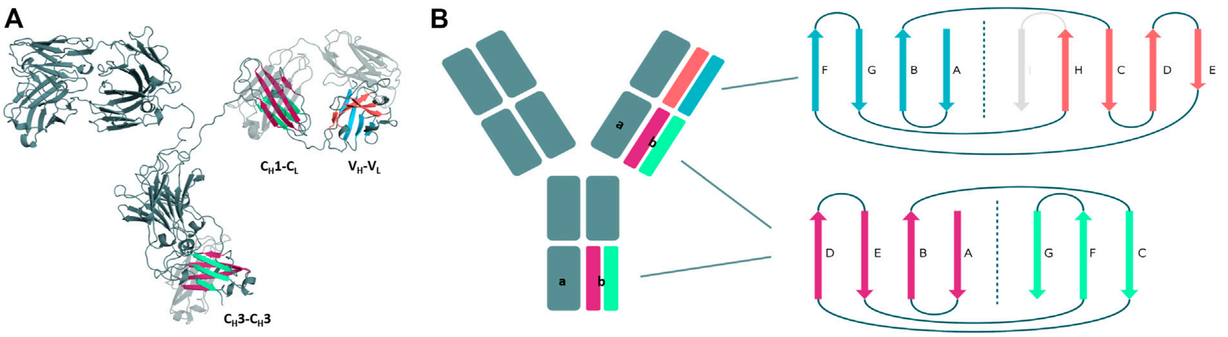


FIGURE 1 | Structural and schematic representation of the modular anatomy of an antibody focusing on the discrete protein domains, which are characterized by the immunoglobulin fold. **(A)** Crystal structure of a whole IgG2 antibody (PDB: 1IGT) highlighting the different interface classes and their respective domain architecture. **(B)** Schematic depiction of the tertiary structure features, i.e., number and organization of the β-strands, for each of the individual antibody domains sharing the immunoglobulin fold. The C_H3 and C_L/C_H1 domains do not only share a similar structure and topology but also contain the same number and arrangement of β-strands. The variable domains on the other hand differ in their number of β-strands and their architecture and are therefore color-coded differently.

(**Figure 1**). All investigated antibody interfaces (23 Fab fragments and 23 C_H3-C_H3 domains), summarized in **Supplementary Table S1**, have been simulated for 1 µs with classical molecular dynamics simulations (extracting 10,000 frames) in explicit solvation to better understand and capture the variability of these interfaces. **Figure 1** shows the comparison of the dimeric antibody interfaces in the antigen-binding fragment (V_H-V_L and C_H1-C_L domains) and in the third constant domain (C_H3-C_H3 domain). All of the presented interfaces share the same immunoglobulin fold, which is characterised by hydrogen bond interactions between the different β-strands. Additionally, we find that the C_H1-C_L and C_H3-C_H3 domains have actually the same number of β-strands, i.e., a 3-stranded sheet packed against a 4-stranded sheet. Also, the relative orientation of the two monomers with respect to each other (approximately 90° observed in X-ray structures) is nearly identical between the C_H3-C_H3 and C_H1-C_L dimers. Thus, the C_H3-C_H3 and C_H1-C_L dimers share a very similar structure and fold. However, we observe structural differences in the overall architecture between the C_H1-C_L/C_H3-C_H3 and the V_H-V_L domains, as the V_H-V_L domains differ in their number of strands (9 β-strands arranged in two sheets of 4 and 5 strands), and in their relative orientation between the V_H and V_L monomers with respect to each other (approximately 50° observed in X-ray structures).

Relative Interdomain Orientations of C_H1-C_L and C_H3-C_H3 Domains

Apart from understanding the structural architecture, the dimeric interfaces are strongly influenced by the relative interdomain orientation and their respective dynamics. To calculate the interface movements, we used the well-established ABangle tool (Dunbar et al., 2013) and a recently presented python tool, called OCD tool (Hoerschinger et al., 2021), which both allow to calculate the interface orientations of different immunoglobulin like domains by defining five angles (one torsion angle (HL/AB)/four tilt angles (LC1, LC2, HC1, HC2/AC1, AC2, BC1, BC2) and one distance (dC). For the C_H1-C_L domains we added the prefix c to the angle names (cHL, cLC1, cLC2, cHC1, cHC2, dC), as the C_H1-C_L dimer forms the constant domain of the Fab fragment and to be able to distinguish them from the variable fragment (Fv) nomenclature. The detailed definition of these angles is presented in the methods section. **Figure 2A** shows a superimposition of the two dimers (C_H1-C_L and C_H3-C_H3), highlighting the high structural similarity of the β-strands, while the loops on the other hand differ between the two dimers (Cα-RMSD 1.8Å). **Figure 2B** depicts the interdomain angle distributions of the relative interdomain orientations for all investigated C_H1-C_L and C_H3-C_H3 simulations and shows significant overlaps in the interface angle (cHL/AB) distributions. However, the C_H1-C_L shows a higher variability in the interface angle, which is reflected in broader angle distributions, compared to the C_H3-C_H3 dimer. Apart from the higher flexibility in the interdomain angle, we also find shifted C_H1-C_L distributions towards lower cHL-Angle values. The torsion angle (cHL) of all C_H1-C_L domains ranges from 65°–110°, while the torsion angle (AB angle) of all C_H3-C_H3 ranges from 85–125° (cHL angle, AB angle). The biggest difference in the relative interdomain orientations can be observed for the Fv torsion angles (HL angle), which range from 35°–80° (**Supplementary Figure S1**).

Structural Characterization of the C_H1-C_L and C_H3-C_H3 Interfaces

To structurally characterize interactions in the C_H1-C_L and C_H3-C_H3 interfaces, we use the GetContacts tool (Stanford University, adate), which calculates the interface contacts in a time-resolved way and depicts them with so-called flareplots (https://getcontacts.github.io/). To better visualize the comparison between the two interfaces we grouped the residues belonging to the same loops and β-strands to obtain coarse grained flareplots. This coarse-grained representation of the C_H1-C_L and C_H3-C_H3 interfaces also allows having a better overview about the regions of these interfaces that actually form key

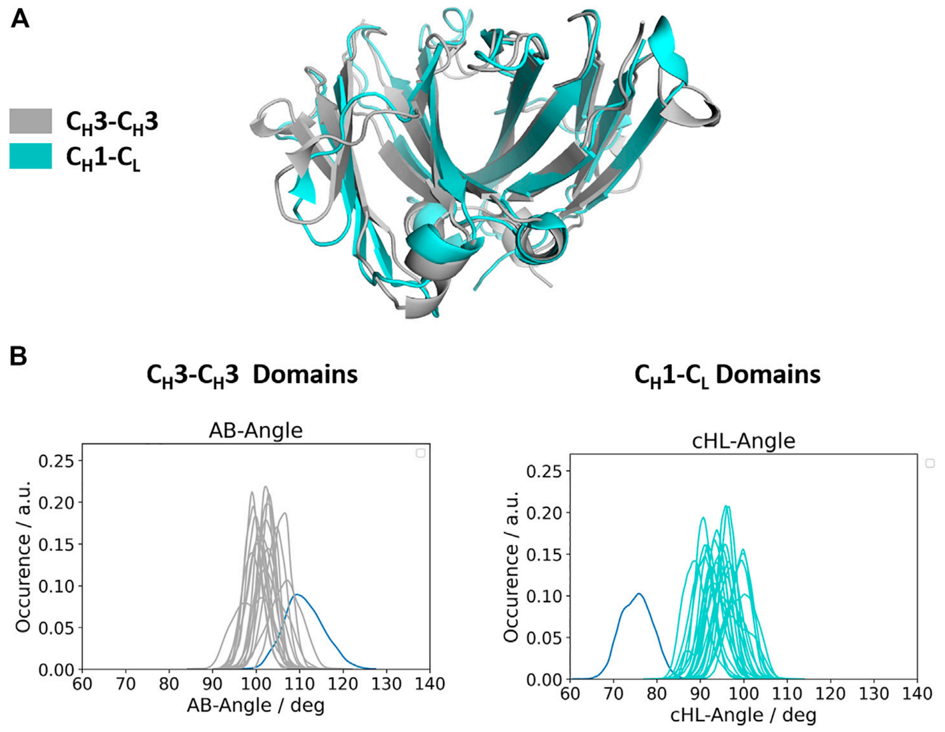


FIGURE 2 | Comparison of the structurally highly similar C_H3-C_H3 and C_H1-C_L domains. **(A)** Structural overlay of a C_H3-C_H3 (grey, PDB: 3AVE) and a C_H1-C_L (cyan, PDB: 5I19) dimer illustrating their identical scaffold, despite having diverging loop structures. **(B)** Distributions of interdomain angles for all C_H3-C_H3 and C_H1-C_L domains, respectively. These angles haven been calculated with the recently published OCD tool and show that the C_H3-C_H3 interfaces cover narrower angle ranges, while the C_H1-C_L, due to their higher number of sequence variations, reveal a larger spread. The highlighted distributions shown in blue correspond to the C_H3-C_H3 heterodimeric DE-KK variant (PDB: 5NSC) and to the λ-light chain antibody (PDB: 1NL0).

interactions, which contribute to their structural integrity and to their stability. The β-strands are labelled with single letters, while the loops are tagged with a two-letter combination of the respective β-strands before and after the loop. To ease the comparison between C_H1-C_L and C_H3-C_H3 we refer both to the C_L domain and one of the C_H3 domains (the domain A) as "a" and to the C_H1 domain and to the other C_H3 domain (the domain B) as "b." The thickness of the lines in the flareplots corresponds to the occurrence of a contact (ratio) over the whole simulation time (10,000 frames/simulation). A representative C_H1-C_L and C_H3-C_H3 structure color-coded and labeled according to the flareplots (right) is depicted in **Figure 3**. The coarse-grained flareplots presented in **Figure 3** show all interdomain contact patterns for both the C_H1-C_L and C_H3-C_H3 interface. While the C_H1-C_L and C_H3-C_H3 domains share common interaction patterns, we also investigated the type of interactions contributing to the formation of the respective interface. The flareplots shown in **Supplementary Figure S2**, **Figures 4**, **5** are just exemplary plots. The barplots on the right quantitatively summarize and compare the contacts observed for all investigated C_H3-C_H3 and C_H1-C_L domains. **Supplementary Figure S2** illustrates representative coarse-grained flareplots showing the interdomain hydrogen bond interactions of both the C_H1-C_L and C_H3-C_H3 domains. While we find overlaps in the hydrogen bond interaction patterns for the C_H3-C_H3 and C_H1-C_L interfaces (**Supplementary Figures S2A,B**), they differ substantially in number and occurrence of interdomain hydrogen bonds between C_H1-C_L and C_H3-C_H3 domains, i.e., the C_H3-C_H3 domains form significantly more hydrogen bonds between the a_E – b_DE, a_DE – b_E, a_A – b_AB, a_B – b_E, a_B – b_B and a_G – b_AB loops/strands (**Supplementary Figure S2**).

In line with these observations, we find that the C_H3-C_H3 interfaces are strongly stabilized by salt bridges (**Figure 4**), while the C_H1-C_L interfaces reveal substantially more hydrophobic interactions (**Figure 5**). Long-lasting salt bridge interactions (>60% of the simulation time) in the C_H3-C_H3 interfaces are formed by the a_E – b_DE, a_DE – b_E, a_D – b_E, a_B – b_AB, a_AB – b_B and a_G – b_AB loops/strands. Salt bridges between the a_AB – b_G and a_DE – b_D loops/strands are present in both C_H1-C_L and C_H3-C_H3 domains (**Figure 4**). While the C_H3-C_H3 domains are characterized by a substantially higher number of charged interactions, the C_H1-C_L domains are stabilized by hydrophobic interactions between the a_B – b_D, a_B – b_A, a_A – b_A, a_B – b_E and a_A – b_B strands. Even though the C_H3-C_H3 interface is strongly stabilized by salt bridge interactions, the hydrophobic interactions between the a_E – b_B, a_D – b_D, a_E – b_E and a_E – b_D strands (**Figure 5C**) are characteristic for the C_H3-C_H3 domains, compared to the C_H1-C_L domains.

Moreover, we find interdomain van der Waals interaction patterns that are present in both the C_H1-C_L and C_H3-C_H3

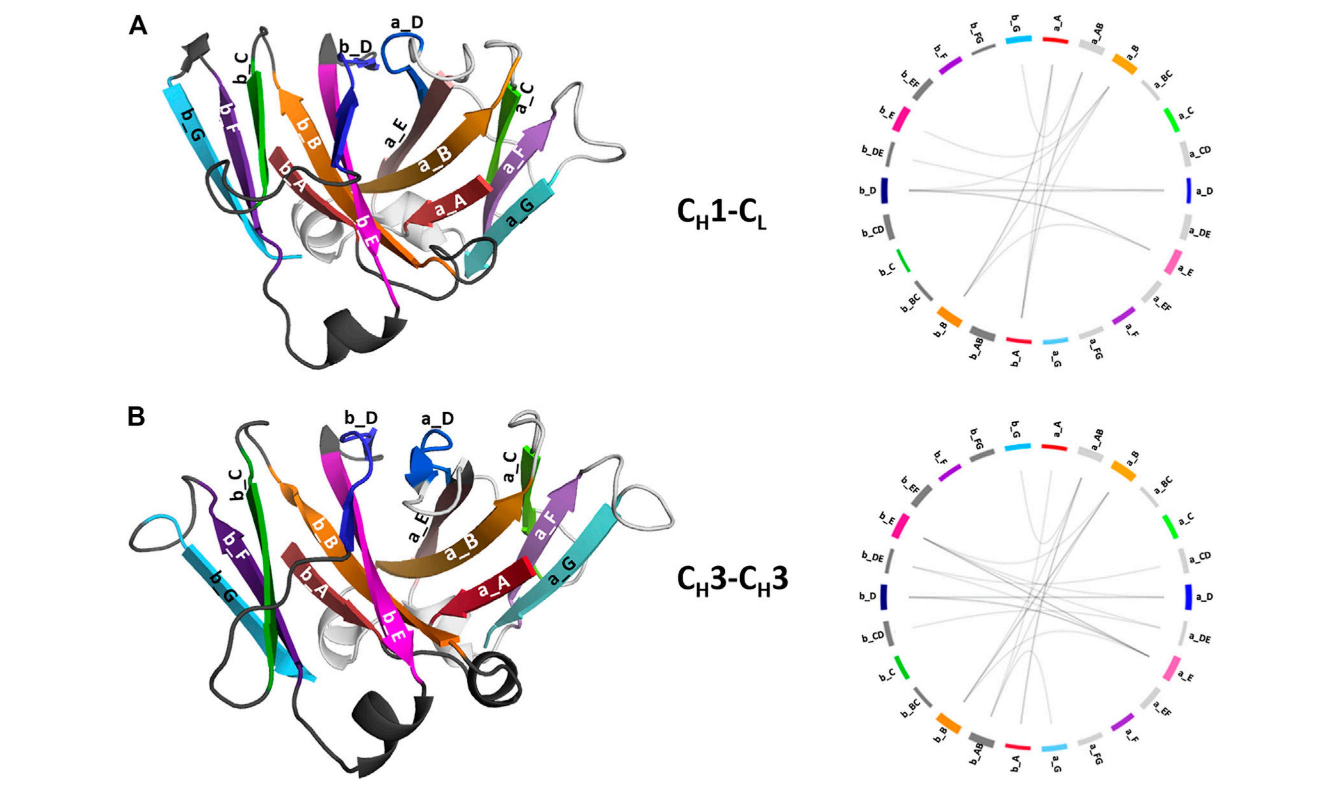


FIGURE 3 | Structural representation of the C_H3-C_H3 and C_H1-C_L domains including a coarse-grained contact analysis of the interactions contributing to the formation and stabilization of the domain interfaces. **(A)** Structure of C_H1-C_L heterodimer (PDB: 5I17) color-coded and labeled according to the coarse-grained flareplots on the right showing the interdomain interactions present in the X-ray structure. We coarse grained residues belonging to the same loops or β-strands. **(B)** Structure of C_H3-C_H3 (PDB: 5DJ0) dimer color-coded and labeled according to the coarse-grained flareplots on the right, which illustrate the interdomain contacts present in the X-ray structure.

domains, e.g., interactions between a_D – b_D strands (**Supplementary Figure S3**). However, also substantial differences between C_H1-C_L and C_H3-C_H3 domains can be identified for the interdomain van der Waals interactions, such as the interactions between the a_E – b_E strands and the a_A – b_AB strand/loop, which are dominantly present in C_H3-C_H3 domains and the a_E – b_D and a_A – b_B strands, which can be found more in C_H1-C_L domains. **Figure 6** illustrates contact maps depicting differences in the number and duration of hydrogen bond, salt bridge and hydrophobic interactions for all investigated antibody fragments. The color bar is normalized according to the most frequent contacts in either of the two interface classes.

Thus, **Figure 6** summarizes the findings shown in **Figures 4**, **5** and **Supplementary Figure S2**, as it clearly displays the substantially higher number of hydrogen bond and salt bridge interactions for the C_H3-C_H3 domains, while the C_H1-C_L interface is dominated by hydrophobic interactions. To quantify this difference even more, we calculated the electrostatic interface interaction energies for all investigated C_H1-C_L and C_H3-C_H3 dimers (**Supplementary Table S2**). The strong difference in the type of interactions between the C_H1-C_L and C_H3-C_H3 are even more pronounced in the electrostatic interface interaction energies, where we find significantly higher electrostatic interaction energies for the C_H3-C_H3 dimer, compared to the C_H1-C_L domains.

Supplementary Figure S4 shows the comparison of three C_H3-C_H3 domains (Dengl et al., 2020) with three engineered C_H1-C_L interfaces (Dillon et al., 2017), which were designed following similar heterodimerization strategies. The goal of redesigning the C_H1/C_L interface was to reduce mispairings by having a stably paired C_H1-C_L interface due to mutations that create incompatibilities towards the binding of wildtype C_H1 or C_L domains (Dillon et al., 2017). Apart from inserting KiH mutations, the interface was redesigned by introducing charge mutations, which co-determine orthogonal heavy and light chain pairing preferences. The first two presented C_H1-C_L domains (**Supplementary Figures S4A,B**) have newly introduced charge pairs and are therefore described as KE (C_H1 S183K interacts with C_L V133E) and EK (C_H1 S183E interacts with C_L V133K) variants (PDB accession codes: 5TDN and 5TDO, respectively). The third C_H1-C_L interface (**Supplementary Figure S4C**) contains mutations at the edge of the interface at position C_L F116A and C_H1 S181M, which introduce more flexibility. Additionally, KiH modifications are introduced at position C_H1 F170S and C_L S176F (PDB accession code: 5TDP). **Supplementary Figures S4A,B** shows strong hydrogen bond networks for the KE and EK variants, especially between the a_E – b_E and a_B – b_E

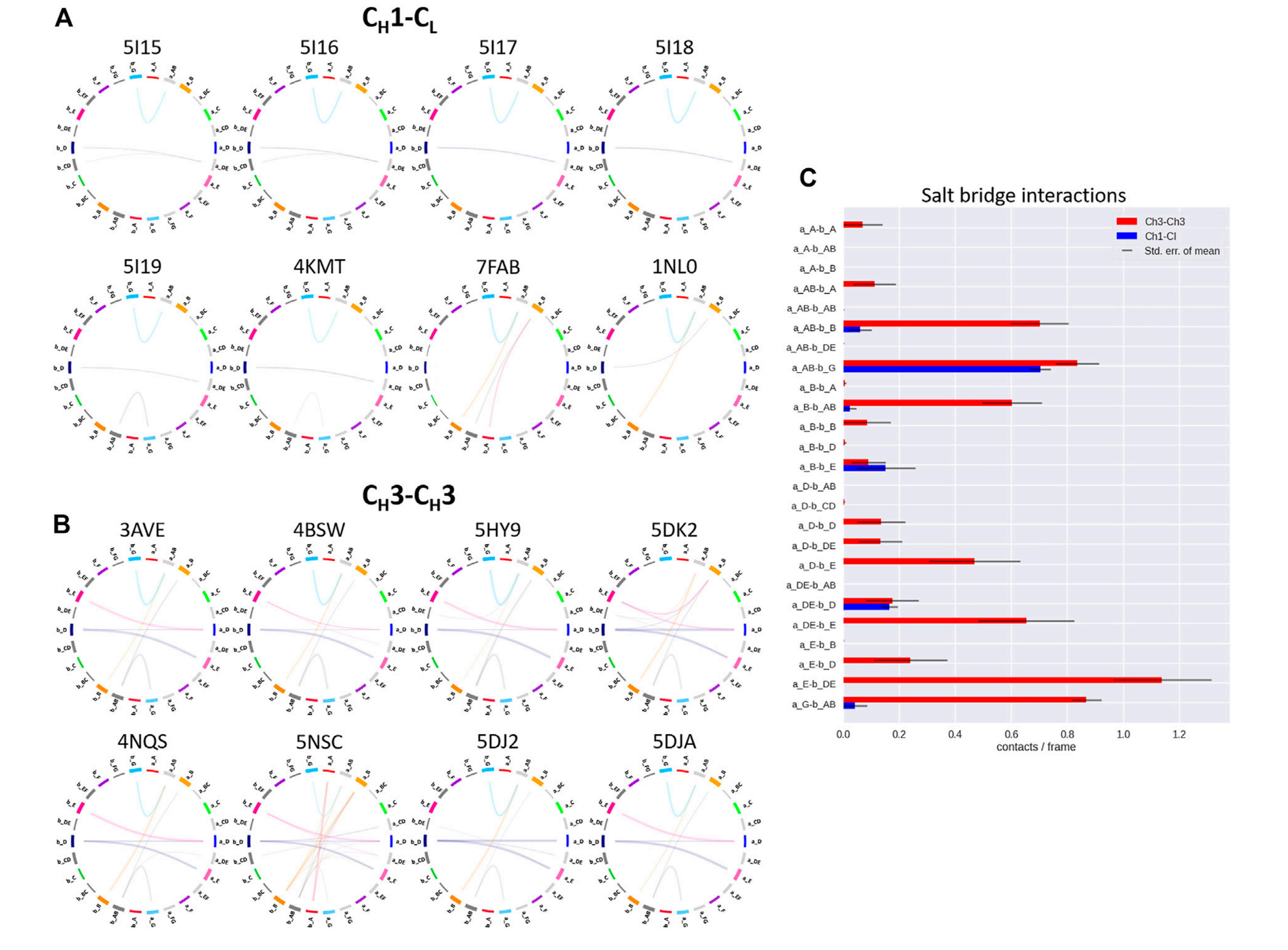


FIGURE 4 | Exemplary coarse-grained flareplots showing the salt bridge interactions formed between the different interdomain β-strands and loops of both **(A)** C_H1-C_L and **(B)** C_H3-C_H3 domains. **(C)** Bar plots quantitatively depicting differences in per strand/loop salt bridge interactions. We compare the two interface classes, i.e., C_H1-C_L (blue) and C_H3-C_H3 (red). Thus, we show averages and standard errors of the mean of all investigated antibodies within the respective class.

strands. Additionally, also strong salt bridge interactions can be observed for both the KE and EK variants between a_B – b_E strands, which cannot be observed in the third variant (**Supplementary Figure S4C**). Differences can also be observed in the hydrophobic contacts between the three engineered C_H1-C_L variants. While hydrophobic contacts between a_A – b_A and a_A – b_B are present in all three variants, the third variant has long-lasting contacts between the a_E – b_E strands (**Supplementary Figure S4C**). Additionally, the two charge optimized C_H1-C_L domains make strong hydrophobic interactions between the a_B – b_D and a_B – b_E strands (**Supplementary Figures S4A,B**). Comparing C_H1-C_L variants with C_H3-C_H3 domains, we find that the EK and KE C_H1-C_L variants (**Supplementary Figures S4A,B**) are able to form salt bridges between the a_B – b_E strands, which we only identified in C_H3-C_H3 domains before and not in other investigated C_H1-C_L domains. The hydrophobic interactions of the KiH designed C_H1-C_L domain (**Supplementary Figure S4C**) also show C_H3-C_H3 specific interactions between a_E – b_E strands, while the EK and KE variants show hydrophobic interaction patterns which are present in both C_H3-C_H3 and C_H1-C_L domains. Panels d–f in **Supplementary Figure S4** illustrate the interdomain interactions of three engineered C_H3-C_H3 variants, which are part of bispecific antibody matrices generated by Format Chain Exchange (FORCE), which enables the screening of the combinatorial format spaces (Dengl et al., 2020). These variants were originally designed by further modifying the 5HY9 KiH structure, which already differs from the 4NQS KiH structure by an additional intermolecular disulfide bridge.

In **Figure 7** we show three exemplary C_H1-C_L and three exemplary C_H3-C_H3 interfaces color-coded according to the number of interdomain salt bridge interactions. To facilitate the visualization of interface interactions, we flip the C_H3 domain A and the C_L domain. In line with the results presented in **Figure 6**, we find that the C_H3-C_H3 interface is dominated by salt bridge interactions, while the C_H1-C_L interface reveals a substantially lower number of ionic interactions,

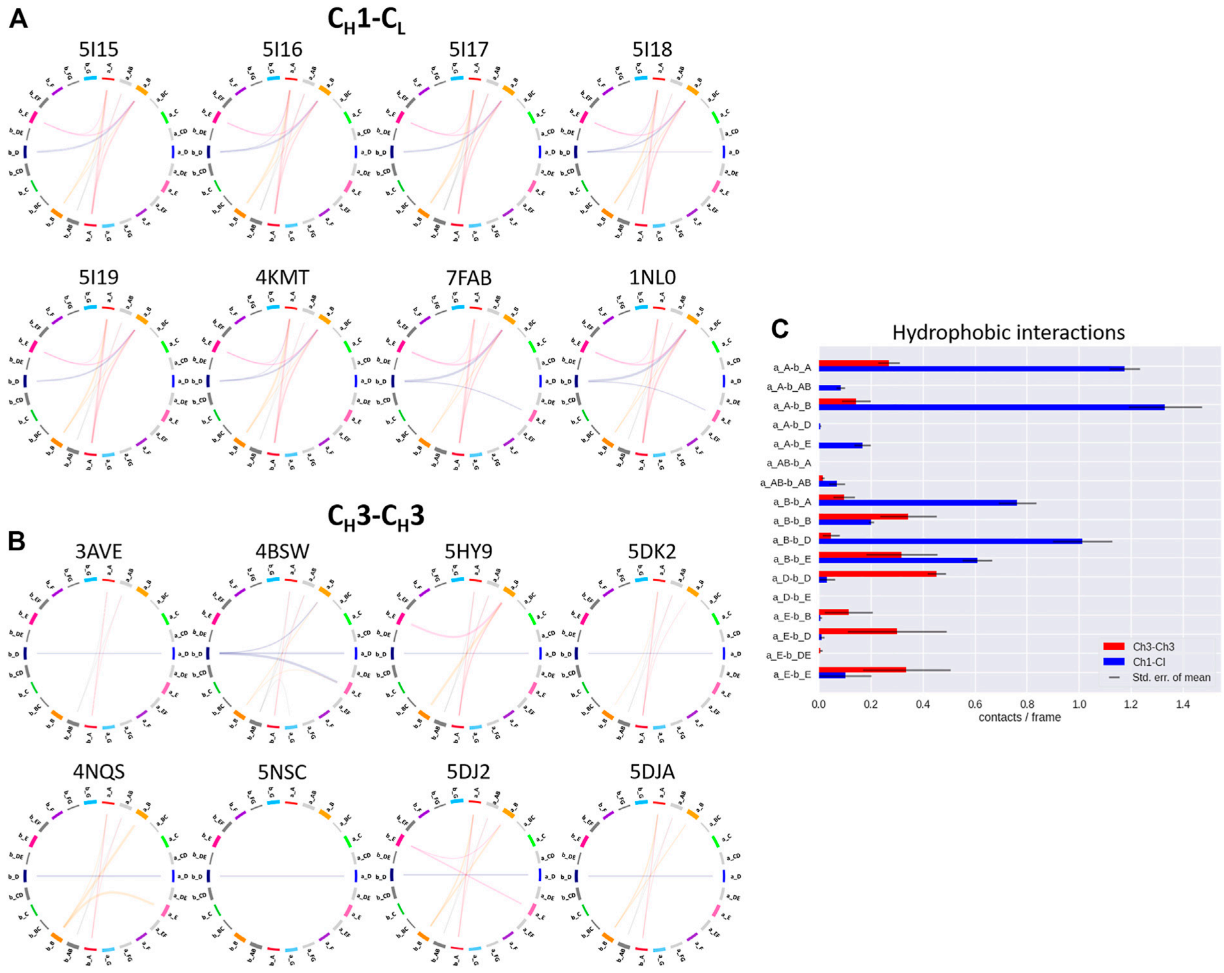


FIGURE 5 | Exemplary coarse-grained flareplots showing the hydrophobic interactions formed between the different interdomain β-strands and loops of both **(A)** C_H1-C_L and **(B)** C_H3-C_H3 domains. **(C)** Bar plots quantitatively depicting differences in per strand/loop hydrophobic interactions. We compare the two interface classes, i.e., C_H1-C_L (blue) and C_H3-C_H3 (red). Thus, we show averages and standard errors of the mean of all investigated antibodies within the respective class.

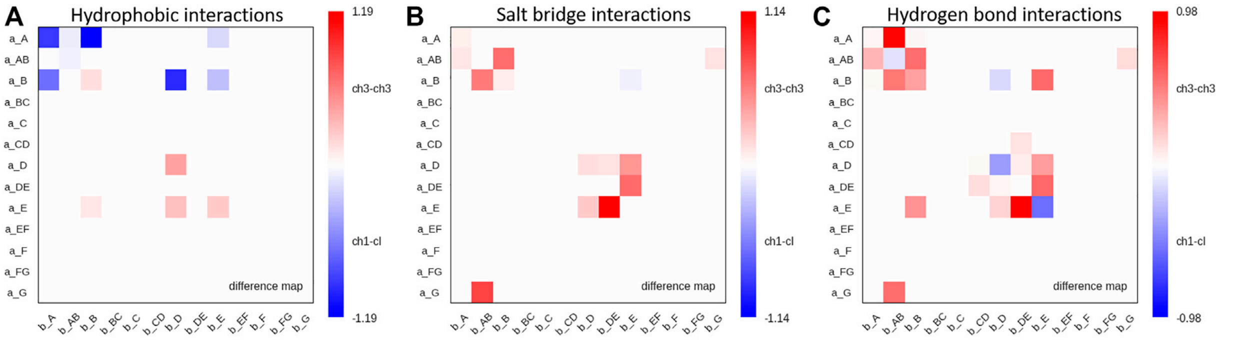


FIGURE 6 | Maps depicting the differences in hydrophobic interactions, salt bridges and hydrogen bonds between C_H3-C_H3 and C_H1-C_L domains. **(A)** Difference in hydrophobic interactions between all investigated C_H3-C_H3 and C_H1-C_L domains (**Supplementary Table S1**) based on the previously defined coarse graining of the residues belonging to the same loops or β-strands. We normalized the colorbar according to the most frequent contact in either of the two interface classes. **(B)** Difference in salt bridge interactions between all investigated C_H3-C_H3 and C_H1-C_L interfaces, showing the substantially higher number of salt bridge interactions dominating the C_H3-C_H3 interface. **(C)** Hydrogen bond difference maps for all investigated C_H3-C_H3 and C_H1-C_L interfaces.

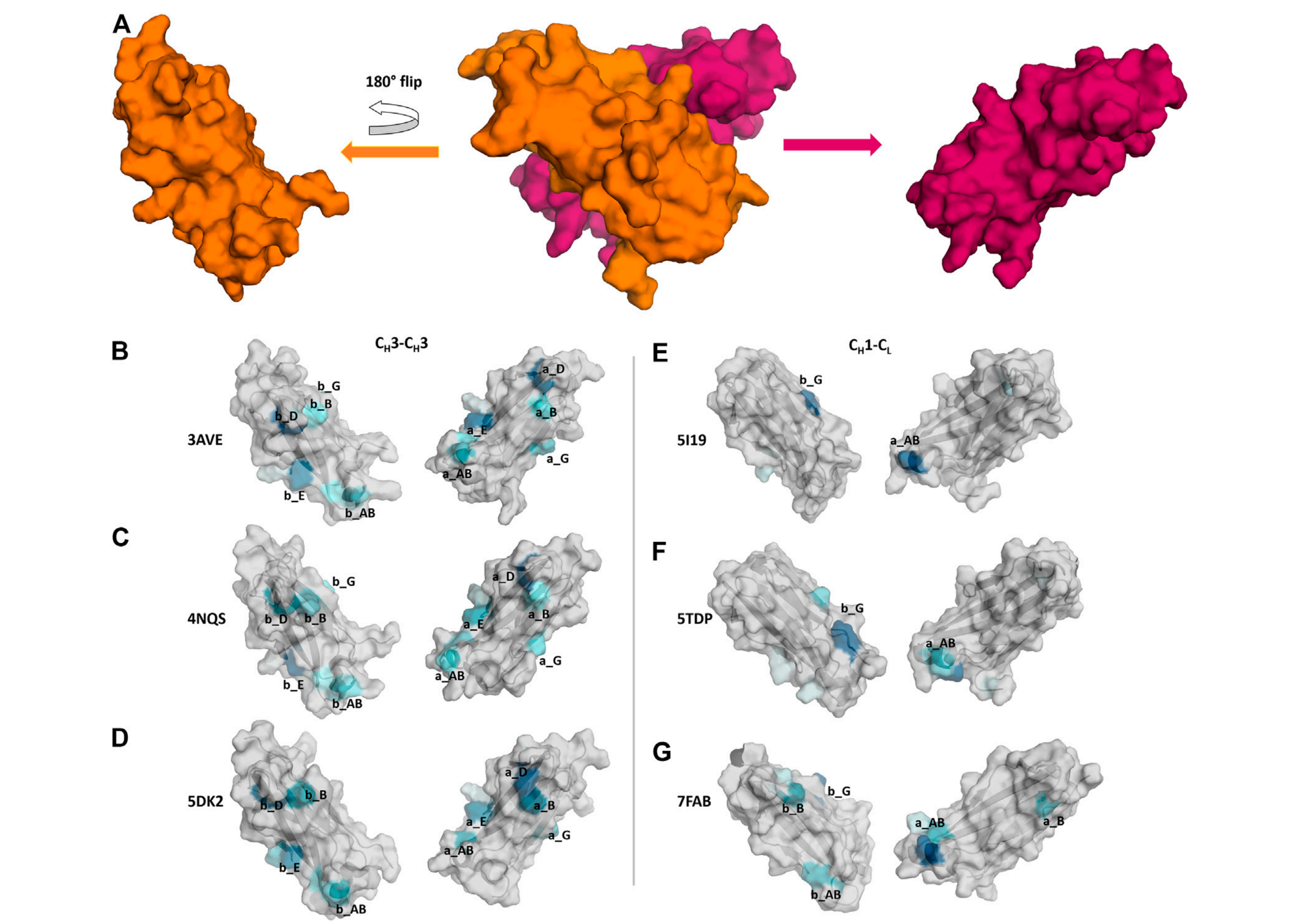


FIGURE 7 | Comparison of C_H3-C_H3 and C_H1-C_L interface interaction patterns by analyzing their salt bridge interdomain interactions. **(A)** Stepwise illustration of the workflow to obtain the "open-book" representation (PDB: 3AVE). **(B–G)** Each individual domain is gradually colored based on the number and duration of interdomain interactions. The color-gradient (grey to blue) corresponds to the number of interdomain salt bridges each residue is forming (the higher the number of contacts, the darker are the shades of blue).

precisely the C_H1-C_L reveals one characteristic salt bridge between loop a_AB and β-strand b_G.

Structural C_H3-C_H3 Interface Characterization

Apart from identifying differences in interface interaction patterns between the structurally highly similar C_H1-C_L and C_H3-C_H3 interfaces, we provide in **Figure 8** an overview of the main interactions stabilizing the homo-and-heterodimeric C_H3-C_H3 interfaces (wildtype, KiH and charge inversion). Already from the panels in **Figure 8** the unique and well-defined organization of the C_H3-C_H3 interface becomes apparent. Together with the hydrophobic core interactions (shown in green), various salt bridge interactions located at the N-terminal and C-terminal charge cluster (highlighted in pink) contribute to the stabilization of the dimeric interface. To characterize interactions and to identify residues that are critical for the interface formation, we analysed the investigated C_H3-C_H3 homo-and-heterodimer simulations in-detail. We find that the interactions in the core of the interface are particularly important for stabilization and formation of the dimer. One of these crucial interactions is the stacking interaction between residues Y407-Y407, which are present in all frames of the simulation in the variants with both interaction partners present (highlighted in **Figure 8**). We observe that especially mutations at the centre of the interface have a strong influence on the hydrophobic and salt bridge interaction network of the whole interface. One example would be the DE-KK variant (PDB accession code: 5NSC) (De Nardis et al., 2017), which introduces two ion pair interactions into the hydrophobic core by substituting L351D and L368E in one domain and L351K and T366K in the other. Even though these introduced residues strongly interact with each other, the mutations result in a change of the overall interdomain interaction patterns, which also differ from all other engineered variants. Particularly interesting is, that this DE-KK variant has the highest variability in the interdomain orientations (dC, AB, AC1, AC2, BC1, BC2) compared to all other investigated variants

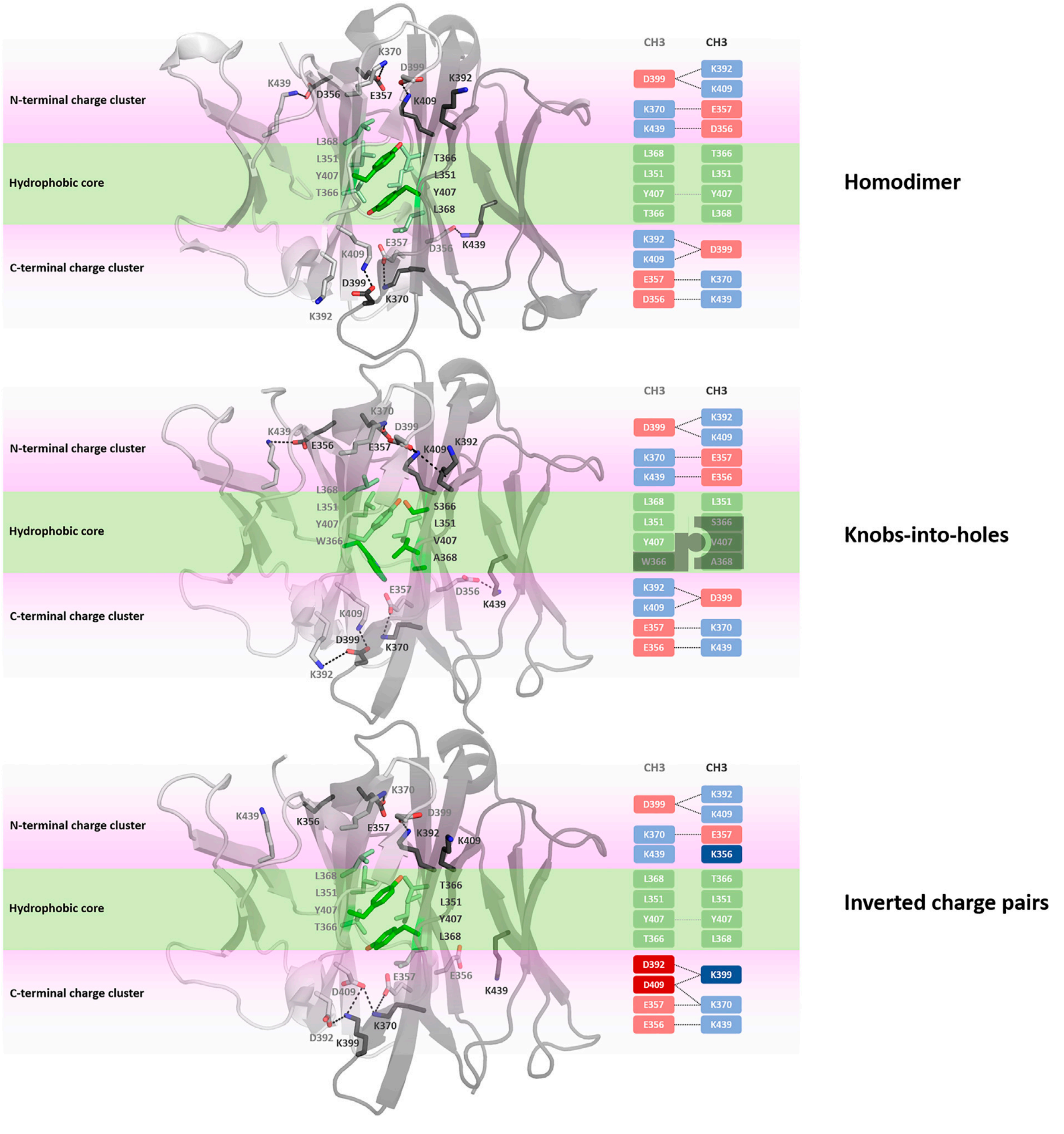


FIGURE 8 | Interface interaction analysis of the homodimer and two heterodimers following different design strategies, i.e., knobs-into-holes (PDB: 4NQS) and charge inversion (PDB: 5DK2). The C_H3-C_H3 dimers consist of a hydrophobic core at the centre of the C_H3-C_H3 interface (illustrated in palegreen) and two highly charged regions (N-terminal charge cluster and C-terminal charge cluster), shown in light pink, that stabilize the interface.

(**Figure 2B**). It also shows a slightly higher distance (dc) between the two domains and bigger variations in the tilt and bend angles, allowing also water molecules to interact with the N-terminal and C-terminal charge clusters. We also find similar results for the DD-KK variant (PDB accession code: 5DK2). The main difference between the DE-KK and the DD-KK variant is the location of the mutations. While the DE-KK disrupts the hydrophobic core interactions at the centre of the interface, the DD-KK variant introduces substitutions in the N-terminal charge cluster and C-terminal charge cluster. Introducing charge reversions in the charge clusters in this example results in an imbalance of positive and negative charges in the respective domains and, i.e., five negative charges in domain A, six positive charges in domain B. In particular the E356K mutation additionally results in a loss of a critical salt bridge interaction situated at the N-terminal charge cluster, which consequently shifts the interdomain tilt angles AC1 and BC1 and thereby increases the conformational variability in the interface. In line with these findings, we observe an increase in flexibility of the core interface residues for the KiH (PDB: 4NQS) and the charge inversion variants (PDB: 5DK2, 5NSC), which is reflected in higher root-mean-square-fluctuation (RMSF) values, compared to the homodimer (**Supplementary Figure S7**).

DISCUSSION

The idea of modifying antibody interfaces to reduce the risk of random assembly of different chains has motivated numerous studies to find variations of the proposed KiH approach, e.g., introducing charge pairs. Inverted charge interactions, instead of steric KiH interactions, were used for example for the design of the C_H3-C_H3 heterodimer DD-KK (K409D, K392D-D399K, E356K) variant (PDB accession code: 5DK2) (Ha et al., 2016). Also, the combination of the KiH interactions with the introduction of charge mutations have been presented in the C_H3-C_H3 heterodimer EW-RVT (K360E, K409W – Q347R, K399V, F405T) variant (PDB accession code: 4X98) (Choi et al., 2015).

One of the most frequent interactions situated in the centre of the homodimeric C_H3-C_H3 interface is the Y407-Y407 pi-stacking contact, residing in the central part of the E strands (**Figure 8**). (Dall'Acqua et al., 1998) Mutational studies confirmed the importance of these residues for the formation of the homodimeric interface. The salt bridge interactions at the N-terminal charge cluster and the C-terminal charge cluster (**Figure 8**) determine the characteristic interaction profile of the C_H3-C_H3 interface and substantially stabilize the dimer. The hydrophobic core in the homodimeric C_H3-C_H3 interface is formed by contacts between residues F405, L368, L351, Y407 and T366. These hydrophobic interactions are often modified following the KiH strategy (Ridgway et al., 1996; Elliott et al., 2014; Kuglstatter et al., 2017). The KiH variant (PDB accession codes: 4NQS, 5HY9, 5DI8) contains a knob in one C_H3 domain (domain A) by mutating residue T366 to the bulkier amino acid tryptophane (**Figure 8**). Three other residues on the other C_H3 domain (domain B) are also exchanged to smaller residues (T366S, Y407A, L368V) to ensure hydrophobic and steric complementarity. The orientation and position of the introduced tryptophane residue, also called "knob," dominates the shape complementary between the two domains.

For the C_H3-C_H3 interfaces investigated in this study, we provide a sequence alignment showing the respective mutations including a classification of the underlying engineering strategies. To connect the sequence variations to our coarse-grained flareplots, we included our color-coded strand/loop definition in the alignment (**Supplementary Figures S5, S6**).

In our simulations of all different C_H3-C_H3 homo-and-heterodimers, we find that if both tyrosine residues are present, the pi-stacking interaction occurs in all frames of the simulation and contributes to stabilizing the interface. Additionally, Y407 forms a stabilizing and conserved hydrogen bond with T366, located in strand B, which occurs on average in 65% of the simulation time. Thus, as these Y407 residues form critical interactions, stabilizing the centre of the C_H3-C_H3 interface, mutating one of these residues can already prevent homodimerization (Ridgway et al., 1996; Von Kreudenstein et al., 2013). Additionally, we observe that introducing charge mutations/inversions at the hydrophobic core, can strongly influence the interface interaction network as shown for the DE-KK variant and result in a different interface formation, which can be accompanied by a decrease in stability. We find that this decrease in stability for the KiH (PDB: 4NQS) and the charge inversion variants (PDB: 5DK2, 5NSC), can result in a higher flexibility of the core interface residues, which is reflected in higher RMSF values (**Supplementary Figure S7**).

To compare the interaction patterns of the structurally highly similar C_H1-C_L and C_H3-C_H3 interfaces, we calculate coarse-grained interdomain interaction maps, which are visualized as flareplots and quantified as barplots. When comparing different C_H3-C_H3 interfaces we find a highly conserved salt bridge between two glutamate residues (E356/E357) located in the a_AB loop with the lysine (K439) located in the b_G strand. These interactions can also be found in the C_H1-C_L interfaces containing a λ light chain (PDB accession codes: 7FAB, 1NL0). Another critical conserved interdomain interaction among C_H1-C_L domains can be found between the a_DE loop and the b_D strand, which is unique for kappa light chain antibodies. Especially for the salt bridges and hydrophobic interactions the patterns between κ and λ light chains differ the most (**Figures 4A**, **5A**). Apart from the conserved contacts among all C_H1-C_L interfaces, salt bridges are formed between the a_AB loops and b_B strands for the λ light chain antibodies. Interestingly, these salt bridges between a_AB loops and b_B strands are actually present in all considered C_H3-C_H3 domains (**Figure 4B**). Furthermore, an additional hydrophobic interaction can be found for the λ light chain antibodies between the a_E - b_D strands, which again can also be found in the C_H3-C_H3 interface (**Figure 5C**). Astonishingly, we observe in **Figure 6** that the C_H3 dimer is not only primarily stabilized by hydrophobic interactions but actually dominated by strong electrostatic interactions. Our observation, that the C_H3-C_H3 domains have a substantially higher number of salt bridges and hydrogen bonds, can also be explained by very frequently occurring interactions between residues D399-K409, D399-K392, E356-K439 and E357-K370, which surround the hydrophobic core. The high number of salt bridge interactions in the C_H3-C_H3 interface are also reflected in the electrostatic interaction energies, which are substantially higher compared to the C_H1-C_L domains (**Supplementary Table S2**). However, there are high fluctuations in the electrostatic energies of the individual C_H3-C_H3 interfaces, which result from repairing salt bridge interactions between different residues. The C_H1-C_L interface on the other hand is formed by mainly hydrophobic contacts.

The difference in electrostatic interaction energy is also reflected in the findings presented in **Figure 7**, which show a comparison of three C_H3-C_H3 and three C_H1-C_L interfaces, illustrated as an "open-book" representation. The surfaces of the individual domains are color-coded according to the number of interdomain salt bridge interactions. We find substantial differences in the interface interaction patterns between the two interface classes. In particular, the C_H1-C_L interface is dominated by one salt bridge between the a_AB loop and the b_G strand (**Figure 7E**). **Figure 7F** shows an engineered and mutated C_H1-C_L interface, which contains mutations at the edge of the interface, which have been discussed to introduce more flexibility and indeed, we find more frequent switches in interdomain salt bridge interactions, which suggests a higher flexibility at the edge of the interface. In **Figure 7G** we depict interdomain salt bridge interactions of a C_H1-C_L interface containing a λ light chain. In agreement with the results in **Figure 4**, we find more salt bridge interactions in C_H1-C_L interface for λ light chain C_H1-C_L domains and thus observe similar interaction patterns compared to the C_H3-C_H3 interfaces.

Apart from a detailed characterization of the C_H3-C_H3 and C_H1-C_L interfaces, we also investigated the relative interdomain orientations during the simulations. In line with previous studies, we find that for the C_H1-C_L, as well as the C_H3-C_H3 domains, the majority of interdomain movements are surprisingly fast and can be captured in the low nanosecond timescale (Fernández-Quintero et al., 2020a; Fernández-Quintero et al., 2020b). Additionally, we observe for the investigated C_H1-C_L domains (both λ and κ) left shifted cHL angle distributions towards lower cHL angles with a broader spread angle in the angle ranges, compared to the C_H3-C_H3 domains. For one λ light chain antibody (PDB accession code: 1NL0) we even observe a substantially shifted angle distribution towards lower cHL angle ranges. This higher variability in these cHL angle distributions is not surprising considering the higher number of sequence variations that occur in C_H1-C_L domains, while the C_H3-C_H3 domains contain solely point mutations.

CONCLUSION

In conclusion, we present a systematic characterization and a structural comparison of different C_H1-C_L and C_H3-C_H3 domains. By using molecular dynamics simulations, we find substantial differences in interaction patterns of the structurally highly similar C_H1-C_L and C_H3-C_H3 interfaces. While C_H1-C_L interfaces are dominated by hydrophobic interactions, we find that the C_H3-C_H3 interfaces are stabilized by numerous salt bridge interactions surrounding the hydrophobic core. Furthermore, we provide quantitative contact maps comparing C_H1-C_L and C_H3-C_H3 domains and highlighting which strands are key determinants for their structural integrity. Apart from the comparison, we also mechanistically discuss different C_H3-C_H3 interface engineering strategies, which provide an extensive understanding of the C_H3-C_H3 interfaces and thereby advance the design of bispecific antibodies.

METHODS

Dataset

The investigated C_H1-C_L and C_H3-C_H3 X-ray structures were chosen to have a representative set of antibodies covering various challenges in antibody engineering and design, as they differ in light chain types and follow different design strategies to reduce the risk of mispairings (**Supplementary Table S1**). (Ha et al., 2016; Teplyakov et al., 2016; Dillon et al., 2017; Dengl et al., 2020) 23 crystal structures of heterodimeric C_H3 IgG1 mutants, as well as the corresponding wildtype were obtained from the PDB. The 23 mutants have been designed following different strategies: knobs-into-holes strategy, complementary electrostatic interactions, format chain exchange platform or by using Multistate Design (MSD), which is a computational sequence optimization tool.

Apart from the 23 C_H3-C_H3 domains, we also simulated 23 Fab crystal structures.

16 germline Fab crystal structures are from the same library (Teplyakov et al., 2016). We chose this dataset as it allows to systematically investigate the influence of different heavy and light chain pairings. The phage library is composed of 4 heavy chain germlines IGHV1-69 (H1-69), IGHV3-23 (H3-23), IGHV5-51 (H5-51) and IGHV3-53 (H5-53) and 4 light chain germlines (all κ) IGKV1-39 (L1-39), IGKV3-11 (L3-11), IGKV3-20 (L3-20) and IGKV4-1 (L4-1). These genes were selected based on the frequency of their use, their cognate canonical structures, which can recognize proteins and peptides and their ability to be expressed in bacteria. Additionally, we included three Fab fragments which were part of a study redesigning the Fab interfaces. Furthermore, we also investigated two λ light chain antibodies and two recently published DutaFab structures, which are characterized by their high stability and their ability to recognize two different antigens (Beckmann et al., 2021). Dual targeting (Duta) Fab molecules contain two independent and spatially separated binding sites within the CDR loops (H-side paratope and L-side paratope) that simultaneously allow to bind two target molecules at the same Fv.

MD Simulation Protocol

All X-ray structures were prepared in MOE (Molecular Operating Environment, Montreal, QC, Canada: 2019) (Chemical Computing Group, 2020) using the Protonate 3D (Labute, 2009) tool. With the tleap tool of the Amber Tools20 package, we explicitly bonded all existing disulphide bridges (**Supplementary Figure S8**) and placed the Fab and C_H3-C_H3 structures into cubic water boxes of TIP3P(Jorgensen et al., 1983) water molecules with a minimum wall distance to the protein of 10 Å (El Hage et al., 2018; Gapsys and de Groot, 2019). Parameters for all antibody simulations were derived from the AMBER force field 14SB (Cornell et al., 1995; Maier et al., 2015). To neutralize the charges, we used uniform background charges (Darden et al., 1993; Salomon-Ferrer et al., 2013; Hub et al., 2014). Each system was carefully equilibrated using a multistep equilibration protocol (Wallnoefer et al., 2010; Wallnoefer et al., 2011).

Molecular dynamics simulations were performed using pmemd.cuda in an NpT ensemble to be as close to the experimental conditions as possible and to obtain the correct density distributions of both protein and water. Bonds involving hydrogen atoms were restrained by applying the SHAKE algorithm (Miyamoto and Kollman, 1992), allowing a timestep of 2.0 fs. Atmospheric pressure of the system was preserved by weak coupling to an external bath using the

Berendsen algorithm (Berendsen et al., 1984). The Langevin thermostat was used to maintain the temperature at 300K during simulations (Adelman and Doll, 1976). The parameter file used to perform all MD simulations is provided at the end of the Supporting Information.

Contacts

To calculate contacts of both C_H1-C_L and C_H3-C_H3 interfaces we used the GetContacts software (Stanford University, adate). This tool can compute interactions within one protein structure, but also between different protein interfaces and allows to monitor the evolution of contacts during the simulation. The development of the contacts during a simulation can be visualized in so-called flareplots. For all available simulations (**Supplementary Table S1**) we calculated all different types of contacts, including hydrogen bonds (sidechain/sidechain, sidechain/backbone, backbone/backbone), salt bridges, hydrophobic and Van der Waals interactions. The contacts are determined based on the default geometrical criteria provided by GetContacts. To recognize interface patterns and to describe the dissociation mechanisms of both the C_H1-C_L and C_H3-C_H3 domains, we coarse grained residues belonging to the same loops or β-strands. The secondary structure assignment has been performed with STRIDE (Frishman and Argos, 1995; Heinig and Frishman, 2004). To quantitively identify systematic differences in the interface interactions of the two interface classes, we evaluated the frequency of different interaction types. Thus, we counted contacts (for each type of interaction) of certain structural elements, e.g., salt bridges between the strand a_A and the loop b_AB. Furthermore, we calculated mean contact frequencies (contact per frame) in the simulations and averaged these frequencies within the interface classes and compared the results. In addition, we quantified the standard error of the mean of these contact frequencies within these classes. This comparison enabled us to find contacts, which, e.g., exist in all the C_H1-C_L interfaces, but not in C_H3-C_H3 interfaces, or vice versa. Apart from visualizing and quantifying the contacts of both C_H1-C_L and C_H3-C_H3 interfaces, we also calculated the linear interaction energies (LIE) by using the LIE tool implemented in cpptraj (Roe and Cheatham, 2013). We calculated the electrostatic interaction energies for all frames of each simulation (10,000 frames/simulation) and provided the simulation-averages of these interaction energies in **Supplementary Table S2**.

Interdomain Orientation Calculations

While computational tools to fully characterize the Fv region of antibodies and TCRs are already available, no such tools were published for other immunoglobulin domain interfaces, such as the C_H3-C_H3 and the C_H1-C_L interface (Dunbar et al., 2013). The OCD approach (Hoerschinger et al., 2021) creates a suitable coordinate system for the characterization of these interfaces for any user-provided reference structure. This allows a straight-forward analysis without the significant demands on previous structural knowledge. Using this tool, a reference coordinate system is created based on user-defined reference structures consisting of an atomic structure and two domain selections over these atoms. To this end, the reference structure for each domain is generated by considering a center axis linking the two centers of mass of the different domains, and the first principal axis P of inertia of each domain corresponding to the lowest eigenvalue of the inertia tensor. Each individual domain is aligned to the world coordinate system by aligning this principal axis to the z unit vector and the center axis as close as possible to the x unit vector, yielding a reference structure for each domain. To map the coordinate system onto a sample structure, the references are aligned to the sample and the alignment transformations are applied to the xyz unit vectors. The transformed z vectors (A1/B1) and y vectors (A2/B2) as well as the center axis are then used to calculate six orientational measures: Two tilt angles for each vector towards the center axis (AC1, AC2, BC1, BC2), the length of the center axis (dC) and a torsion angle (AB) between the two intersecting planes composed of A1, the centre axis and B1. To better visualize the relative interdomain orientations we performed the Gaussian kernel density estimation (KDE) on the HL angles, to obtain the probability density distributions. To calculate the KDE we used the recently published implementation of KDE in C++ (Kraml et al., 2021). We used 10,000 frames of each MD simulation (1μs) to calculate and plot the relative interdomain orientations.

Relative V_H and V_L Orientations Using ABangle

ABangle is a computational tool (Dunbar et al., 2013; Bujotzek et al., 2015; Bujotzek et al., 2016; Fernández-Quintero et al., 2020b) to characterize the relative orientations between the antibody variable domains (V_H and V_L) using six measurements (five angles and a distance). A plane is projected on each of the two variable domains. Between these two planes, a distance vector C is defined. The six measures are then two tilt angles between each plane (HC1, HC2, LC1, LC2) and a torsion angle (HL) between the two planes along the distance vector C (dC). The ABangle script can calculate these measures for an arbitrary Fv region by aligning the consensus structures to the found core set positions and fitting the planes and distance vector from this alignment. This online available tool was combined with an in-house python script to reduce computational effort and to visualize our simulation data over time. The in-house script makes use of ANARCI(Dunbar and Deane, 2016) for fast local annotation of the Fv region and pytraj from the AmberTools package (Case et al., 2020) for rapid trajectory processing.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

MF and PQ performed research and wrote the manuscript. FW, CS, NP, KK, JL and VH analysed data. AB, GG, HK and KL advised and supervised the research. All authors contributed to writing the manuscript.

FUNDING

This work was supported by Austrian Science Fund (P30565, P30737, P34518 and DOC 30). Furthermore, this project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 764958. The computational results presented have been achieved in part using the Vienna Scientific Cluster (VSC). European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 847476 to NP.

ACKNOWLEDGMENTS

We acknowledge PRACE for awarding us access to Piz Daint at CSCS, Switzerland.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.812750/full#supplementary-material

REFERENCES

Adachi, M., Kurihara, Y., Nojima, H., Takeda-Shitaka, M., Kamiya, K., and Umeyama, H. (2003). Interaction between the Antigen and Antibody Is Controlled by the Constant Domains: normal Mode Dynamics of the HEL-HyHEL-10 Complex. *Protein Sci.* 12, 2125–2131. doi:10.1110/ps.03100803

Addis, P. W., Hall, C. J., Bruton, S., Veverka, V., Wilkinson, I. C., Muskett, F. W., et al. (2014). Conformational Heterogeneity in Antibody-Protein Antigen Recognition: Implications for High Affinity Protein Complex Formation. *J. Biol. Chem.* 289, 7200–7210. doi:10.1074/jbc.m113.492215

Adelman, S. A., and Doll, J. D. (1976). Generalized Langevin Equation Approach for Atom/solid-Surface Scattering: General Formulation for Classical Scattering off Harmonic Solids. *J. Chem. Phys.* 64, 2375–2388. doi:10.1063/1.432526

Beckmann, R., Jensen, K., Fenn, S., Speck, J., Krause, K., Meier, A., et al. (2021). DutaFabs Are Engineered Therapeutic Fab Fragments that Can Bind Two Targets Simultaneously. *Nat. Commun.* 12, 708. doi:10.1038/s41467-021-20949-3

Berendsen, H. J. C., van Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984). Molecular-Dynamics with Coupling to an External Bath. *J. Chem. Phys.* 81, 3684. doi:10.1063/1.448118

Bönisch, M., Sellmann, C., Maresch, D., Halbig, C., Becker, S., Toleikis, L., et al. (2017). Novel CH1:CL Interfaces that Enhance Correct Light Chain Pairing in Heterodimeric Bispecific Antibodies. *Protein Eng. Des. Selection* 30, 685–696. doi:10.1093/protein/gzx044

Brinkmann, U., and Kontermann, R. E. (2017). The Making of Bispecific Antibodies. *mAbs* 9, 182–212. doi:10.1080/19420862.2016.1268307

Bujotzek, A., Dunbar, J., Lipsmeier, F., Schäfer, W., Antes, I., Deane, C. M., et al. (2015). Prediction of VH-VL Domain Orientation for Antibody Variable Domain Modeling. *Proteins* 83, 681–695. doi:10.1002/prot.24756

Bujotzek, A., Lipsmeier, F., Harris, S. F., Benz, J., Kuglstatter, A., and Georges, G. (2016). VH-VL Orientation Prediction for Antibody Humanization Candidate Selection: A Case Study. *mAbs* 8, 288–305. doi:10.1080/19420862.2015.1117720

Case, D. A., Belfon, K., Ben-Shalom, I. Y., Brozell, S. R., Cerutti, D. S., Cheatham, T. E., III, et al. (2020). *AMBER 2020*. San Francisco: University of California.

Chemical Computing Group (2020). *Molecular Operating Environment (MOE). 1010 Sherbrooke St. West, Suite #910*. Montreal, QC, Canada: Chemical Computing Group, H3A. 2R7.

Chiu, M. L., Goulet, D. R., Teplyakov, A., and Gilliland, G. L. (2019). Antibody Structure and Function: The Basis for Engineering Therapeutics. *Antibodies* 8, 55. doi:10.3390/antib8040055

Choi, H.-J., Seok, S.-H., Kim, Y.-J., Seo, M.-D., and Kim, Y.-S. (2015). Crystal Structures of Immunoglobulin Fc Heterodimers Reveal the Molecular Basis for Heterodimer Formation. *Mol. Immunol.* 65, 377–383. doi:10.1016/j.molimm.2015.02.017

Colman, P. M. (1988). "Structure of Antibody-Antigen Complexes: Implications for Immune Recognition," in *Advances in Immunology*. Editor F. J. Dixon (Cambridge, UK: Academic Press). doi:10.1016/s0065-2776(08)60364-8

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., et al. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* 117, 5179–5197. doi:10.1021/ja00124a002

Dall'Acqua, W., Simon, A. L., Mulkerrin, M. G., and Carter, P. (1998). Contribution of Domain Interface Residues to the Stability of Antibody CH3 Domain Homodimers. *Biochemistry* 37, 9266–9273.

Darden, T., York, D., and Pedersen, L. (1993). Particle Mesh Ewald: AnN·Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* 98, 10089–10092. doi:10.1063/1.464397

Davies, D. R., and Chacko, S. (1993). Antibody Structure. *Acc. Chem. Res.* 26, 421–427. doi:10.1021/ar00032a005

De Nardis, C., Hendriks, L. J. A., Poirier, E., Arvinte, T., Gros, P., Bakker, A. B. H., et al. (2017). A New Approach for Generating Bispecific Antibodies Based on a Common Light Chain Format and the Stable Architecture of Human Immunoglobulin G1. *J. Biol. Chem.* 292, 14706–14717. doi:10.1074/jbc.m117.793497

Dengl, S., Mayer, K., Bormann, F., Duerr, H., Hoffmann, E., Nussbaum, B., et al. (2020). Format Chain Exchange (FORCE) for High-Throughput Generation of Bispecific Antibodies in Combinatorial Binder-Format Matrices. *Nat. Commun.* 11, 4974. doi:10.1038/s41467-020-18477-7

Dillon, M., Yin, Y., Zhou, J., McCarty, L., Ellerman, D., Slaga, D., et al. (2017). Efficient Production of Bispecific IgG of Different Isotypes and Species of Origin in Single Mammalian Cells. *mAbs* 9, 213–230. doi:10.1080/19420862.2016.1267089

Dunbar, J., and Deane, C. M. (2016). ANARCI: Antigen Receptor Numbering and Receptor Classification. *Bioinformatics* 32, 298–300. doi:10.1093/bioinformatics/btv552

Dunbar, J., Fuchs, A., Shi, J., and Deane, C. M. (2013). ABangle: Characterising the VH-VL Orientation in Antibodies. *Protein Eng. Des. Selection* 26, 611–620. doi:10.1093/protein/gzt020

El Hage, K., Hédin, F., Gupta, P. K., Meuwly, M., and Karplus, M. (2018). Valid Molecular Dynamics Simulations of Human Hemoglobin Require a Surprisingly Large Box Size. *eLife* 7, e35560. doi:10.7554/eLife.35560

Elliott, J. M., Ultsch, M., Lee, J., Tong, R., Takeda, K., Spiess, C., et al. (2014). Antiparallel Conformation of Knob and Hole Aglycosylated Half-Antibody Homodimers Is Mediated by a CH2-CH3 Hydrophobic Interaction. *J. Mol. Biol.* 426, 1947–1957. doi:10.1016/j.jmb.2014.02.015

Feige, M. J., Grawert, M. A., Marcinowski, M., Hennig, J., Behnke, J., Auslander, D., et al. (2014). The Structural Analysis of Shark IgNAR Antibodies Reveals Evolutionary Principles of Immunoglobulins. *Proc. Natl. Acad. Sci.* 111, 8155–8160. doi:10.1073/pnas.1321502111

Fernández-Quintero, M. L., Hoerschinger, V. J., Lamp, L. M., Bujotzek, A., Georges, G., and Liedl, K. R. (2020). VH-VL Interdomain Dynamics Observed by Computer Simulations and NMR. *Proteins: Struct. Funct. Bioinformatics* 88, 830–839. doi:10.1002/prot.25872

Fernández-Quintero, M. L., Kroell, K. B., Heiss, M. C., Loeffler, J. R., Quoika, P. K., Waibl, F., Bujotzek, A., et al. (2020). Surprisingly Fast Interface and Elbow

Angle Dynamics of Antigen-Binding Fragments. *Front. Mol. Biosciences* 7, 339. doi:10.3389/fmolb.2020.609088

Fernández-Quintero, M. L., Loeffler, J. R., Waibl, F., Kamenik, A. S., Hofer, F., and Liedl, K. R. (2020c). Conformational Selection of Allergen-Antibody Complexes—Surface Plasticity of Paratopes and Epitopes. *Protein Eng. Des. Selection* 32, 513–523. doi:10.1093/protein/gzaa014

Frishman, D., and Argos, P. (1995). Knowledge-based Protein Secondary Structure Assignment. *Proteins* 23, 566–579. doi:10.1002/prot.340230412

Fudenberg, H. H., Drews, G., and Nisonoff, A. (1964). Serologic Demonstration of Dual Specificity of Rabbit Bivalent Hybrid Antibody. *J. Exp. Med.* 119, 151–166. doi:10.1084/jem.119.1.151

Gapsys, V., and de Groot, B. L. (2019). Comment on 'Valid Molecular Dynamics Simulations of Human Hemoglobin Require a Surprisingly Large Box Size'. *Elife* 8, 563064. doi:10.7554/eLife.44718

Ha, J.-H., Kim, J.-E., and Kim, Y.-S. (2016). Immunoglobulin Fc Heterodimer Platform Technology: From Design to Applications in Therapeutic Antibodies and Proteins. *Front. Immunol.* 7, 394. doi:10.3389/fimmu.2016.00394

Heinig, M., and Frishman, D. (2004). STRIDE: a Web Server for Secondary Structure Assignment from Known Atomic Coordinates of Proteins. *Nucleic Acids Res.* 32, W500–W502. doi:10.1093/nar/gkh429

Hoerschinger, V. J., Fernández-Quintero, M. L., Waibl, F., Kraml, J., Bujotzek, A., Georges, G., et al. (2021). OCD.py - Characterizing Immunoglobulin Interdomain Orientations. *bioRxiv 2021* 03 (15), 435379. doi:10.1101/2021.03.15.435379

Hub, J. S., de Groot, B. L., Grubmüller, H., and Groenhof, G. (2014). Quantifying Artifacts in Ewald Simulations of Inhomogeneous Systems with a Net Charge. *J. Chem. Theor. Comput.* 10, 381–390. doi:10.1021/ct400626b

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* 79, 926–935. doi:10.1063/1.445869

Jost Lopez, A., Quoika, P. K., Linke, M., Hummer, G., and Köfinger, J. (2020). Quantifying Protein-Protein Interactions in Molecular Simulations. *J. Phys. Chem. B* 124, 4673–4685. doi:10.1021/acs.jpcb.9b11802

Kaplon, H., Muralidharan, M., Schneider, Z., and Reichert, J. M. (2020). Antibodies to Watch in 2020. *mAbs* 12, 1703531. doi:10.1080/19420862.2019.1703531

Kaplon, H., and Reichert, J. M. (2021). Antibodies to Watch in 2021. *mAbs* 13, 1860476. doi:10.1080/19420862.2020.1860476

Kraml, J., Hofer, F., Quoika, P. K., Kamenik, A. S., and Liedl, K. R. (2021). X-entropy: A Parallelized Kernel Density Estimator with Automated Bandwidth Selection to Calculate Entropy. *J. Chem. Inf. Model.* 61, 1533–1538. doi:10.1021/acs.jcim.0c01375

Kuglstatter, A., Stihle, M., Neumann, C., Müller, C., Schaefer, W., Klein, C., et al. Roche Pharmaceutical Research and Early Development (2017). Structural Differences between Glycosylated, Disulfide-Linked Heterodimeric Knob-Into-Hole Fc Fragment and its Homodimeric Knob–Knob and Hole–Hole Side Products. *Protein Eng. Des. Selection* 30, 649–656. doi:10.1093/protein/gzx041

Labute, P. (2009). Protonate3D: Assignment of Ionization States and Hydrogen Coordinates to Macromolecular Structures. *Proteins* 75, 187–205. doi:10.1002/prot.22234

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theor. Comput.* 11, 3696–3713. doi:10.1021/acs.jctc.5b00255

Miyamoto, S., and Kollman, P. A. (1992). Settle: An Analytical Version of the SHAKE and RATTLE Algorithm for Rigid Water Models. *J. Comput. Chem.* 13, 952–962. doi:10.1002/jcc.540130805

Moore, G. L., Bernett, M. J., Rashid, R., Pong, E. W., Pong, D.-H. T., Jacinto, J., et al. (2019). A Robust Heterodimeric Fc Platform Engineered for Efficient Development of Bispecific Antibodies of Multiple Formats. *Methods* 154, 38–50. doi:10.1016/j.ymeth.2018.10.006

Nisonoff, A., and Rivers, M. M. (1961). Recombination of a Mixture of Univalent Antibody Fragments of Different Specificity. *Arch. Biochem. Biophys.* 93, 460–462. doi:10.1016/0003-9861(61)90296-x

Regula, J. T., Imhof-Jung, S., Mølhøj, M., Benz, J., Ehler, A., Bujotzek, A., et al. (2018). Variable Heavy-Variable Light Domain and Fab-Arm CrossMabs with Charged Residue Exchanges to Enforce Correct Light Chain Assembly. *Protein Eng. Des. Selection* 31, 289–299. doi:10.1093/protein/gzy021

Ridgway, J. B. B., Presta, L. G., and Carter, P. (1996). 'Knobs-into-holes' Engineering of Antibody CH3 Domains for Heavy Chain Heterodimerization. *Protein Eng. Des. Sel* 9, 617–621. doi:10.1093/protein/9.7.617

Roe, D. R., and Cheatham, T. E. (2013). PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theor. Comput.* 9, 3084–3095. doi:10.1021/ct400341p

Rose, R. J., van Berkel, P. H. C., van den BremerLabrijn, E. T. J., Labrijn, A. F., Vink, T., Schuurman, J., et al. (2013). Mutation of Y407 in the CH3 Domain Dramatically Alters Glycosylation and Structure of Human IgG. *MAbs* 5, 219–228. doi:10.4161/mabs.23532

Röthlisberger, D., Honegger, A., and Plückthun, A. (2005). Domain Interactions in the Fab Fragment: A Comparative Evaluation of the Single-Chain Fv and Fab Format Engineered with Variable Domains of Different Stability. *J. Mol. Biol.* 347, 773–789. doi:10.1016/j.jmb.2005.01.053

Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S., and Walker, R. C. (2013). Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theor. Comput.* 9, 3878–3888. doi:10.1021/ct400314y

Sedykh, S., Prinz, V., Buneva, V., and Nevinsky, G. (2018). Bispecific Antibodies: Design, Therapy, Perspectives. *Dddt* 12, 195–208. doi:10.2147/dddt.s151282

Stanfield, R. L., Zemla, A., Wilson, I. A., and Rupp, B. (2006). Antibody Elbow Angles Are Influenced by Their Light Chain Class. *J. Mol. Biol.* 357, 1566–1574. doi:10.1016/j.jmb.2006.01.023

Stanford University (adate). GetContacts. Available at: https://getcontacts.github.io/

Teplyakov, A., Obmolova, G., Malia, T. J., Luo, J., Muzammil, S., Sweet, R., et al. (2016). Structural Diversity in a Human Antibody Germline Library. *mAbs* 8, 1045–1063. doi:10.1080/19420862.2016.1190060

Teplyakov, A., Zhao, Y., Malia, T. J., Obmolova, G., and Gilliland, G. L. (2013). IgG2 Fc Structure and the Dynamic Features of the IgG CH2-CH3 Interface. *Mol. Immunol.* 56, 131–139. doi:10.1016/j.molimm.2013.03.018

Vanhove, M., Usherwood, Y.-K., and Hendershot, L. M. (2001). Unassembled Ig Heavy Chains Do Not Cycle from BiP *In Vivo* but Require Light Chains to Trigger Their Release. *Immunity* 15, 105–114. doi:10.1016/s1074-7613(01)00163-7

Von Kreudenstein, T. S., Escobar-Carbrera, E., Lario, P. I., D'Angelo, I., Brault, K., Kelly, J. F., et al. (2013). Improving Biophysical Properties of a Bispecific Antibody Scaffold to Aid Developability. *mAbs* 5, 646–654. doi:10.4161/mabs.25632

Wallnoefer, H. G., Handschuh, S., Liedl, K. R., and Fox, T. (2010). Stabilizing of a Globular Protein by a Highly Complex Water Network: A Molecular Dynamics Simulation Study on Factor Xa. *J. Phys. Chem. B* 114, 7405–7412. doi:10.1021/jp101654g

Wallnoefer, H. G., Liedl, K. R., and Fox, T. (2011). A Challenging System: Free Energy Prediction for Factor Xa. *J. Comput. Chem.* 32, 1743–1752. doi:10.1002/jcc.21758

Conflict of Interest: Authors AB, GG and HK were employed by the company Roche.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Molecular Biosciences

TYPE Technology and Code
PUBLISHED 09 May 2023
DOI 10.3389/fmolb.2023.1178035

Check for updates

OPEN ACCESS

EDITED BY
Serdal Kirmizialtin,
New York University Abu Dhabi, United Arab Emirates

REVIEWED BY
Jian Wang,
The Pennsylvania State University, United States
Jose A. Gavira,
Spanish National Research Council (CSIC), Spain

*CORRESPONDENCE
Georg Künze,
✉ georg.kuenze@uni-leipzig.de

RECEIVED 02 March 2023
ACCEPTED 26 April 2023
PUBLISHED 09 May 2023

CITATION
Engelberger F, Zakary JD and Künze G (2023), Guiding protein design choices by per-residue energy breakdown analysis with an interactive web application. *Front. Mol. Biosci.* 10:1178035.
doi: 10.3389/fmolb.2023.1178035

Guiding protein design choices by per-residue energy breakdown analysis with an interactive web application

Felipe Engelberger, Jonathan D. Zakary and Georg Künze*

Institute for Drug Discovery, Leipzig University, Leipzig, Germany

Recent developments in machine learning have greatly facilitated the design of proteins with improved properties. However, accurately assessing the contributions of an individual or multiple amino acid mutations to overall protein stability to select the most promising mutants remains a challenge. Knowing the specific types of amino acid interactions that improve energetic stability is crucial for finding favorable combinations of mutations and deciding which mutants to test experimentally. In this work, we present an interactive workflow for assessing the energetic contributions of single and multi-mutant designs of proteins. The energy breakdown guided protein design (ENDURE) workflow includes several key algorithms, including per-residue energy analysis and the sum of interaction energies calculations, which are performed using the Rosetta energy function, as well as a residue depth analysis, which enables tracking the energetic contributions of mutations occurring in different spatial layers of the protein structure. ENDURE is available as a web application that integrates easy-to-read summary reports and interactive visualizations of the automated energy calculations and helps users selecting protein mutants for further experimental characterization. We demonstrate the effectiveness of the tool in identifying the mutations in a designed polyethylene terephthalate (PET)-degrading enzyme that add up to an improved thermodynamic stability. We expect that ENDURE can be a valuable resource for researchers and practitioners working in the field of protein design and optimization. ENDURE is freely available for academic use at: http://endure.kuenzelab.org.

KEYWORDS

protein design, energy calculation, amino acid interaction, web application (app), machine learning

1 Introduction

The design of proteins with improved stability and activity is a critical aspect of research in biotechnology and related fields. It holds the potential to revolutionize a wide range of applications (Arnold, 2018), from developing enzymes for industrial processes (Chen and Arnold, 2020), antibodies and antivirals for medicine (Sevy and Meiler, 2014; Willis et al., 2015) to molecular switches and biosensors (Stein and Alexandrov, 2015; Quijano-Rubio et al., 2021). Protein design has been demonstrated to play a crucial role in many different areas of biotechnology (Castro et al., 2022; Habibi et al., 2022; Reetz, 2022).

Two widely used protein design approaches are directed evolution and computer-aided protein design. The former approach mimics the natural gene diversification and selection process and involves iterative rounds of mutagenesis, which create a library of mutants, and selection of mutants with desired functions (Arnold, 2018). Computer-aided protein design typically involves algorithms that suggest mutations for experimental testing (Pan and Kortemme, 2021). These algorithms may be based on in-depth molecular modeling, e.g., with the Rosetta software suite (Leman et al., 2020), or machine learning predictions (Dauparas et al., 2022). The experimental testing of the designed proteins is time-, cost-, and labor-intensive. Thus, prioritizing the most probable candidates for experimental testing is necessary. To facilitate mutant selection, it can be informative to determine the specific types of amino acid interactions that contribute to protein stability and assess the energetic impact of mutations (Goldenzweig et al., 2016).

Machine learning algorithms have revolutionized the field of protein design, enabling researchers to generate novel proteins with improved properties more efficiently (Dauparas et al., 2022; Ferruz and Höcker, 2022). Two current state-of-the-art methods are the evolutionary and structure-based design method PROSS (Goldenzweig et al., 2016; Weinstein et al., 2021), and the deep learning method ProtMPNN (Dauparas et al., 2022). Both methods can yield tens to hundreds of candidate structures or sequences. Selecting the best candidates for experimental testing is a rather tricky task, particularly for designed proteins with multiple mutations, as the effect of each mutation is dependent upon the presence of other mutations—a phenomenon referred to as epistasis (Starr and Thornton, 2016). Yet the selection process determines the success of the overall design process. Thus, it is essential to have reliable, comprehensive, and easy-to-use methods for evaluating and selecting the most probable designs, based on the energetic magnitude and type of interactions (hydrogen bonds, salt bridges, etc.) introduced by the mutations.

Some existing tools for protein design and mutant selection include Rosetta (Leman et al., 2020), HotSpotWizard3.0 (Sumbalova et al., 2018), ProteinSolver (Strokach et al., 2021), and FoldX (Schymkowitz et al., 2005). Scoring functions and modeling algorithms from Rosetta were previously tested for the prediction of protein stability and affinity changes (ΔΔG). Kellogg et al. (2011) investigated the role of conformational sampling in computing mutation-induced changes in protein stability and compared the predictions to experimental ΔΔG values. Barlow et al. (2018) developed the Flex ddG method using Rosetta ensemble-based estimation of changes in protein-protein binding affinity upon mutation. Frenz et al. (2020) focused on improving the prediction of protein mutational free energy using Cartesian coordinate minimization. However, all these tools often require extensive knowledge of the software, leaving non-expert users without easy access to analyze the outcomes of their protein design experiments. In the field of *de novo* protein design, DE-STRESS (Stam and Wood, 2021) has been developed to help non-expert users evaluate the plausibility of such designs. Unfortunately, equivalent tools for assessing the results of sequence design experiments operated on a provided structure are lacking. To address this gap, we have developed ENergy breakDown gUided pRotein dEsign (ENDURE), a modular web application that provides an interactive and user-friendly interface for analyzing the energetic contributions of protein designs. ENDURE integrates easy-to-read summary tables and interactive visualizations of automated energy calculations, helping users to explore and reveal mutational hotspots—which confer stabilization or destabilization—and compare the specific types of interactions that a particular mutation is introducing. In that way, ENDURE helps selecting the best protein mutants for further experimental characterization.

The application workflow (Figure 1) integrates several key algorithms, which analyze the protein structure using the Rosetta energy function, including per-residue energy breakdown and the sum of interaction energies calculations. Additionally, the tool provides a residue depth analysis, which enables users to track the energetic contributions of mutations occurring at different spatial layers of the protein structure, thus easily shedding light on the particular strategies that a design pipeline might have used and assessing its particular energetic impact. We demonstrate the use of ENDURE in assessing a previously designed version of a polyethylene terephthalate (PET)-hydrolyzing enzyme from *Ideonella sakaiensis* (*Is*PETase), called DuraPETase (Cui et al., 2021), carrying ten mutations compared to the wildtype *Is*PETase (Yoshida et al., 2016). We expect that ENDURE will be a valuable resource for protein designers, filling a crucial gap in assessing and explaining the outcomes of protein design calculations.

2 Methods

The architecture and interface of the ENDURE web-app are designed in such a way that users are guided through the process of analyzing the energetic contributions of their protein designs in an intuitive and user-friendly manner. As shown in Figure 1, the whole workflow consists of several pages implemented in Streamlit (https://streamlit.io), a web application framework for Python. Starting from the File Upload page, the user uploads the PDB files of the designed protein and the reference protein structure. Alternatively, users can choose to provide amino acid sequences of the reference and mutant protein, and ENDURE will automatically predict their structures using ESMFold (Rives et al., 2021). The app then guides the user through three analysis steps, including pairwise interaction analysis, residue depth analysis, and inspection of all pairwise interaction changes by means of an energy difference heatmap. The results of these analyses are presented to the user in an interactive way that can be easily adapted to suit their specific needs using the controls on the left sidebar. The resulting tables can also be exported and downloaded as CSV file. Streamlit abstracts both front-end and back-end programming, making the application easily extensible by users with Python programming knowledge. Finally, the source code of ENDURE is packaged in a Docker container for easy installation and replication of the tool across different compute environments such as servers or HPC clusters. More details on the technological implementation of the app and the computational algorithms are provided in the following paragraphs.

2.1 Front end/backend in streamlit

The application's front end consists of the user interface (UI), which includes the welcome page, file upload section, and interactive visualizations. The UI is designed to be user-friendly and easy to navigate, allowing users to upload their protein structure file,

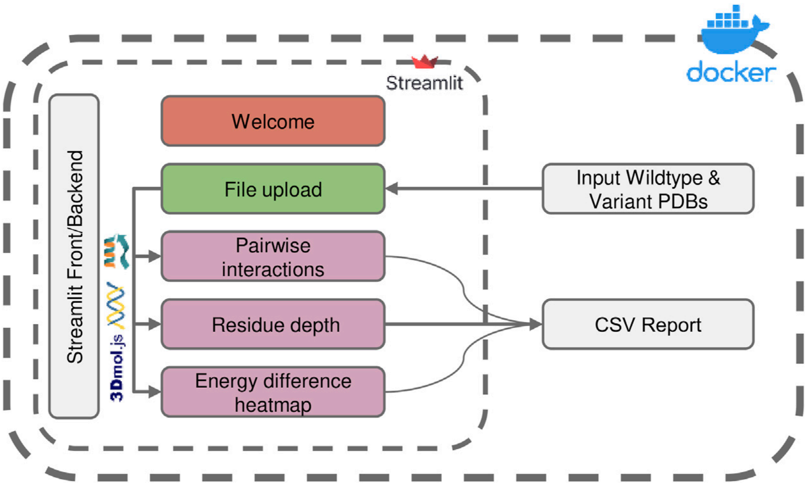


FIGURE 1
Overview of the ENDURE application architecture. Each colored box represents a subpage that performs a specific task, such as structure pre-processing and Rosetta calculations (File Upload page), pairwise interactions and residue depth analyses, and lastly visualization of an energy difference heatmap. The symbols represent the tools and libraries used for analysis and visualization. The Rosetta software is used for residue interaction analysis, the Biopython library is used for residue depth calculation, and the 3Dmol.js library is used for generating the visualization of the 3D structures. The process starts with uploading the protein structures, and then proceeds to processing and analysis of the structures. Pairwise interactions and residue depth are analyzed, and a CSV report can be generated. Finally, an energy difference heatmap is created.

prepare the structure, and run the energy breakdown and residue depth calculations. Since Streamlit is our web development framework, the front and backend are managed under the hood. We do include several analysis modules, which are either pure Python functions or Python functions that call other auxiliary executables. The names of important internal functions in the ENDURE app are written in `typewriter font` in the following sections.

2.2 Processing structures with ENDURE

The protocol consists of two main stages: 1) structure pre-processing actions and 2) structure analysis actions. Those actions are run on the File Upload page.

The pre-processing is an important step in the ENDURE web tool as it ensures that the uploaded protein structure is in the correct format and ready for analysis. The pre-processing actions include cleaning of the PDB files to remove any ligands, ions, or water molecules, relaxing of the protein structures to remove any steric clashes or unfavorable interactions, and determining the mutations for the designed protein relative to the reference protein. Additionally, the PDB file is renumbered in this step so that the first residue in the file is at position 1. Relaxing the protein structures is the second important action in preparing the protein structures for analysis. This ensures that the protein structure is in a low energy state according to the Rosetta energy function (Alford et al., 2017), which can help to minimize false positive results. The ENDURE web tool uses RosettaScripts (Fleishman et al., 2011) to perform a single iteration of FastRelax (Khatib et al., 2011) by default, taking a previously minimized protein structure and optimizing its energy landscape further. Overall, the input preparation protocol is crucial for ensuring that the protein structure is in the correct format and ready for analysis. (See Supplementary Material for the specific relax and energy breakdown commands).

The analysis actions include the energy breakdown (EB) calculations, which provide information about the energetics of each residue's interaction in the two analyzed protein structures, and the residue depth (RD) calculations. These algorithms are powered by the Rosetta energy function (Alford et al., 2017) and the Biopython (Cock et al., 2009) library, respectively. The EB calculations are performed using the Rosetta EB executable. In short, EB determines the one-body and two-body energies for each residue and decomposes them further into individual score term contributions, thus allowing the simultaneous exploration of, e.g., sidechain and backbone interactions. By clicking on the Calculate Energy button on the File Upload page, the Rosetta EB calculation is run in a subprocess. Internally, the `run` function is launched with four parameters as input: the input PDB file name, the location to save the result file, the location to save the log file, and the file path of the Rosetta executable. The output of the protocol is converted to a downloadable CSV file using the `convert_outfile` function, which saves the CSV file as a dictionary in the current session state.

These actions are followed by the RD calculation, which uses the MSMS algorithm (Sanner et al., 1996) from Biopython to calculate the distance of each residue to the molecular surface. The MSMS software computes the solvent-excluded surface from a set of spheres, representing the atoms in a protein structure. The reduced surface is calculated, and an analytical description of the solvent-excluded surface is derived from it. The calculation is done in the `calculate_depth` function.

The processing and analysis actions launched from the File Upload page are run in the background to prevent the UI from freezing, and their results are integrated into the interactive visualizations in the front end.

2.3 Analysis of ENDURE outputs: pairwise interactions analysis

The Interaction Analysis page allows users carrying out a comprehensive analysis of the energetic changes in pairwise interactions of single- and multi-mutant protein designs. The user can select and analyze interactions from different categories (A to F explained below) and from different physical interaction types (salt bridges, etc.) through different control parameters on the side bar. Individual residue pairs, affected directly or indirectly by the mutations, can be selected and displayed in interactive 3D visualizations.

We defined six categories of residue pair interactions to identify the regions in the protein structures most affected by the mutations.

Category A: residue pairs that are interacting in the reference structure have different interaction energy in the mutant, even though neither of the two residues were mutated.

Category B: residue pairs that are interacting in the reference structure have one member replaced, resulting in a different interaction energy in the mutant.

Category C: residue pairs that are interacting in the reference structure no longer interact in the mutant, even though neither of the two residues was mutated.

Category D: residue pairs that are interacting in the reference structure no longer interact in the mutant because one member was mutated.

Category E: residue pairs that are not interacting in the reference structure interact in the mutant, even though neither of the two residues was mutated.

Category F: residue pairs that are not interacting in the reference structure interact in the mutant because one member was mutated.

In order for ENDURE to detect the different interaction categories A-F and different physical interaction types, there are several functions implemented on the page that users can execute by clicking the Start Calculations button (after having run all the pre-processing actions on the File Upload page). Such functions perform a post-processing and filtering of the scorefile generated by the Rosetta EB calculations.

The `energy_calc` function is used to identify the essential changes in interaction energies between the mutant and the reference protein structure. It takes the outputs of the residue EB computation performed for the mutant and reference structures, the list of mutations between the reference and mutant, and a streamlit progress bar. The function calls several sub-functions to perform various interaction energy comparisons for each interaction category and physical interaction type. The former is managed by the `interaction_analysis` function and the latter is managed by the following functions: `salt_bridges`, `disulfide_bonds`, and `hydrogen_bonds`. As the names suggest, these functions calculate the energy differences for different types of interactions, such as salt bridges, disulfide bonds, and hydrogen bonds.

The `interaction_analysis` function calculates the difference in interaction energies between a mutant and a reference protein structure for a given list of mutations. The `interaction_analysis` function processes the outputs of the per-residue EB performed on the reference and mutant structures. It calculates the differences in all single-body (i.e., within a single residue) and two-body (i.e., between two residues) energies for all categories (A to F) between the two proteins.

In analyzing the interaction energy changes between the wildtype and mutant protein, it is important to distinguish between total energy changes and significant energy changes. The former is the sum of all energy changes, including those with a small value. However, since thousands of small changes can occur, it is possible for insignificant changes to mask chemically important changes. To overcome this issue, the significant energy change is calculated, which is the sum of only those interaction energy changes that exceed a minimum magnitude. In ENDURE changes that are larger than +1.0 Rosetta Energy Units (REU) or smaller than −1.0 REU are considered significant. By focusing only on significant energy changes, the mutations that are likely to have a significant impact can be more easily detected.

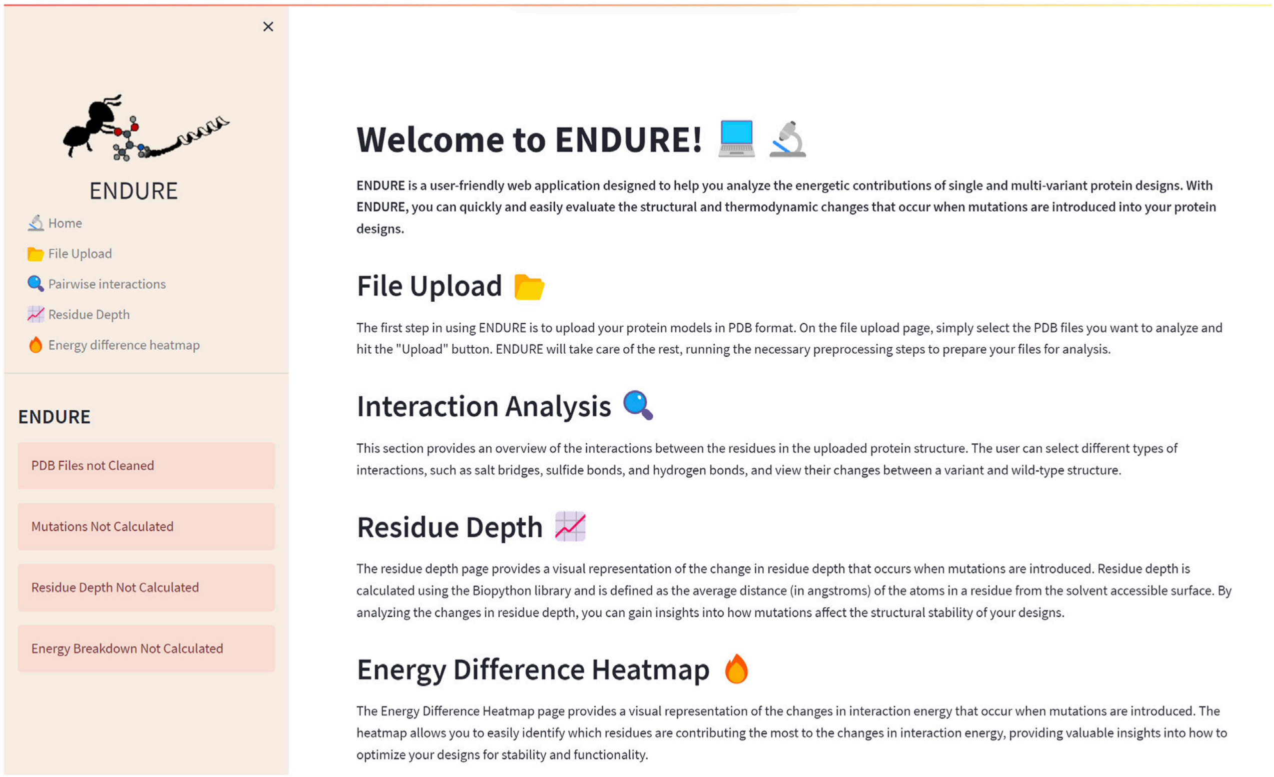
For the REF2015 scoring function for soluble proteins, there is an approximate 1:1 correspondence of REU and kcal/mol (Alford et al., 2017). However, for other Rosetta scoring functions, which include statistically derived potentials, the correspondence of Rosetta score units to thermodynamic energies is convoluted. For compatibility with other Rosetta scoring functions, which we plan to add to ENDURE in the future, we decided to report energy values on the ENDURE web page in REU.

The total energy change for all interactions and the sum of the subset of significant changes, is calculated using the `total_energy_changes` and `significant_changes` functions, respectively. These functions operate on a dictionary returned by `interaction_analysis` and sum all as well as the significant energy changes for each interaction category specified in an interaction list. Interactions from categories A-F can be selected from a list and the change for a given interaction type (salt bridges, disulfide bonds, sidechain-sidechain hydrogen bonds, sidechain-backbone hydrogen bonds, backbone-backbone short-range hydrogen bonds, backbone-backbone long-range hydrogen bonds, and all interactions) can be furthered inspected.

2.4 Analysis of ENDURE outputs: residue depth analysis

Mutation-induced energetic changes can have different effects on different layers of the protein structure. Therefore, we implemented a residue depth analysis and combined it with the per-residue energy breakdown analysis to distinguish changes occurring on the protein surface from those occurring in buried regions of the protein structure.

The Residue Depth page of the ENDURE app allows the user to analyze and compare the effect of mutations on the energy and spatial location of residues. Specifically, the user can select a residue pair that displays a strong negative energetic contribution and visualize the interaction in the protein structure. By adjusting the threshold slider, the user can see which residues have a significant impact on stability and select the most promising mutations for further analysis. Once the user selects a particular mutation by clicking on a point in the scatter plot, the app displays a side-by-side 3D visual comparison of the residue in the mutant and reference structures, which allows for a direct comparison of the effect of mutations on the protein structure. This analysis provides valuable insights into the structural and energetic changes resulting from mutations.



Welcome to ENDURE! 💻 🔬

ENDURE is a user-friendly web application designed to help you analyze the energetic contributions of single and multi-variant protein designs. With ENDURE, you can quickly and easily evaluate the structural and thermodynamic changes that occur when mutations are introduced into your protein designs.

File Upload 📁

The first step in using ENDURE is to upload your protein models in PDB format. On the file upload page, simply select the PDB files you want to analyze and hit the "Upload" button. ENDURE will take care of the rest, running the necessary preprocessing steps to prepare your files for analysis.

Interaction Analysis 🔍

This section provides an overview of the interactions between the residues in the uploaded protein structure. The user can select different types of interactions, such as salt bridges, sulfide bonds, and hydrogen bonds, and view their changes between a variant and wild-type structure.

Residue Depth 📈

The residue depth page provides a visual representation of the change in residue depth that occurs when mutations are introduced. Residue depth is calculated using the Biopython library and is defined as the average distance (in angstroms) of the atoms in a residue from the solvent accessible surface. By analyzing the changes in residue depth, you can gain insights into how mutations affect the structural stability of your designs.

Energy Difference Heatmap 🔥

The Energy Difference Heatmap page provides a visual representation of the changes in interaction energy that occur when mutations are introduced. The heatmap allows you to easily identify which residues are contributing the most to the changes in interaction energy, providing valuable insights into how to optimize your designs for stability and functionality.

FIGURE 2
Welcome page. On the left-hand side, the side bar containing the subpages and status bars is given. The latter indicate, by turning from red to green, if the pre-processing and analysis actions on the File Upload page are completed. The main section of the Welcome page contains a brief description of the tool and of each subpage.

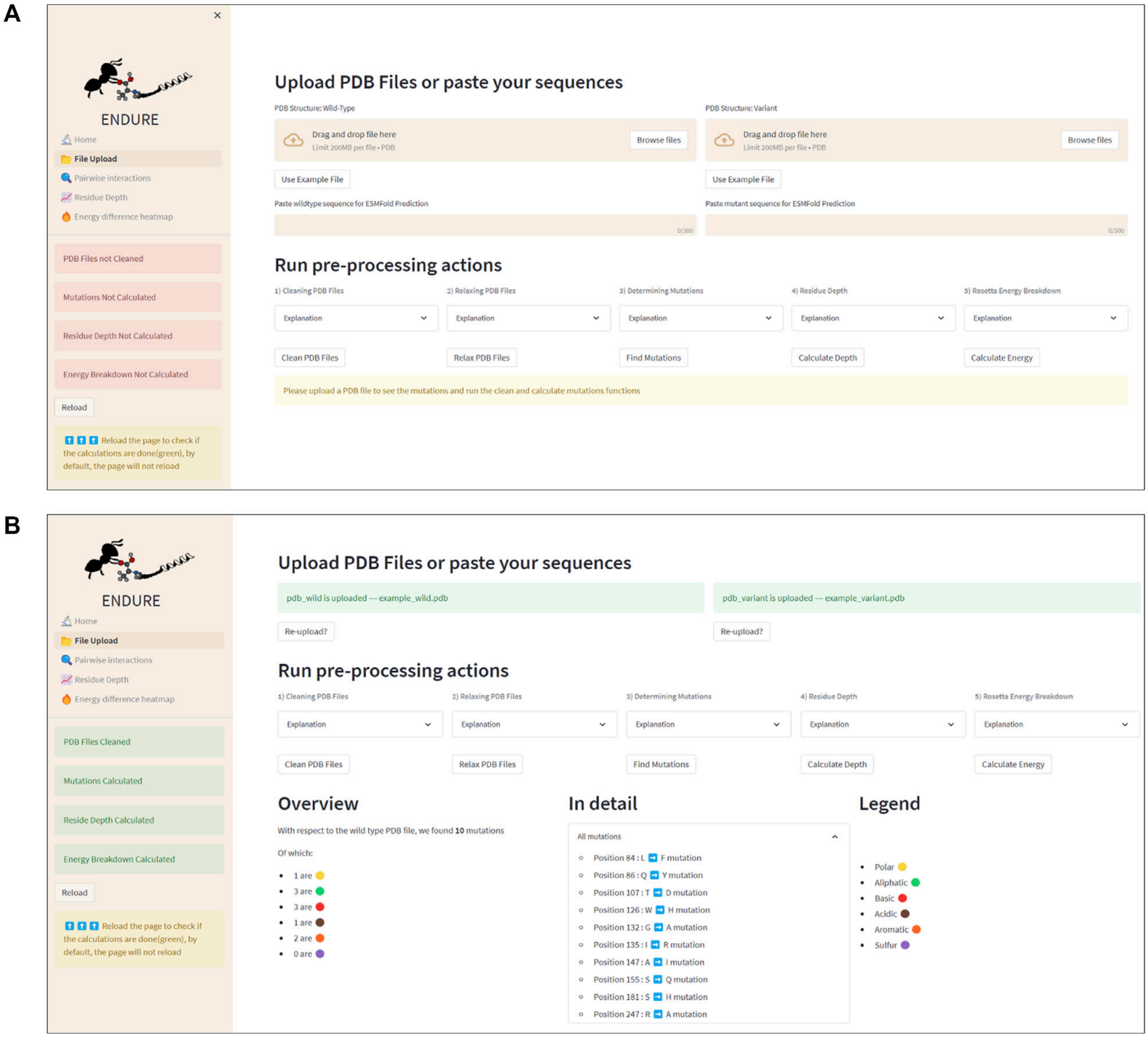
3 Results

The Welcome page (Figure 2) is the first page users will encounter when accessing the web application. This page provides an overview of the tool and its functionality, allowing new users to quickly familiarize themselves with the interface. The Welcome page includes a brief description of the purpose of ENDURE and the steps involved in using the tool to analyze protein designs. The Welcome page is designed to be user-friendly and intuitive, providing a clear and concise introduction to the tool and its capabilities, making it easier for users to get started and use the tool most effectively for their particular research questions.

Figure 3 presents the File Upload section, where users can upload their protein structure files in PDB format. The interface is designed to be intuitive and user-friendly, with options for either selecting the file from the local file system or dragging and dropping the file into the designated area. After uploading the file, the user is prompted to run five important pre-processing actions: 1) cleaning PDBs, which ensures that the files are correctly parsed, and the residues are renumbered so that the first residue is at position 1. 2) Relaxing PDB files, which prepares the files for the analysis. 3) Determining mutations by identifying the amino acid differences between the reference and mutant sequences. This information is crucial for many components of the analysis, as it allows tracking the position of mutations. 4) Calculating residue depth determines the average distance of residues from the solvent-accessible surface. This is a key factor in understanding the energetic contributions of mutations. The calculation is performed using the Biopython library and is executed in a separate thread to avoid hanging the GUI. 5) Creating energy breakdown files, which provide a detailed breakdown of the energy contributions of individual residues. This calculation is performed using the Rosetta EB executable, as explained above, and is run in the background to prevent the GUI from being frozen.

3.1 ENDURE use case: analysis of an *in silico-designed* PET-degrading enzyme

To demonstrate the workflow and scope of application of ENDURE, we present here the results obtained for a previously reported designed PET-degrading enzyme, called DuraPETase (PDB ID: 6KY5) (Cui et al., 2021), which has higher thermostability than the wildtype *Is*PETase enzyme (PDB ID: 5XJH) (Yoshida et al., 2016; Joo et al., 2018). As can be seen in Figure 4A, ENDURE confirms that the particular mutant (carrying ten mutations) has favorable total and significant energy changes (−5.8 REU), indicating improved stability. This result is in line with the experimentally determined increase in the apparent melting temperature (T_m) of DuraPETase by 31°C compared the wildtype *Is*PETase (Cui et al., 2021). In addition, we calculated the significant



A

Upload PDB Files or paste your sequences

PDB Structure: Wild-Type

Drag and drop file here
Limit 200MB per file • PDB

Paste wildtype sequence for ESMFold Prediction

0/300

PDB Structure: Variant

Drag and drop file here
Limit 200MB per file • PDB

Paste mutant sequence for ESMFold Prediction

0/300

Run pre-processing actions

1) Cleaning PDB Files
2) Relaxing PDB Files
3) Determining Mutations
4) Residue Depth
5) Rosetta Energy Breakdown

Please upload a PDB file to see the mutations and run the clean and calculate mutations functions

PDB Files not Cleaned

Mutations Not Calculated

Residue Depth Not Calculated

Energy Breakdown Not Calculated

Reload the page to check if the calculations are done(green), by default, the page will not reload

B

Upload PDB Files or paste your sequences

pdb_wild is uploaded --- example_wild.pdb

pdb_variant is uploaded --- example_variant.pdb

Run pre-processing actions

1) Cleaning PDB Files
2) Relaxing PDB Files
3) Determining Mutations
4) Residue Depth
5) Rosetta Energy Breakdown

Overview

With respect to the wild type PDB file, we found **10** mutations

Of which:

- 1 are 🟡
- 3 are 🟢
- 3 are 🔴
- 1 are 🟤
- 2 are 🟠
- 0 are 🟣

In detail

- Position 84 : L ➡ F mutation
- Position 86 : Q ➡ Y mutation
- Position 107 : T ➡ D mutation
- Position 126 : W ➡ H mutation
- Position 132 : G ➡ A mutation
- Position 135 : I ➡ R mutation
- Position 147 : A ➡ I mutation
- Position 155 : S ➡ Q mutation
- Position 181 : S ➡ H mutation
- Position 247 : R ➡ A mutation

Legend

- Polar 🟡
- Aliphatic 🟢
- Basic 🔴
- Acidic 🟤
- Aromatic 🟠
- Sulfur 🟣

PDB Files Cleaned

Mutations Calculated

Reside Depth Calculated

Energy Breakdown Calculated

Reload the page to check if the calculations are done(green), by default, the page will not reload

FIGURE 3
File Upload page. **(A)** Interface before uploading PDB files or using the example files. The latter action can be activated by clicking the "Use Example File" button. **(B)** Interface after running all pre-processing and analysis actions. Note that the color of the status boxes turned from red to green.

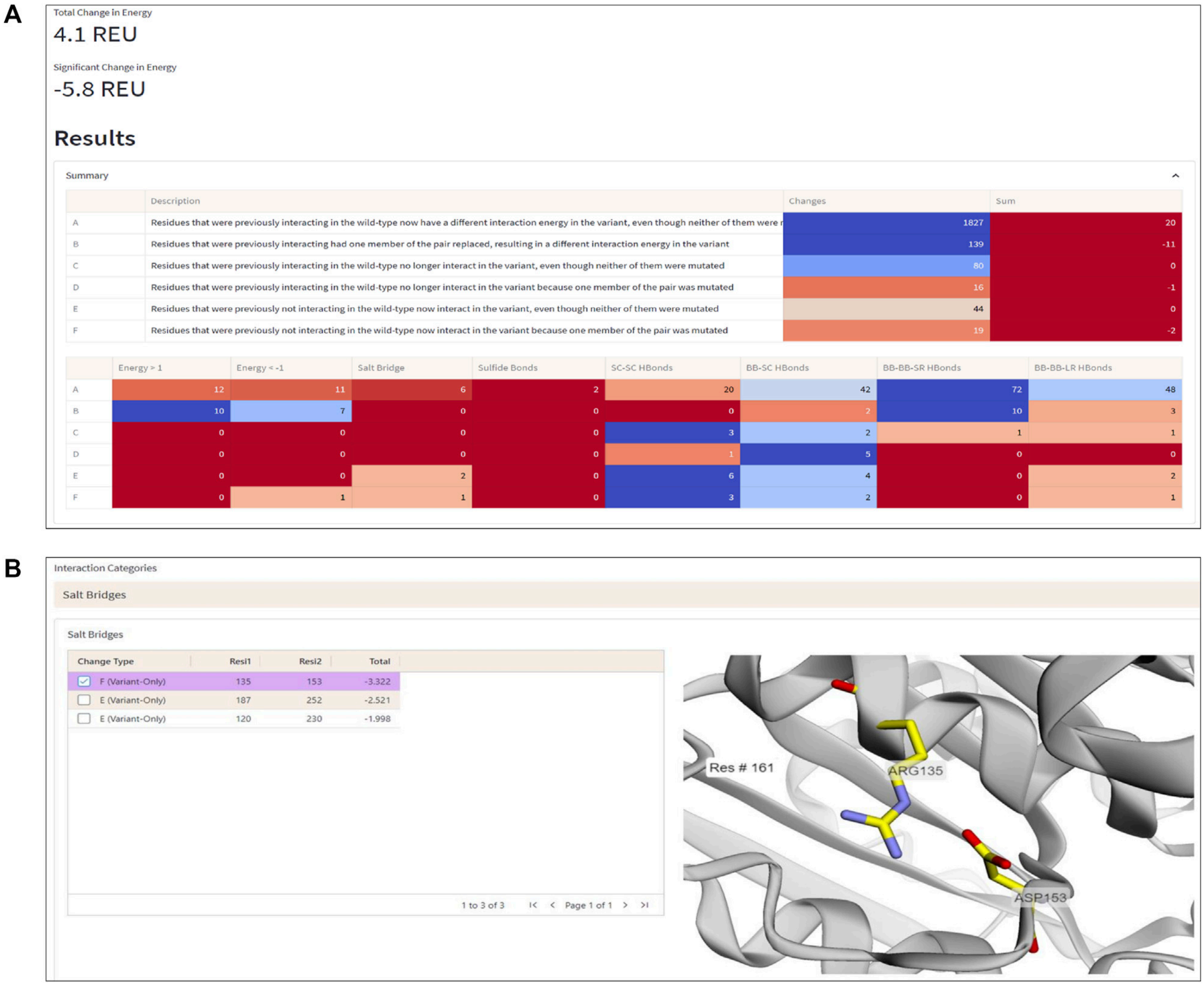
energy changes for two other designed PET hydrolases, FastPETase (Lu et al., 2022) and HotPETase (Bell et al., 2022), using ENDURE (Supplementary Table S1). The favorable energy changes for FastPETase (−10.2 REU) and HotPETase (−27.6 REU) confirmed their higher stability of T_m = 67°C and T_m = 82°C, respectively, compared to wildtype *Is*PETase. This shows the utility of ENDURE in estimating overall stability changes.

In addition to the overall energy comparison, ENDURE also provides detailed information about the specific amino acid interactions that contribute to the improved stability of the selected mutant. The interaction analysis feature allows focusing on and visualizing specific residue interactions in the protein structure, which help to rationalize the underlying molecular mechanisms contributing to the improved stability of the selected mutant. These features will be described next.

3.1.1 Interaction analysis—Changes in pairwise interactions

The Interaction Analysis page of ENDURE enables performing a detailed examination of the energetic changes for residue pairs of different types (see Figure 4B). With this help, the user can identify the particular interactions contributing to improved or impaired energetic stability.

For example, Figure 4B shows a salt bridge interaction between residues Arg135 and Asp153, which causes a significant energy improvement of −3.32 REU in DuraPETase compared to wildtype *Is*PETase. This salt bridge isn't present in the wildtype protein but is in DuraPETase—i.e., it belongs to interaction category F. *Is*PETase has an isoleucine (Ile135) at the same position, which cannot form a salt bridge. In addition, as shown in the left-hand table in the screenshot in Figure 4B, there are two more salt bridges detected by



Total Change in Energy

4.1 REU

Significant Change in Energy

-5.8 REU

Results

Summary

	Description	Changes	Sum
A	Residues that were previously interacting in the wild-type now have a different interaction energy in the variant, even though neither of them were r	1827	20
B	Residues that were previously interacting had one member of the pair replaced, resulting in a different interaction energy in the variant	139	-11
C	Residues that were previously interacting in the wild-type no longer interact in the variant, even though neither of them were mutated	80	0
D	Residues that were previously interacting in the wild-type no longer interact in the variant because one member of the pair was mutated	16	-1
E	Residues that were previously not interacting in the wild-type now interact in the variant, even though neither of them were mutated	44	0
F	Residues that were previously not interacting in the wild-type now interact in the variant because one member of the pair was mutated	19	-2

	Energy > 1	Energy < -1	Salt Bridge	Sulfide Bonds	SC-SC HBonds	BB-SC HBonds	BB-BB-SR HBonds	BB-BB-LR HBonds
A	12	11	6	2	20	42	72	48
B	10	7	0	0	0	2	10	3
C	0	0	0	0	3	2	1	1
D	0	0	0	0	1	5	0	0
E	0	0	2	0	6	4	0	2
F	0	1	1	0	3	2	0	1

Interaction Categories

Salt Bridges

Salt Bridges

Change Type	Resi1	Resi2	Total
☑ F (Variant-Only)	135	153	-3.322
☐ E (Variant-Only)	187	252	-2.521
☐ E (Variant-Only)	120	230	-1.998

1 to 3 of 3 |< < Page 1 of 1 > >|

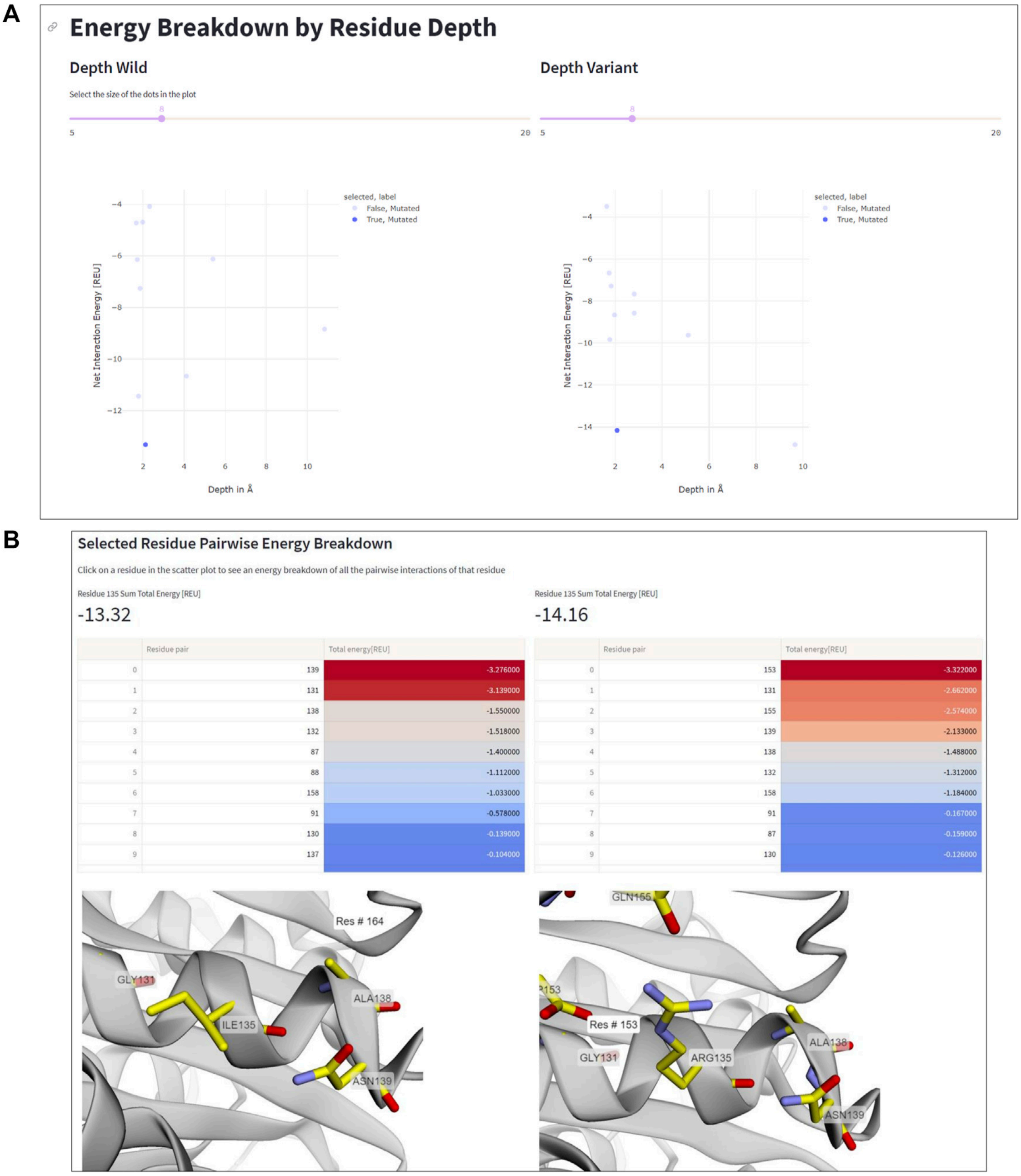
FIGURE 4
Interaction Analysis page. **(A)** Summary table of the number of significant energy changes for residue pairs belonging to different interaction categories (A-F, explained in the text) and types of physical interactions in the DuraPETase enzyme. **(B)** 3D visualization of a salt bridge interaction (Arg135-Asp153) in DuraPETase from interaction category F (i.e., significant energy improvement due to mutation of one residue in the pairwise interaction). Other salt bridges with improved energy in DuraPETase compared to the wildtype *Is*PETase are listed in the table on the left side. The selected Arg135-Asp153 salt bridge is visually represented in the structure viewer window on the right side, highlighting the selected residue pair in sticks. Residue pairs from other physical interaction types (hydrogen bonds between backbone or side chain atoms, disulfide bonds) can also be selected, as explained in the text.

ENDURE, which contribute significant energy changes in DuraPETase: Asp187-Arg252 (−2.52 REU) and Arg120-Asp230 (−1.99 REU). These two interactions belong to category E, i.e., they do not involve a mutated residue but are probably an indirect effect of nearby mutations in the environment of the four residues. Information like this provides valuable insight into protein structure and can help guide the user in their efforts to design a better-functioning mutant protein.

The different categories of interaction changes (A to F) allow the user to quickly identify the changes in interactions that have occurred and to focus their attention on the most important ones. For example, if a user observes that most changes are of type E or F, this might suggest that new interactions have formed, which could also significantly affect function. By providing this information, the Interaction Analysis page helps the user to quickly understand and prioritize the changes in interactions that have occurred and guide their efforts in protein design.

3.1.2 Residue depth analysis

The residue depth analysis feature allows determining the depth of each residue in the protein structure, which reflects its accessibility to the solvent. By analyzing the energetic changes of mutations occurring in different spatial layers of the protein structure, the location of mutations that improve or impair stability can be determined. In Figure 5, the aforementioned mutation Ile135Arg is displayed as an example. In wildtype *Is*PETase, Ile135 is located on the protein surface, indicated by a low residue depth value of ~2Å (Figure 5A). This surface-exposed



Energy Breakdown by Residue Depth

Selected Residue Pairwise Energy Breakdown

Click on a residue in the scatter plot to see an energy breakdown of all the pairwise interactions of that residue

Residue 135 Sum Total Energy [REU]

-13.32

	Residue pair	Total energy[REU]
0	139	-3.276000
1	131	-3.139000
2	138	-1.550000
3	132	-1.518000
4	87	-1.400000
5	88	-1.112000
6	158	-1.033000
7	91	-0.578000
8	130	-0.139000
9	137	-0.104000

Residue 135 Sum Total Energy [REU]

-14.16

	Residue pair	Total energy[REU]
0	153	-3.322000
1	131	-2.662000
2	155	-2.574000
3	139	-2.133000
4	138	-1.488000
5	132	-1.312000
6	158	-1.184000
7	91	-0.167000
8	87	-0.159000
9	130	-0.126000

FIGURE 5
Residue depth analysis. **(A)** Net interaction energy versus residue depth plots for wildtype *Is*PETA (left) and DuraPETase (right). The blue data point in the lower left corner of the plot, corresponding to Ile135 in *Is*PETase or Arg135 in DuraPETase, respectively, has been selected. **(B)** Side-by-side comparison of the location and surrounding amino acids of Ile135 and Arg135, respectively. The interacting residues and their energy contributions are listed above the 3D viewer.

location is unfavorable for a hydrophobic amino acid. By contrast, the Arg135 residue in DuraPETase can form favorable interactions on the protein surface. In addition to the already mentioned Arg135-Asp153 salt bridge, the side chain of Arg135 interacts with the side chain of Gln155 through a hydrogen bond (Figure 5B). The table in Figure 5B, which lists the interacting residues for Ile135 or Arg135, respectively, and highlights their respective energy contributions (blue: high energy, red: low energy), confirms these visual observations. These extra interactions, can explain the lower net energy of −14.16 REU for Arg135 in DuraPETase compared to −13.32 REU for Ile135 in *Is*PETase (Figure 5A).

Information like this is important because it can help decide which residues to mutate in order to improve protein stability. For example, mutations in the protein core may have a greater impact on stability than those located near the surface, and thus targeting core residues for mutagenesis can considerable impact the stability of the designed protein. Additionally, the residue depth analysis feature can be used to select the most promising mutants for further

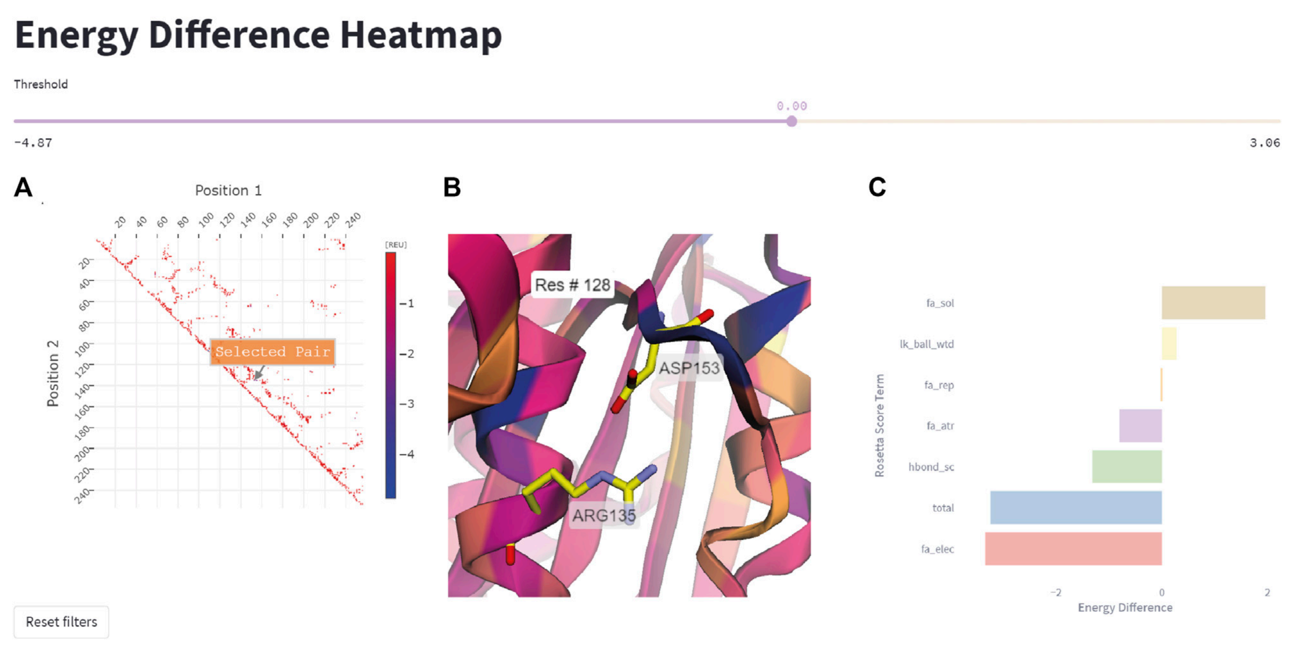


FIGURE 6
Energy difference heatmap. **(A)** 2D matrix of residue pair energy changes between *Is*PETase and DuraPETase. Negative values indicate that an interaction has a lower (more negative) energy in DuraPETase. The 2D matrix is illustrated as a heatmap with larger negative changes colored blue and smaller negative energy changes colored red. Users can zoom in the 2D matrix and interactively select a residue pair for further analysis. **(B)** 3D visualization of the Arg135-Asp153 residue pair after selecting it in the heatmap in **(A)**. The structure is colored according to the positional energy difference. **(C)** Breakdown of the energy change for the selected pairwise interaction into individual Rosetta score terms, representing different physical interactions.

characterization, by identifying mutations that have an improved net interaction energy and are located in favorable positions within the protein structure.

3.1.3 Energy difference heatmap

This feature allows the user to quickly represent all residue pairs that have a significantly changed interaction energy in the mutant protein compared to the wildtype, and selectively track specific pair interactions. The user can select interactively from the residue pair interaction matrix on the left side in Figure 6 a pair of residues that display a large negative energetic contribution. The relevance threshold for the energy can be adjusted with the threshold slider present above. Once the user has selected a residue pair (Figure 6A), the corresponding pair (Arg135-Asp153 in this example) will be highlighted in the structure viewer in the middle of the page (Figure 6B), and a breakdown of the interaction energy change into individual score terms from the Rosetta energy function will appear (Figure 6C). For visual clarity only none-zero score terms are displayed to the user.

4 Discussion

The ENDURE web application provides a user-friendly interface for analyzing protein structures to facilitate mutant selection in protein design workflows. The study has shown that the application can accurately and efficiently process PDB files, clean and renumber them, determine mutations, calculate residue depth, and generate energy breakdown files. The interaction analysis section allows users to view changes in pairwise interactions between residues in the wildtype and mutant structures by providing visual representations of the significant and total energy changes.

One of the key innovations of the ENDURE web application is its ability to group changes in residue pairwise interactions into different types and categories, ranging from residues that are interacting in the reference and mutant protein structure with different interaction energies (category A), to residues that aren't interacting in the reference structure but make interactions in the mutant due to a mutation (category F). These different categories of changes provide a useful way for the user to identify which mutations have the strongest impact on the protein structure, and consequently focus design efforts on specific areas of the protein structure.

Compared to previous research, our web application offers a user-friendly and accessible solution for analyzing protein structures and interaction changes. Prior research in this field has typically been focused on developing computational tools for protein structure prediction and analysis (Schymkowitz et al., 2005; Leman et al., 2020; Stam and Wood, 2021) or the analysis of the effects of single point mutations (Yin et al., 2007). For instance, the Eris web server (Yin et al., 2007) is an estimator of protein stability that primarily focuses on single mutations. In contrast, our ENDURE server is specifically designed to handle multiple mutations simultaneously while assessing the introduction of significant pairwise interactions. ENDURE provides a unique solution by incorporating advanced Rosetta modeling and analysis algorithms into a user-friendly interface and making them more broadly accessible.

It is important to note that the tool is limited by the accuracy of the underlying computational tools and algorithms used for protein structure prediction and analysis. Additionally, the interaction analysis section is based on a static protein structure analysis. Accuracy in predicting mutation-induced energy changes could benefit from a model ensemble approach (Peccati et al., 2023), in which protein dynamic changes like loop rearrangements can be considered. These limitations should be taken into consideration when interpreting the results of the analysis.

Despite these limitations, it represents a significant advancement in the field of protein structure and interaction analysis. The integration of computational tools into a user-friendly interface makes it possible for scientists outside the field of computational structural biology to quickly and efficiently analyze protein structures and identify potential areas for improvement. In future versions, ENDURE could be expanded to include additional features and improvements, such as the ability to examine other kinds of proteins, including membrane proteins and proteins with noncanonical amino acid modifications, through the incorporation of different energy functions. It could also be enhanced to consider ligand molecules in the design analysis, allowing the identification of designs with improved binding affinity. These features are currently planned for the next version of ENDURE, which will be released in the future. Additionally, the tool could be further developed to incorporate machine learning techniques to improve the accuracy of the analysis. With these advancements, ENDURE could become an even more powerful tool for protein design and analysis.

In conclusion, the ENDURE web application provides a unique and accessible solution for analyzing protein structure and interaction changes and, in that way, represents a significant advancement for the field of protein design. Categorizing changes in pairwise interactions for different interaction types provides a straightforward way for the user to guide their protein design strategies. Integrating computational tools into a user-friendly interface makes it possible for a broader audience to quickly and efficiently analyze protein structures. The future direction of the research will focus on further developing the application to incorporate analysis on protein dynamics, support for non-standard amino acid residues, and application of machine learning techniques. Furthermore, a command line interface integrated in the front end is planned, which will help further customize some of the analyses.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. The source code for ENDURE can be found on https://github.com/kuenzelab/ENDURE. Further inquiries can be directed to the corresponding author.

Author contributions

Conceptualization, FE and JZ; methodology, FE and JZ; software, FE and JZ; investigation, FE and JZ; data curation, FE; writing—original draft preparation, FE; writing—review and editing, FE, JZ, and GK; visualization, FE; supervision, GK; project administration, GK. All authors have read and agreed to the published version of the manuscript.

Funding

This work was funded by the German Research Foundation (CRC1052, project Z6) and the Open Access Publishing Fund of Leipzig University, supported by the German Research Foundation within the program Open Access Publication.

Acknowledgments

JZ acknowledges the Max Kade foundation for receiving a summer exchange scholarship. The authors thank Ivan Ivanikov for his technical support in installing the ENDURE webserver. The authors further acknowledge the Leipzig University Scientific Computing Center and the Institute for Drug Discovery for providing the computational resources.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2023.1178035/full#supplementary-material

References

Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., et al. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J Chem Theory Comput* 13, 3031–3048. doi:10.1021/acs.jctc.7b00125

Arnold, F. H. (2018). Directed Evolution: Bringing New Chemistry to Life. *Angewandte Chemie International Edition* 57, 4143–4148. doi:10.1002/anie.201708408

Barlow, K. A., S, O. C., Thompson, S., Suresh, P., Lucas, J. E., Heinonen, M., et al. (2018). Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity upon Mutation. *J Phys Chem B* 122, 5389–5399. doi:10.1021/acs.jpcb.7b11367

Bell, E. L., Smithson, R., Kilbride, S., Foster, J., Hardy, F. J., Ramachandran, S., et al. (2022). Directed evolution of an efficient and thermostable PET depolymerase. *Nat Catal* 5, 673–681. doi:10.1038/s41929-022-00821-3

Castro, K. M., Scheck, A., Xiao, S., and Correia, B. E. (2022). Computational design of vaccine immunogens. *Current Opinion in Biotechnology* 78, 102821. doi:10.1016/j.copbio.2022.102821

Chen, K., Arnold, F. H., Cheng, R., Fisher, M., Zhang, B. H., Di Maggio, M., et al. (2020). Facial Recognition Neural Networks Confirm Success of Facial Feminization Surgery. *Nat Catal* 3, 203–209. doi:10.1097/PRS.0000000000006342

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi:10.1093/bioinformatics/btp163

Cui, Y., Chen, Y., Liu, X., Dong, S., Tian, Y., Qiao, Y., et al. (2021). Computational Redesign of a PETase for Plastic Biodegradation under Ambient Condition by the GRAPE Strategy. *ACS Catal.* 11, 1340–1350. doi:10.1021/acscatal.0c05126

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., et al. (2022). Robust deep learning–based protein sequence design using ProteinMPNN. *Science* 378, 49–56. doi:10.1126/science.add2187

Ferruz, N., and Höcker, B. (2022). Controllable protein design with language models. *Nat Mach Intell* 4, 521–532. doi:10.1038/s42256-022-00499-z

Fleishman, S. J., Leaver-Fay, A., Corn, J. E., Strauch, E. M., Khare, S. D., Koga, N., et al. (2011). RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite. *PloS one* 6, e20161. doi:10.1371/journal.pone.0020161

Frenz, B., Lewis, S. M., King, I., DiMaio, F., Park, H., and Song, Y. (2020). Prediction of Protein Mutational Free Energy: Benchmark and Sampling Improvements Increase Classification Accuracy. *Front. Bioeng. Biotechnol.* 8. 558247, doi:10.3389/fbioe.2020.558247

Goldenzweig, A., Goldsmith, M., Hill, S. E., Gertman, O., Laurino, P., Ashani, Y., et al. (2016). Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Molecular Cell* 63, 337–346. doi:10.1016/j.molcel.2016.06.012

Habibi, N., Mauser, A., Ko, Y., and Lahann, J. (2022). Protein Nanoparticles: Uniting the Power of Proteins with Engineering Design Approaches. *Advanced Science* 9, 2104012. doi:10.1002/advs.202104012

Joo, S., Cho, I. J., Seo, H., Son, H. F., Sagong, H.-Y., Shin, T. J., et al. (2018). Structural insight into molecular mechanism of poly(ethylene terephthalate) degradation. *Nat Commun* 9, 382. doi:10.1038/s41467-018-02881-1

Kellogg, E. H., Leaver-Fay, A., and Baker, D. (2011). Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* 79, 830–8. doi:10.1002/prot.22921

Khatib, F., Cooper, S., Tyka, M. D., Xu, K., Makedon, I., Popovic, Z., et al. (2011). Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences of the United States of America* 108, 18949, 18953. doi:10.1073/pnas.1115898108

Leman, J. K., Weitzner, B. D., Lewis, S. M., Adolf-Bryfogle, J., Alam, N., Alford, R. F., et al. (2020). Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods* 17, 665–680. doi:10.1038/s41592-020-0848-2

Lu, H., Diaz, D. J., Czarnecki, N. J., Zhu, C., Kim, W., Shroff, R., et al. (2022). Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* 604, 662–667. doi:10.1038/s41586-022-04599-z

Pan, X., and Kortemme, T. (2021). Recent advances in de novo protein design: Principles, methods, and applications. *Journal of Biological Chemistry* 296, 100558. doi:10.1016/j.jbc.2021.100558

Peccati, F., Alunno-Rufini, S., and Jiménez-Osés, G. (2023). Accurate Prediction of Enzyme Thermostabilization with Rosetta Using AlphaFold Ensembles. *J. Chem. Inf. Model.* 63, 898–909. doi:10.1021/acs.jcim.2c01083

Quijano-Rubio, A., Yeh, H.-W., Park, J., Lee, H., Langan, R. A., Boyken, S. E., et al. (2021). De novo design of modular and tunable protein biosensors. *Nature* 591, 482–487. doi:10.1038/s41586-021-03258-z

Reetz, M. (2022). Making Enzymes Suitable for Organic Chemistry by Rational Protein Design. *ChemBioChem* 23, e202200049. doi:10.1002/cbic.202200049

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA* 118, e2016239118. doi:10.1073/pnas.2016239118

Sanner, M. F., Olson, A. J., and Spehner, J.-C. (1996). Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers* 38, 305–320. doi:10.1002/(SICI)1097-0282(199603)38:3<305::AID-BIP4>3.0.CO;2-Y

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Research* 33, W382–W388. doi:10.1093/nar/gki387

Sevy, A. M., and Meiler, J. (2014). Antibodies: Computer-Aided Prediction of Structure and Design of Function. *Microbiology spectrum* 2. doi:10.1128/microbiolspec.AID-0024-2014

Stam, M. J., and Wood, C. W. (2021). DE-STRESS: a user-friendly web application for the evaluation of protein designs. *Protein Engineering, Design and Selection* 34, gzab029. doi:10.1093/protein/gzab029

Starr, T. N., and Thornton, J. W. (2016). Epistasis in protein evolution. *Protein Science* 25, 1204–1218. doi:10.1002/pro.2897

Stein, V., and Alexandrov, K. (2015). Synthetic protein switches: design principles and applications. *Trends in Biotechnology* 33, 101–110. doi:10.1016/j.tibtech.2014.11.010

Strokach, A., Becerra, D., Corbi-Verge, C., Perez-Riba, A., and Kim, P. M. (2021). Computational generation of proteins with predetermined three-dimensional shapes using ProteinSolver. *STAR Protocols* 2, 100505. doi:10.1016/j.xpro.2021.100505

Sumbalova, L., Stourac, J., Martinek, T., Bednar, D., and Damborsky, J. (2018). HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information. *Nucleic Acids Research* 46, W356–W362. doi:10.1093/nar/gky417

Weinstein, J. J., Goldenzweig, A., Hoch, S., and Fleishman, S. J. (2021). PROSS 2: a new server for the design of stable and highly expressed protein variants. *Bioinformatics* 37, 123–125. doi:10.1093/bioinformatics/btaa1071

Willis, J. R., Sapparapu, G., Murrell, S., Julien, J. P., Singh, V., King, H. G., et al. (2015). Redesigned HIV antibodies exhibit enhanced neutralizing potency and breadth. *The Journal of clinical investigation* 125, 2523–31. doi:10.1172/JCI80693

Yin, S., Ding, F., and Dokholyan, N. V. (2007). Eris: an automated estimator of protein stability. *Nat Methods* 4, 466–467. doi:10.1038/nmeth0607-466

Yoshida, S., Hiraga, K., Takehana, T., Taniguchi, I., Yamaji, H., Maeda, Y., et al. (2016). A bacterium that degrades and assimilates poly(ethylene terephthalate). *Science* 351, 1196–1199. doi:10.1126/science.aad6359

frontiers | Frontiers in Molecular Biosciences

TYPE Mini Review
PUBLISHED 07 July 2023
DOI 10.3389/fmolb.2023.1214424

EDITED BY
Traian Sulea,
National Research Council Canada
(NRC), Canada

REVIEWED BY
Monica Lisa Fernández-Quintero,
University of Innsbruck, Austria
Andreas Evers,
Merck, Germany

*CORRESPONDENCE
Konrad Krawczyk,
konrad@naturalantibody.com

[†]These authors have contributed equally to this work and share first authorship

RECEIVED 29 April 2023
ACCEPTED 12 June 2023
PUBLISHED 07 July 2023

CITATION
Jaszczyszyn I, Bielska W, Gawlowski T, Dudzic P, Satława T, Kończak J, Wilman W, Janusz B, Wróbel S, Chomicz D, Galson JD, Leem J, Kelm S and Krawczyk K (2023), Structural modeling of antibody variable regions using deep learning—progress and perspectives on drug discovery.
Front. Mol. Biosci. 10:1214424.
doi: 10.3389/fmolb.2023.1214424

Structural modeling of antibody variable regions using deep learning—progress and perspectives on drug discovery

Igor Jaszczyszyn[1,2†], Weronika Bielska[1,3†], Tomasz Gawlowski[1], Pawel Dudzic[1], Tadeusz Satława[1], Jarosław Kończak[1], Wiktoria Wilman[1], Bartosz Janusz[1], Sonia Wróbel[1], Dawid Chomicz[1], Jacob D. Galson[4], Jinwoo Leem[4], Sebastian Kelm[5] and Konrad Krawczyk[1]*

[1]NaturalAntibody, Kraków, Poland, [2]Medical University of Warsaw, Warsaw, Poland, [3]Medical University of Lodz, Lodz, Poland, [4]Alchemab Therapeutics Ltd., London, United Kingdom, [5]UCB, Slough, United Kingdom

AlphaFold2 has hallmarked a generational improvement in protein structure prediction. In particular, advances in antibody structure prediction have provided a highly translatable impact on drug discovery. Though AlphaFold2 laid the groundwork for all proteins, antibody-specific applications require adjustments tailored to these molecules, which has resulted in a handful of deep learning antibody structure predictors. Herein, we review the recent advances in antibody structure prediction and relate them to their role in advancing biologics discovery.

KEYWORDS

deep learning, structural modeling, drug discovery, antibody therapeutics, antibody structure prediction

1 Introduction

Antibodies are the largest class of biotherapeutics, with more than 100 approved molecules (Kaplon et al., 2023). The antibody drug market is rapidly growing, and it is predicted to reach approximately $300 billion by 2025 (Lu et al., 2020). As a result, there is much interest in streamlining antibody discovery methods by tapping into recent computational advances in deep learning.

One of the most striking computational advances has taken place in structure prediction, with the development of tools such as AlphaFold2 (Jumper et al., 2021). For antibodies, the determination of the proper antibody structure is key to many downstream drug discovery tasks, such as developability annotation (Raybould et al., 2019) or antibody–antigen docking (Krawczyk et al., 2014; Schneider et al., 2021). Though AlphaFold2 works well for general proteins, it falls short on the specific case of antibodies (Ruffolo et al., 2022a; Abanades et al., 2022b; Cohen et al., 2022), prompting the development of antibody-specific modeling protocols.

In this review, we describe the methods which contribute to the improvement of computational structure modeling for antibodies and provide context to the role they play in designing antibody-based therapeutics.

2 Antibody structure in the context of 3D modeling

Antibody structure prediction is primarily focused on the variable domains of the heavy chain (Vh) and the light chain (Vl) (Figure 1A). Each domain is relatively small, comprising ~110 residues each. There are two major hurdles within the overall antibody structure prediction problem: determining the relative orientation of the two domains (Figure 1B) and predicting the complementarity-determining region (CDR) loop structures. The two domains can be juxtaposed differently, which affects the overall shape of the antibody binding site. For this reason, orientating the multimer of the heavy and light chains is crucial (Dunbar et al., 2013; Bujotzek et al., 2015).

The CDR prediction problem can be further subdivided into classifying the "canonical" CDRs (CDR-L1, CDR-L2, CDR-L3, CDRH1, and CDR-H2) or modeling the CDR-H3. The canonical CDRs have reasonably conserved folds (Nowak et al., 2016; Kelow et al., 2022) (Figure 1C). The latter problem is arguably the most difficult and critical, as the CDR-H3 is the most variable (Figure 1D), and also plays the major role in binding (Marks and Deane, 2017; Regep et al., 2017; Ruffolo et al., 2020; Abanades et al., 2022a).

There is a diversity of methods to approach any of these sub-problems individually, or predicting the entire multimeric gamut of variable domains. However, attention is often focused around CDR-H3 prediction accuracy given its central role in binding and function. Compilation of the available antibody structure prediction methods that leverage recent advances in machine learning are listed in Table 1.

3 Current machine learning methods tackling antibody structure prediction

3.1 What data fuel the models?

Antibody-based deep learning methods require antibody structures for training and validation which are typically downloaded from the Protein Data Bank (PDB). At the time of writing, there were approximately 7,000 redundant antibody structures. Although this sample of the antibody sequence space represents a small subset of all possible antibody molecules (>10^{15}), it can still be used to model most naturally occurring antibodies (Krawczyk et al., 2018).

Databases such as AbDb (Ferdous and Martin, 2018) and SAbDab (Dunbar et al., 2014) curate such antibody-specific information. Most of the antibody structure prediction tools use these two resources that facilitate the creation of training, validation, and test datasets.

3.2 How is the antibody model quality assessed?

In the original AlphaFold2 work and CASP competition in general, the structural accuracy is calculated using GDT_TS (Kryshtafovych et al., 2021). This score is a measure of structural alignment between the model and native structure that is capable of indicating fold similarities. All antibodies are already of the same fold and one needs to account for differences in single loops (e.g., CDR-H3), where the RMSD is more suitable.

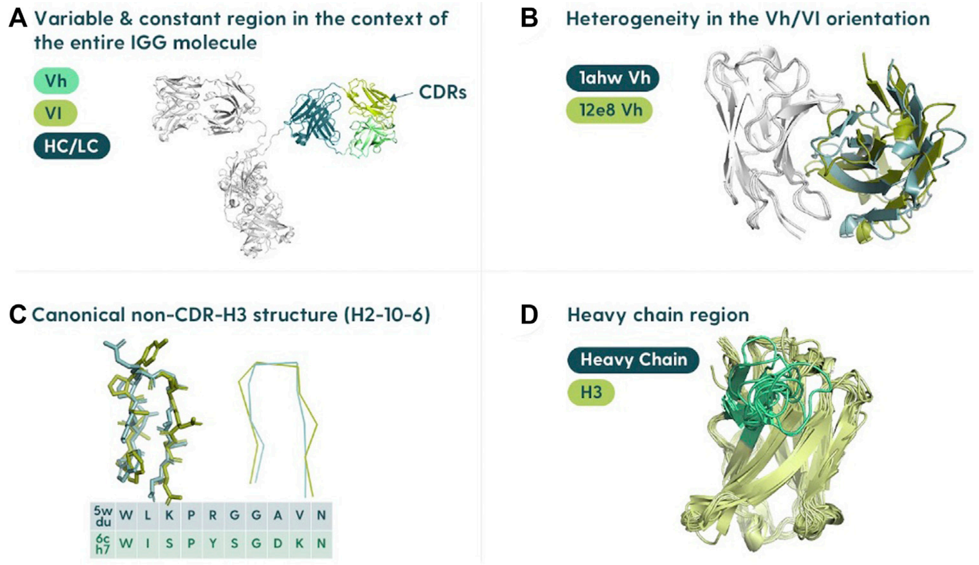

A Variable & constant region in the context of the entire IGG molecule
Vh
Vl
HC/LC
CDRs
B Heterogeneity in the Vh/Vl orientation
1ahw Vh
12e8 Vh
C Canonical non-CDR-H3 structure (H2-10-6)
5w du W L K P R G G A V N
6c h7 W I S P Y S G D K N
D Heavy chain region
Heavy Chain
H3

FIGURE 1
Specifics of the antibody structure in the context of modeling. **(A)** Variable region in the context of the entire antibody structure. The antibody binding site is located in the variable region composed of the variable heavy (Vh) and variable light (Vl) polypeptide chains associated with the constant portions (HC/LC). **(B)** Heavy/light chain orientation. The orientation of the Vh and Vl is not constant, and differing angles can affect the shape of the binding site. **(C)** Canonical structures of CDRs. Most of the binding residues (the paratope) are found in the complementarity-determining regions (CDRs). There are three CDRs on each of the heavy and light chains. All the CDRs except the CDR-H3 cluster into a set of "canonical shapes" depending on residues in key positions. **(D)** Heterogeneity of CDR-H3. CDR-H3 is not only the most variable of the regions but also usually the most important for antigen binding.

TABLE 1 Compilation of the available antibody structure prediction methods that leverage recent advances in machine learning. For each method, we describe the general goal (e.g., CDR prediction or whole variable region prediction), the accuracy of the most difficult region, the CDR-H3, its code/server availability, and the source paper. Please note that the CDR-H3 root mean square deviations (RMSDs) are not directly comparable as they could have been obtained from a different test set and are sometimes calculated in a different fashion, e.g., based on Cα or main chain heavy atom positions. As a baseline and reference point, we also include the AlphaFold2 predictions since many methods report values with respect to that method.

Method	Problem addressed	Model characteristic	CDR-H3 prediction accuracy	Corresponding AlphaFold2 accuracy	Availability	Source
DeepH3	CDR prediction	Residual neural network	2.2 Å backbone atoms are used	N/a	https://github.com/Graylab/deepH3-distances-orientations	Ruffolo et al. (2020)
Quaternion and Euler angle combined method	CDR prediction	Graph neural network	SAbDab benchmark: 2.29 Å	N/a	N/a	Son et al. (2022)
ABlooper	CDR prediction	Graph neural network based	RosettaAntibody benchmark: 2.49 Å; SAbDab latest structures: 2.72 Å. Backbone atoms were used	RosettaAntibody benchmark: 2.87 Å	https://github.com/oxpig/ABlooper	Abanades et al. (2022a)
DeepSCAb	Antibody side chain prediction	Residual neural network	Not applicable (side chain prediction)	N/a	https://github.com/Graylab/DeepSCAb	Akpinaroglu et al. (2022)
NanoNet	Heavy chain prediction	Residual network	RosettaAntibody benchmark: 2.38 Å; Nanobodies: 3.16 Å. Backbone atoms were used	Nanobodies: 2.88 Å	https://github.com/dina-lab3D/NanoNet	Cohen et al. (2022)
AbodyBuilder2	Variable region prediction	Based on AlphaFold2 structural module	AbodyBuilder2 benchmark: 2.81 Å. Backbone atoms were used	AbodyBuilder2 benchmark: 2.90 Å	https://github.com/oxpig/ImmuneBuilder	Abanades et al. (2022b)
EquiFold	Variable region prediction	SE(3)-equivariant neural network	IgFold benchmark: 2.86 Å (only N, Cα, C, and O RMSD)	IgFold benchmark: 3.02 Å	https://github.com/Genentech/equifold	Lee et al. (2022)
tfold-Ab	Variable region prediction	Based on AlphaFold2, using language models in the place of Evoformer	IgFold benchmark: 2.74 Å; SAbDab-22H1-Ab benchmark: 3.03 Å. Backbone atoms were used	IgFold benchmark: 3.02 Å; SAbDab-22H1-Ab benchmark: 3.18 Å	https://drug.ai.tencent.com/en	Wu et al. (2022)
xTrimoABfold	Variable region prediction	Based on AlphaFold2, using language models in place of Evoformer	1.25 Å (Cα only)	1.47 Å	N/a	Wang et al. (2022)
IgFold	Variable region prediction	Graph transformer using language model AntiBERTy	IgFold benchmark: 2.99 Å (backbone heavy atoms)	IgFold benchmark: 3.02Å	https://github.com/Graylab/IgFold	Ruffolo et al. (2022a)
AbFold	Variable region prediction	Based on AlphaFold2	AbFold benchmark: 3.04 Å, (backbone heavy atoms)	AbFold benchmark: 3.14 Å (backbone heavy atoms)	N/a	Peng et al. (2023)
AbBERT-HMPN	Sequence and structure generation	Deep graph neural network employing language models with generative capabilities	2.38 Å backbone atoms were used	N/a	N/a	Gao et al. (2022)
RefineGNN	CDR prediction and design	Graph neural network with generative capabilities	2.50 Å (Cα only)	N/a	https://github.com/wengong-jin/RefineGNN	Jin et al. (2021)
AbDockGen	CDR-H3 prediction, design, and antigen docking	Graph neural network-based with generative capabilities	Not applicable (docking scores reported)	N/a	https://github.com/wengong-jin/abdockgen	Jin et al. (2022)
DiffAb	Antibody sequence and the structure design	Diffusion method	Test set of 19 complexes: 3.246 Å (Cα only)	N/a	https://github.com/luost26/diffab	Luo et al. (2022)
DeepAb	Variable region prediction	Residual neural network	RosettaAntibody benchmark: 2.33 Å; therapeutics: 2.52 Å. Backbone heavy atoms were used	N/a	https://github.com/RosettaCommons/DeepAb	Ruffolo et al. (2022b)

Methods that attempt the modeling of the entire variable region report the entire chain RMSD, further dividing it into the individual CDR RMSDs. Nevertheless, here, the gains in structure prediction accuracy are typically small as such predictions are already of very good quality, excluding the CDR-H3.

Since the CDR-H3 is the most difficult to predict, it is the benchmark point of reference for accuracy across different models. Methods typically report the RMSD of the prediction versus the native structure. RMSD can be calculated using two different approaches. Typically, RMSD is calculated based on the backbone atoms (N, C, CA, and O) or C-alpha (Cα) carbons only, with the latter always being lower. RMSD can also be calculated after aligning the entire variable region or only after aligning the CDR-H3 atoms. The latter method leads to a slightly lower reported RMSD, as it causes bias in the structural alignment for a better fit.

3.3 What methods and techniques are used for modeling individual antibody loops and individual chains?

Due to the importance of the CDR-H_3 loop, many methods focus exclusively on modeling this region. For instance, DeepH_3 and ABlooper were designed for CDR-H_3 loop prediction, rather than addressing the entire variable region. DeepH_3 is based on a residual network architecture that receives one-hot encoding of the sequence to be predicted as input. In terms of residual network size, it is much smaller than RaptorX (Källberg et al., 2012) on which it is based (6 1D + 60 2D) with only 3 1D + 25 2D blocks. It operates by predicting discretized inter-residue distances (assigning distances into equally spaced intervals) and orientation angles which are employed for full structure reconstruction by RosettaAntibody. In contrast, ABlooper is based on equivariant graph neural networks [E(G)NNs], which are equivariant to translations and rotations in 3D space (Satorras et al., 2021). ABlooper allows for coordinate uncertainty estimation by calculating the agreement between five independently trained neural network models. The chief advantage of ABlooper is speed, as it can produce hundreds of structures within seconds as opposed to previously available homology modeling methods that required around a minute.

Beyond CDR-H_3-focused predictions, one needs to contextualize this loop to the rest of the heavy chain. One of the early machine learning models that could perform whole chain predictions is NanoNet. Originally designed as a predictor of single-chain antibodies (Deszyński et al., 2022), it can also predict heavy chains of canonical antibodies. Similar to DeepH_3, it is a residual neural network (ResNet) that relies on one-dimensional convolutions to map sequence elements to three-dimensional coordinates. Unlike DeepH_3, which operates on invariant features (residue distances and orientation angles), NanoNet operates on a single frame of reference by aligning all PDB heavy chains to a single reference structure. Owing to this, the predictions of the NanoNet are 3D coordinates, not requiring further translation into the structure as is the case with invariant features. In the context of the entire heavy chain prediction, authors report 2.38 Å accuracy for CDR-H_3 (solutions in the region of 1 Å can be considered near-native). Similar to ABlooper, NanoNet is rapid, allowing for predicting thousands of structures in a matter of seconds. However, the predicted structures are often of bad physical quality [e.g., atomic clashes, D-amino acids, etc. (Fernández-Quintero et al., 2023)], requiring refinement.

3.4 What architectures and techniques are currently used to predict the entire antibody variable region structure?

Prediction of the entire variable region requires modeling and multimeric assembly of both heavy and light chains. Herein, DeepAb is a network that predicts discretized residue distances and orientation angle bins that are then passed for structure realization using Rosetta. The chief innovation of DeepAb is the usage of a language model as an input to the network. Employing embedding (internal efficient representations of input antibody sequences) for prediction offers an opportunity for the network to perform prediction on more efficient features extracted by the language model. Furthermore, the network employs attention mechanisms that allow tracking of which residues contribute to each other's predictive signal.

Residual neural networks provide limited ways to abstract invariant three-dimensional information. Representing the entire variable region structure as a graph (as was the case with ABlooper) offers a solution to this problem. For instance, one can encode amino acids as nodes (using features such as amino acid and position) and draw edges between nodes/residues in proximity (e.g., within 8 Å heavy atom distance). Graph neural networks (GNNs) have increasingly gained popularity; this is hallmarked by ABlooper, IgFold, and EquiFold. The authors of EquiFold employed a coarse-grained representation for nodes to demonstrate its power within the framework of a SE (3) (special Euclidean (3) group ensuring rotation and translation equivariance) equivariant network. Ensuring geometric equivariance helps the network in learning features that can be rotated and translated. A more abstract representation using quaternions and Euler angles to encode the amino acids as invariant representations and as an extension of RefineGNN residues has been shown to achieve CDR-H3 predictions in the region of 2.5 Å. IgFold is another GNN-based method that also employs embeddings from AntiBERTy, which is trained on 500-m antibody sequences to supplement its prediction of the entire variable region.

Three key components have contributed to the success of AlphaFold2: the Evoformer, invariant point attention (IPA), and recycling. First, AlphaFold2 infers spatial constraints between amino acids by extracting evolutionary information embedded within multiple sequence alignments (MSAs) using its Evoformer module. In parallel, this information is fed into a structural module that leverages IPA to predict coordinates. IPA is a novel attention mechanism designed to be invariant to rotations and translations by aligning the feature vectors based on the geometric relationship between the residues without changing their 3D positions. It has been shown that it improves the accuracy of protein structure prediction by enabling the network to better capture the complex spatial relationships between residues in a protein. Finally, the whole workflow is repeated or "recycled" three times to refine the prediction.

While AlphaFold2 was designed for predicting any arbitrary protein sequence, its main components have influenced the design of antibody-specific tools. There are variations in the implementation of each of the aforementioned three components. For example, IgFold uses separated weights for each IPA layer and gradient propagation through rotations. xTrimoFoldAb and tfold-Ab use language model embeddings to replace the Evoformer, before applying the learned constraints into the structural module. Other methods, such as ABodyBuilder2, demonstrated that one can use only the structural module without resorting to antibody-specific embeddings or modified Evoformers. The antibody-focused methods are more accurate than AlphaFold2, but these improvements are limited. One major advantage of antibody-specific methods is their efficient running time. For instance, ABodyBuilder2 achieves predictions in a matter of seconds, compared to tens of minutes for AlphaFold2. AlphaFold2's running time is comparatively long because of the MSA search step, which is unnecessary for antibody-specific methods.

The loss function drives the training of a model as it penalizes wrong predictions and rewards better ones. It is extremely important as one of the chief innovations of AlphaFold2 was the introduction of the frame aligned point error (FAPE) loss. This component exposes the model to information related to physicochemical constraints, such as proper chemical bond distances and angles, as well as penalizing atom clashes and other structural violations, and is also used in some of the antibody-specific models. However, because of the skewed difficulty in structure prediction, applying the same loss to each antibody region is not an ideal approach. For instance, xTrimoABFold employs focal loss focused on CDRs that are more difficult to predict. On the other hand, ABodyBuilder2 treats framework and CDR regions differently, clamping framework regions at a FAPE loss of 30 Å and the CDRs' FAPE loss at 10 Å.

3.5 How do networks approach fine-structural details beyond the backbone?

Despite the progress in predictions, a seemingly trivial problem faced by the networks is the physical plausibility of the produced models (Fernández-Quintero et al., 2022). It was observed in AlphaFold2 that the structure module can produce predictions violating physical constraints, such as atomic clashes. This is not only a problem of AlphaFold2-based methods, and methods such as NanoNet and EquiFold are also afflicted. Methods such as ABodyBuilder2 and IgFold employed OpenMM and Rosetta, respectively, to reduce the number of physical clashes in the model produced by a neural network. The number of non-physical distances can also be reduced by introducing various physical constraints at the training time (Eguchi et al., 2022; Kończak et al., 2022).

Although structure prediction is typically evaluated on its ability to recapitulate the backbone, the determination of the side chains is important for fine-grained modeling of molecular function, such as binding affinity. Methods such as ABodyBuilder2 and IgFold produce the backbone structures annotated with side chains. Other methods such as EquiFold use a novel coarse-graining scheme where atoms are mapped to coarse-grained "superatom" prior to structural modeling and then reverse-mapped to the individual atoms once the backbone is constructed (Akpinaroglu et al., 2022). Other methods such as NanoNet only produce the backbone. Side chains are typically added by algorithms such as SCWRL (Krivov et al., 2009) or PEARS (Leem et al., 2018), but recently an antibody-specific side chain prediction mechanism using convolutional neural networks has been introduced—DeepSCAb (Akpinaroglu et al., 2022). Altogether, fast and accurate prediction of all-atom models is key to using the antibody structures for practical drug discovery purposes.

4 Drug discovery perspective of antibody structure prediction

Antibodies are a well-established drug format, with the structure as a key component in aiding their discovery and development, paving ground for real-world applications of 3D modeling.

Antibody structures provide rich information that can be used to improve various prediction features, such as molecular recognition (Oh et al., 2021), liability detection (Irudayanathan et al., 2022), or developability screening (Jain et al., 2023). These models can complement wet-lab antibody discovery methods, such as immunization or phage display, to ultimately improve the selection of binders. For instance, the identification of antigen-specific antibodies was typically tackled using clonotype/sequence clustering methods. Machine learning has shown alternative ways to group these molecules such as by embeddings from variational autoencoders (VAEs) (Friedensohn et al., 2020), predicting paratope residues (Richardson et al., 2021), or clustering structures (Robinson et al., 2021). In particular, structural clustering can provide a highly translational interpretation of antibody binding. The methods described in this Review are highly scalable, making it possible to group thousands of structures.

The optimization of biologics is the process of improving an existing molecule, which already displays a variety of desirable properties, with regard to a set of physicochemical properties. Structural features can be employed to guide the optimization process. A trivial example would be to focus existing liability removal (e.g., deamidation) protocols on only surface-exposed residues, which can be identified based on a reasonably accurate three-dimensional model (Irudayanathan et al., 2022). Structural features can be indicative of successful therapeutics (Raybould et al., 2019; Ahmed et al., 2021), with some differences in the calculated results based on the underlying modeling method (Jain et al., 2023). In some cases, such as antibody–antigen docking, good quality models are needed to reach the results achieved by docking crystal structures (Schneider et al., 2021).

The most ambitious use of antibody structure prediction is for the *de novo* antibody design, where the goal is to computationally define an antibody sequence that can bind to a given target epitope. One approach to the *de novo* design that relies on structural predictions is "virtual screening," a methodology that has been practiced in small molecule drug discovery for decades but has only been recently applied to antibodies. This can involve the modeling of and selection from millions of antibody molecules, which are then funneled into a molecular docking approach (Schneider et al., 2021; Jin et al., 2022) or alternative binding site design methods (Rangel et al., 2022). The quality of the models is a key consideration as subtle changes in Vh/Vl arrangement, backbone, or side chain orientation can affect the quality of the predictions (Fernández-Quintero et al., 2022). In addition, any such efforts hinge on linking the antibody structural predictions to paratope–epitope interaction prediction. In this context, "zero-shot" predictions require the models to propose sequences binding a specific epitope without observing it, or any close variants of it, in the training/test sets.

Another approach to the *de novo* design is using generative methods. Herein, the latent space of the input (e.g., antibody sequences) is learned, providing a way to sample novel elements. Producing novel sequences based on transformer models has already been shown in general proteins (Rives et al., 2021) as well as in the antibody world (Melnyk et al., 2021; Saka et al., 2021; Shin et al., 2021; Shuai et al., 2021). Autoregressive methods such as IgLM (Shuai et al., 2021) offer a way to generate new binder sequences based on millions of sequences from natural repertoires. Such generation can also be biased toward sequences with certain biophysical properties by GANs (Amimeur et al., 2020). Most such methods, however, are currently sequence-driven but not structure-driven.

Structure holds the potential to provide more information than sequence alone (Kovaltsuk et al., 2017). Encoding the structural space, in the form of torsional angles using VAEs, has shown potential in generating novel 3D shapes (Eguchi et al., 2022). Leveraging structural information for generating paratopes to specific antigens should produce better results than using sequence alone (Jin et al., 2022). Higher quality structural models have the potential to inform better structure-generation methods, leading to more accurate emulation of molecular space than sequence alone. Embeddings generated by the inverse-folding of general proteins have already shown potential to be useful for B-cell epitope prediction (Hsu et al., 2022; Høie et al., 2023).

In the context of structure-conditioned generative methods, RefineGNN, AbDockGen, AbBERT-HMPN, and DiffAb go a step further than the modeling methods described in this review. They also provide a "compatibility" score for the structure and designed sequence. RefineGNN, AbDockGen, and AbBERT-HMPN are based on the iterative refinement of latent representations from graph neural networks, whereas DiffAb samples via a denoising diffusion model. The integration of structure prediction and sequence design is the next intuitive step superseding structure prediction, which holds the promise to enhance antibody-based drug discovery.

5 Conclusion

Advances in protein structure prediction have practical application in the discovery of new antibody drugs.

In general, accuracy increasing with respect to the pioneer in ML-based accurate structure prediction, AlphaFold2, is noticeable, but stay within an order of magnitude. Predictions of the CDR-H3 structure in particular appear to be "stuck" in the 2–3 Å heavy atom backbone RMSD interval. Difficulty in the prediction of CDR-H3 conformation could stem from the loops' flexibility (Wong et al., 2011; Fernández-Quintero et al., 2018; Jeliazkov et al., 2018) as well as the possible influence of the Vh/Vl orientation (Marze et al., 2016; Boucher et al., 2023). With only several thousand antibody structures at hand (Dunbar et al., 2014), it is challenging to study any flexibility or allosteric effects, but perhaps with a larger number of better quality cryo-EM structures we will increase the volume of structural information available. Efforts in improving antibody structure prediction might take the flexibility into account by scoring the CDR-H3 conformational ensemble rather than single "best structure" produced.

The main advantage of the antibody-specific methods with respect to AlphaFold2 is the speed. The antibody sequence space in a single individual [~10^9–10^{11} (Briney et al., 2019)] easily surpasses the human proteome (~20 k). The speed of antibody modeling methods is of utmost importance, as it directly translates to the mapping of the available antibody sequence space (Kovaltsuk et al., 2018; Olsen et al., 2022), antibody virtual screening (Schneider et al., 2021; Rangel et al., 2022), and the development of novel generative models (Eguchi et al., 2022).

Given the number of currently available antibody-specific structure predictions, it might be suitable to take stock of the state of the field and devote efforts into benchmarking the different methods as was the case with the two rounds of the Antibody Modeling Assessment competition (Almagro et al., 2011; Almagro et al., 2014). In the field of antibody discovery specifically, we could use the tools not only to test by a single measure of RMSD but also to assess how useful the structural predictions are for therapeutically minded tasks, such as lead optimization, docking, epitope, or paratope prediction.

Altogether the accuracy, speed, and accessibility of the current antibody modeling methods make it possible to apply structural information to various aspects of biologics discovery pipelines today. An incremental improvement to existing discovery approaches using structure-guided computational methods appears entirely feasible, while the field continues to move ever forward toward the "holy grail" of the true *de novo* antibody design.

Author contributions

IJ, WB, and KK contributed to the conception and design of the study. IJ and WB wrote the first draft of the manuscript. IJ, WB, PD, TS, JK, WW, BJ, SW, DC, JG, JL, SK, and KK wrote sections of the manuscript. TG prepared the figures. All authors contributed to the article and approved the submitted version.

Funding

This work was co-financed by the European Regional Development Fund within the Smart Growth Operational Programme 2014–2020 POIR.01.01.01-00-0962/21 (to KK).

Conflict of interest

IJ, WB, TG, PD, TS, JK, WW, BJ, SW, DC, and KK are employees of NaturalAntibody that develops data, software and machine learning solutions for the therapeutic antibody industry. JG and JL are employees of Alchemab. SK is an employee of UCB Pharma.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abanades, B., Georges, G., Alexander, B., and Deane, C. M. (2022a). ABlooper: Fast accurate antibody CDR loop structure prediction with accuracy estimation. *Bioinformatics* 38 (7), 1877–1880. doi:10.1093/bioinformatics/btac016

Abanades, B., Wong, W. K., Boyles, F., Georges, G., Alexander, B., and Charlotte, M. D. (2022b). ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins. Available at: https://www.biorxiv.org/content/10.1101/2022.11.04.514231v1 (Accessed November 4, 2022).

Ahmed, L., Gupta, P., Martin, K. P., Scheer, J. M., Nixon, A. E., and Kumar, S. (2021). Intrinsic physicochemical profile of marketed antibody-based biotherapeutics. *Proc. Natl. Acad. Sci.* 118 (37), 577118. doi:10.1073/pnas.2020577118

Akpinaroglu, D., Ruffolo, J. A., Mahajan, S. P., and Gray, J. J. (2022). Simultaneous prediction of antibody backbone and side-chain conformations with deep learning. *PloS One* 17 (6), 0258173. doi:10.1371/journal.pone.0258173

Almagro, J. C., Beavers, M. P., Hernandez-Guzman, F., Maier, J., Shaulsky, J., Butenhof, K., et al. (2011). Antibody modeling assessment. *Proteins* 79 (11), 3050–3066. doi:10.1002/prot.23130

Almagro, J. C., Teplyakov, A., Luo, J., Sweet, R. W., Kodangattil, S., Hernandez-Guzman, F., et al. (2014). Second antibody modeling assessment (AMA-II). *Proteins* 82 (8), 1553–1562. doi:10.1002/prot.24567

Amimeur, T., Shaver, J. M., Ketchem, R. R., Taylor, J. A., Clark, R. H., Smith, J., et al. (2020). Designing feature-controlled humanoid antibody discovery libraries using generative adversarial networks. Available at: https://www.biorxiv.org/content/10.1101/2020.04.12.024844v1 (Accessed April 13, 2020).

Boucher, L. E., Prinslow, E. G., Feldkamp, M., Yi, F., Nanjunda, R., Wu, S.-J., et al. (2023). 'Stapling' scFv for multispecific biotherapeutics of superior properties. *mAbs* 15 (1), 2195517. doi:10.1080/19420862.2023.2195517

Briney, B., Inderbitzin, A., Joyce, C., and Burton, D. R. (2019). Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* 566 (7744), 393–397. doi:10.1038/s41586-019-0879-y

Bujotzek, A., Dunbar, J., Lipsmeier, F., Schäfer, W., Antes, I., Deane, C. M., et al. (2015). Prediction of VH-vl domain orientation for antibody variable domain modeling. *Proteins* 83 (4), 681–695. doi:10.1002/prot.24756

Cohen, T., Halfon, M., and Schneidman-Duhovny, D. (2022). NanoNet: Rapid and accurate end-to-end nanobody modeling by deep learning. *Front. Immunol.* 13, 958584. doi:10.3389/fimmu.2022.958584

Deszyński, P., Młokosiewicz, J., Adam, V., Jaszczyszyn, I., Castellana, N., Bonissone, S., et al. (2022). INDI—Integrated nanobody database for immunoinformatics. *Nucleic Acids Res.* 50 (1), D1273–D1281. doi:10.1093/nar/gkab1021

Dunbar, J., Fuchs, A., Shi, J., and CharlotteDeane, M. (2013). ABangle: Characterising the VH-VL orientation in antibodies. *Protein Eng. Des. Sel. PEDS* 26 (10), 611–620. doi:10.1093/protein/gzt020

Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., et al. (2014). SAbDab: The structural antibody database. *Nucleic Acids Res.* 42, D1140–D1146. doi:10.1093/nar/gkt1043

Eguchi, R. R., Choe, C. A., and Huang, P.-S. (2022). Ig-VAE: Generative modeling of protein structure by direct 3D coordinate generation. *PLoS Comput. Biol.* 18 (6), 1010271. doi:10.1371/journal.pcbi.1010271

Ferdous, S., and Martin, A. C. R. (2018). AbDb: Antibody structure database-a database of PDB-derived antibody structures. *Database J. Biol. Databases Curation* 2018, bay040. doi:10.1093/database/bay040

Fernández-Quintero, M. L., Kokot, J., Waibl, F., Fischer, A. M., Quoika, P. K., Deane, C. M., et al. (2023). Challenges in antibody structure prediction. *mAbs* 15 (1), 2175319. doi:10.1080/19420862.2023.2175319

Fernández-Quintero, M. L., Kokot, J., Franz, W., Fischer, A.-L. M., Quoika, P. K., Deane, C. M., et al. (2022). Challenges in antibody structure prediction. Available at: https://www.biorxiv.org/content/10.1101/2022.11.09.515600v1 (Accessed November 9, 2022).

Fernández-Quintero, M. L., Loeffler, J. R., Kraml, J., Kahler, U., Kamenik, A. S., and Liedl, K. R. (2018). Characterizing the diversity of the CDR-H3 loop conformational ensembles in relationship to antibody binding properties. *Front. Immunol.* 9, 3065. doi:10.3389/fimmu.2018.03065

Friedensohn, S., Neumeier, D., Khan, T. A., Csepregi, L., Parola, C., Arthur, R. G. D. V., et al. (2020). Convergent selection in antibody repertoires is revealed by deep learning. Available at:https://www.biorxiv.org/content/10.1101/2020.02.25.965673v1 (Accessed February 26, 2020).

Gao, K., Wu, L., Zhu, J., Peng, T., Xia, Y., Liang, H., et al. (2022). Incorporating pre-training paradigm for antibody sequence-structure Co-design. Available at: https://arxiv.org/abs/2211.08406 (Accessed October 26, 2022).

Høie, M. H., Gade, F. S., Johansen, J. M., Würtzen, C., Winther, O., Nielsen, M., et al. (2023). DiscoTope-3.0 - improved B-cell epitope prediction using AlphaFold2 modeling and inverse folding latent representations. Available at: https://www.biorxiv.org/content/10.1101/2023.02.05.527174v1 (Accessed February 5, 2023).

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Tom, S., et al. (2022). Learning inverse folding from millions of predicted structures. Available at: https://www.biorxiv.org/content/10.1101/2022.04.10.487779v1 (Accessed April 10, 2022).

Irudayanathan, F. J., Zarzar, J., Lin, J., and Izadi, S. (2022). Deciphering deamidation and isomerization in therapeutic proteins: Effect of neighboring residue. *mAbs* 14 (1), 2143006. doi:10.1080/19420862.2022.2143006

Jain, T., Todd, B., and Vásquez, M. (2023). Identifying developability risks for clinical progression of antibodies using high-throughput *in vitro* and *in silico* approaches. *mAbs* 15 (1), 2200540. doi:10.1080/19420862.2023.2200540

Jeliazkov, J. R., Sljoka, A., Kuroda, D., Tsuchimura, N., Katoh, N., Tsumoto, K., et al. (2018). Repertoire analysis of antibody CDR-H3 loops suggests affinity maturation does not typically result in rigidification. *Front. Immunol.* 9, 413. doi:10.3389/fimmu.2018.00413

Jin, W., Barzilay, D. R., and Jaakkola, T. (2022). "Antibody-antigen docking and design via hierarchical structure refinement," in Proceedings of the 39th International Conference on Machine Learning, Baltimore, MA, July 17-23, 2022 (Proceedings of Machine Learning Research. PMLR).

Jin, W., Wohlwend, J., Barzilay, R., and Jaakkola, T. (2021). Iterative refinement graph neural network for antibody sequence-structure Co-design. Available at: http://arxiv.org/abs/2110.04624 (Accessed October 9, 2021).

Jumper, J., Evans, R., Alexander, P., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596 (7873), 583–589. doi:10.1038/s41586-021-03819-2

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., et al. (2012). Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.* 7 (8), 1511–1522. doi:10.1038/nprot.2012.085

Kaplon, H., Crescioli, S., Visweswaraiah, J., and Reichert, J. M. (2023). Antibodies to watch in 2023. *mAbs* 15 (1), 2153410. doi:10.1080/19420862.2022.2153410

Kelow, S., Faezov, B., Xu, Q., Parker, M., Adolf-Bryfogle, J., and Roland, L. D. (2022). A penultimate classification of canonical antibody CDR conformations. Available at: https://www.biorxiv.org/content/10.1101/2022.10.12.511988v1 (Accessed October 16, 2022).

Kończak, J., Janusz, B., Młokosiewicz, J., Satława, T., Wróbel, S., Dudzic, P., et al. (2022). Structural pre-training improves physical accuracy of antibody structure prediction using deep learning. Available at: https://www.biorxiv.org/content/10.1101/2022.12.06.519288v1 (Accessed December 9, 2022).

Kovaltsuk, A., Krawczyk, K., Galson, J. D., Kelly, D. F., Deane, C. M., and Trück, J. (2017). How B-cell receptor repertoire sequencing can Be enriched with structural antibody Data. *Front. Immunol.* 8, 1753. doi:10.3389/fimmu.2017.01753

Kovaltsuk, A., Leem, J., Kelm, S., James, S., Deane, C. M., and Krawczyk, K. (2018). Observed antibody space: A resource for Data mining next-generation sequencing of antibody repertoires. *J. Immunol.* 201 (8), 2502–2509. doi:10.4049/jimmunol.1800708

Krawczyk, K., Kelm, S., Kovaltsuk, A., Galson, J. D., Kelly, D., Trück, J., et al. (2018). Structurally mapping antibody repertoires. *Front. Immunol.* 9, 1698. doi:10.3389/fimmu.2018.01698

Krawczyk, K., Liu, X., Baker, T., Shi, J., and CharlotteDeane, M. (2014). Improving B-cell epitope prediction and its application to global antibody-antigen docking. *Bioinformatics* 30 (16), 2288–2294. doi:10.1093/bioinformatics/btu190

Krivov, G. G., Shapovalov, M. V., and Dunbrack, R. L., Jr. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* 77 (4), 778–795. doi:10.1002/prot.22488

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2021). Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins* 89 (12), 1607–1617. doi:10.1002/prot.26237

Lee, J. H., Yadollahpour, P., Watkins, A., Frey, N. C., Leaver-Fay, A., Stephen, R., et al. (2022). EquiFold: Protein structure prediction with a novel coarse-grained structure representation. Available at: https://www.biorxiv.org/content/10.1101/2022.10.07.511322v1 (Accessed October 8, 2022).

Leem, J., Georges, G., Shi, J., and CharlotteDeane, M. (2018). Antibody side chain conformations are position-dependent. *Proteins* 86 (4), 383–392. doi:10.1002/prot.25453

Lu, R.-M., Hwang, Y.-C., Liu, I -J., Lee, C.-C., Tsai, H. Z., Li, H. J., et al. (2020). Development of therapeutic antibodies for the treatment of diseases. *J. Biomed. Sci.* 27 (1), 1. doi:10.1186/s12929-019-0592-z

Luo, S., Su, Y., Peng, X., Wang, S., Peng, J., and Ma, J. (2022). Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. Available at: https://www.biorxiv.org/content/10.1101/2022.07.10.499510v1 (Accessed July 11, 2022).

Marks, C., and Deane, C. M. (2017). Antibody H3 structure prediction. *Comput. Struct. Biotechnol. J.* 15, 222–231. doi:10.1016/j.csbj.2017.01.010

Marze, N. A., Lyskov, S., and Gray, J. J. (2016). Improved prediction of antibody VL-VH orientation. *Protein Eng. Des. Sel. PEDS* 29 (10), 409–418. doi:10.1093/protein/gzw013

Melnyk, I., Das, P., Chenthamarakshan, V., and Lozano, A. (2021). Benchmarking deep generative models for diverse antibody sequence design. Available at:http://arxiv.org/abs/2111.06801 (Accessed November 12, 2021).

Nowak, J., Baker, T., Georges, G., Kelm, S., Klostermann, S., Shi, J., et al. (2016). Length-independent structural similarities enrich the antibody CDR canonical class model. *mAbs* 8, 751–760. doi:10.1080/19420862.2016.1158370

Oh, L., Dai, B., and Bailey-Kellogg, C. (2021). "A multi-resolution graph convolution network for contiguous epitope prediction," in Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics BCB '21 38, Chicago, IL, August 7-10, 2022 (Association for Computing Machinery).

Olsen, T. H., Boyles, F., and Deane, C. M. (2022). Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Sci. A Publ. Protein Soc.* 31 (1), 141–146. doi:10.1002/pro.4205

Peng, C., Wang, Z., Zhao, P., Ge, W., and Huang, C. (2023). AbFold - an AlphaFold based transfer learning model for accurate antibody structure prediction. Available at: https://www.biorxiv.org/content/10.1101/2023.04.20.537598v1 (Accessed April 21, 2023).

Rangel Aguilar, M., Bedwell, A., Costanzi, E., Taylor, R. J., Russo, Rosaria, Bernardes, G. J. L., et al. (2022). Fragment-based computational design of antibodies targeting structured epitopes. *Sci. Adv.* 8 (45), eabp9540. doi:10.1126/sciadv.abp9540

Raybould, M. I. J., Marks, C., Krawczyk, K., Taddese, B., Nowak, J., Claire, M., et al. (2019). Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl. Acad. Sci. U. S. A.* 116 (10), 4025–4030. doi:10.1073/pnas.1810576116

Regep, C., Georges, G., Shi, J., Popovic, B., and Charlotte, M. D. (2017). The H3 loop of antibodies shows unique structural characteristics. *Proteins* 85 (7), 1311–1318. doi:10.1002/prot.25291

Richardson, E., Galson, J. D., Paul, K., Kelly, D. F., Anne Palser, S. E. S., Watson, S., et al. (2021). A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-pertussis toxoid antibodies. *mAbs* 13 (1), 1869406. doi:10.1080/19420862.2020.1869406

Rives, A., Meier, J., Tom, S., Goyal, S., Lin, Z., Liu, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* 118 (15), 239118. doi:10.1073/pnas.2016239118

Robinson, S. A., Raybould, M. I. J., Schneider, C., Wong, W. K., Marks, C., and Deane, C. M. (2021). Epitope profiling using computational structural modelling demonstrated on coronavirus-binding antibodies. *PLoS Comput. Biol.* 17 (12), 1009675. doi:10.1371/journal.pcbi.1009675

Ruffolo, J. A., Chu, L.-S., Mahajan, S. P., and Gray, J. J. (2022a). Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. Available at: https://www.biorxiv.org/content/10.1101/2022.04.20.488972v1 (Accessed April 21, 2022).

Ruffolo, J. A., Guerra, C., Mahajan, S. P., Sulam, J., and Gray, J. J. (2020). Geometric potentials from deep learning improve prediction of CDR H3 loop structures. *Bioinformatics* 36 (1), i268–i275. doi:10.1093/bioinformatics/btaa457

Ruffolo, J. A., Sulam, J., and Gray, J. J. (2022b). Antibody structure prediction using interpretable deep learning. *Patterns (New York, N.Y.)* 3 (2), 100406. doi:10.1016/j.patter.2021.100406

Saka, K., Kakuzaki, T., Metsugi, S., Kashiwagi, D., Yoshida, K., Wada, M., et al. (2021). Antibody design using LSTM based deep generative model from phage display library for affinity maturation. *Sci. Rep.* 11 (1), 5852. doi:10.1038/s41598-021-85274-7

Satorras, V. G., Hoogeboom, E., and Welling, M. (2021). "E(n) equivariant graph neural networks," in Proceedings of the 38th International Conference on Machine Learning, July 18-24, 2021 (Proceedings of Machine Learning Research. PMLR).

Schneider, C., Buchanan, A., Taddese, B., and CharlotteDeane, M. (2021). DLAB-deep learning methods for structure-based virtual screening of antibodies. *Bioinformatics* 38, 377–383. doi:10.1093/bioinformatics/btab660

Shin, J.-E., Riesselman, A. J., Kollasch, A. W., McMahon, C., Simon, E., Sander, C., et al. (2021). Protein design and variant prediction using autoregressive generative models. *Nat. Commun.* 12 (1), 2403–2411. doi:10.1038/s41467-021-22732-w

Shuai, R. W., Ruffolo, J. A., and Gray, J. J. (2021). Generative Language modeling for antibody design. Available at: https://www.biorxiv.org/content/10.1101/2021.12.13.472419v2 (Accessed December 20, 2022).

Son, Y.-H., Shin, D.-H., Han, J.-W., Won, S.-H., and Kam, T.-E. (2022). "GNN-based antibody structure prediction using quaternion and euler angle combined representation," in 2022 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), Yeosu, South Korea, October 26-28, 2022, 1–4.

Wang, Y., Gong, X., Li, S., Yang, B., Sun, Y., Shi, C., et al. (2022). xTrimoABFold: De novo antibody structure prediction without MSA. Available at: http://arxiv.org/abs/2212.00735 (Accessed November 30, 2022).

Wong, S. E., Sellers, B. D., and Jacobson, M. P. (2011). Effects of somatic mutations on CDR loop flexibility during affinity maturation. *Proteins* 79 (3), 821–829. doi:10.1002/prot.22920

Wu, J., Wu, F., Jiang, B., Liu, W., and Zhao, P. (2022). tFold-Ab: Fast and accurate antibody structure prediction without sequence homologs. Available at: https://www.biorxiv.org/content/10.1101/2022.11.10.515918v1 (Accessed November 13, 2022).

TYPE Review
PUBLISHED 07 August 2023
DOI 10.3389/fmolb.2023.1221626

OPEN ACCESS

EDITED BY
Shubhra Ghosh Dastidar,
Bose Institute, India

REVIEWED BY
Moutusi Manna,
Central Salt and Marine Chemicals Research Institute (CSIR), India
Frederic Cadet,
INSERM Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), France

*CORRESPONDENCE
Sandeep Kumar,
✉ Sandeep.Kumar@modernatx.com

†PRESENT ADDRESS
Sandeep Kumar, Computational Protein Design and Modeling, Computational Sciences, Moderna Therapeutics, Cambridge, MA, United States

RECEIVED 12 May 2023
ACCEPTED 10 July 2023
PUBLISHED 07 August 2023

CITATION
Bauer J, Rajagopal N, Gupta P, Gupta P, Nixon AE and Kumar S (2023), How can we discover developable antibody-based biotherapeutics?
Front. Mol. Biosci. 10:1221626.
doi: 10.3389/fmolb.2023.1221626

How can we discover developable antibody-based biotherapeutics?

Joschka Bauer[1,2], Nandhini Rajagopal[2,3], Priyanka Gupta[2,3], Pankaj Gupta[2,3], Andrew E. Nixon[3] and Sandeep Kumar[2,3]*†

[1]Early Stage Pharmaceutical Development Biologicals, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach/Riss, Germany, [2]In Silico Team, Boehringer Ingelheim, Hannover, Germany, [3]Biotherapeutics Discovery, Boehringer Ingelheim Pharmaceuticals Inc., Ridgefield, CT, United States

Antibody-based biotherapeutics have emerged as a successful class of pharmaceuticals despite significant challenges and risks to their discovery and development. This review discusses the most frequently encountered hurdles in the research and development (R&D) of antibody-based biotherapeutics and proposes a conceptual framework called biopharmaceutical informatics. Our vision advocates for the syncretic use of computation and experimentation at every stage of biologic drug discovery, considering developability (manufacturability, safety, efficacy, and pharmacology) of potential drug candidates from the earliest stages of the drug discovery phase. The computational advances in recent years allow for more precise formulation of disease concepts, rapid identification, and validation of targets suitable for therapeutic intervention and discovery of potential biotherapeutics that can agonize or antagonize them. Furthermore, computational methods for *de novo* and epitope-specific antibody design are increasingly being developed, opening novel computationally driven opportunities for biologic drug discovery. Here, we review the opportunities and limitations of emerging computational approaches for optimizing antigens to generate robust immune responses, *in silico* generation of antibody sequences, discovery of potential antibody binders through virtual screening, assessment of hits, identification of lead drug candidates and their affinity maturation, and optimization for developability. The adoption of biopharmaceutical informatics across all aspects of drug discovery and development cycles should help bring affordable and effective biotherapeutics to patients more quickly.

KEYWORDS

biotherapeutics, drug discovery and development, developability, biopharmaceutical informatics, machine learning, computational biophysics

1 Introduction

Since the inception of hybridoma technology, which facilitated large-scale monoclonal antibody (mAb) production, biotherapeutics have experienced significant growth (Koehler and Milstein, 1975). The Food and Drug Administration's (FDA) approval of the pioneering mAb therapeutic, muromonab or Orthoclone OKT3, in 1986 (Smith, 1996), set the stage for numerous groundbreaking developments in biotherapeutics. As of 2022, over 110 approved mAbs and more than 65 mAbs in phase-2/3 and phase-3 clinical trials have emerged (Kaplon et al., 2022). Clinically, mAbs have demonstrated their efficacy in treating serious conditions such as neurodegenerative diseases, autoimmune diseases, and diverse types of cancers (Reichert et al., 2009; Lu et al., 2020).

Despite the promising trajectory of biotherapeutics, the biopharmaceutical industry faces mounting pressure due to decreasing productivity and increasing research and development (R&D) costs. The average R&D cost surged from $1.2 billion in 2007 (adjusted United States dollar value of $1.6 billion in 2020) to $2.8 billion in 2016 (equivalent to $3.1 billion in 2020) (DiMasi and Grabowski, 2007; DiMasi et al., 2016; Farid et al., 2020). Concurrently, the success rate of phase-1 to approval dropped from 30% in 2007 to 12% or lower in 2016 (Farid et al., 2020). These trends suggest the presence of several challenges along various stages of discovery and development of novel biological therapeutics. A lack of detailed understanding of disease biology, the inability of model systems to reliably predict human diseases and outcomes of therapeutic interventions, the lack of efficacy, target-mediated toxicity and other safety issues, and suboptimal developability profiles are among the major reasons that may contribute to drug failures during clinical trials (Mehta et al., 2017; Fogel, 2018). The identification of new targets presents additional challenges toward development of novel therapeutic concepts and discovery of multi-specific biotherapeutics, resulting in low approval rates despite high development costs (Swinney and Anthony, 2011). Next-generation biotherapeutics such as nanobodies, bi- and multi-specific antibodies, and T-cell receptor mimetics are broadening clinical applications (Strohl, 2018); however, these novel formats are often more challenging to develop into marketed biologic drug products (Runcie et al., 2018; Wang et al., 2019; Sawant et al., 2020). Furthermore, as the biopharmaceutical industry shifts its focus toward patient convenience, drug product development processes must be tailored to emerging routes of drug administration such as subcutaneous or intravitreal delivery, necessitating high-concentration protein formulations (HCPFs) (Garidel et al., 2017). These requirements introduce additional challenges to the manufacturability and developability of novel drugs. Integrating developability early in the drug discovery process can help avoid costly delays or failures at later stages and potentially increase the likelihood of success during clinical trials and approvals. Numerous technological advancements have been made since the approval of the first mAb to overcome challenges in the R&D pipelines and accelerate novel drug discovery and development (Martin et al., 2023). However, every new technology comes with associated risks and limitations (Gray A. C. et al., 2020; González-Fernández et al., 2020).

In silico techniques have been well established in small-molecule drug discovery (Shaker et al., 2021). Over the past decade, considerable progress has been made toward developing *in silico* strategies for the discovery and development of biologic drugs as well. In fact, developability has emerged as a key concept for biologic drugs over this time (Jarasch et al., 2015; Kumar and Singh, 2015; Bailly et al., 2020; Garripelli et al., 2020; Khetan et al., 2022; Mieczkowski et al., 2023). A variety of computational tools and procedures are now employed across various stages of drug development, such as hit selection, lead identification, optimization, affinity maturation, and early developability assessment. However, a significant potential of *in silico* technologies toward the discovery of biotherapeutics still remains untapped. As collaborative academic and industrial initiatives continue to demonstrate the viability of *in silico* antibody discovery techniques, it is important to acknowledge that the nascent nature of these methods often results in a lack of historical evidence to support their success and therefore requires a cultural shift toward proactive adoption of innovation to continually improve drug discovery and development processes. To address these challenges and enhance the success rate of novel targets, there is an urgent requirement for an integrated vision to create a platform that streamlines biotherapeutic discovery and development via syncretic use of experimentation and computation. Such a vision would not only accelerate the development of new biotherapeutics and reduce costs but also expand the druggable target space.

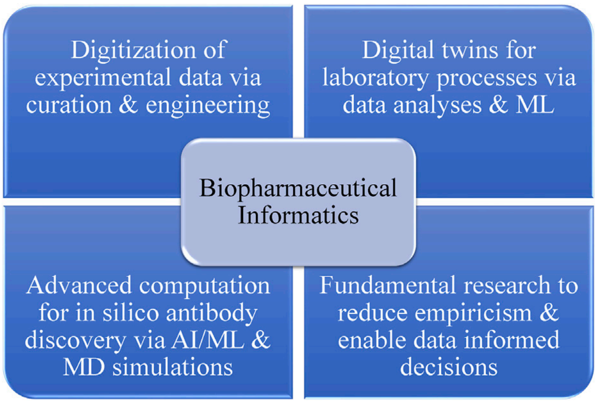

Digitization of experimental data via curation & engineering
Digital twins for laboratory processes via data analyses & ML
Biopharmaceutical Informatics
Advanced computation for in silico antibody discovery via AI/ML & MD simulations
Fundamental research to reduce empiricism & enable data informed decisions

FIGURE 1
Strategic components for the vision of biopharmaceutical informatics. The digital transformation of the biopharmaceutical industry, achieved through capturing and curing experimental data, can enable the development and continuous improvement of digital twins for laboratory processes and prediction of experimental results before their execution. Fundamental research connecting molecular sequences, structures, and dynamics of biologic drug candidates can enhance our understanding of experimental observations, reduce empiricism, and enable more data-informed decision-making at various project stages. Moreover, the integration of computational learning technologies with principles of molecular modeling and simulations can potentially facilitate the *in silico* discovery of biotherapeutics. It is important to note that the key to biopharmaceutical informatics lies in the syncretic use of experimentation and computation, with a shared goal of making the discovery and development of biotherapeutics more efficient.

2 Biopharmaceutical informatics: integrating drug discovery and development

In the realm of biotherapeutics, it is crucial for drug candidates to be both developable and functional. Biotherapeutic drug candidates often encounter developability challenges related to manufacturing, safety, immunogenicity, efficacy, pharmacology, and drug product heterogeneity. Many of these risks can be linked to the inherent physicochemical properties of a biologic drug candidate, as determined by its protein sequence, three-dimensional structure, and molecular dynamics (MD) (Xu et al., 2018). Considering the intrinsic physicochemical properties of a biotherapeutic drug candidate, which are encoded in its amino acid sequence and structure, early in the discovery and development can help identify and mitigate risks associated with various developability issues, such as chemical, conformational, colloidal, and physical instabilities. Moreover, by employing the innovative approach of biopharmaceutical informatics, these sequence–structural attributes can be modified for improved developability as described previously by Kumar et al. (2018a). Figure 1 outlines the primary components of biopharmaceutical informatics. This interdisciplinary field advocates for the digital transformation of the biopharmaceutical industry by converting experimental data collected during drug discovery and development phases into FAIR (findable, accessible, interoperable, and reusable) information systems. These systems can be leveraged by data scientists to create predictive tools such as digital twins of actual laboratory processes. Additionally, the field promotes the increased use of AI/ML (artificial intelligence/machine learning) and computational biophysics to address fundamental challenges in drug discovery and development through research. Biopharmaceutical informatics seeks to enable data-driven decision-making at every stage of biologic drug discovery and development. Developability is a key aspect of biopharmaceutical informatics, encompassing both *in silico* tools and experimental studies such as developability assessments. Rooted in the energy landscape theory, the concept of developability posits that the conformational ensembles and potential energy landscapes of large macromolecules, like mAbs, change with their environment (e.g., pH, temperature, and physicochemical state) (Onuchic, 1997; Ma et al., 2000; Kumar et al., 2009). As a result, the physicochemical properties of conformational ensembles of biotherapeutics under a given set of environmental conditions dictate their biophysical experiment outcomes. If proteins with the same size and fold are analyzed under identical conditions using standardized experiments, differences in the results should be attributable to sequence–structural variations among the proteins. The ability to predict experimental outcomes by analyzing the sequence–structural characteristics of biotherapeutic drug candidates is a primary goal for biopharmaceutical informatics, as part of the development of computational methods that facilitate discovery of antibodies *in silico* (DAbI).

Optimal synergies and benefits can be achieved by integrating cost-effective, rapid computational methods with standardized biophysical experimental studies, which are characteristic of current developability assessments in biologic drug discovery and early-stage product development (Zurdo, 2013; Jarasch et al., 2015; Xu et al., 2018). Late-stage development approaches typically focus on assessing the changing conditions of a single molecule in the drug manufacturing process using quantitative unit operation models (Smiatek et al., 2020), while early-stage approaches require analyzing a diverse set of molecules under identical conditions. Biopharmaceutical informatics plays a pivotal role in bridging the gap between biologic drug discovery and development by improving the understanding of the relationship between macromolecular sequence–structure–function and developability.

A key challenge in biopharmaceutical informatics is correlating the “macroscopic” experimentally determined properties of a biologic with its “microscopic” sequence–structure features computed *in silico*. Uncovering these correlations can guide molecular sequence optimization strategies, proactively addressing potential obstacles in drug product development by predicting the performance of the final drug candidate in the streamlined platform processes used during development stages. This process necessitates combining data from standardized biophysical experiments with descriptors computed from molecular modeling and simulations in a common database. Various statistical and machine learning approaches can be employed to develop mathematical models that predict the solution behavior of mAbs based solely on their sequence–structure information, depending on the available data (Tomar et al., 2016; Jain et al., 2017b; Chiu et al., 2019; Hebditch and Warwicker, 2019; Lecerf et al., 2019; Raybould et al., 2019; Starr and Tessier, 2019; Kuroda and Tsumoto, 2020; Zhang et al., 2020). As a result, the interdisciplinary field of biopharmaceutical informatics aims to seamlessly integrate techniques from computational and experimental biophysics, information technology, and data science to provide data-driven inputs for the decision-making framework for all stages of biologic drug discovery and development.

3 Opportunities for computation at various stages of biotherapeutic discovery and early development

There are numerous opportunities to collaboratively apply computational and experimental tools to facilitate faster and more efficient drug engineering and development. In this review article, we present a diverse set of use cases at various stages of biotherapeutics discovery and development projects that could benefit with increased use of computation in synchrony with the experiments to demonstrate the practical feasibility of our vision. The major challenges faced at distinct stages of biotherapeutic discovery and early drug development are described in Table 1 along with potential computational opportunities to address them. The pros and cons of these computational opportunities are also presented in Table 1. It is important to note that the field has not matured uniformly across all stages of discovery and development cycles for biotherapeutics. For example, computational approaches to developability assessments and lead optimization (LO) are currently more advanced than *in silico* antibody discovery and *in silico* formulation development. Moreover, there are also opportunities to modify the workflows and transitions between the different discoveries and development stages in view of the rapidly growing capabilities of computation. These opportunities are described in the following sections.

3.1 Antigen optimization

The discovery of antibody-based biotherapeutics adheres to a stepwise approach once a target antigen or multiple antigens for simultaneous targeting in a multi-specific format have been identified. The initial phase entails producing enough target antigens to enable animal immunization, *in vitro* selection of antigen-specific antibodies, and functional activity characterization. However, some antigens exhibit favorable expression *in vivo* but encounter conformational stability and solubility issues *in vitro*, outside the cellular context (Qing et al.,

TABLE 1 Opportunities for the expanded use of computational approaches throughout the discovery and development process of biotherapeutics.

Process stage	Typical problems	Potential applications of computational approaches	Pros	Cons
In vitro synthesis of immunogens/antigens to generate corresponding antibodies	1. Availability of structural models for immunogens and accurate definition of epitope(s) of therapeutic interest 2. Aggregation tendency, protein insolubility, and reduced conformational stability may result in limited material availability for immunization experiments 3. Epitope(s) of therapeutic interest might not be immunodominant	1. Protein structure prediction and precise definition of epitope(s) of therapeutic interest 2. Sequence/structure-based optimization for improved solubility via APR disruption, supercharging; and increased conformational stability via residue scan can help improve quantity as well as quality of material needed for immunization 3. Strategies for disruption or masking of immunodominant but therapeutically irrelevant epitopes to improve chance of antibody binders to the therapeutically relevant epitopes	1. Protein structure is crucial for structure-based approaches to drug discovery defining epitopes of therapeutic interest. The emergence of AI-based protein structure prediction methods has enhanced the structural definition of immunogens in recent years 2. Judiciously selected mutations at single or multiple sites can significantly improve the availability of immunogen material in the laboratory	1. Confidence levels in different regions of the structure should be considered, as flexible regions are typically predicted with lower confidence levels 2. Defining the epitope(s) of therapeutic interest and avoiding mutations in and around them is important 3. Implementing site-directed mutagenesis of immunogens to improve material availability also requires a cultural shift among experimental scientists
Antibody generation	1. Animal immunizations can be time-consuming, expensive, and may yield inconsistent results 2. The lead antibody molecule identified through animal immunization may necessitate humanization and developability enhancements 3. Humanized mice and display technologies do not entirely capture the complete human immunome 4. Phage and yeast display technologies can quickly identify high-affinity binders, but these may require further optimization for developability	1. Generative AI can aid in designing antigen-specific and agnostic libraries with incorporated developability features 2. Virtual screening of antibody libraries against given antigen(s)/epitope(s), followed by docking and structure-based affinity enhancements 3. Utilizing computational methods to design phage and yeast display libraries for enhanced developability and/or affinity 4. Employing computational approaches to redesign antibody CDRs for altered specificities	1. Adopting computational methods can reduce timelines and costs associated with antibody discovery 2. Expanded druggable antigen space 3. Opportunities to explore a broader sequence diversity, thereby maximizing the odds for antibody discovery compared to conventional methods 4. Addressing developability during library design can help reduce time required for lead optimization	1. Emerging technology 2. Necessitates more extensive validation and experimental demonstration of its capabilities before routine project use 3. Requires a cultural shift from experimentally driven antibody discovery to computationally driven approaches
Hit selection and lead identification	1. Sequencing of identified hits 2. Epitope mapping of the hits to ensure the desired therapeutic effect in the absence of structural models for the antigen-antibody complex 3. Experimental evaluation of several hundreds of candidates for functionality and developability can be time and resource-intensive	1. Establishment of suitable sequencing pipelines 2. Computational prediction of epitopes and paratopes for epitope mapping purposes 3. In-silico evaluations of candidates for developability and manufacturability can facilitate the selection of developable hits and identification of lead candidate(s) with favorable developability characteristics 4. Development of digital twins for biophysical processes via computational biophysics and data science	1. Incorporation of computational assessments can aid in guiding hit selection for experimental testing 2. Proactive consideration of developability can help reduce costs and efforts to identify lead molecules 3. Opportunities to enhance our understanding of the connection between molecular sequence-structural properties and experimental outcomes	1. Greater availability of data is needed to connect ’microscopic’ sequence-structural features of antibodies with the ’macroscopic’ biophysical outcomes 2. Lack of digitization and digital transformation present significant challenges 3. A cultural shift from protecting experimental data to sharing it with computational scientists is required among discovery scientists
Lead optimization	Lead candidates may require humanization, affinity optimization, and elimination of physicochemical liabilities in the CDRs for enhanced developability	1. Structure-based modeling of the lead candidates can assist in their humanization, affinity maturation, and identification of potential sequence/structural motifs that may contribute to their physicochemical degradation. Access to this information can help direct protein engineering strategies for lead optimization 2. Assessment of the optimized lead candidates for their drug likeness	1. Computational guidance for lead optimization efforts can decrease timelines and costs 2. This aspect represents the most developed application of computational protein design in biotherapeutic drug discovery 3. Numerous well-developed computational solutions are available	1. There remains cultural resistance to the adoption of computational protein design for lead optimization among industrial scientists 2. Greater dissemination of successful case studies, where computational protein design makes a difference, is needed to raise awareness

(Continued on following page)

TABLE 1 (*Continued*) Opportunities for the expanded use of computational approaches throughout the discovery and development process of biotherapeutics.

Process stage	Typical problems	Potential applications of computational approaches	Pros	Cons
Early stage developability assessments	1. Assessing molecular stability and compatibility of drug candidates, identified during drug discovery, with platform processes utilized in drug development 2. Adapting to multiple product development goals such routes of administration and product presentations	1. Structure prediction of full length antibodies and novel formats 2. In-silico development of formulations 3. Employing multi-scale simulations to anticipate platform compatibility and evaluate molecular responses to stresses encountered during manufacturing, storage, and transportation 4. Utilization of predictive algorithms to determine suitable bioprocess conditions 5. Establishing digital twins for various facets of drug development	1. Developing full-length models of the drug substance can facilitate improved prediction of molecular origins of dominant degradation routes during manufacturing, storage, and shipping 2. Accelerating formulation process development and saving costs of drug development can be achieved through pH and buffer screening of antibody formulations via in-silico characterization of molecular integrity of the drug substance 3. Resource savings can be realized with the development of digital twins	1. Computationally intensive calculations 2. Need for improved correlations between experimental results and molecular simulations 3. Consistent availability of development data across different projects 4. Requirement for greater investments in the digitalization of drug development data

2022). Producing recombinant antigens can be particularly challenging for certain target classes, such as membrane proteins (G protein–coupled receptors and ion channels) (Bill et al., 2011). If antigen binding is impacted by the *in vitro* conformational stability and/or solubility of the antigen, then these issues may hinder the entire antibody discovery strategy and functional validation of the antibody hits.

Computational methods can aid in the redesign of antigens with enhanced conformational stability and solubility when a threedimensional crystal structure or model is available. Bioinformatic tools can enable crystal structure refinement, modeling of breaks and gaps, loop modeling, energy minimization and molecular dynamics simulations to support antigen redesign. When the crystal structure of an antigen is unavailable, protein structure prediction techniques can often estimate it (Nimrod et al., 2018). For example, homology-based structure modeling can be employed using crystal homologs. A sequence identity of at least 30% between the protein of interest and its crystal homologue is typically sufficient for structure generation through homology modeling. However, some novel targets may not have homologs with existing crystal structures. This can be due to the inherent difficulty in obtaining crystal structures of membrane-associated proteins, which often have poor solubility. Membrane proteins represent a significant class of drug targets, and the discovery pipeline frequently proceeds without knowledge of the antigen structure. In such challenging cases, recent groundbreaking advances in *de novo* protein structure prediction techniques have achieved remarkable success and accuracy by leveraging machine learning and deep learning algorithms (AlQuraishi, 2019; Gao et al., 2020; Pereira et al., 2021; Jones and Thornton, 2022). Deep learning–based structure prediction methods, such as AlphaFold2 and RoseTTAFold, combined with physical modeling, have outperformed numerous conventional approaches (Baek et al., 2021; Jumper et al., 2021; Pereira et al., 2021; Jones and Thornton, 2022). Understanding of the antigen's three-dimensional structure can be crucial for accurately assessing its stability and solubility, computationally. This knowledge can also help enhance solubility without sacrificing stability and functional activity, allowing for the extraction of crystal structures, and facilitating experimental assays that measure target binding. Care should be taken, however, to minimize the impact of such mutations on the overall molecular structure of the target antigen and preserve its potential to generate adequate immune response to epitopes of therapeutic interest. Bioinformatics can also support rational strategies to immunize only therapeutically relevant epitopes on the antigen surface. This means epitopes that may be immune-dominant but are of no therapeutic interest or relevance can be either eliminated or masked to facilitate the immunization of the desired epitopes of therapeutic importance.

3.2 Antibody generation

Immunization strategies have long been employed to generate high-affinity antibodies, using previously expressed and purified antigens to establish immune reactions in animals (typically laboratory mice, humanized/transgenic mice, or other animals like chickens, rabbits, or cows). Antibody binding to specific antigens can be obtained through techniques such as hybridoma (Koehler and Milstein, 1975), single B cells (Yu et al., 2008), or screening natural and/or synthetic antibody libraries via display technologies using phage or yeast (Benatuil et al., 2010; Chen and Sidhu, 2014; Alfaleh et al., 2020; Gray A. et al., 2020; Nagano and Tsutsumi, 2021; Ledsgaard et al., 2022; Valldorf et al., 2022). Promising candidates are selected and validated using antigen-binding assays that align with the research target profile. Currently used methods in the biopharmaceutical industry for antibody generation are almost exclusively experimental, and depending on the techniques used, it can take several months before an initial set of antibody-based binders is available for further investigation and lead identification. Fully synthetic human antibody libraries containing Fabs chosen for their biophysically favorable development characteristics have been developed using experimental means (Valldorf et al., 2022). Special emphasis has been placed on selecting molecules with enhanced chemical, conformational, and colloidal stabilities (Tiller et al., 2013). The availability of such libraries can significantly help accelerate the discovery of antibody-based biotherapeutics by pre-paying for developability.

The concept of optimized antibody libraries for generating developable antibodies can be integrated with *de novo* computational databases containing an immense variety of human-like light- and heavy-chain combinations (Pan and Kortemme, 2021; Akbar et al., 2022). Targeted mutations at specific sequence positions [e.g., complementarity-determining regions (CDRs)] in the antibody sequences could further broaden the library, either to recognize different antigens or to optimize binding affinity toward a specific antigen (Ledsgaard et al., 2022). Recently, a generative adversarial network was successfully employed to create a diverse library of novel antibodies that emulate somatically hypermutated human repertoire responses (Amimeur et al., 2020). This *in silico* method further revealed residue diversity throughout the variable region, which could be useful for additional computational tools like CDR redesign. CDR redesign utilizes a highly developable antibody framework and modifies the original CDRs, or paratope, to recognize a new antigen. In recent years, noteworthy progress has been made in designing not only thermodynamically stable but also biologically functional antibodies (Baran et al., 2017).

Computational technologies, initially developed for small-molecule drug discovery, can also be applied to antibody-based drug discovery. Once fully developed and implemented, these computational methods will provide additional means to generate diverse antibody binders against a target antigen. These methods will not only help reduce animal use in biologic drug discovery but also decrease reliance on experimental trial and error for finding initial hits. Initial case studies describing such methods are beginning to emerge in the literature (bioRxiv.org for preprints) (Sever et al., 2019; Wilman et al., 2022). Additionally, it becomes feasible to find potential binders to difficult targets, thereby expanding the druggable target space for antibody-based biotherapeutics.

Figure 2 provides an overall conceptual roadmap for Discovery of antibodies in silico (DAbI). The proposed roadmap encompasses three major parts where each part can have multiple stages depending upon the project in hand. In the first part, the key is to use different computational algorithms to generate medicine-like human antibody sequence libraries *in silico*. These libraries can be either antigen-specific or antigen-agnostic and are of orthogonal utilities. For example, creation of antigen- or epitope-specific antibody libraries via machine learning can help us achieve early success in each antibody discovery project by facilitating a focused path to the discovery of lead candidates toward the antigen and support the therapeutic concept. A biological analog of such libraries shall be the sequence repertoires obtained from immunized animals, hybridomas, or the results obtained by panning the display libraries against a specific antigen. However, such libraries have to be generated repeatedly for each different antigen or epitope. Antigen-/epitope-agnostic libraries on the other hand can be incredibly useful toward supporting multiple drug discovery projects simultaneously. Such libraries can be thought of as naive B-cell repertoires obtained from humanized animals prior to immunization with specific antigens. The computationally generated naive antibody repertoires can potentially capture greater sequence diversities than those feasible from humanized animals, display technologies, or observable B-cell repertoires. Within a discovery organization, such libraries have to be constructed only once and be potentially useful toward pre-computation of binders for all the targets of interest to the organization. These pre-computed antibody binder libraries can potentially accelerate early antibody discovery projects because now the discovery process does not have to wait for availability of target reagent in the laboratory. Therefore, such libraries can be particularly useful toward difficult to express and purify targets such as membrane proteins. Irrespective of the purpose of *in silico* generated antibody libraries, it is important to generate structural models of (at least) the variable regions of the antibodies sampled from these libraries. The generated structures can then be used for assessing their medicine-likeness and developability. Early elimination of non–medicine-like antibodies from such libraries can improve their utility and differentiate them from those generated using the experimental means solely. The structural models can also be used for predicting antibody paratopes. Many computational methods are currently available for the structural prediction of antibodies. The major challenges in this field include prediction of HCDR3 conformation and pairing of the light- and heavy-chain variable regions (Fernández-Quintero et al., 2023).

In addition to the design of the *in silico* antibody libraries, currently available computational methods also provide an opportunity to design single or a few human antibody variable regions against specific antigen epitopes *de novo* (Chowdhury et al., 2018; Nimrod et al., 2018). The design process can also commence with a structural model of an antigen:antibody (Ag:Ab) complex, generated using molecular docking of the antigen and antibody structures (Nimrod et al., 2018). Subsequently, the affinity of the antigen toward the antibody can be either altered by randomly introducing sequence variations or selectively re-designing interfaces using structure-based approaches (Nimrod et al., 2018). For example, interfacial residues in the Ag:Ab complexes that significantly contribute to their stability and instability can be identified through computational alanine (Ala) scanning. In the following step, the identified residue positions can be scanned for mutations that either increase or decrease the stability of the Ag:Ab complex and enhance or reduce the affinity of the antibody toward its cognate antigen (Sheng et al., 2022), depending on the project requirements. Another appealing alternative for rational antibody design involves hotspot grafting with CDR loop swapping, which only requires information about interactions with the antigen (Liu et al., 2017).

The goal of epitope-driven antibody generation is to design an antibody variable region with a paratope that complements the given epitope. Since CDRs make up most of the paratope, initial efforts to design epitope-specific antibodies have focused on *ab initio* CDR redesign and modeling. OptCDR (Pantazes and Maranas, 2010), used in conjunction with Rosetta Antibody Modeler, generates epitope-specific high-affinity CDRs by selecting the most feasible canonical loop conformations followed by iterative model optimization and improvements in binding energy. This method enables the generation of a focused library of antibody binders, quite like hit sequences obtained from experiments. OptCDR was later optimized (OptMAVEn) to consider the entire fragment variable (Fv) region rather than just CDRs as the starting point for generating antibody binders (Li et al., 2014), allowing for the incorporation of humanness at the antibody generation stage through careful selection of human framework region residues. Further advances have incorporated MD simulations for accurate evaluation of

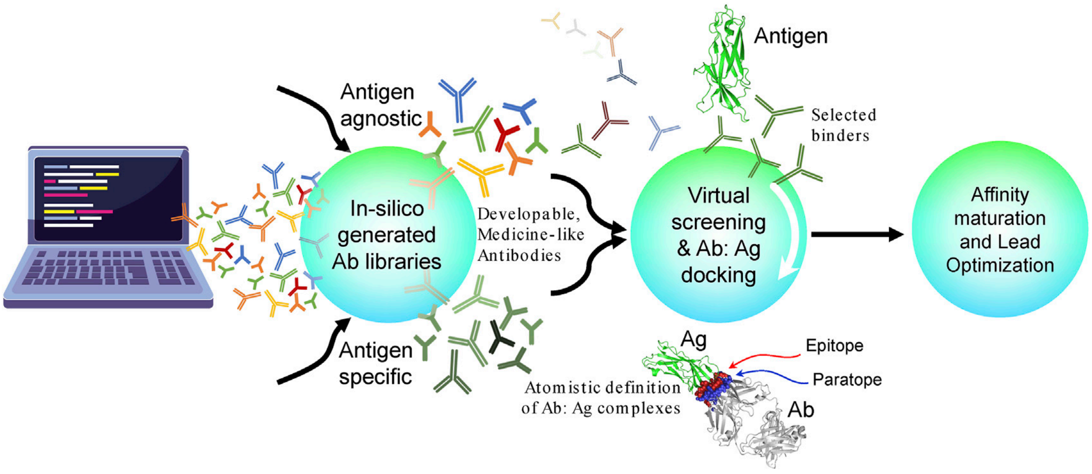

Antigen agnostic
Antigen specific
In-silico generated Ab libraries
Developable, Medicine-like Antibodies
Antigen
Selected binders
Virtual screening & Ab: Ag docking
Affinity maturation and Lead Optimization
Ag
Epitope
Paratope
Ab
Atomistic definition of Ab: Ag complexes

FIGURE 2
Conceptual roadmap for the discovery of antibodies *in silico* (DAbI). This conceptual roadmap can be divided into three major parts that can be developed either independently or in synchrony. The first part focuses on the *in silico* generation of medicine-like, antigen-agnostic, or specific antibody sequence libraries. Several machine learning algorithms are currently being developed to facilitate the *in silico* generation of antibodies. In the second part, these *in silico* generated antibodies and their structural models can be used to screen against a given antigen or an epitope on an antigen via virtual screening, docking, or other computational chemistry-based algorithms. Conversely, a large set of potential antigens can also be pre-screened against the antibody libraries using the same computational technologies. In both cases, the goal is to obtain atomistic definitions of putative antibody–antigen complexes. At this stage, it is preferable to virtually screen a larger number of antibodies (e.g., 1–10 million) and then select a much smaller number (e.g., 10–100) for docking simulations. This will help speed up the calculations and save computational resources. It is also important to quantitatively assess the quality of modeled antibody–antigen complexes by comparing them against crystal structures of other antigen–antibody complexes. A third option is to convert the whole or portions of the *in silico* generated antibody libraries into molecular libraries suitable for phage or yeast display and then pan them against a diverse panel of desired antigens. In the third part, the structural models of the putative antibody–antigen complexes obtained previously can be used to identify potential lead antibody candidates and modify their binding affinities to the desired levels via single- or multi-residue mutations in the paratope regions through computational protein design. These structural models can also be used to impart cross-reactivity to homologous antigens from other non-human species and/or to even create surrogate antibodies. Care should be taken to avoid introducing residues susceptible to physicochemical degradation and therefore reducing the developability of the lead candidates. It is important to note that DAbI will require changing the discovery workflows because it is pre-paying for developability and may therefore require significantly reduced effort during lead optimization (LO).

binding energetics (Chowdhury et al., 2018). A one-to-one residue matching method called epitoping, which starts from antibody structures with basic shape complementarity, was developed to obtain an accurate epitope–paratope binding match (Nimrod et al., 2018). Although this process requires a pre-identified approximate match, it can be considered for lead optimization to improve binding.

Recent advancements in generative deep learning and the availability of approximately 2,000 solved crystal structures of the antibody–antigen complexes have opened possibilities for structure-based *de novo* antibody generation. A proof-of-concept study utilizing a variational autoencoder (VAE)–based generative algorithm demonstrated the capability to directly generate 3D coordinates of antibody backbones that complement a specific epitope (Eguchi et al., 2022). Additionally, another deep learning algorithm was developed to learn the 3D features of antibodies from 1D sequences, enabling the generation of antibody sequences with desired structural characteristics (Akbar et al., 2022). Although the proof-of-concept study primarily aimed to achieve high-affinity binder antibody sequences for a given epitope, the method holds potential for encoding additional features, allowing the model to be tailored to produce highly developable sequences. As stated previously, generation of epitope-specific antibodies or libraries thereof has immediate applications for individual drug discovery projects, since the knowledge of epitopes is often required for defining novel therapeutic concepts.

3.3 Early screening for developability of *in silico* generated antibody libraries

Once the *in silico* antibody sequence libraries have been generated, it is worth assessing the generated antibody sequences for developability and advancing highly developable sequences to further stages of discovery. The developability assessment tools to be employed here can be ported over easily from those used at the hit selection and lead identification, lead optimization, and early development stages in the conventional biotherapeutic discovery and development workflows.

Lipinski’s “rule-of-five” revolutionized the discovery and development of small molecules by providing guidelines for improving their solubility and permeability (Lipinski, 2000). However, establishing similar rules for new biological entities (NBEs) has proven challenging due to their complex structures. In response, researchers have turned to biophysical evaluations and computational approaches to better understand these entities and overcome inherent obstacles. Biophysical evaluations of clinical-stage antibodies have contributed to the empirical definition of analogous boundaries, offering valuable insights for NBE development (Jain et al., 2017b; Raybould et al., 2019; Jain et al., 2023). Additionally, marketed antibodies have been profiled using calculated physicochemical descriptors, in an approach known as the DEvelopability Navigator *In Silico* (DENIS) (Ahmed et al., 2021;

Licari et al., 2022). These advances have significantly contributed to our understanding of NBEs and their development processes.

Biotherapeutics can undergo various levels of conformational changes over time, which presents significant challenges regarding conformational stability during manufacturing, shipping, and storage. This is because the environment of a biotherapeutic drug candidate can influence its structure, highlighting the importance of understanding these complex molecules in more detail. To address this, biophysical analysis employs a variety of techniques, such as thermodynamic, spectroscopic, and hydrodynamic methods, for characterizing protein-based drug candidates. These techniques are routinely used during the discovery phase to guide the identification and characterization of the lead drug candidates. Some properties commonly assessed during biophysical analysis include post-translational modifications (e.g., glycosylation, deamidation, isomerization, oxidation, and fragmentation), aggregation, self-association, hydrophobicity, molecule pI, and viscosity for high-concentration liquid formulations. While these techniques are well established, they can be time- and resource-consuming and demand expert knowledge and advanced instrumentation. This has driven researchers to seek more efficient and accessible methods for obtaining critical data. *In silico* tools can predict the intrinsic biophysical properties of drug candidates along with identifying their degradation routes, whose knowledge is important for establishing appropriate formulation strategies. These tools demonstrate significant relationships between the Fv domain sequences and physicochemical properties that define antibody developability. For example, post-translational modification sites, such as deamidation, aspartate isomerization, oxidation, and fragmentation can be identified using computational approaches (Irudayanathan et al., 2022; Vatsa, 2022). Similarly, hydrophobic interaction chromatography (HIC) retention times have been successfully correlated with sequence and structure features through diverse methods such as quantitative structure–property relationship (QSPR) modeling and machine learning (Jain et al., 2017a; Jetha et al., 2018; Karlberg et al., 2020). Although solution and colloidal state properties are challenging to predict due to multiple influencing factors, computational tools like SOLpro and PROSO II have demonstrated their ability to predict solubility upon expression with an accuracy of ~75% (Magnan et al., 2009; Smialowski et al., 2012). The isoelectric point (pI) is a crucial physicochemical property for mAbs. It has been associated with specific developability aspects such as thermostability, viscosity, and resistance to high molecular weight species formation at low pH. Tools like MassLynx, Vector NTI, and EMBOSS (Rice et al., 2000) calculate pI based on sequence data, achieving results within a 15% range of experimentally determined values (Goyon et al., 2017). Tools that predict the pI based on protein structure can provide a more accurate result, since the underlying residue p*K*a values are calculated by considering the residual microenvironments. Viscosity is also a critical factor in the colloidal stability of biologics and is influenced by electrostatics and hydrophobicity, which are in turn determined by the Fv sequence and structure. The *in silico* tool, spatial charge map (SCM), can identify highly viscous antibodies based on the mAb structure (Agrawal et al., 2015). Biomolecule aggregation is related to sequence and structural characteristics, such as the presence of aggregation-prone regions, hydrophobicity (Münch and Bertolotti, 2010), electrostatics (Buell et al., 2013), and dipole moments (Tartaglia et al., 2004), which enable both sequence- and structure-based computational predictions. Various *in silico* tools play a significant role in guiding mAb candidate design with high colloidal stability by predicting the impact of single or multiple amino acid exchanges on aggregation propensity. Alternative tools such as TANGO, PASTA, FoldAmyloid, SALSA, and AggreRATE-Pred can detect aggregation-prone regions based on the physicochemical principles of secondary structure elements, particularly the ability to form intermolecular cross-β-structures (Fernandez-Escamilla et al., 2004; Trovato et al., 2007; Zibaee et al., 2007; Garbuzynskiy et al., 2010; Walsh et al., 2014; Rawat et al., 2019). In summary, these *in silico* tools can effectively predict various biophysical properties of biotherapeutics. Their high-throughput capabilities make them particularly attractive for biophysical assessments during various stages of the drug discovery process.

3.4 Hit selection and lead identification

Following the production of antigen-binding antibodies through immunized animals, hybridoma cells, or phage and yeast display techniques, the variable regions of the antibodies are sequenced, and the binders are validated in the conventional workflows adapted by the biopharmaceutical industry. The immunization methods, strength and diversity of the immune responses, and sequencing technologies used can yield numerous unique hits, particularly via B-cell cloning and repertoire sequencing. Subsequently, these diverse hits must be prioritized to identify the most promising lead candidates, necessitating extensive resources to experimentally test each hit and confirm antigen binding.

Several bioinformatic techniques can aid in prioritizing and selecting hits for *in vitro* confirmation of antigen binding and lead identification (Figure 3). A common strategy involves clustering hits into high-, medium-, and low-binding bins based on the initial estimates, analyzing each bin for heavy- and light-chain germline diversity, and then examining CDR diversity to select multiple representatives from each germline pair in each bin for experimental testing. Alternatively, hits can be binned based on the germline pair and CDR diversity, with selections made according to their estimated antigen binding. At this stage of hit selection, developability aspects can also be considered using computational tools introduced in the previous section. In a basic application, heavy- (HC) and light-chain (LC) sequences of hits can be scored based on the presence of potential chemical degradation motifs, aggregation-prone regions (APRs), and T-cell immune epitopes present in or overlapping with the CDRs of the heavy and light chains. The scoring schemes can be further optimized by assigning different weights based on which CDRs contain these motifs and whether they are in the Vernier zones or middle of the CDRs.

Structure-based approaches require accurate three-dimensional antibody fold information, typically generated via homology modeling. This process includes 1) identifying a high-identity structural template for framework (FW) regions, 2) loop

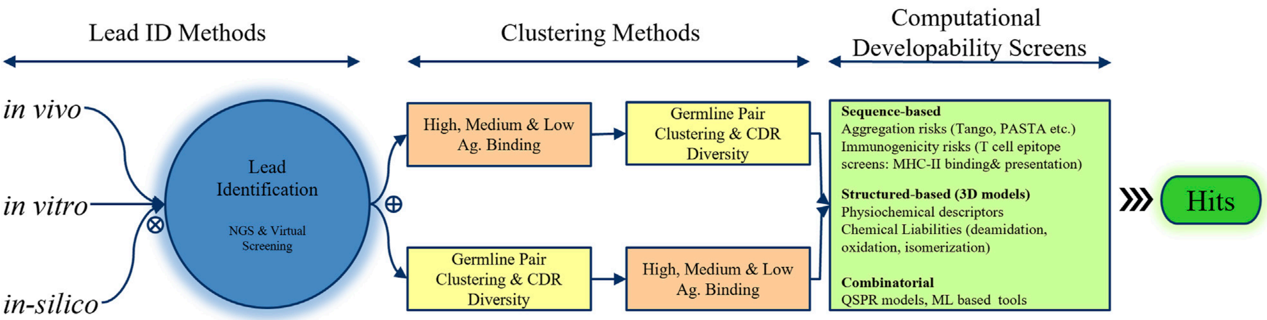

Lead ID Methods
Clustering Methods
Computational Developability Screens
in vivo
in vitro
in-silico
Lead Identification
NGS & Virtual Screening
High, Medium & Low Ag. Binding
Germline Pair Clustering & CDR Diversity
Germline Pair Clustering & CDR Diversity
High, Medium & Low Ag. Binding
Sequence-based
Aggregation risks (Tango, PASTA etc.)
Immunogenicity risks (T cell epitope screens: MHC-II binding& presentation)
Structured-based (3D models)
Physiochemical descriptors
Chemical Liabilities (deamidation, oxidation, isomerization)
Combinatorial
QSPR models, ML based tools
Hits

FIGURE 3
Integration of *in vivo*, *in vitro*, and *in silico* approaches for hit selection in the discovery phase of the pharmaceutical industry. Next-generation screening and virtual screening methods are employed to identify promising leads, which are then prioritized using clustering techniques based on 1) antigen binding and 2) a combination of germline pair clustering and CDR diversity. Finally, computational developability screens that analyze the amino acid sequence, structure, and combinatorial methods such as QSPR or machine learning are performed to select the most promising hits.

modeling of LCDR1-3 and HCDR1-2 using canonical loop conformations, 3) HCDR3 loop modeling and optimization of the orientation of heavy-chain variable region (VH) and light-chain variable region (VL), and 4) sidechain packing and refinement. The key challenges involve obtaining high-resolution templates with optimal VH-VL orientations and accurately modeling loops, particularly the HCDR3 loop. Recent progress in Fv structure modeling has led to advanced tools, such as RosettaAntibody (Weitzner et al., 2017; Adolf-Bryfogle et al., 2018; Schoeder et al., 2021), AbPredict2 (Lapidoth et al., 2018), ABodyBuilder (Leem et al., 2016), LYRA (Klausen et al., 2015), MoFvAb (Bujotzek et al., 2015), and Kotai Antibody Builder (Yamashita et al., 2014), which demonstrate high performance in the AMA-II benchmark test. Commercial packages like Molecular Operating Environment (MOE) and BioLuminate are popular for high-throughput full-length Fv structure modeling. A detailed discussion of recent advancements in Fv structure modeling tools can be found in focused reviews (Fernández-Quintero et al., 2023). Additionally, tools like FREAD, H3LoopPred, SPHINX, MODELER, PLOP, SCWRL, BetaSCPWeb, Chothia canonical assignment, and SCALOP have significantly contributed to full-length Fv region three-dimensional structure modeling. Tools such as TopModel efficiently examine the structure for cis-amide bonds, D-amino acids, and steric clashes, allowing for rapid evaluation of model quality and accuracy prior to conducting further analysis (Norman et al., 2019; Wilman et al., 2022; Fernández-Quintero et al., 2023). The generated three-dimensional structural models of all or a subset of hits can then be analyzed regarding their physicochemical descriptors, such as pI, charge, dipole moment, and solvent-exposed hydrophobic and ionic patches. These physicochemical properties have been demonstrated to potentially influence the chemical, conformational, colloidal, and physical stabilities of antibodies, and consequently their developability. In subsequent studies, a few of the best hits are rigorously tested in the laboratory for biological function, cross-reactivity across species, non-specific binding, and pharmacological indicators, such as serum stability. This process results in the identification of one or more lead candidates.

3.5 Virtual screening and docking as potential alternatives to *in vitro* hit selection and lead identification

Identification of potential binders through immunization campaigns can be accomplished using bioinformatics tools for paratope and epitope prediction, followed by rapid virtual screening, as outlined in Part 2 of the *in silico* roadmap, we call DAbI (Figure 2). This approach involves three-dimensional structure modeling of a diverse antibody sequence library and screening it against a given antigen by taking advantage of the shape and charge complementarity between the epitopes and paratopes. The antibody libraries to be screened can be endowed with the biophysical characteristics desired from a developability perspective as described previously.

Small-molecule drug discovery has successfully employed virtual screening to identify binders from a library of drug candidates (Gorgulla et al., 2020; Maia et al., 2020; Yan et al., 2020). Typically, millions of small-molecule drug candidates undergo structural and energetic screening processes through docking, pharmacophore-, or ligand-based approaches. Modern techniques involving computer vision, image-based, and geometric learning–based algorithms have reached advanced stages of validation and are now well established among the marketed small-molecule drugs designed using *in silico* methods (Eguida and Rognan, 2020; Gorgulla et al., 2020; Yan et al., 2020). Similarly, a curated and modeled antibody library may be treated as a potential set of drugs to be screened against a given antigen. However, directly applying these techniques may not be feasible due to the significant structural and functional differences between small-molecule drugs and large antibodies, with size (molecular weights, 500–1,000 Da versus approximately 25,000 Da for the Fv) being a primary concern even when considering only the Fv regions. Additionally, given the estimated theoretical diversity of B-cell repertoire (BCR) based on V(D)J recombination, which is about 10^{13}–10^{20} unique sequences, it is crucial to consider large antibody libraries to allow screening over a highly diverse sample space of paratopes.

Hypothetically speaking, we consider an antibody library of 1 million Fv sequences and assume a screening time of 1 min per Fv for a given antigen, the total runtime would amount to approximately 695 days (close to 23 months) for screening a single antigen against the library, which consists of only a small fraction of BCR diversity. Currently existing docking methods have runtimes of several minutes per complex. On the bright side, rapid virtual screening may not necessarily require rigorous energy-based binding evaluations employed in modern docking programs. Sacrificing the accuracy afforded by pose refinement can allow for greater speed in the screening process. Consequently, novel techniques have to be developed to enable the screening of large antibody libraries by considering the key aspects of the structural and chemical complementarity of the antigen:antibody interfaces and ensuring high-throughput rapid execution. An ideal *in silico* antibody virtual screening process could narrow down the potential binding hits to the order of 10^1–10^2, meaning that virtual screening would enable identifying binders at least as accurately as about one in a thousand to a few thousand sequences from the library, significantly impacting the discovery pipeline.

While *in silico* virtual screening does not replicate the generation of antigen binders via experimental methods in terms of binding affinity or functional efficacy, it can allow for comprehensive screening of the antibody library to identify all possible structural matches of epitopes and paratopes. Iterative refinement of these matches can help discover antibody binders to a given antigen with a diverse set of binding affinities and therefore suitable for antagonist as well as agonist function. Novel techniques, such as image-based and graph-based deep learning algorithms, have been proposed for identifying complementary paratope/epitope interfaces. These approaches can be further accelerated through pre-identified or predicted paratope and epitope information (Gainza et al., 2020; Pittala and Bailey-Kellogg, 2020; Akbar et al., 2021; Ripoll et al., 2021). Schneider et al. (2021) proposed a structure-based virtual screening method using voxel representation of the interfacing surface atom groups in their screening method called Deep Learning for AntiBodies (DLAB), adapted and extended from its small-molecule counterpart (Imrie et al., 2018). Recently proposed image fingerprinting–based approaches, with analogous applications in small molecules, show promising potential for protein interface matching and could be further expanded to predict paratope/epitope binders for hit selection (Gainza et al., 2020; Ripoll et al., 2021). More recently, a geometric deep learning method called ScanNet has been introduced to predict protein–protein and protein–antibody binding interfaces through geometric deep learning of three-dimensional structural features (Tubiana et al., 2022). Moreover, some of the paratope/epitope prediction methods involving deep learning of interfacial interactions may be extrapolated to interface screening and predicting binders.

The *in silico* virtual screening of antibodies against a given antigen can also borrow techniques such as fragment-based drug design (Sormanni et al., 2015; Sormanni et al., 2018) and pharmacophore modeling from the realm of small-molecule drug discovery. By facilitating the identification of binding sites, improving antibody–antigen docking, and enabling more accurate structure-based virtual screening, these methods can accelerate the development of novel therapeutic antibodies and enhance our ability to target a wider range of diseases and conditions.

Recent molecular docking protocols feature highly robust, energy-based scoring functions for evaluating and ranking protein–protein or protein–antibody binding partners. This offers a suitable toolkit for further optimization of hits identified through virtual screening of target antigens against an antibody library. Docking methods have demonstrated accurate prediction of protein-binding interfaces; however, speed has not been a priority for molecular docking programs. Although the current speed of implementation poses a bottleneck, rapid advancements in the field of protein–protein docking have spurred the development of new methods utilizing advanced machine learning algorithms and hybrid physics and learning-based technologies, promising faster docking methods soon. Moreover, such advancements may bridge the gap between virtual screening and docking, further accelerating *in silico* antibody screening, hit selection, and lead identification processes altogether.

Antibody–antigen docking has often been considered with paratope/epitope prediction and improving CDR modeling accuracy. SnugDock combines docking with accurate modeling prediction of the paratope (CDR loop construction), where the Rosetta Antibody Modeler operates alongside the docking protocol, iteratively improving docking and model prediction (Sircar and Gray, 2010; Jeliazkov et al., 2021). Additionally, methods employing more rigorous energy-defined constructs to evaluate multiple docking poses through the MM-GBSA (molecular mechanics—generalized Born solvent accessibility) method have shown promising outcomes (Shimba et al., 2016). Information-driven docking methods depend on a set of data to reduce the number of decoys, thus saving prediction time. Interface prediction-based methods, such as Antibody i-patch and EpiPred, focus on refining docking poses through paratope/epitope interface prediction (Krawczyk et al., 2013) By contrast, proABC adopts a more site-directed approach driven by the interface (paratope) (Olimpieri et al., 2013; Krawczyk et al., 2014). Advances in machine learning and deep learning algorithms have significantly contributed to enhancing docking prediction methods.

Other widely employed programs such as ClusPro, LightDock, ZDOCK, and HADDOCK, coupled with CDR and binding epitope information for directed/biased docking approaches, have shown promising results, with HADDOCK demonstrating notable performance improvement (Ambrosetti et al., 2020a). Pro-ABC-2, another information-driven docking approach and an updated version of Pro-ABC, utilizes deep learning convolutional neural networks (CNNs) for paratope prediction to assist in docking (Ambrosetti et al., 2020b). Such information-driven methods may also be applicable in pipelines using commercial docking techniques offered by MOE from Chemical Computing Group, PIPER from Schrodinger, and others with additional efforts.

Several other methods that employ deep learning through CNNs, recurrent neural networks (RNNs), or graph-based learning have demonstrated promise in predicting binding interfaces, consequently improving docking accuracy (Liberis et al., 2018; Deac et al., 2019; Pittala and Bailey-Kellogg, 2020; Lu et al., 2021; Myung et al., 2021; Vecchio et al., 2021; Davila et al., 2022). Additionally, research groups have been exploring the exceptional modeling performance of AlphaFold2 in docking prediction. The accelerated advancements in AI related to AlphaFold and other docking methods offer significant potential for the development of faster and more accurate docking programs in the future.

3.6 *In silico* affinity maturation of lead candidates

In a conventional discovery workflow, the lead candidates identified may have to be optimized for affinity, cross-reactivity, and developability. Among these, the focus is often on developability of the lead candidates. By contrast, DAbI may yield developable lead candidates already since the sequence and structural features that support good developability are already included in the library design (Part 1 of DAbI, see Figure 2). Depending on the library choice (antigen-agnostic or antigen-specific), the *in silico* generated lead candidate may have to be optimized for binding affinity and any residual physicochemical developability issues, particularly from the CDRs. For these reasons, the third part of our conceptual roadmap, DAbI (Figure 2), envisages an ability to adjust the binding affinity as per the project requirements. Depending upon the novel therapeutic concept (NTC), both enhancement (affinity maturation) and decrease (affinity de-maturation) in binding affinities may be required. However, affinity maturation may be required more often than de-maturation, particularly when the lead antibody binders have been derived from antigen-agnostic libraries. In our conceptual roadmap, both affinity maturation and de-maturation begin with a structural representation of the atomic interaction between two proteins, namely, the antigen and the antibody. The methodology's reliability depends on accurately analyzing the interacting sites. Therefore, co-crystallized antibody–antigen complexes are typically preferred over structure-based homology models or AI predictions, which may lead to less reliable results if CDRs are not precisely modeled. The *in silico* affinity maturation relies on accurate molecular interactions for free energy or MM-GBSA–based calculations (Comeau et al., 2023; Thorsteinson et al., 2023), highlighting the importance of improving antibody–antigen complex predictions and the implicit incorporation of multiple conformational ensembles to enhance the effectiveness of *in silico* calculations and optimize library design. Despite this limitation, these methods have been already applied to predicted antibody–antigen complexes (Rangel et al., 2022), facilitating the generation of *in silico* affinity maturation libraries (Conti et al., 2022; Thorsteinson et al., 2023).

The *in silico* scanning of the individual paratope residues yields potential mutations and estimates of the corresponding free energy changes in binding to the target. The subsequent challenge involves designing a combinatorial assembly of these mutations into a library suitable for phage/yeast display. This is because the *in silico* affinity maturation often involves computationally expensive calculations that tend to be more accurate at identifying the single point mutations rather than combinations thereof (Comeau et al., 2023; Thorsteinson et al., 2023). The physical display libraries built using computational guidance can be used to pan combinatorial mutations. Therefore, this part of DAbI requires an understanding of the limitations associated with the library size and panning methodology (Tsumoto and Kuroda, 2022). It is also in consonance with the spirit of biopharmaceutical informatics which calls for taking advantage of the strengths of computation and experiments in a synergistic manner. When combined with library technologies like phage display, computational tools have proven particularly powerful in guiding the design of affinity maturation libraries (Tiller et al., 2017; Nelson et al., 2018; Wang et al., 2018; Thorsteinson et al., 2023). Incorporation of additional considerations along with the binding affinity can help narrow down the mutations for experimental testing and therefore the size of the display libraries. At this stage, the mutations that enhance specificity, humanness, and CDR germlining along with developability can be considered by incorporating relevant physicochemical properties and stability criteria (Khan et al., 2023; Svilenov et al., 2023). Consequently, the selection of lead antibody candidates with high binding affinity and favorable biophysical properties can be achieved simultaneously. In-house, we successfully improved binding affinities of the antibody drug candidates 10- to 1,000-fold in multiple proprietary projects using this strategy.

Several studies have demonstrated the computational design of functional antibodies using multiple structural models supported by statistical or machine learning models (Nimrod et al., 2018; Liu et al., 2019; Amimeur et al., 2020). Upon selecting an initial antibody scaffold, mutations to enhance complementarity with a given epitope can be designed to obtain specific antibody binders to an antigen. For example, the generative adversarial network (GAN) model was trained on over 400,000 light- and heavy-chain human antibody sequences to learn the rules of human antibody formation (Amimeur et al., 2020). The resulting model outperforms common *in silico* techniques, generating diverse libraries of novel antibodies mimicking somatically hypermutated human repertoire responses. Through transfer learning, the GAN can generate molecules with improved stability, developability, lower predicted major histocompatibility complex class II binding, and specific CDR characteristics. In-house, we could independently train the GAN on a much smaller set of approximately 31,500 paired antibody sequences belonging to the VH3-VK1 germline pair and format them as single chain variable regions (ScFvs). These sequences were selected based on their high percent humanness, low incidence of chemical liabilities in the CDRs, and high medicine-likeness. The in-house developed GAN model was then used to generate 100,000 unique antibody ScFv sequences and a small yet highly diverse subset of them was produced in the laboratory as immunoglobulin G1K (IgG1K) antibodies. The initial experimental characterization showed that most of the generated antibodies showed desirable attributes for expression, purification, thermal stability, and colloidal stabilities that compare favorably with those of trastuzumab, a biotherapeutic well known for its good developability profile (unpublished results). In summary, these *in silico* approaches enable the control of pharmaceutical properties for antibodies, potentially offering a more rapid and cost-effective screening, docking, and binding affinity maturation against a given target antigen.

3.7 Humanization and optimization of lead candidates

During the conventional discovery workflow, lead optimization (LO) is carried out as soon as one or more lead candidates have been identified and revalidated for function. The Fv regions may require humanization if the lead molecule is from a non-human source, the removal of post-translational modification (PTM) sites, optimization of affinity, and ideally, improvement of developability (Figure 4). When all parts of DAbI are fully enabled, time and efforts required for LO may be significantly reduced, if not eliminated completely as stated earlier. However, for now, humanization and optimization of the functional lead candidates remain an integral part of biotherapeutic drug discovery. The following describes how computation can support every aspect of the LO process for therapeutic antibodies.

Humanization optimizes the amino acid sequence of non-human Fv regions, decreasing immunogenicity and anti-drug antibodies (ADAs) (Roguska et al., 1994; Townsend et al., 2015). Computational protein design methods can efficiently increase antibody humanness while maintaining structural stability (Choi et al., 2015). State-of-the-art software like MOE (ULC, 2021) enables CDR grafting and humanness optimization through *in silico* calculations (Abhinandan and Martin, 2007; Lazar et al., 2007; Gao et al., 2013; Seeliger, 2013; Olimpieri et al., 2015; Choi et al., 2017; Kuroda and Tsumoto, 2020). Bioinformatic studies have also revealed structural differences between the lambda (VL) and kappa (VK) isotypes, which must be considered during re-engineering (van der Kant et al., 2019). Structure-guided approaches can aid in enhancing the biophysical properties of a therapeutic mAb by transitioning from a problematic lambda framework (FWR) region to a more stable kappa FWR (Lehmann et al., 2015).

The humanized sequences progress to liability engineering campaigns. Pre-formulation assessments, forced degradation studies, and *in silico* evaluations are incorporated into the engineering design plan. Phage display or other screening technologies can be employed to screen a large panel of variants. *In silico* tools monitor and guide the redesign of candidates' individual liabilities (see Figure 4), and medicine-likeness can be estimated by comparing molecular characteristics with marketed antibodies (Ahmed et al., 2021).

Computational tools have successfully guided antibody optimization campaigns, improving solubility, viscosity, self-association, colloidal stability, and binding specificity (Yadav et al., 2011, 2012; Nichols et al., 2015; Kumar et al., 2018b; Shan et al., 2018; Zhang et al., 2018; Navarro and Ventura, 2019; Sakhnini et al., 2019; Bauer et al., 2020). *In silico*–guided LO campaigns have demonstrated single amino acid residue exchanges that can improve multiple chemistry, manufacturing, and control (CMC) properties, such as expression titer, yield, purity, and colloidal stability (Bauer et al., 2020). A case study enhanced antibody developability using a multi-stage approach, starting with *in silico* screening for mutations addressing liabilities while preserving thermodynamic stability, followed by production and characterization of stable candidates (Sakhnini et al., 2019). An alternative hybrid method combined computational and experimental alanine scans to identify CDR positions for mutagenesis, maintaining antigen binding and creating antibody libraries (Tiller et al., 2017). Structure-based computational designs have been effectively employed to improve the affinity and specificity of therapeutic antibodies by pinpointing the key residues in the paratope for site-directed single, double, or even triple mutations (Kiyoshi et al., 2014; Grossman et al., 2016; Kumar et al., 2018b; Chiba et al., 2020). Computational methods offer conformational stability predictions for humanization or LO (Dehouck et al., 2011; Baets et al., 2015; Folkman et al., 2016; Quan et al., 2016; Pandurangan et al., 2017; Cao et al., 2019; Leman et al., 2020), with some tools using ML on experimental data (Pandurangan et al., 2017; Cao et al., 2019). Furthermore, glycoengineering reduces aggregation propensity and enhances conformational stability of biotherapeutics (Hristodorov et al., 2013; Courtois et al., 2015).

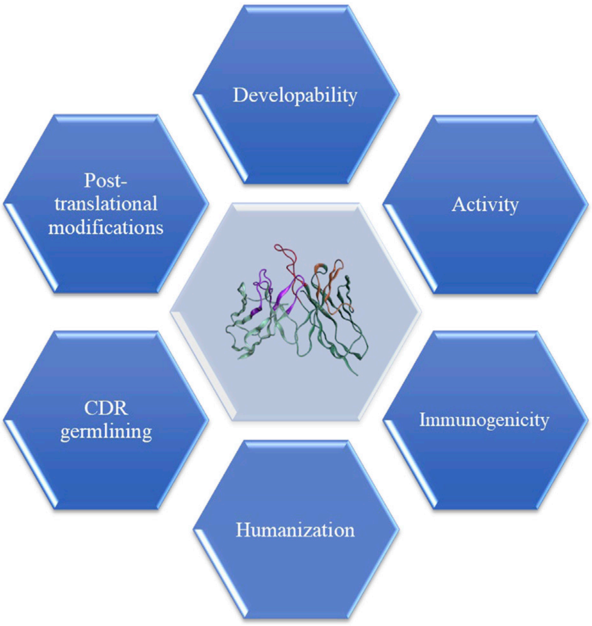


FIGURE 4
Lead humanization and optimization involve converting non-human sequences to human-like sequences while maintaining critical key attributes. *In vitro* binding affinity, which acts as surrogate for function, is the paramount criteria for accepting the mutations. Furthermore, *in silico* tools can be used to identify potential T-cell reactive epitopes, resulting in leads with lowest potential for immunogenicity and high percentage human content by germlining of the CDRs. Another aspect of optimization includes developability, which involves identifying leads with desirable biophysical properties and avoiding incidence of the post-translational modification sites such as N-linked glycosylation, unpaired cysteines, oxidation, deamidation, or aspartate isomerization, particularly in the CDRs.

Recommendations for amino acid substitutions help design a customized humanization and optimization strategy for the lead mAb candidate. The top lead optimized candidates (3–6) are selected for large-scale production and biophysical characterizations. These processes can be extended to multi-specific antibodies, with additional engineering for optimizing Fv or ScFv domains and identifying optimal multi-specific formats.

3.8 Formatting of conventional and next-generation antibodies

After optimizing Fv regions, biotherapeutic engineering proceeds with formatting Fvs into the desired antibody format, combining Fv with the chosen IgG Fc isotype. Fc engineering may be required to adjust receptor-mediated functions like antibody-dependent cell-mediated cytotoxicity (ADCC), antibody-dependent cellular phagocytosis (ADCP), complement-dependent cytotoxicity (CDC), and endosomal recycling (Mimoto et al., 2016). For next-generation biotherapeutics like bi- and multi-specific antibodies, an intermediate formatting step assesses compatibility and developability properties. Structure-based engineering supports antibody formatting, as demonstrated in a study where TGFβ1 (transforming growth factor β1) binder affinity was restored after converting from ScFv to IgG (Lord et al., 2018). Similar approaches can support formatting complex next-generation antibodies.

In the discovery process's final step, top-performing lead variants undergo pre-formulation studies before transferring to development for cell line generation and early developability assessments (Bailly et al., 2020). The research phase concludes with the final candidate selection, after which conventional and DAbI-enabled workflows for antibody discovery are identical.

3.9 *In silico* assessments in early development

The initial stages of drug substance and drug product development are resource intensive, with full development programs justified only for the final candidate. At the time of selecting the final lead candidate, experimental data are often scarce due to material limitations. The sequence of the final lead candidate becomes locked at the start of development. This decision puts product development at a disadvantage, as real-time stability data are typically unavailable but crucial for meeting regulatory requirements concerning shelf-life, Critical Quality Attributes (CQA), and product heterogeneity. There is significant demand for early, rapid, and reliable stability predictions addressed through hybrid approaches combining *in vitro* and *in silico* techniques. Computational approaches can help estimate a molecule's fit to specific platform processes and tailor subsequent development programs to the biologic candidate's inherent liabilities and characteristics (Figure 5). Conversely, platform processes continuously gather data for new molecules, improving existing and developing novel bioinformatic predictions.

One platform step is the ultrafiltration/diafiltration (UF/DF), typically employed to process the antibody into the desired formulation. Recently, *in silico* models have demonstrated that protein charge can predict common UF/DF effects, such as Gibbs–Donnan and volume-exclusion phenomena (Kannan et al., 2023). After antibody formulation, certain stability aspects become most relevant for evaluating the developability of the final lead candidates using hybridized assessments.

Conformational stability is generally not an issue for conventional mAbs but can pose a significant challenge for next-generation biologics like ScFvs and multi-specific antibodies (Bailly et al., 2020). Numerous bioinformatics tools have been developed to calculate conformational stability, mostly applicable during LO for analyzing stability changes upon point mutations (Koenig et al., 2017; Pandurangan et al., 2017; Steinbrecher et al., 2017; Cao et al., 2019; Kuroda and Tsumoto, 2020; Leman et al., 2020; Harmalkar et al., 2023). Prediction accuracy heavily relies on the quality of the underlying structure or homology model, allowing comparisons between similar sequence variants.

Recent advancements in homology modeling and MD-based free energy calculations offer potential for enhancing thermal stability prediction (Kuhlman and Bradley, 2019; Berner et al., 2021; Tomar et al., 2021; Ko et al., 2022; Licari et al., 2022). Soon, these simulation approaches will extend from antibody fragments to full-length structures (Tomar et al., 2021). MD-derived predictions will improve by considering formulation aspects influencing conformational stability (Somani et al., 2021; Blanco, 2022; Saurabh et al., 2022; Shmool et al., 2022). High-throughput (HTP) screening of biologics' thermal stabilities in platform formulations enables AI, ML, and neural networks to train computational tools to predict the thermal stabilities of diverse candidates (Gentiluomo et al., 2019a; Cao et al., 2019; Wei, 2019; Bailly et al., 2020; Harmalkar et al., 2023). The pharmaceutical industry will benefit from bioinformatic tools predicting optimal formulation composition for specific candidates or identifying the best-suited candidate for a given formulation.

Predicting colloidal stability and aggregation propensity of drug products is critical, with bioinformatics offering significant advantages in development efforts. First, real-time stability studies may take years, allowing bioinformatics to reduce development time and risk of late-stage failure. Second, stability studies require large material amounts, particularly for HCPF, increasing the cost of failures. Third, extrapolations from accelerated stability studies often inaccurately reflect molecular behavior under storage conditions. Simplified approaches using conformational stability to estimate aggregation propensity only account for non-native aggregation (Brader et al., 2015), neglecting self-association and aggregation of natively folded mAbs. Fourth, analytical techniques like HIC, dynamic light scattering (DLS), self-interaction nanoparticle spectroscopy (SINS), size exclusion chromatography (SEC), and micro-flow imaging (MFI) partially characterize colloidal instability and aggregation, often necessitating a comprehensive analytical panel (Kopp et al., 2020). Last, colloidal instability and aggregation can be triggered by various intrinsic (molecule-related) (Alam et al., 2019; Gentiluomo et al., 2019b; Lai et al., 2022) and extrinsic (process-related) factors, following complex mechanisms. Conventional methods struggle to accurately predict shelf-life, leading to resource-intensive development studies and troubleshooting efforts when the development success is at risk.

A thorough understanding of molecular behavior is essential for addressing self-association, aggregation, or particulate formation issues. Computational approaches have been developed to estimate mechanistic and kinetic characteristics for better comprehension and prediction of colloidal instability and aggregation. Mechanistic tools aid in screening and minimizing APRs during the discovery phase (Kuhn et al., 2017; Prabakaran et al., 2017, 2020; van der Kant et al., 2017; Gil-Garcia et al., 2018; Rawat et al., 2018; Bauer et al., 2020; Ebo et al., 2020; Shahfar et al., 2022), while kinetic predictors

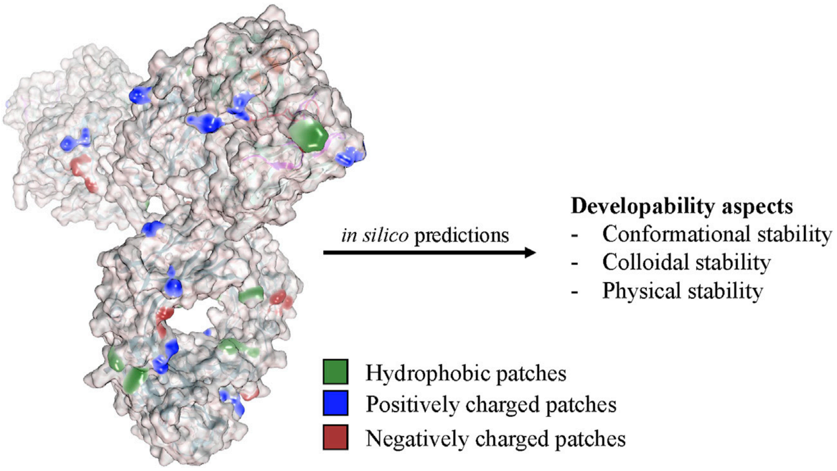


FIGURE 5
Computational approaches analyze the physicochemical properties of the antibody structure to predict various developability aspects and stability factors. These *in silico* methods evaluate factors such as aggregation propensity, conformational stability, colloidal stability, and post-translational modifications and help to select candidates with improved developability and reduced risk of immunogenicity or manufacturing challenges.

estimate aggregation rates, crucial for liquid formulation development meeting regulatory requirements for shelf life (Rawat et al., 2019; Yang et al., 2019; Santos et al., 2020). Machine Learning (ML) can train kinetic models using extensive data sets with experimental and sequence/structure information (Rawat et al., 2019; Yang et al., 2019), facilitating prediction of optimal formulation compositions (pH, salt, excipients) for minimal kinetics.

In the final development stage, creating liquid drug products with stable physical properties is vital. Manufacturing, processing, and administration of highly concentrated antibody formulations often face viscosity challenges. Viscosity is linked to surface charge and hydrophobicity of the mAb (Tomar et al., 2018; Apgar et al., 2020; Lai et al., 2021; Blanco, 2022; Han et al., 2022; Lai, 2022). Studies have shown computational ability to predict viscosity profiles at platform conditions using mAb sequence and structure (Tilegenova et al., 2019; Bauer et al., 2020; Thorsteinson et al., 2021; Han et al., 2022; Lai et al., 2022; Rosace et al., 2022). A recent deep learning approach utilized a 3D convolutional neural network to predict high-concentration viscosity of therapeutic antibodies (Rai et al., 2023). Feature attribution analysis identified key biophysical drivers of viscosity, such as the electrostatic potential surface. The predictor was successfully trained despite limited data. Early integration of viscosity predictors enables addressing viscosity issues and adjusting platform formulations and technologies before finalizing the development strategy.

4 Discussion and conclusion

In this review, we have presented numerous opportunities for computation to play a greater role in biotherapeutics discovery and development. However, the excitement around computation's enhanced role should be tempered with pragmatism. Machine learning experts often lack practical experience in biotherapeutics discovery and development and *vice versa*. Thus, a strong collaboration between bench scientists and data scientists is recommended. Computational biophysics and antibody structure–function–developability relationship experts should work with machine learning and artificial intelligence experts, as well as experimentalists, to fully enable biopharmaceutical informatics. Additionally, technical limitations exist in emerging technologies like machine learning and artificial intelligence. For instance, deep learning model performance often depends on size and diversity within training data sets (Wittmund et al., 2022), posing challenges in sparse or less diverse data settings. Moreover, the lack of insights into the latent space and interpretability of AI models in terms of the underlying physicochemical rules hinders our ability to better understand the models and extend their applicability beyond the tasks they have been trained for. For example, AI-based methods have transformed protein structure prediction, but contrary to popular belief, they have not solved the protein folding problem (Chen et al., 2023), as they do not provide insights into protein folding processes, such as initial building blocks, intermediate states, energy landscapes, and pathways.

In the specific context of protein engineering, the complexity of prediction tasks is escalated by non-additive mutational interactions or epistatic effects, which can significantly alter the impact of single or multiple mutational outcomes (Reetz, 2013; Miton and Tokuriki, 2016; Cadet et al., 2022). A further layer of complication is presented by the dynamic interplay between mutated amino acids and the subsequent establishment of intramolecular interaction networks, which can alter the protein function (Acevedo-Rocha et al., 2021). The situation is exacerbated by the limitations of tools such as AlphaFold2 or ProteinMPNN, which may struggle to predict how individual amino acid changes affect protein structure due to their heavy reliance on evolutionary perspectives and variant sequences (Eisenstein, 2021; Dauparas et al., 2022). Deep learning methods offer a way to investigate protein attributes, such as stability, solubility, aggregation, and binding affinity. However, these methods operate within the confines of the training data.

Although this does not eliminate the possibility of identifying beneficial protein variations within these parameters, it may fail to recognize or accurately predict variants exhibiting fitness values outside the learned range. This means that while beneficial variants can be identified, the optimal variant, particularly if it is an epistatic variant, might be overlooked. Against this backdrop, the use of deep learning models in conjunction with conventional neural network architectures is being explored as a solution for these challenges. By representing numerical quantities as individual neurons without non-linearity, these models can learn to perform systematic numerical computation, enabling them to handle data that lie outside the range used during training (Trask et al., 2018). The adaptability of these models across various task domains augments their potential to tackle challenges encountered in antibody therapeutics. Importantly, the ability to harness epistatic effects and predict mutational outcomes could significantly enhance the design of therapeutic antibodies. Moreover, other studies have indicated the potency of a Machine learning (ML) approach focused exclusively on sequences in accurately predicting epistatic phenomena (Cadet et al., 2018). Unlike most ML and deep learning methodologies that predominantly capture low-order non-linear interactions and predict the additive effects of mutations, this innovative strategy comprehensively encapsulates both low- and high-order non-linear interactions. By utilizing ML in tandem with digital signal processing such as Fourier transform, case studies have demonstrated a significant improvement in the resistance of proteins to unfavorable unfolding and aggregation. Crucially, this method unveils the correlation between epistatic mutational interactions and protein resilience, offering unique, predictive insights beyond those provided by conventional machine learning or deep learning approaches (Li et al., 2021). This approach has considerably enhanced precision, reduced overfitting, and surpassed conventional methods without increasing complexity (Medina-Ortiz et al., 2022). Understanding the rules underlying these interactions could contribute to a more efficient model design and a more predictive performance, thereby bolstering the success of deep learning in the realm of biopharmaceutical informatics.

In conclusion, this review article aims to broaden our strategic perspective on biopharmaceutical informatics. Initially, we emphasized the syncretic use of computation and experimentation for the drug product development of antibody-based biotherapeutics (Kumar et al., 2015; Kumar et al., 2018a). Subsequently, Khetan et al. (2022) demonstrated its feasibility by spelling out different methods and published studies already available in the public domain to support our vision. Here, we propose a more generalized vision of biopharmaceutical informatics by including DAbI and digital transformation. It is widely agreed that digital transformation is essential for modernizing the biopharmaceutical industry's work processes, leading to more judicious use of resources and reduced costs in biotherapeutics discovery and development. Recent advancements in AI and ML, along with the availability of large-scale antibody sequencing data in the public domain, have fueled excitement for DAbI. When fully embraced by the biopharmaceutical industry, DAbI will revolutionize the way biotherapeutic drugs are discovered and developed. Current drug discovery processes and workflows are dominated by experimental trials and errors, with computation playing an assistive role at the best. DAbI can support the start of projects even before the availability of antigen material for *in vitro* experimental studies. This is particularly attractive when the antigens involved are difficult to express and purify. DAbI can also accelerate discovery projects by pre-paying for developability and therefore save on resources and time required to fix these issues at the later stages. These two features may eventually lead to situations where computation plays an equal, if not greater, role alongside experimentation in supporting biotherapeutics discovery and development projects. Therefore, our vision of biopharmaceutical informatics points to an exciting future where we can better serve patients by addressing unmet medical needs through more successful, faster, and affordable discovery and development of biotherapeutics. Additionally, the discovery and development of antibody-based biotherapeutics are rapidly becoming industrialized, with several aspects becoming more uniform (e.g., discovery processes and drug formulations), while multiple options are being explored for others, such as molecular formats, routes of administration, and dosing options (Martin et al., 2023). Biopharmaceutical informatics contributes toward accelerating this industrialization and helping to improve human health.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Acknowledgments

All authors thank numerous colleagues in Biotherapeutics Discovery and Development Biologicals for excellent discussions and support.

Conflict of interest

JB was employed by the company Boehringer Ingelheim Pharma GmbH & Co. KG; NR, PrG, PaG, AN, and SK were employed by the company Boehringer Ingelheim Pharmaceuticals Inc.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, editors, and reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abhinandan, K. R., and Martin, A. C. R. (2007). Analyzing the "degree of humanness" of antibody sequences. *J. Mol. Biol.* 369, 852–862. doi:10.1016/j.jmb.2007.02.100

Acevedo-Rocha, C. G., Li, A., D'Amore, L., Hoebenreich, S., Sanchis, J., Lubrano, P., et al. (2021). Pervasive cooperative mutational effects on multiple catalytic enzyme traits emerge via long-range conformational dynamics. *Nat. Commun.* 12, 1621. doi:10.1038/s41467-021-21833-w

Adolf-Bryfogle, J., Kalyuzhniy, O., Kubitz, M., Weitzner, B. D., Hu, X., Adachi, Y., et al. (2018). RosettaAntibodyDesign (RAbD): A general framework for computational antibody design. *Plos Comput. Biol.* 14, e1006112. doi:10.1371/journal.pcbi.1006112

Agrawal, N. J., Helk, B., Kumar, S., Mody, N., Sathish, H. A., Samra, H. S., et al. (2015). Computational tool for the early screening of monoclonal antibodies for their viscosities. *Mabs* 8, 43–48. doi:10.1080/19420862.2015.1099773

Ahmed, L., Guptaa, P., Martin, K. P., Scheer, J. M., Nixon, A. E., and Kumar, S. (2021). Intrinsic physicochemical profile of marketed antibody-based biotherapeutics. *PNAS* 118, e2020577118. doi:10.1073/pnas.2020577118

Akbar, R., Robert, P. A., Pavlović, M., Jeliazkov, J. R., Snapkov, I., Slabodkin, A., et al. (2021). A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Rep.* 34, 108856. doi:10.1016/j.celrep.2021.108856

Akbar, R., Robert, P. A., Weber, C. R., Widrich, M., Frank, R., Pavlović, M., et al. (2022). *In silico* proof of principle of machine learning-based antibody design at unconstrained scale. *Mabs* 14, 2031482. doi:10.1080/19420862.2022.2031482

Alam, M. E., Barnett, G. V., Slaney, T. R., Starr, C. G., Das, T. K., and Tessier, P. M. (2019). Deamidation can compromise antibody colloidal stability and enhance aggregation in a pH-dependent manner. *Mol. Pharm.* 16, 1939–1949. doi:10.1021/acs.molpharmaceut.8b01311

Alfaleh, M. A., Alsaab, H. O., Mahmoud, A. B., Alkayyal, A. A., Jones, M. L., Mahler, S. M., et al. (2020). Phage display derived monoclonal antibodies: From bench to bedside. *Front. Immunol.* 11, 1986. doi:10.3389/fimmu.2020.01986

AlQuraishi, M. (2019). AlphaFold at CASP13. *Bioinformatics* 35, 4862–4865. doi:10.1093/bioinformatics/btz422

Ambrosetti, F., Jiménez-García, B., Roel-Touris, J., and Bonvin, A. M. J. J. (2020a). Modeling antibody-antigen complexes by information-driven docking. *Structure* 28, 119–129.e2. doi:10.1016/j.str.2019.10.011

Ambrosetti, F., Olsen, T. H., Olimpieri, P. P., Jiménez-García, B., Milanetti, E., Marcatilli, P., et al. (2020b). proABC-2: PRediction Of AntiBody Contacts v2 and its application to information-driven docking. *Bioinformatics* 36, 5107–5108. doi:10.1093/bioinformatics/btaa644

Amimeur, T., Shaver, J. M., Ketchem, R. R., Taylor, J. A., Clark, R. H., Smith, J., et al. (2020). Designing feature-controlled humanoid antibody discovery libraries using generative adversarial networks. *Biorxiv.* doi:10.1101/2020.04.12.024844

Apgar, J. R., Tam, A. S. P., Sorm, R., Moesta, S., King, A. C., Yang, H., et al. (2020). Modeling and mitigation of high-concentration antibody viscosity through structure-based computer-aided protein design. *Plos One* 15, e0232713. doi:10.1371/journal.pone.0232713

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373, 871–876. doi:10.1126/science.abj8754

Baets, G. D., Durme, J. V., van der Kant, R., Schymkowitz, J., and Rousseau, F. (2015). Solubis: Optimize your protein. *Bioinformatics* 31, 2580–2582. doi:10.1093/bioinformatics/btv162

Bailly, M., Mieczkowski, C., Juan, V., Metwally, E., Tomazela, D., Baker, J., et al. (2020). Predicting antibody developability profiles through early stage discovery screening. *Mabs* 12, 1743053. doi:10.1080/19420862.2020.1743053

Baran, D., Pszolla, M. G., Lapidoth, G. D., Norn, C., Dym, O., Unger, T., et al. (2017). Principles for computational design of binding antibodies. *Proc. Natl. Acad. Sci.* 114, 10900–10905. doi:10.1073/pnas.1707171114

Bauer, J., Mathias, S., Kube, S., Otte, K., Garidel, P., Gamer, M., et al. (2020). Rational optimization of a monoclonal antibody improves the aggregation propensity and enhances the CMC properties along the entire pharmaceutical process chain. *Mabs* 12, 1787121. doi:10.1080/19420862.2020.1787121

Benatuil, L., Perez, J. M., Belk, J., and Hsieh, C.-M. (2010). An improved yeast transformation method for the generation of very large human antibody libraries. *Protein Eng. Des. Sel.* 23, 155–159. doi:10.1093/protein/gzq002

Berner, C., Menzen, T., Winter, G., and Svilenov, H. L. (2021). Combining unfolding reversibility studies and molecular dynamics simulations to select aggregation-resistant antibodies. *Mol. Pharm.* 18, 2242–2253. doi:10.1021/acs.molpharmaceut.1c00017

Bill, R. M., Henderson, P. J. F., Iwata, S., Kunji, E. R. S., Michel, H., Neutze, R., et al. (2011). Overcoming barriers to membrane protein structure determination. *Nat. Biotechnol.* 29, 335–340. doi:10.1038/nbt.1833

Blanco, M. A. (2022). Computational models for studying physical instabilities in high concentration biotherapeutic formulations. *Mabs* 14, 2044744. doi:10.1080/19420862.2022.2044744

Brader, M. L., Estey, T., Bai, S., Alston, R. W., Lucas, K. K., Lantz, S., et al. (2015). Examination of thermal unfolding and aggregation profiles of a series of developable therapeutic monoclonal antibodies. *Mol. Pharm.* 12, 1005–1017. doi:10.1021/mp400666b

Buell, A. K., Hung, P., Salvatella, X., Welland, M. E., Dobson, C. M., and Knowles, T. P. J. (2013). Electrostatic effects in filamentous protein aggregation. *Biophys. J.* 104, 1116–1126. doi:10.1016/j.bpj.2013.01.031

Bujotzek, A., Fuchs, A., Qu, C., Benz, J., Klostermann, S., Antes, I., et al. (2015). MoFvAb: Modeling the Fv region of antibodies. *Mabs* 7, 838–852. doi:10.1080/19420862.2015.1068492

Cadet, F., Fontaine, N., Li, G., Sanchis, J., Chong, M. N. F., Pandjaitan, R., et al. (2018). A machine learning approach for reliable prediction of amino acid interactions and its application in the directed evolution of enantioselective enzymes. *Sci. Rep.* 8, 16757. doi:10.1038/s41598-018-35033-y

Cadet, X. F., Gelly, J. C., van Noord, A., Cadet, F., and Acevedo-Rocha, C. G. (2022). "Learning strategies in protein directed EvolutionDirected evolution (DE)," in *Methods in molecular biology* (New York, NY: Springer US), 225–275. doi:10.1007/978-1-0716-2152-3_15

Cao, H., Wang, J., He, L., Qi, Y., and Zhang, J. Z. (2019). DeepDDG: Predicting the stability change of protein point mutations using neural networks. *J. Chem. Inf. Model* 59, 1508–1514. doi:10.1021/acs.jcim.8b00697

Chen, G., and Sidhu, S. S. (2014). Design and generation of synthetic antibody libraries for phage display. *Methods Mol. Biol.* 1131, 113–131. doi:10.1007/978-1-62703-992-5_8

Chen, S.-J., Hassan, M., Jernigan, R. L., Jia, K., Kihara, D., Kloczkowski, A., et al. (2023). Opinion: Protein folds vs. protein folding: Differing questions, different challenges. *Proc. Natl. Acad. Sci.* 120, e2214423119. doi:10.1073/pnas.2214423119

Chiba, S., Tanabe, A., Nakakido, M., Okuno, Y., Tsumoto, K., and Ohta, M. (2020). Structure-based design and discovery of novel anti-tissue factor antibodies with cooperative double-point mutations, using interaction analysis. *Sci. Rep-uk* 10, 17590. doi:10.1038/s41598-020-74545-4

Chiu, M. L., Goulet, D. R., Teplyakov, A., and Gilliland, G. L. (2019). Antibody structure and function: The basis for engineering therapeutics. *Antibodies* 8, 55. doi:10.3390/antib8040055

Choi, Y., Hua, C., Sentman, C. L., Ackerman, M. E., and Bailey-Kellogg, C. (2015). Antibody humanization by structure-based computational protein design. *Mabs* 7, 1045–1057. doi:10.1080/19420862.2015.1076600

Choi, Y., Verma, D., Griswold, K. E., and Bailey-Kellogg, C. (2017). "EpiSweep: Computationally driven reengineering of therapeutic proteins to reduce immunogenicity while maintaining function computational protein design," in *Methods in molecular biology.* Editor I. Samish doi:10.1007/978-1-4939-6637-0_20

Chowdhury, R., Allan, M. F., and Maranas, C. D. (2018). OptMAVEn-2.0: De novo design of variable antibody regions against targeted antigen epitopes. *Antibodies* 7, 23. doi:10.3390/antib7030023

Comeau, S. R., Thorsteinson, N., and Kumar, S. (2023). "Structural considerations in affinity maturation of antibody-based biotherapeutic candidates," in *Methods in molecular biology* (New York, NY: Springer US), 309–321. doi:10.1007/978-1-0716-2609-2_17

Conti, S., Lau, E. Y., and Ovchinnikov, V. (2022). On the rapid calculation of binding affinities for antigen and antibody design and affinity maturation simulations. *Antibodies* 11, 51. doi:10.3390/antib11030051

Courtois, F., Agrawal, N. J., Lauer, T. M., and Trout, B. L. (2015). Rational design of therapeutic mAbs against aggregation through protein engineering and incorporation of glycosylation motifs applied to bevacizumab. *Mabs* 8, 99–112. doi:10.1080/19420862.2015.1112477

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., et al. (2022). Robust deep learning–based protein sequence design using ProteinMPNN. *Science* 378, 49–56. doi:10.1126/science.add2187

Davila, A., Xu, Z., Li, S., Rozewicki, J., Wilamowski, J., Kotelnikov, S., et al. (2022). AbAdapt: An adaptive approach to predicting antibody–antigen complex structures from sequence. *Bioinform Adv.* 2, vbac015. doi:10.1093/bioadv/vbac015

Deac, A., VeliČković, P., and Sormanni, P. (2019). Attentive cross-modal paratope prediction. *J. Comput. Biol.* 26, 536–545. doi:10.1089/cmb.2018.0175

Dehouck, Y., Kwasigroch, J. M., Gilis, D., and Rooman, M. (2011). PoPMuSiC 2.1: A web server for the estimation of protein stability changes upon mutation and sequence optimality. *Bmc Bioinforma.* 12, 151. doi:10.1186/1471-2105-12-151

DiMasi, J. A., and Grabowski, H. G. (2007). The cost of biopharmaceutical R&D: Is biotech different? *Manag. Decis. Econ.* 28, 469–479. doi:10.1002/mde.1360

DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *J. Health Econ.* 47, 20–33. doi:10.1016/j.jhealeco.2016.01.012

Ebo, J. S., Guthertz, N., Radford, S. E., and Brockwell, D. J. (2020). Using protein engineering to understand and modulate aggregation. *Curr. Opin. Struc Biol.* 60, 157–166. doi:10.1016/j.sbi.2020.01.005

Eguchi, R. R., Choe, C. A., and Huang, P.-S. (2022). Ig-VAE: Generative modeling of protein structure by direct 3D coordinate generation. *Plos Comput. Biol.* 18, e1010271. doi:10.1371/journal.pcbi.1010271

Eguida, M., and Rognan, D. (2020). A computer vision approach to align and compare protein cavities: Application to fragment-based drug design. *J. Med. Chem.* 63, 7127–7142. doi:10.1021/acs.jmedchem.0c00422

Eisenstein, M. (2021). Artificial intelligence powers protein-folding predictions. *Nature* 599, 706–708. doi:10.1038/d41586-021-03499-y

Farid, S. S., Baron, M., Stamatis, C., Nie, W., and Coffman, J. (2020). Benchmarking biopharmaceutical process development and manufacturing cost contributions to R&D. *Mabs* 12, 1754999. doi:10.1080/19420862.2020.1754999

Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J., and Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* 22, 1302–1306. doi:10.1038/nbt1012

Fernández-Quintero, M. L., Kokot, J., Waibl, F., Fischer, A.-L. M., Quoika, P. K., Deane, C. M., et al. (2023). Challenges in antibody structure prediction. *mAbs* 15, 2175319. doi:10.1080/19420862.2023.2175319

Fogel, D. B. (2018). Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. *Contemp. Clin. Trials Commun.* 11, 156–164. doi:10.1016/j.conctc.2018.08.001

Folkman, L., Stantic, B., Sattar, A., and Zhou, Y. (2016). EASE-MM: Sequence-Based prediction of mutation-induced stability changes with feature-based multiple models. *J. Mol. Biol.* 428, 1394–1405. doi:10.1016/j.jmb.2016.01.012

Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M. M., et al. (2020). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* 17, 184–192. doi:10.1038/s41592-019-0666-6

Gao, S. H., Huang, K., Tu, H., and Adler, A. S. (2013). Monoclonal antibody humanness score and its applications. *BMC Biotechnol.* 55, 55. doi:10.1186/1472-6750-13-55

Gao, W., Mahajan, S. P., Sulam, J., and Gray, J. J. (2020). Deep learning in protein structural modeling and design. *Patterns* 1, 100142. doi:10.1016/j.patter.2020.100142

Garbuzynskiy, S. O., Lobanov, M. Y., and Galzitskaya, O. V. (2010). FoldAmyloid: A method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics* 26, 326–332. doi:10.1093/bioinformatics/btp691

Garidel, P., Kuhn, A. B., Schäfer, L. V., Karow-Zwick, A. R., and Blech, M. (2017). High-concentration protein formulations: How high is high? *Eur. J. Pharm. Biopharm.* 119, 353–360. doi:10.1016/j.ejpb.2017.06.029

Garripelli, V. K., Wu, Z., and Gupta, S. (2020). Developability assessment for monoclonal antibody drug candidates: A case study. *Pharm. Dev. Technol.* 1, 11–20. doi:10.1080/10837450.2020.1829641

Gentiluomo, L., Roessner, D., Augustijn, D., Svilenov, H., Kulakova, A., Mahapatra, S., et al. (2019a). Application of interpretable artificial neural networks to early monoclonal antibodies development. *Eur. J. Pharm. Biopharm.* 141, 81–89. doi:10.1016/j.ejpb.2019.05.017

Gentiluomo, L., Roessner, D., Streicher, W., Mahapatra, S., Harris, P., and Frieß, W. (2019b). Characterization of native reversible self-association of a monoclonal antibody mediated by Fab-Fab interaction. *J. Pharm. Sci.* 109, 443–451. doi:10.1016/j.xphs.2019.09.021

Gil-Garcia, M., Bañó-Polo, M., Varejão, N., Jamroz, M., Kuriata, A., Díaz-Caballero, M., et al. (2018). Combining structural aggregation propensity and stability predictions to redesign protein solubility. *Mol. Pharm.* 15, 3846–3859. doi:10.1021/acs.molpharmaceut.8b00341

González-Fernández, Á., Silva, F. J. B., López-Hoyos, M., Cobaleda, C., Montoliu, L., Val, M. D., et al. (2020). Non-animal-derived monoclonal antibodies are not ready to substitute current hybridoma technology. *Nat. Methods* 17, 1069–1070. doi:10.1038/s41592-020-00977-5

Gorgulla, C., Boeszoermenyi, A., Wang, Z.-F., Fischer, P. D., Coote, P. W., Das, K. M. P., et al. (2020). An open-source drug discovery platform enables ultra-large virtual screens. *Nature* 580, 663–668. doi:10.1038/s41586-020-2117-z

Goyon, A., Excoffier, M., Janin-Bussat, M.-C., Bobaly, B., Fekete, S., Guillarme, D., et al. (2017). Determination of isoelectric points and relative charge variants of 23 therapeutic monoclonal antibodies. *J. Chromatogr. B* 1065, 119–128. doi:10.1016/j.jchromb.2017.09.033

Gray, A., Bradbury, A. R. M., Knappik, A., Plückthun, A., Borrebaeck, C. A. K., and Dübel, S. (2020). Animal-free alternatives and the antibody iceberg. *Nat. Biotechnol.* 38, 1234–1239. doi:10.1038/s41587-020-0687-9

Gray, A. C., Bradbury, A. R. M., Knappik, A., Plückthun, A., Borrebaeck, C. A. K., and Dübel, S. (2020). Animal-derived-antibody generation faces strict reform in accordance with European Union policy on animal use. *Nat. Methods* 17, 755–756. doi:10.1038/s41592-020-0906-9

Grossman, I., Ilani, T., Fleishman, S. J., and Fass, D. (2016). Overcoming a species-specificity barrier in development of an inhibitory antibody targeting a modulator of tumor stroma. *Protein Eng. Des. Sel.* 29, 135–147. doi:10.1093/protein/gzv067

Han, X., Shih, J., Lin, Y., Chai, Q., and Cramer, S. M. (2022). Development of QSAR models for *in silico* screening of antibody solubility. *Mabs* 14, 2062807. doi:10.1080/19420862.2022.2062807

Harmalkar, A., Rao, R., Xie, Y. R., Honer, J., Deisting, W., Anlahr, J., et al. (2023). Toward generalizable prediction of antibody thermostability using machine learning on sequence and structure features. *Mabs* 15, 2163584. doi:10.1080/19420862.2022.2163584

Hebditch, M., and Warwicker, J. (2019). Charge and hydrophobicity are key features in sequence-trained machine learning models for predicting the biophysical properties of clinical-stage antibodies. *Peerj* 7, e8199. doi:10.7717/peerj.8199

Hristodorov, D., Fischer, R., Joerissen, H., Müller-Tiemann, B., Apeler, H., and Linden, L. (2013). Generation and comparative characterization of glycosylated and aglycosylated human IgG1 antibodies. *Mol. Biotechnol.* 53, 326–335. doi:10.1007/s12033-012-9531-x

Imrie, F., Bradley, A. R., van der Schaar, M., and Deane, C. M. (2018). Protein family-specific models using deep neural networks and transfer learning improve virtual screening and highlight the need for more data. *J. Chem. Inf. Model* 58, 2319–2330. doi:10.1021/acs.jcim.8b00350

Irudayanathan, F. J., Zarzar, J., Lin, J., and Izadi, S. (2022). Deciphering deamidation and isomerization in therapeutic proteins: Effect of neighboring residue. *Mabs* 14, 2143006. doi:10.1080/19420862.2022.2143006

Jain, T., Boland, T., Lilov, A., Burnina, I., Brown, M., Xu, Y., et al. (2017a). Prediction of delayed retention of antibodies in hydrophobic interaction chromatography from sequence using machine learning. *Bioinformatics* 33, 3758–3766. doi:10.1093/bioinformatics/btx519

Jain, T., Sun, T., Durand, S., Hall, A., Houston, N. R., Nett, J. H., et al. (2017b). Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl. Acad. Sci.* 114, 944–949. doi:10.1073/pnas.1616408114

Jain, T., Boland, T., and Vásquez, M. (2023). Identifying developability risks for clinical progression of antibodies using high-throughput *in vitro* and *in silico* approaches. *Mabs* 15, 2200540. doi:10.1080/19420862.2023.2200540

Jarasch, A., Koll, H., Regula, J. T., Bader, M., Papadimitriou, A., and Kettenberger, H. (2015). Developability assessment during the selection of novel therapeutic antibodies. *J. Pharm. Sci.* 104, 1885–1898. doi:10.1002/jps.24430

Jeliazkov, J. R., Frick, R., Zhou, J., and Gray, J. J. (2021). Robustification of RosettaAntibody and Rosetta SnugDock. *Plos One* 16, e0234282. doi:10.1371/journal.pone.0234282

Jetha, A., Thorsteinson, N., Jmeian, Y., Jeganathan, A., Giblin, P., and Fransson, J. (2018). Homology modeling and structure-based design improve hydrophobic interaction chromatography behavior of integrin binding antibodies. *Mabs* 10, 890–900. doi:10.1080/19420862.2018.1475871

Jones, D. T., and Thornton, J. M. (2022). The impact of AlphaFold2 one year on. *Nat. Methods* 19, 15–20. doi:10.1038/s41592-021-01365-3

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2

Kannan, A., Chinn, M., Izadi, S., Maier, A., Dvornicky, J., Fedesco, M., et al. (2023). Predicting formulation conditions during ultrafiltration and dilution to drug substance using a donnan model with homology-model based protein charge. *J. Pharm. Sci.* 112, 820–829. doi:10.1016/j.xphs.2022.10.028

Kaplon, H., Chenoweth, A., Crescioli, S., and Reichert, J. M. (2022). Antibodies to watch in 2022. *Mabs* 14, 2014296. doi:10.1080/19420862.2021.2014296

Karlberg, M., de Souza, J. V., Fan, L., Kizhedath, A., Bronowska, A. K., and Glassey, J. (2020). QSAR implementation for HIC retention time prediction of mAbs using fab structure: A comparison between structural representations. *Int. J. Mol. Sci.* 21, 8037. doi:10.3390/ijms21218037

Kemmish, H., Fasnacht, M., and Yan, L. (2017). Fully automated antibody structure prediction using BIOVIA tools: Validation study. *Plos One* 12, e0177923. doi:10.1371/journal.pone.0177923

Khan, A., Cowen-Rivers, A. I., Grosnit, A., Deik, D.-G.-X., Robert, P. A., Greiff, V., et al. (2023). Toward real-world automated antibody design with combinatorial Bayesian optimization. *Cell Rep. Methods* 3, 100374. doi:10.1016/j.crmeth.2022.100374

Khetan, R., Curtis, R., Deane, C. M., Hadsund, J. T., Kar, U., Krawczyk, K., et al. (2022). Current advances in biopharmaceutical informatics: Guidelines, impact and challenges in the computational developability assessment of antibody therapeutics. *Mabs* 14, 2020082. doi:10.1080/19420862.2021.2020082

Kiyoshi, M., Caaveiro, J. M. M., Miura, E., Nagatoishi, S., Nakakido, M., Soga, S., et al. (2014). Affinity improvement of a therapeutic antibody by structure-based computational design: Generation of electrostatic interactions in the transition state stabilizes the antibody-antigen complex. *Plos One* 9, e87099. doi:10.1371/journal.pone.0087099

Klausen, M. S., Anderson, M. V., Jespersen, M. C., Nielsen, M., and Marcatili, P. (2015). LYRA, a webserver for lymphocyte receptor structural modeling. *Nucleic Acids Res.* 43, W349–W355. doi:10.1093/nar/gkv535

Ko, S. K., Berner, C., Kulakova, A., Schneider, M., Antes, I., Winter, G., et al. (2022). Investigation of the pH-dependent aggregation mechanisms of GCSF using low resolution protein characterization techniques and advanced molecular dynamics simulations. *Comput. Struct. Biotechnol. J.* 20, 1439–1455. doi:10.1016/j.csbj.2022.03.012

Koehler, G., and Milstein, C. (1975). Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* 256, 495–497. doi:10.1038/256495a0

Koenig, P., Lee, C. V., Walters, B. T., Janakiraman, V., Stinson, J., Patapoff, T. W., et al. (2017). Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding. *Proc. Natl. Acad. Sci.* 114, E486–E495. doi:10.1073/pnas.1613231114

Kopp, M. R. G., Pérez, A.-M. W., Zucca, M. V., Palmiero, U. C., Friedrichsen, B., Lorenzen, N., et al. (2020). An accelerated surface-mediated stress assay of antibody instability for developability studies. *Mabs* 12, 1815995. doi:10.1080/19420862.2020.1815995

Krawczyk, K., Baker, T., Shi, J., and Deane, C. M. (2013). Antibody i-Patch prediction of the antibody binding site improves rigid local antibody–antigen docking. *Protein Eng. Des. Sel.* 26, 621–629. doi:10.1093/protein/gzt043

Krawczyk, K., Liu, X., Baker, T., Shi, J., and Deane, C. M. (2014). Improving B-cell epitope prediction and its application to global antibody-antigen docking. *Bioinformatics* 30, 2288–2294. doi:10.1093/bioinformatics/btu190

Kuhlman, B., and Bradley, P. (2019). Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Bio* 20, 681–697. doi:10.1038/s41580-019-0163-x

Kuhn, A. B., Kube, S., Karow-Zwick, A. R., Seeliger, D., Garidel, P., Blech, M., et al. (2017). Improved solution-state properties of monoclonal antibodies by targeted mutations. *J. Phys. Chem. B* 121, 10818–10827. doi:10.1021/acs.jpcb.7b09126

Kumar, S., and Singh, S. K. (2015). in *Developability of biotherapeutics: Computational approaches*. Editors S. Kumar and S. K. Singh

Kumar, S., Singh, S. K., and Gromiha, M. M. (2009). "Temperature-dependent molecular adaptations, microbial proteins," in *Encyclopedia of industrial biotechnology*, 1–22. doi:10.1002/9780470054581.eib516

Kumar, S., Robins, R. H., Buck, P. M., Hickling, T. P., Thangakani, A. M., Li, L., et al. (2015). *Biopharmaceutical informatics: Applications of computation in biologic drug development.* New York, NY: CRC Press, 3–34.

Kumar, S., Plotnikov, N. V., Rouse, J. C., and Singh, S. K. (2018a). Biopharmaceutical informatics: Supporting biologic drug development via molecular modelling and informatics. *J. Pharm. Pharmacol.* 70, 595–608. doi:10.1111/jphp.12700

Kumar, S., Roffi, K., Tomar, D. S., Cirelli, D., Luksha, N., Meyer, D., et al. (2018b). Rational optimization of a monoclonal antibody for simultaneous improvements in its solution properties and biological activity. *Protein Eng. Des. Sel.* 31, 313–325. doi:10.1093/protein/gzy020

Kuroda, D., and Tsumoto, K. (2020). Engineering stability, viscosity, and immunogenicity of antibodies by computational design. *J. Pharm. Sci.* 109, 1631–1651. doi:10.1016/j.xphs.2020.01.011

Lai, P.-K., Fernando, A., Cloutier, T. K., Gokarn, Y., Zhang, J., Schwenger, W., et al. (2021). Machine learning applied to determine the molecular descriptors responsible for the viscosity behavior of concentrated therapeutic antibodies. *Mol. Pharm.* 18, 1167–1175. doi:10.1021/acs.molpharmaceut.0c01073

Lai, P.-K., Gallegos, A., Mody, N., Sathish, H. A., and Trout, B. L. (2022). Machine learning prediction of antibody aggregation and viscosity for high concentration formulation development of protein therapeutics. *Mabs* 14, 2026208. doi:10.1080/19420862.2022.2026208

Lai, P.-K. (2022). DeepSCM: An efficient convolutional neural network surrogate model for the screening of therapeutic antibody viscosity. *Comput. Struct. Biotechnol. J.* 20, 2143–2152. doi:10.1016/j.csbj.2022.04.035

Lapidoth, G., Parker, J., Prilusky, J., and Fleishman, S. J. (2018). AbPredict 2: A server for accurate and unstrained structure prediction of antibody variable domains. *Bioinformatics* 35, 1591–1593. doi:10.1093/bioinformatics/bty822

Lazar, G. A., Desjarlais, J. R., Jacinto, J., Karki, S., and Hammond, P. W. (2007). A molecular immunology approach to antibody humanization and functional optimization. *Mol. Immunol.* 44, 1986–1998. doi:10.1016/j.molimm.2006.09.029

Lecerf, M., Kanyavuz, A., Lacroix-Desmazes, S., and Dimitrov, J. D. (2019). Sequence features of variable region determining physicochemical properties and polyreactivity of therapeutic antibodies. *Mol. Immunol.* 112, 338–346. doi:10.1016/j.molimm.2019.06.012

Ledsgaard, L., Ljungars, A., Rimbault, C., Sørensen, C. V., Tulika, T., Wade, J., et al. (2022). Advances in antibody phage display technology. *Drug Discov. Today* 27, 2151–2169. doi:10.1016/j.drudis.2022.05.002

Leem, J., Dunbar, J., Georges, G., Shi, J., and Deane, C. M. (2016). ABodyBuilder: Automated antibody structure prediction with data–driven accuracy estimation. *Mabs* 8, 1259–1268. doi:10.1080/19420862.2016.1205773

Lehmann, A., Wixted, J. H. F., Shapovalov, M. V., Roder, H., Dunbrack, R. L., and Robinson, M. K. (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. *Mabs* 7, 1058–1071. doi:10.1080/19420862.2015.1088618

Leman, J. K., Weitzner, B. D., Lewis, S. M., Adolf-Bryfogle, J., Alam, N., Alford, R. F., et al. (2020). Macromolecular modeling and design in Rosetta: Recent methods and frameworks. *Nat. Methods* 17, 665–680. doi:10.1038/s41592-020-0848-2

Li, T., Pantazes, R. J., and Maranas, C. D. (2014). OptMAVEn – a New Framework for the de novo Design of Antibody Variable Region Models Targeting Specific Antigen Epitopes. *Plos One* 9, e105954. doi:10.1371/journal.pone.0105954

Li, G., Qin, Y., Fontaine, N. T., Chong, M. N. F., Maria-Solano, M. A., Feixas, F., et al. (2021). Machine learning enables selection of epistatic enzyme mutants for stability against unfolding and detrimental aggregation. *ChemBioChem* 22, 904–914. doi:10.1002/cbic.202000612

Liberis, E., Velickovic, P., Sormanni, P., Vendruscolo, M., and Liò, P. (2018). Parapred: Antibody paratope prediction using convolutional and recurrent neural networks. *Bioinform Oxf Engl.* 34, 2944–2950. doi:10.1093/bioinformatics/bty305

Licari, G., Martin, K. P., Crames, M., Mozdzierz, J., Marlow, M. S., Karow-Zwick, A. R., et al. (2022). Embedding dynamics in intrinsic physicochemical profiles of market-stage antibody-based biotherapeutics. *Mol. Pharm.* 20, 1096–1111. doi:10.1021/acs.molpharmaceut.2c00838

Lipinski, C. A. (2000). Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol.* 44, 235–249. doi:10.1016/s1056-8719(00)00107-6

Liu, X., Taylor, R. D., Griffin, L., Coker, S.-F., Adams, R., Ceska, T., et al. (2017). Computational design of an epitope-specific Keap1 binding antibody using hotspot residues grafting and CDR loop swapping. *Sci. Rep-uk* 7, 41306. doi:10.1038/srep41306

Liu, G., Zeng, H., Mueller, J., Carter, B., Wang, Z., Schilz, J., et al. (2019). Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics* 36, 2126–2133. doi:10.1093/bioinformatics/btz895

Lord, D. M., Bird, J. J., Honey, D. M., Best, A., Park, A., Wei, R. R., et al. (2018). Structure-based engineering to restore high affinity binding of an isoform-selective anti-TGFβ1 antibody. *Mabs* 10, 444–452. doi:10.1080/19420862.2018.1426421

Lu, R. M., Hwang, Y. C., Liu, I. J., Lee, C. C., Tsai, H. Z., Li, H. J., et al. (2020). Development of therapeutic antibodies for the treatment of diseases. *J. Biomed. Sci.* 27, 1. doi:10.1186/s12929-019-0592-z

Lu, S., Li, Y., Wang, F., Nan, X., and Zhang, S. (2021). Leveraging sequential and spatial neighbors information by using CNNs linked with GCNs for paratope prediction. *Ieee Acm Trans. Comput. Biol. Bioinform* 19, 68–74. doi:10.1109/tcbb.2021.3083001

Ma, B., Kumar, S., Tsai, C. J., Hu, Z., and Nussinov, R. (2000). Transition-state ensemble in enzyme catalysis: Possibility, reality, or necessity? *J. Theor. Biol.* 203, 383–397. doi:10.1006/jtbi.2000.1097

Magnan, C. N., Randall, A., and Baldi, P. (2009). SOLpro: Accurate sequence-based prediction of protein solubility. *Bioinformatics* 25, 2200–2207. doi:10.1093/bioinformatics/btp386

Maia, E. H. B., Assis, L. C., de Oliveira, T. A., da Silva, A. M., and Taranto, A. G. (2020). Structure-based virtual screening: From classical to artificial intelligence. *Front. Chem.* 8, 343. doi:10.3389/fchem.2020.00343

Martin, K. P., Grimaldi, C., Grempler, R., Hansel, S., and Kumar, S. (2023). Trends in industrialization of biotherapeutics: A survey of product characteristics of 89 antibody-based biotherapeutics. *Mabs* 15, 2191301. doi:10.1080/19420862.2023.2191301

Medina-Ortiz, D., Contreras, S., Amado-Hinojosa, J., Torres-Almonacid, J., Asenjo, J. A., Navarrete, M., et al. (2022). Generalized property-based encoders and digital signal processing facilitate predictive tasks in protein engineering. *Front. Mol. Biosci.* 9, 898627. doi:10.3389/fmolb.2022.898627

Mehta, D., Jackson, R., Paul, G., Shi, J., and Sabbagh, M. (2017). Why do trials for alzheimer's disease drugs keep failing? A discontinued drug perspective for 2010-2015. *Expert Opin. Inv Drug* 26, 735–739. doi:10.1080/13543784.2017.1323868

Mieczkowski, C., Zhang, X., Lee, D., Nguyen, K., Lv, W., Wang, Y., et al. (2023). Blueprint for antibody biologics developability. *Mabs* 15, 2185924. doi:10.1080/19420862.2023.2185924

Mimoto, F., Kuramochi, T., Katada, H., Igawa, T., and Hattori, K. (2016). Fc engineering to improve the function of therapeutic antibodies. *Curr. Pharm. Biotechnol.* 17, 1298–1314. doi:10.2174/1389201017666160824161854

Miton, C. M., and Tokuriki, N. (2016). How mutational epistasis impairs predictability in protein evolution and design. *Protein Sci.* 25, 1260–1272. doi:10.1002/pro.2876

Münch, C., and Bertolotti, A. (2010). Exposure of hydrophobic surfaces initiates aggregation of diverse ALS-causing superoxide dismutase-1 mutants. *J. Mol. Biol.* 399, 512–525. doi:10.1016/j.jmb.2010.04.019

Myung, Y., Pires, D. E. V., and Ascher, D. B. (2021). CSM-AB: Graph-based antibody–antigen binding affinity prediction and docking scoring function. *Bioinformatics* 38, 1141–1143. doi:10.1093/bioinformatics/btab762

Nagano, K., and Tsutsumi, Y. (2021). Phage display technology as a powerful platform for antibody drug discovery. *Viruses* 13, 178. doi:10.3390/v13020178

Navarro, S., and Ventura, S. (2019). Computational re-design of protein structures to improve solubility. *Expert Opin. Drug Dis.* 14, 1077–1088. doi:10.1080/17460441.2019.1637413

Nelson, B., Adams, J., Kuglstatter, A., Li, Z., Harris, S. F., Liu, Y., et al. (2018). Structure-guided combinatorial engineering facilitates affinity and specificity optimization of anti-CD81 antibodies. *J. Mol. Biol.* 430, 2139–2152. doi:10.1016/j.jmb.2018.05.018

Nichols, P., Li, L., Kumar, S., Buck, P. M., Singh, S. K., Goswami, S., et al. (2015). Rational design of viscosity reducing mutants of a monoclonal antibody: Hydrophobic versus electrostatic inter-molecular interactions. *Mabs* 7, 212–230. doi:10.4161/19420862.2014.985504

Nimrod, G., Fischman, S., Austin, M., Herman, A., Keyes, F., Leiderman, O., et al. (2018). Computational design of epitope-specific functional antibodies. *Cell Rep.* 25, 2121–2131.e5. doi:10.1016/j.celrep.2018.10.081

Norman, R. A., Ambrosetti, F., Bonvin, A. M. J. J., Colwell, L. J., Kelm, S., Kumar, S., et al. (2019). Computational approaches to therapeutic antibody design: Established methods and emerging trends. *Brief. Bioinform* 21, 1549–1567. doi:10.1093/bib/bbz095

Olimpieri, P. P., Chailyan, A., Tramontano, A., and Marcatili, P. (2013). Prediction of site-specific interactions in antibody-antigen complexes: The proABC method and server. *Bioinformatics* 29, 2285–2291. doi:10.1093/bioinformatics/btt369

Olimpieri, P. P., Marcatili, P., and Tramontano, A. (2015). Tabhu: Tools for antibody humanization. *Bioinformatics* 31, 434–435. doi:10.1093/bioinformatics/btu667

Onuchic, J. N., Luthey-Schulten, Z., and Wolynes, P. G. (1997). Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.* 48, 545–600. doi:10.1146/annurev.physchem.48.1.545

Pan, X., and Kortemme, T. (2021). Recent advances in de novo protein design: Principles, methods, and applications. *J. Biol. Chem.* 296, 100558. doi:10.1016/j.jbc.2021.100558

Pandurangan, A. P., Ochoa-Montaño, B., Ascher, D. B., and Blundell, T. L. (2017). SDM: A server for predicting effects of mutations on protein stability. *Nucleic Acids Res.* 45, W229–W235. doi:10.1093/nar/gkx439

Pantazes, R. J., and Maranas, C. D. (2010). OptCDR: A general computational method for the design of antibody complementarity determining regions for targeted epitope binding. *Protein Eng. Des. Sel.* 23, 849–858. doi:10.1093/protein/gzq061

Pereira, J., Simpkin, A. J., Hartmann, M. D., Rigden, D. J., Keegan, R. M., and Lupas, A. N. (2021). High-accuracy protein structure prediction in CASP14. *Proteins Struct. Funct. Bioinform* 89, 1687–1699. doi:10.1002/prot.26171

Pittala, S., and Bailey-Kellogg, C. (2020). Learning context-aware structural representations to predict antigen and antibody binding interfaces. *Bioinform Oxf Engl.* 36, 3996–4003. doi:10.1093/bioinformatics/btaa263

Prabakaran, R., Goel, D., Kumar, S., and Gromiha, M. M. (2017). Aggregation prone regions in human proteome: Insights from large-scale data analyses. *Proteins Struct. Funct. Bioinform* 85, 1099–1118. doi:10.1002/prot.25276

Prabakaran, R., Rawat, P., Kumar, S., and Gromiha, M. M. (2020). ANuPP: A versatile tool to predict aggregation nucleating regions in peptides and proteins. *J. Mol. Biol.* 433, 166707. doi:10.1016/j.jmb.2020.11.006

Qing, R., Hao, S., Smorodina, E., Jin, D., Zalevsky, A., and Zhang, S. (2022). Protein design: From the aspect of water solubility and stability. *Chem. Rev.* 122, 14085–14179. doi:10.1021/acs.chemrev.1c00757

Quan, L., Lv, Q., and Zhang, Y. (2016). STRUM: Structure-based prediction of protein stability changes upon single-point mutation. *Bioinformatics* 32, 2936–2946. doi:10.1093/bioinformatics/btw361

Rai, B. K., Apgar, J. R., and Bennett, E. M. (2023). Low-data interpretable deep learning prediction of antibody viscosity using a biophysically meaningful representation. *Sci. Rep-uk* 13, 2917. doi:10.1038/s41598-023-28841-4

Rangel, M. A., Bedwell, A., Costanzi, E., Taylor, R. J., Russo, R., Bernardes, G. J. L., et al. (2022). Fragment-based computational design of antibodies targeting structured epitopes. *Sci. Adv.* 8, eabp9540. doi:10.1126/sciadv.abp9540

Rawat, P., Kumar, S., and Gromiha, M. M. (2018). An *in-silico* method for identifying aggregation rate enhancer and mitigator mutations in proteins. *Int. J. Biol. Macromol.* 118, 1157–1167. doi:10.1016/j.ijbiomac.2018.06.102

Rawat, P., Prabakaran, R., Kumar, S., and Gromiha, M. M. (2019). AggreRATE-pred: A mathematical model for the prediction of change in aggregation rate upon point mutation. *Bioinformatics* 31, 1439–1444. doi:10.1093/bioinformatics/btz764

Raybould, M. I. J., Marks, C., Krawczyk, K., Taddese, B., Nowak, J., Lewis, A. P., et al. (2019). Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl. Acad. Sci.* 116, 4025–4030. doi:10.1073/pnas.1810576116

Reetz, M. T. (2013). The importance of additive and non-additive mutational effects in protein engineering. *Angew. Chem. Int. Ed.* 52, 2658–2666. doi:10.1002/anie.201207842

Reichert, J. M., Rosensweig, C. J., Faden, L. B., and Dewitz, M. C. (2009). Monoclonal antibody successes in the clinic. *Nat. Biotechnol.* 23, 1073–1078. doi:10.1038/nbt0905-1073

Rice, P., Longden, I., and Bleasby, A. (2000). Emboss: The European molecular biology open software suite. *Trends Genet.* 16, 276–277. doi:10.1016/s0168-9525(00)02024-2

Ripoll, D. R., Chaudhury, S., and Wallqvist, A. (2021). Using the antibody-antigen binding interface to train image-based deep neural networks for antibody-epitope classification. *Plos Comput. Biol.* 17, e1008864. doi:10.1371/journal.pcbi.1008864

Roguska, M. A., Pedersen, J. T., Keddy, C. A., Henry, A. H., Searle, S. J., Lambert, J. M., et al. (1994). Humanization of murine monoclonal antibodies through variable domain resurfacing. *Proc. Natl. Acad. Sci. U. S. A.* 91, 969–973. doi:10.1073/pnas.91.3.969

Rosace, A., Bennett, A., Oeller, M., Mortensen, M. M., Sakhnini, L., Lorenzen, N., et al. (2022). Automated optimisation of solubility and conformational stability of antibodies and proteins. *Nat. Commun.* 14, 1937. doi:10.1038/s41467-023-37668-6

Runcie, K., Budman, D. R., John, V., and Seetharamu, N. (2018). Bi-specific and tri-specific antibodies-the next big thing in solid tumor therapeutics. *Mol. Med.* 24, 50. doi:10.1186/s10020-018-0051-4

Sakhnini, L. I., Greisen, P. J., Wiberg, C., Bozoky, Z., Lund, S., Perez, A.-M. W., et al. (2019). Improving the developability of an antigen binding fragment by aspartate substitutions. *Biochemistry-us* 58, 2750–2759. doi:10.1021/acs.biochem.9b00251

Santos, J., Pujols, J., Pallarès, I., Iglesias, V., and Ventura, S. (2020). Computational prediction of protein aggregation: Advances in proteomics, conformation-specific algorithms and biotechnological applications. *Comput. Struct. Biotechnol. J.* 18, 1403–1413. doi:10.1016/j.csbj.2020.05.026

Saurabh, S., Kalonia, C., Li, Z., Hollowell, P., Waigh, T., Li, P., et al. (2022). Understanding the stabilizing effect of histidine on mAb aggregation: A molecular dynamics study. *Mol. Pharm.* 19, 3288–3303. doi:10.1021/acs.molpharmaceut.2c00453

Sawant, M. S., Streu, C. N., Wu, L., and Tessier, P. M. (2020). Toward drug-like multispecific antibodies by design. *Int. J. Mol. Sci.* 21, 7496. doi:10.3390/ijms21207496

Schneider, C., Buchanan, A., Taddese, B., and Deane, C. M. (2021). DLAB: Deep learning methods for structure-based virtual screening of antibodies. *Bioinformatics* 38, 377–383. doi:10.1093/bioinformatics/btab660

Schoeder, C. T., Schmitz, S., Adolf-Bryfogle, J., Sevy, A. M., Finn, J. A., Sauer, M. F., et al. (2021). Modeling immunity with Rosetta: Methods for antibody and antigen design. *Biochemistry-us* 60, 825–846. doi:10.1021/acs.biochem.0c00912

Seeliger, D. (2013). Development of scoring functions for antibody sequence assessment and optimization. *Plos One* 8, e76909. doi:10.1371/journal.pone.0076909

Sever, R., Roeder, T., Hindle, S., Sussman, L., Black, K.-J., Argentine, J., et al. (2019). bioRxiv: the preprint server for biology. *Biorxiv*, 833400. doi:10.1101/833400

Shahfar, H., Du, Q., Parupudi, A., Shan, L., Esfandiary, R., and Roberts, C. J. (2022). Electrostatically driven protein–protein interactions: Quantitative prediction of second osmotic virial coefficients to aid antibody design. *J. Phys. Chem. Lett.* 13, 1366–1372. doi:10.1021/acs.jpclett.1c03669

Shaker, B., Ahmad, S., Lee, J., Jung, C., and Na, D. (2021). *In silico* methods and tools for drug discovery. *Comput. Biol. Med.* 137, 104851. doi:10.1016/j.compbiomed.2021.104851

Shan, L., Mody, N., Sormani, P., Rosenthal, K. L., Damschroder, M. M., and Esfandiary, R. (2018). Developability assessment of engineered monoclonal antibody variants with a complex self-association behavior using complementary analytical and *in silico* tools. *Mol. Pharm.* 15, 5697–5710. doi:10.1021/acs.molpharmaceut.8b00867

Sheng, Z., Bimela, J. S., Katsamba, P. S., Patel, S. D., Guo, Y., Zhao, H., et al. (2022). Structural basis of antibody conformation and stability modulation by framework somatic hypermutation. *Front. Immunol.* 12, 811632. doi:10.3389/fimmu.2021.811632

Shimba, N., Kamiya, N., and Nakamura, H. (2016). Model building of antibody–antigen complex structures using GBSA scores. *J. Chem. Inf. Model* 56, 2005–2012. doi:10.1021/acs.jcim.6b00066

Shmool, T. A., Martin, L. K., Matthews, R. P., and Hallett, J. P. (2022). Ionic liquid-based strategy for predicting protein aggregation propensity and thermodynamic stability. *Jacs Au* 2, 2068–2080. doi:10.1021/jacsau.2c00356

Sircar, A., and Gray, J. J. (2010). SnugDock: Paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *Plos Comput. Biol.* 6, e1000644. doi:10.1371/journal.pcbi.1000644

Smialowski, P., Doose, G., Torkler, P., Kaufmann, S., and Frishman, D. (2012). PROSO II – A new method for protein solubility prediction. *Febs J.* 279, 2192–2200. doi:10.1111/j.1742-4658.2012.08603.x

Smiatek, J., Jung, A., and Bluhmki, E. (2020). Towards a digital bioprocess replica: Computational approaches in biopharmaceutical development and manufacturing. *Trends Biotechnol.* 38, 1141–1153. doi:10.1016/j.tibtech.2020.05.008

Smith, S. (1996). Ten years of Orthoclone OKT3 (muromonab-CD3): A review. *J. Transpl. Coord.* 6, 109–119; quiz 120-121. doi:10.7182/prtr.1.6.3.8145l3u185493182

Somani, S., Jo, S., Thirumangalathu, R., Rodrigues, D., Tanenbaum, L. M., Amin, K., et al. (2021). Toward biotherapeutics formulation composition engineering using site-identification by ligand competitive saturation (SILCS). *J. Pharm. Sci.* 110, 1103–1110. doi:10.1016/j.xphs.2020.10.051

Sormanni, P., Aprile, F. A., and Vendruscolo, M. (2015). Rational design of antibodies targeting specific epitopes within intrinsically disordered proteins. *Proc. Natl. Acad. Sci.* 112, 9902–9907. doi:10.1073/pnas.1422401112

Sormanni, P., Aprile, F. A., and Vendruscolo, M. (2018). Third generation antibody discovery methods: In silico rational design. *Chem. Soc. Rev.* 47, 9137–9157. doi:10.1039/c8cs00523k

Starr, C. G., and Tessier, P. M. (2019). Selecting and engineering monoclonal antibodies with drug-like specificity. *Curr. Opin. Biotech.* 60, 119–127. doi:10.1016/j.copbio.2019.01.008

Steinbrecher, T., Zhu, C., Wang, L., Abel, R., Negron, C., Pearlman, D., et al. (2017). Predicting the effect of amino acid single-point mutations on protein stability—large-scale validation of MD-based relative free energy calculations. *J. Mol. Biol.* 429, 948–963. doi:10.1016/j.jmb.2016.12.007

Strohl, W. R. (2018). Current progress in innovative engineered antibodies. *Protein Cell* 9, 86–120. doi:10.1007/s13238-017-0457-8

Svilenov, H. L., Arosio, P., Menzen, T., Tessier, P., and Sormanni, P. (2023). Approaches to expand the conventional toolbox for discovery and selection of antibodies with drug-like physicochemical properties. *Mabs* 15, 2164459. doi:10.1080/19420862.2022.2164459

Swinney, D. C., and Anthony, J. (2011). How were new medicines discovered? *Nat. Rev. Drug Discov.* 10, 507–519. doi:10.1038/nrd3480

Tartaglia, G. G., Cavalli, A., Pellarin, R., and Caflisch, A. (2004). The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein Sci.* 7, 1939–1941. doi:10.1110/ps.04663504

Thorsteinson, N., Gunn, J. R., Kelly, K., Long, W., and Labute, P. (2021). Structure-based charge calculations for predicting isoelectric point, viscosity, clearance, and profiling antibody therapeutics. *Mabs* 13, 1981805. doi:10.1080/19420862.2021.1981805

Thorsteinson, N., Comeau, S. R., and Kumar, S. (2023). Structure-based optimization of antibody-based biotherapeutics for improved developability: A practical guide for molecular modelers. *Methods Mol. Biol.* doi:10.1007/978-1-0716-2609-2

Tilegenova, C., Izadi, S., Yin, J., Huang, C. S., Wu, J., Ellerman, D., et al. (2019). Dissecting the molecular basis of high viscosity of monospecific and bispecific IgG antibodies. *Mabs* 12, 1692764. doi:10.1080/19420862.2019.1692764

Tiller, T., Schuster, I., Deppe, D., Siegers, K., Strohner, R., Herrmann, T., et al. (2013). A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. *Mabs* 5, 445–470. doi:10.4161/mabs.24218

Tiller, K. E., Chowdhury, R., Li, T., Ludwig, S. D., Sen, S., Maranas, C. D., et al. (2017). Facile affinity maturation of antibody variable domains using natural diversity mutagenesis. *Front. Immunol.* 8, 986. doi:10.3389/fimmu.2017.00986

Tomar, D. S., Kumar, S., Singh, S. K., Goswami, S., and Li, L. (2016). Molecular basis of high viscosity in concentrated antibody solutions: Strategies for high concentration drug product development. *Mabs* 8, 216–228. doi:10.1080/19420862.2015.1128606

Tomar, D. S., Singh, S. K., Li, L., Broulidakis, M. P., and Kumar, S. (2018). *In silico* prediction of diffusion interaction parameter (kD), a key indicator of antibody solution behaviors. *Pharm. Res.* 35, 193. doi:10.1007/s11095-018-2466-6

Tomar, D. S., Licari, G., Bauer, J., Singh, S. K., Li, L., and Kumar, S. (2021). Stress-dependent flexibility of a full-length human monoclonal antibody: Insights from molecular dynamics to support biopharmaceutical development. *J. Pharm. Sci.* 111, 628–637. doi:10.1016/j.xphs.2021.10.039

Townsend, S., Fennell, B. J., Apgar, J. R., Lambert, M., McDonnell, B., Grant, J., et al. (2015). Augmented Binary Substitution: Single-pass CDR germ-lining and stabilization of therapeutic antibodies. *Proc. Natl. Acad. Sci.* 112, 15354–15359. doi:10.1073/pnas.1510944112

Trask, A., Hill, F., Reed, S., Rae, J., Dyer, C., and Blunsom, P. (2018). Neural arithmetic logic units. *arXiv.* doi:10.48550/arxiv.1808.00508

Trovato, A., Seno, F., and Tosatto, S. C. E. (2007). The PASTA server for protein aggregation prediction. *Protein Eng. Des. Sel.* 20, 521–523. doi:10.1093/protein/gzm042

Tsumoto, K., and Kuroda, D. (2022). in *Computer-aided antibody design*. Editors K. Tsumoto and D. Kuroda (Springer-Verlag New York Inc).

Tubiana, J., Schneidman-Duhovny, D., and Wolfson, H. J. (2022). ScanNet: An interpretable geometric deep learning model for structure-based protein binding site prediction. *Nat. Methods* 19, 730–739. doi:10.1038/s41592-022-01490-7

ULC, C. C. G. (2021). *Molecular operating environment (MOE)*. Montreal, QC, Canada. 2019.01. 1010 Sherbooke St. West, Suite #910, H3A 2R7

Valldorf, B., Hinz, S. C., Russo, G., Pekar, L., Mohr, L., Klemm, J., et al. (2022). Antibody display technologies: Selecting the cream of the crop. *Biol. Chem.* 403, 455–477. doi:10.1515/hsz-2020-0377

van der Kant, R., Karow-Zwick, A. R., Durme, J. V., Blech, M., Gallardo, R., Seeliger, D., et al. (2017). Prediction and reduction of the aggregation of monoclonal antibodies. *J. Mol. Biol.* 429, 1244–1261. doi:10.1016/j.jmb.2017.03.014

van der Kant, R., Bauer, J., Karow-Zwick, A. R., Kube, S., Garidel, P., Blech, M., et al. (2019). Adaption of human antibody λ and κ light chain architectures to CDR repertoires. *Protein Eng. Des. Sel.* 32, 109–127. doi:10.1093/protein/gzz012

Vatsa, S. (2022). *In silico* prediction of post-translational modifications in therapeutic antibodies. *Mabs* 14, 2023938. doi:10.1080/19420862.2021.2023938

Vecchio, A. D., Deac, A., Liò, P., and Veličković, P. (2021). Neural message passing for joint paratope-epitope prediction. *Arxiv.*

Walsh, I., Seno, F., Tosatto, S. C. E., and Trovato, A. (2014). PASTA 2.0: An improved server for protein aggregation prediction. *Nucleic Acids Res.* 42, W301–W307. doi:10.1093/nar/gku399

Wang, Q., Chung, C., Chough, S., and Betenbaugh, M. J. (2018). Antibody glycoengineering strategies in mammalian cells. *Biotechnol. Bioeng.* 115, 1378–1393. doi:10.1002/bit.26567

Wang, Q., Chen, Y., Park, J., Liu, X., Hu, Y., Wang, T., et al. (2019). Design and production of bispecific antibodies. *Antibodies* 8, 43. doi:10.3390/antib8030043

Wei, G.-W. (2019). Protein structure prediction beyond AlphaFold. *Nat. Mach. Intell.* 1, 336–337. doi:10.1038/s42256-019-0086-4

Weitzner, B. D., Jeliazkov, J. R., Lyskov, S., Marze, N., Kuroda, D., Frick, R., et al. (2017). Modeling and docking of antibody structures with Rosetta. *Nat. Protoc.* 12, 401–416. doi:10.1038/nprot.2016.180

Wilman, W., Wróbel, S., Bielska, W., Deszynski, P., Dudzic, P., Jaszczyszyn, I., et al. (2022). Machine-designed biotherapeutics: Opportunities, feasibility and advantages of deep learning in computational antibody discovery. *Brief. Bioinform* 23, bbac267. doi:10.1093/bib/bbac267

Wittmund, M., Cadet, F., and Davari, M. D. (2022). Learning epistasis and residue coevolution patterns: Current trends and future perspectives for advancing enzyme engineering. *ACS Catal.* 12, 14243–14263. doi:10.1021/acscatal.2c01426

Xu, Y., Wang, D., Mason, B., Rossomando, T., Li, N., Liu, D., et al. (2018). Structure, heterogeneity and developability assessment of therapeutic antibodies. *Mabs* 11, 239–264. doi:10.1080/19420862.2018.1553476

Yadav, S., Sreedhara, A., Kanai, S., Liu, J., Lien, S., Lowman, H., et al. (2011). Establishing a link between amino acid sequences and self-associating and viscoelastic behavior of two closely related monoclonal antibodies. *Pharm. Res.* 28, 1750–1764. doi:10.1007/s11095-011-0410-0

Yadav, S., Laue, T. M., Kalonia, D. S., Singh, S. N., and Shire, S. J. (2012). The influence of charge distribution on self-association and viscosity behavior of monoclonal antibody solutions. *Mol. Pharm.* 9, 791–802. doi:10.1021/mp200566k

Yamashita, K., Ikeda, K., Amada, K., Liang, S., Tsuchiya, Y., Nakamura, H., et al. (2014). Kotai antibody builder: Automated high-resolution structural modeling of antibodies. *Bioinformatics* 30, 3279–3280. doi:10.1093/bioinformatics/btu510

Yan, X. C., Sanders, J. M., Gao, Y.-D., Tudor, M., Haidle, A. M., Klein, D. J., et al. (2020). Augmenting hit identification by virtual screening techniques in small molecule drug discovery. *J. Chem. Inf. Model* 60, 4144–4152. doi:10.1021/acs.jcim.0c00113

Yang, W., Tan, P., Fu, X., and Hong, L. (2019). Prediction of amyloid aggregation rates by machine learning and feature selection. *J. Chem. Phys.* 151, 084106. doi:10.1063/1.5113848

Yu, X., Tsibane, T., McGraw, P. A., House, F. S., Keefer, C. J., Hicar, M. D., et al. (2008). Neutralizing antibodies derived from the B cells of 1918 influenza pandemic survivors. *Nature* 455, 532–536. doi:10.1038/nature07231

Zhang, C., Samad, M., Yu, H., Chakroun, N., Hilton, D., and Dalby, P. A. (2018). Computational design to reduce conformational flexibility and aggregation rates of an antibody fab fragment. *Mol. Pharm.* 15, 3079–3092. doi:10.1021/acs.molpharmaceut.8b00186

Zhang, Y., Wu, L., Gupta, P., Desai, A. A., Smith, M. D., Rabia, L. A., et al. (2020). Physicochemical rules for identifying monoclonal antibodies with drug-like specificity. *Mol. Pharm.* 17, 2555–2569. doi:10.1021/acs.molpharmaceut.0c00257

Zibaee, S., Makin, O. S., Goedert, M., and Serpell, L. C. (2007). A simple algorithm locates beta-strands in the amyloid fibril core of alpha-synuclein, Abeta, and tau using the amino acid sequence alone. *Protein Sci.* 16, 906–918. doi:10.1110/ps.062624507

Zurdo, J. (2013). Developability assessment as an early de-risking tool for biopharmaceutical development. *Pharm. Bioprocess* 1, 29–50. doi:10.4155/pbp.13.3

Glossary

Ab	Antibody
ADA	Anti-drug antibodies
ADCC	Antibody-dependent cell-mediated cytotoxicity
ADCP	Antibody-dependent cellular phagocytosis
Ag	Antigen
AI	Artificial intelligence
Ala	Alanine
APR	Aggregation-prone region
BCR	B-cell repertoire
CDC	Complement-dependent cytotoxicity
CDR	Complementarity determining region
CMC	Chemistry, manufacturing, and control
CNN	Convolutional neural networks
CQA	Clinical Quality Attributes
DAbI	Discovery of antibodies *in silico*
DENIS	DEvelopability Navigator *In Silico*
DLAB	Deep Learning for AntiBodies
DLS	Dynamic light scattering
Fab	Fragment antigen binding
FAIR	Findable, accessible, interoperable, and reusable
FDA	Food and Drug Administration
Fv	Fragment variable
FW	Framework
GAN	Generalized adversarial network
HC	Heavy chain
HCDR1-3	Heavy-chain complementarity determining regions 1–3
HCPF	High-concentration protein formulations
HIC	Hydrophobic interaction chromatography
HTTP	High-throughput
IgG	Immunoglobulin G
LC	Light chain
LCDR1-3	Light-chain complementarity determining regions 1–3
LO	Lead optimization
mAb	Monoclonal antibody
MaSIF	Molecular surface interaction fingerprints
MD	Molecular dynamics
MFI	Micro-flow imaging
MHC	Major histocompatibility complex
ML	Machine learning
MM-GBSA	Molecular mechanics—generalized Born solvent accessibility
MOE	Molecular Operating Environment
NBE	New biologic entity
NTC	Novel therapeutic concept
pI	Isoelectric point
PTM	Post-translational modification
QSAR	Quantitative structure–activity relationship
QSPR	Quantitative structure–property relationship
R&D	Research and development
RNN	Recurrent neural networks
RTP	Research target profile
ScFv	Single-chain fragment variable
SCM	Spatial charge map
SEC	Size exclusion chromatography
SINS	Self-interaction nanoparticle spectroscopy
TGFβ1	Transforming growth factor β1
UF/DF	Ultrafiltration/diafiltration
VH	Heavy-chain variable region
VK	Light-chain variable region (kappa isotype)
VL	Light-chain variable region (lambda isotype)

TYPE Original Research
PUBLISHED 24 August 2023
DOI 10.3389/fmolb.2023.1253689

Exploring rigid-backbone protein docking in biologics discovery: a test using the DARPin scaffold

Francis Gaudreault[1], Jason Baardsnes[1], Yuliya Martynova[1], Aurore Dachon[1], Hervé Hogues[1], Christopher R. Corbeil[1], Enrico O. Purisima[1], Mélanie Arbour[1] and Traian Sulea[1,2]*

[1]Human Health Therapeutics Research Centre, National Research Council Canada, Montreal, QC, Canada, [2]Institute of Parasitology, McGill University, Montreal, QC, Canada

Check for updates

OPEN ACCESS

EDITED BY
F. Javier Luque,
University of Barcelona, Spain

REVIEWED BY
Pablo Chacon,
Spanish National Research Council (CSIC), Spain
Baldomero Oliva,
Pompeu Fabra University, Spain

*CORRESPONDENCE
Traian Sulea,
✉ traian.sulea@nrc-cnrc.gc.ca

RECEIVED 05 July 2023
ACCEPTED 14 August 2023
PUBLISHED 24 August 2023

CITATION
Gaudreault F, Baardsnes J, Martynova Y, Dachon A, Hogues H, Corbeil CR, Purisima EO, Arbour M and Sulea T (2023), Exploring rigid-backbone protein docking in biologics discovery: a test using the DARPin scaffold.
Front. Mol. Biosci. 10:1253689.
doi: 10.3389/fmolb.2023.1253689

Accurate protein-protein docking remains challenging, especially for artificial biologics not coevolved naturally against their protein targets, like antibodies and other engineered scaffolds. We previously developed ProPOSE, an exhaustive docker with full atomistic details, which delivers cutting-edge performance by allowing side-chain rearrangements upon docking. However, extensive protein backbone flexibility limits its practical applicability as indicated by unbound docking tests. To explore the usefulness of ProPOSE on systems with limited backbone flexibility, here we tested the engineered scaffold DARPin, which is characterized by its relatively rigid protein backbone. A prospective screening campaign was undertaken, in which sequence-diversified DARPins were docked and ranked against a directed epitope on the target protein BCL-W. In this proof-of-concept study, only a relatively small set of 2,213 diverse DARPin interfaces were selected for docking from the huge theoretical library from mutating 18 amino-acid positions. A computational selection protocol was then applied for enrichment of binders based on normalized computed binding scores and frequency of binding modes against the predefined epitope. The top-ranked 18 designed DARPin interfaces were selected for experimental validation. Three designs exhibited binding affinities to BCL-W in the nanomolar range comparable to control interfaces adopted from known DARPin binders. This result is encouraging for future screening and engineering campaigns of DARPins and possibly other similarly rigid scaffolds against targeted protein epitopes. Method limitations are discussed and directions for future refinements are proposed.

KEYWORDS

binding affinity, protein-protein docking, rigid backbone, DARPin, ProPOSE

1 Introduction

Biologics have witnessed a tremendous growth in the past decades, with antibody-based therapeutics leading the way and recombinant proteins forming another important market segment (DeFrancesco, 2019; Lu et al., 2020; Kaplon et al., 2023). Advances in computational methods have spurred the idea that in the not-so-distant future, novel biologics can be discovered entirely *in silico*, complementing current wet-lab methods such as immunization and display technologies. This emerging field is dubbed *de novo* discovery of biologics with a particular emphasis on *de novo* antibody engineering (Fischman and Ofran, 2018).

Central to this *de novo* discovery approach is the ability to dock and score large libraries of biologic variants on the three-dimensional (3D) structure of a target protein (e.g., the antigen in the case of antibodies). Artificial intelligence/machine learning (AI/ML)-based methods like AlphaFold2 (Jumper et al., 2021), which have recently demonstrated a tremendous success in predicting protein structures and complexes of biologically co-evolved proteins, unfortunately are not applicable to docking and scoring of antibodies and artificially designed proteins (Yin et al., 2022). This limitation is due to co-evolution data being essential to AI/ML's success in protein-protein docking (Evans et al., 2022; Gao et al., 2022). Compounding the docking and scoring challenge is the difficulty to predict 3D structures of antibody libraries. While there has been some recent success in modeling antibodies with AI/ML methods without co-evolutionary information, there are still challenges in predicting the conformation of the hypervariable CHR-H3 loop (Abanades et al., 2022; Cohen et al., 2022; Ruffolo et al., 2022). Due to technical limitations from the high dimensionality of the CDR-H3 conformational space, the applicability of *de novo* antibody discovery efforts based on docking modeled antibody libraries to an antigen structure was met with limited success, as reported with several classical approaches (Adolf-Bryfogle et al., 2018; Chowdhury et al., 2018; Warszawski et al., 2019; Wood, 2021). Instead, applications on biologics displaying limited amounts of flexibility should be explored for increased likelihood of success (Youn et al., 2017; Radom et al., 2019).

We previously developed ProPOSE, an exhaustive direct protein-protein docker with full atomistic details (Hogues et al., 2018). By allowing side-chain rearrangements upon docking, ProPOSE delivers the current leading-edge performance in both general protein-protein docking and the specific case of antibody-antigen docking, when the backbone conformations of the interacting partners in the complex are *a priori* known. More specifically, ProPOSE maintains a strong performance even when side-chain flexibility is of concern. However, the docking accuracy was lower when backbone atoms experienced significant displacements between the bound and unbound states. We anticipated that despite its limitations, ProPOSE should be able to show utility in *de novo* biologics discovery when there is limited backbone flexibility upon binding and when reasonable models of backbone conformations can be inferred for the library of potential binders.

Hence, in this proof-of-concept study, we turned away from antibodies and towards the well-known engineered scaffold called DARPin (Designed Ankyrin Repeat Protein) (Binz et al., 2003). The DARPin scaffold has been refined over the years and has proven its value for the discovery of molecules with various medical and engineering applications, for example, as biotherapeutics, diagnostic agents, biosensors, molecular probes and crystallization helpers (Pluckthun, 2015; Rothenberger et al., 2022; Strittmatter et al., 2022). Compared to antibodies, DARPins are generally considered to be more rigid due to their smaller size and more defined structure. The repeating ankyrin unit (a β-turn followed by two anti-parallel α-helices) confers rigidity and stability to their structure (Kramer et al., 2010; Schilling et al., 2022). Such a limited backbone flexibility thus appears suitable for modeling DARPin substitution variants relatively reliably starting from available DARPin template structures.

Hence, the exploratory prospective study described here was centered around applying ProPOSE rigid-backbone docking to the DARPin scaffold exhibiting relative backbone rigidity. A computational flow was devised to generate a relatively small library of diverse DARPin interfaces for directed docking to a known epitope on the structure of the protein target, BCL-W. A selection procedure was further devised to establish a score threshold that captured self-consistent positive controls generated within the same computational procedure. Prospective computational designs were then subjected to experimental testing. Testing of 18 top-ranked hits demonstrated that half of them had detected binding to the target. Comparative analysis of computational and experimental data prompted to several limitations and areas for future improvements of the rigid-docking based approach for *de novo* biologics discovery.

2 Materials and methods

2.1 Computational methods

The sequence-based and structure-based computational design process (Figure 1) consisted of 6 steps which are described in the following sub-sections.

2.1.1 Defining the DARPin common framework sequence

Hundreds of DARPin structures with various topologies were published in the literature and are accessible in the PDB, among which many have 4 or 5 repeated ankyrin motifs. Two DARPins evolved through ribosome display to bind BCL-W, and corresponding to PDB entries 4k5a and 4k5b (Schilling et al., 2014b), were used as known binders in this study. These known binders engage the target in a binding mode which is typical for DARPins, which consists of interactions made by the concave paratope formed by their 5 repeated ankyrin motifs (Binz et al., 2003; Kramer et al., 2010; Pluckthun, 2015; Schilling et al., 2022). By inspecting the sequences and structures of these known binders and other DARPins with available crystal structures in PDB, a common framework sequence was defined for further library expansion. The main features considered during the selection of a DARPin common framework sequence were: 1) 157 amino acids starting with DLGKK and ending with LQKAA sequences; 2) conserved regions at these N- and C-terminal ends; 3) consensus residues deemed essential for the stability of the overall fold along repeated ankyrin motifs; and 4) key residues contributing to binding along repeated ankyrin motifs. These criteria led to a single DARPin common framework sequence, which corresponded to the DARPin of chain F in the PDB entry 4drx (the nomenclature 4drx [F] is used) (Pecqueur et al., 2012).

2.1.2 Expanding the framework sequence into a DARPin library

A set of 18 amino-acid positions within the defined DARPin common framework sequence were manually selected and allowed to vary (referred to as variable positions). These positions, which have high-frequency rates of mutation as observed from sequence alignments of many DARPins from the literature, are: 45, 46, 48, 56, 57, 78, 79, 81, 89, 90, 111, 112, 114, 122, 123, 144, 145 and 147 (standard DARPin numbering is applied). Amino-acid side chains at these positions are lining the concave face of the DARPin scaffold by

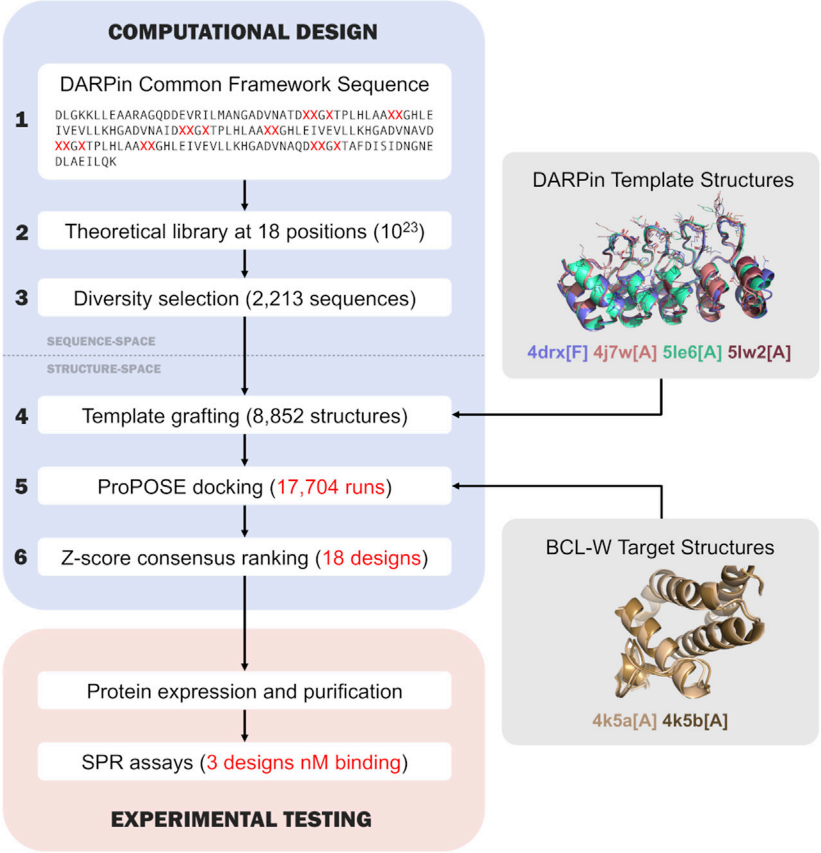

COMPUTATIONAL DESIGN
1
DARPin Common Framework Sequence
DLGKKLLEAARAGQDDEVRILMANGADVNATDXXGXTPLHLAAXXGHLE
IVEVLLKHGADVNAIDXXGXTPLHLAAXXGHLEIVEVLLKHGADVNAVD
XXGXTPLHLAAXXGHLEIVEVLLKHGADVNAQDXXGXTAFDISIDNGNE
DLAEILQK
2
Theoretical library at 18 positions (10^{23})
3
Diversity selection (2,213 sequences)
SEQUENCE-SPACE
STRUCTURE-SPACE
4
Template grafting (8,852 structures)
5
ProPOSE docking (17,704 runs)
6
Z-score consensus ranking (18 designs)
Protein expression and purification
SPR assays (3 designs nM binding)
EXPERIMENTAL TESTING
DARPin Template Structures
4drx[F] 4j7w[A] 5le6[A] 5lw2[A]
BCL-W Target Structures
4k5a[A] 4k5b[A]

FIGURE 1
Flowchart of the overall computational design and experimental testing. The first three steps of the computational design are in the sequence space, while the last three steps are in the 3D-structure space and inherit structural knowledge from the Protein Data Bank (Berman et al., 2000). The main steps of the computational design are numbered outside the boxes and described in the text.

being located within the β-turn loops and following short α-helices of the ankyrin repeats (Figure 2).

Two positive controls having the 18 variable positions corresponding to known DARPin binders of BCL-W, having PDB entries 4k5a [B] and 4k5b [B], were also built manually into the library. It is important to note that these constructed positive controls share the common framework of the designed library described above and thus differ at several positions from the frameworks of the originating known binders (Figure 2).

2.1.3 Selecting a DARPin sub-library of diverse sequences

An alphabet was created to group amino acids by chemical properties. The following five groups excluding Gly, Cys and Pro were defined: positively-charged (Arg, His, Lys); negatively-charged (Asp, Glu); polar (Asn, Gln, Ser, Thr); non-polar (Ala, Ile, Leu, Met, Val); and aromatic (Phe, Trp, Tyr). Equal probability was given to each group to be selected when mutating sequences. Similarly, amino acids within a group were given equal probability.

The designs were generated using a stochastic procedure in which variable amino-acid positions were mutated either through point mutations or through permutations of amino acids. Multiple starting points in the sequence space were used to generate the designs. The set of mutated designs (M-set) were generated starting from the 4drx [F] sequence chosen as common framework. All variable amino acids were forced to be mutated in this set. To be included in the library, a design sequence had to be sufficiently distant to the designs comprised within the same set. A threshold distance of 10 was fixed which required at least 10 alphabet group changes. The set of permutated designs (P-set) were generated starting from the sequences of the two positive controls. No change in the alphabet group was imposed for this set. A threshold distance of 13 was set, requiring at least 13 amino-acid changes.

2.1.4 Grafting DARPin sequences onto template structures

The designed sub-library sequences were grafted onto four DARPin template structures followed by side-chain repacking using SCWRL4 (Krivov et al., 2009). The last two alanine residues at the C-terminus of the template sequence were truncated for modeling purposes. Only those side-chains that are different at a given side-chain were mutated and repacked to preserve the structural integrity of the original crystal structures of the DARPin templates. The entire structure was then allowed to be repacked. The DARPin templates from the following PDB entries were used in this study: 4drx [F], 4j7w [A], 5lw2 [A] and 5le6 [A] (Figure 1). The backbone structures of these templates are distinct

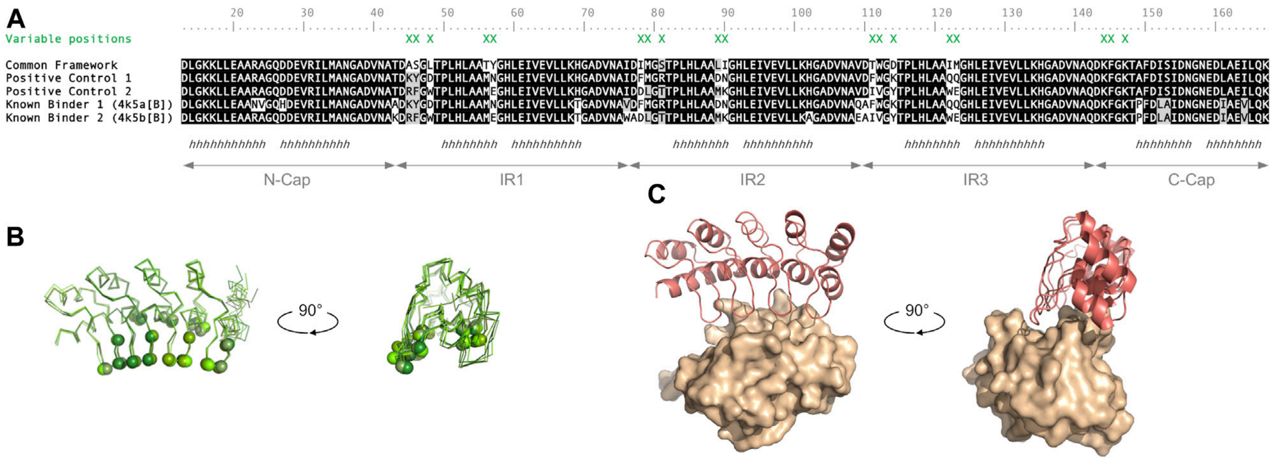

A
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
Variable positions
XX X XX XX X XX XX X XX XX X
Common Framework DLGKKLLEAARAGQDDEVRILMANGADVNATDASGLTPLHLAATYGHLEIVEVLLKHGADVNAIDIMGSTPLHLAALIGHLEIVEVLLKHGADVNAVDTWGDTPLHLAAIMGHLEIVEVLLKHGADVNAQDKFGKTAFDISIDNGNEDLAEILQK
Positive Control 1 DLGKKLLEAARAGQDDEVRILMANGADVNATDKYGDTPLHLAAMNGHLEIVEVLLKHGADVNAIDFMGRTPLHLAADNGHLEIVEVLLKHGADVNAVDFWGKTPLHLAAQQGHLEIVEVLLKHGADVNAQDKFGKTAFDISIDNGNEDLAEILQK
Positive Control 2 DLGKKLLEAARAGQDDEVRILMANGADVNATDRFGWTPLHLAAMEGHLEIVEVLLKHGADVNAIDDLGTTPLHLAAMKGHLEIVEVLLKHGADVNAVDIVGYTPLHLAAWEGHLEIVEVLLKHGADVNAQDKFGKTAFDISIDNGNEDLAEILQK
Known Binder 1 (4k5a[B]) DLGKKLLEAANVGQHDEVRILMANGADVNAADKYGDTPLHLAAMNGHLEIVEVLLKTGADVNAVDFMGRTPLHLAADNGHLEIVEVLLKHGADVNAQAFWGKTPLHLAAQQGHLEIVEVLLKHGADVNAQDKFGKTPFDLAIDNGNEDIAEVLQK
Known Binder 2 (4k5b[B]) DLGKKLLEAARAGQDDEVRILMANGADVNAKDRFGWTPLHLAAMEGHLEIVEVLLKTGADVNAWADLGTTPLHLAAMKGHLEIVEVLLKAGADVNAEAIVGYTPLHLAAWEGHLEIVEVLLKHGADVNAQDKFGKTPFDLAIDNGNEDIAEVLQK
hhhhhhhhhhh hhhhhhhhhh hhhhhhhh hhhhhhhhhh hhhhhhhh hhhhhhhhhh hhhhhhhh hhhhhhhhhh hhhhhhhh hhhhhhhh
N-Cap
IR1
IR2
IR3
C-Cap
B
90°
C
90°

FIGURE 2
Variable positions on the DARPin scaffold docked onto BCL-W target epitope. **(A)** Sequence alignment between the common framework sequence (4drx [F]), known binders (4k5a [B] and 4k5b [B]), and the positive controls (PC1 and PC2) grafting the interface of the known binders onto the common template sequence, at the 18 variable positions (marked by green Xs). The conventional DARPin sequence numbering scheme is used, *h* denotes α-helix, IR1 to IR3 delineate internal ankyrin repeats 1-3, and N-Cap and C-Cap are the terminal ankyrin repeats. **(B)** Location of the 18 variable positions (spheres) on the 4 DARPin template structures (Cα-traces with different shades of green). **(C)** Location of the docking site on the BCL-W target protein indicated by the crystal structure (4k5b) of a known DARPin binder (red cartoon) complexed with the BCL-W target (molecular surface).

from those of the two known DARPin binders of BCL-W, 4k5a [B] and 4k5b [B], which were purposely excluded as structural templates to avoid the cognate-docking bias. The selected DARPin template structures underwent the following preparation procedure: 1) addition of missing side chain atoms (no repacking); 2) addition of missing hydrogen atoms and assignment of standard protonation states at pH 7; 3) optimization of the hydrogen-bond network the minH program (Hogues et al., 2014); and 4) AMBER force-field (Cornell et al., 1995; Hornak et al., 2006) energy minimization of added hydrogen atoms and any newly added side-chain atoms with harmonic restraints on all the other heavy atoms of 1,000 kcal/mol/$\mathring{A}^2$ followed by energy minimization of the entire structure with harmonic restraints of 10 kcal/mol/$\mathring{A}^2$ on backbone heavy atoms, 1 kcal/mol/$\mathring{A}^2$ on side-chain heavy atoms, and no restraints on hydrogen atoms.

2.1.5 DARPin docking protocols

The BCL-W docking-based screening of the DARPin library was performed using the exhaustive docking engine ProPOSE version 1.03 (Hogues et al., 2018). ProPOSE was run with default parameters using the HITSET flag to force binding towards the set of residues involved in binding BCL-W. Initially, no binding location (or epitope) was defined on the target protein BCL-W and exhaustive docking was performed all around the BCL-W structure. Two BCL-W structures were employed for docking, with PDB entries 4k5a [A] and 4k5b [C] (Figure 1), which correspond to the BCL-W complexed with the two known DARPin binders. For each DARPin library sequence, the four DARPin structural templates carrying the grafted designed sequence were docked against the two BCL-W target structure, resulting in 8 docking experiments. In this study, only the top-1 scored pose generated by ProPOSE was considered for a given complex given its accuracy in pose recovery as top-1 when the protein backbone conformation is known, without the need for rescoring (Hogues et al., 2018). On average, a single docking run took 30 min to execute when parallelized on an Intel Xeon Gold 5,218 using 6 cores.

Epitope restriction on the BCL-W target was introduced after all docking calculations were completed. In this proof-of-concept study, we elected to target the same BCL-W epitope and the DARPin binding mode observed for the two known BCL-W DARPin binders (PDB entries 4k5a and 4k5b) (Schilling et al., 2014b). The similarity of predicted docked poses of designed sequences relative to these known structures was based on CAPRI classification (Lensink et al., 2017). Predictions were compared on the basis of: 1) the backbone RMSD of the ligand upon target superposition; 2) the backbone RMSD of the interface upon superposition of interface atoms; and 3) the fraction of preserved contacts (f_{con}). The ligand and target were DARPin and BCL-W, respectively. Noteworthy, f_{con} was used rather than the standard f_{nat} from CAPRI that is derived from the comparison to a native structure. Moreover, f_{con} is a position-dependent (amino acid-independent) measure allowing designs with different sequences to be compared. Two predictions were declared as having high, medium or acceptable quality, or as incorrect otherwise, with thresholds defined by the CAPRI classification (Lensink et al., 2017).

2.1.6 Ranking docked DARPin structures

The number of top-1 scored poses, N_{pose}, docked at the targeted epitope from the 8 docking runs was used to retain only those designs that have at least 2 poses docked at the target epitope. To this end, the predicted poses for a given DARPin were grouped using a greedy clustering algorithm with a tolerance of at least medium quality between cluster representatives. In geometric terms, for poses to be considered bound at the targeted epitope occupied by one of the known binders, they were required to have acceptable quality criteria, i.e., 1) f_{con} of at least 30% with a ligand backbone

RMSD >5.0Å and interface backbone RMSD >2.0 Å; or alternatively, 2) f_{con} between 10% and 30% while having a ligand backbone RMSD <10.0 Å or an interface backbone RMSD <4.0 Å. For poses to be part of the same cluster, they were required to have medium quality criteria, i.e., 1) f_{con} of at least 50% with ligand backbone RMSD >1.0 Å and interface backbone RMSD >1.0 Å; or alternatively, 2) f_{con} between 30% and 50% while having ligand backbone RMSD <5.0 Å or an interface backbone RMSD <2.0 Å. No cut-off in score was applied for the clustering.

For each design with $N_{pose} > 1$, a consensus score was derived as the arithmetic average over the docking scores of the poses binding to the targeted epitope. Consensus scores over the designs with $N_{pose} > 1$ were also normalized into Z-scores to better inform the selection of a top-ranked population based on a minimum number of standard deviations away from the mean calculated from the distribution of all DARPins combining the P-set designs, M-set designs and the positive controls.

2.1.7 Other software and data availability

Structure visualization was performed in PyMOL (The PyMOL Molecular Graphics System, Version 2.0, Schrödinger, LLC). Statistical analyzes were run in R (R Development Core Team, 2011). ClustalW2 was used to run the multiple sequence alignments (Larkin et al., 2007).

The sequence datasets generated for this study have been made available as a MongoDB with example scripts that can be found at the GitHub repository https://github.com/gaudreaultfnrc/Darpins.

2.2 Experimental methods

2.2.1 Protein expression and purification

Each DARPin design included a N-terminus tag (MRGSHHHHHHGS) and two alanines at their C-terminus as described in (Schilling et al., 2014a). The protein sequences were optimized for *Escherichia coli* expression using a multifactor algorithm (https://www.genscript.com/tools/gensmart-codon-optimization), then synthesized by GenScript. After inserting each gene in pET24a (+) via NdeI and NotI restriction enzyme sites, the final plasmids were transformed into NRC *E. coli* BL21-T7 strain (*rhaB lacZ*::P*tac*-T7 RNAP). For each clone, a 2.8-L Fernbach baffled flask containing 500 mL Animal-Product Free (APF) LB Miller (Athena Enzyme Systems Cat. 0133) plus 50 µg/mL kanamycin was inoculated with an overnight preculture to get an initial OD_{600nm} of 0.1. The flasks were incubated at 37°C, 200–250 rpm until an OD_{600nm} between 0.8 and 1.0 were reached. To induce protein expression 1 mM isopropyl β-d-1-thiogalactopyranoside (IPTG) was added and the culture incubated for another 4 h at 37°C, 200–250 rpm. The cultures were harvested, and the cell pellets stored at −80°C.

Before purification, a cell pellet was resuspended in Lysis buffer 50 mM $NaPO_4$, 300 mM NaCl, 10 mM imidazole, pH 7.4 with cOmplete protease inhibitors EDTA-free (Millipore Sigma Cat. 11836170001) and lysed by two passages on a French Pressure Cell Disruptor. Finally, the cell lysate was clarified by centrifugation at 10,000 x *g*, 4°C, for 15 min and filtration on 0.45 µm filter. A fraction of the clarified lysate (15 mL) was applied on a 3 mL HisPur Cobalt Spin Column (Thermo Fisher Cat. 89969) and the column was washed with 20 mM $NaPO_4$, pH 7.5, 500 mM NaCl, 0.3 mM TCEP, 15 mM imidazole. Elution was done with 20 mM $NaPO_4$, pH 7.5, 500 mM NaCl, 0.3 mM TCEP, 100 mM imidazole and pooled after visualization on SDS-PAGE. For some of the proteins, the purification was repeated to increase purity. Buffer exchange for DPBS (Thermo Fisher Cat. 14190144) was done with PD-10 desalting columns (Cytiva Cat. 17085101) and final concentration measured by Qubit Protein Assay (Thermo Fisher Cat. Q33211).

The design of BCL-W was based on (Schilling et al., 2014a) with an N-terminal Avi-tag followed by a bacteriophage lambda protein D fusion tag to improve protein solubility (Forrer and Jaussi, 1998) (see Supplementary Data). A 6xHis tag was added to the C-terminus of BCL-W for purification. Gene optimization, synthesis and cloning in pET24a (+) vector was done as described above for the DARPins. To allow *in vitro* biotinylation, the NRC *E. coli* BL21-T7 strain (*rhaB lacZ*::P*tac*-T7 RNAP) was first transformed with pBirAcm (Avidity), a plasmid expressing biotin ligase under *tac* promoter (IPTG inducible). After growing a chloramphenicol resistant colony in APF LP Miller medium containing 10 µg/mL chloramphenicol, electrocompetent cells were prepared using standard procedures. The plasmid pET24a (+)-BCL-W was then transformed in BL21-T7/pBirAcm strain and selected on APF LB Miller agar containing 50 µg/mL kanamycin and 10 µg/mL chloramphenicol.

Expression of BCL-W, cell lysis and clarification were done as described for the DARPins with some exceptions. Both antibiotics, kanamycin and chloramphenicol, were used, and biotin was added to a final concentration of 5 mM during the culture (25 mL). The cells were lysed in a buffer containing 50 mM $NaPO_4$, 300 mM NaCl, 10 mM imidazole, pH 8.0 (plus cOmplete EDTA-free protease inhibitors). The clarified lysate (2.5 mL) was applied on a 0.2 mL HisPur Cobalt Spin Column (Thermo Fisher Cat. 90090) and the column was washed with 20 mM $NaPO_4$, pH 7.5, 500 mM NaCl, 0.3 mM TCEP, 20 mM imidazole. Elution was done with 20 mM $NaPO_4$, pH 7.5, 500 mM NaCl, 0.3 mM TCEP, 300 mM imidazole and pooled after visualization on SDS-PAGE. Buffer exchange for DPBS (Thermo Fisher Cat. 14190144) was done with G-25 MiniTrap desalting columns (Cytiva Cat. 28918007) and final concentration measured by Qubit Protein Assay (Thermo Fisher Cat. Q33211). Purity levels are given in Supplementary Table S1 and SDS-PAGE gels are provided as Supplementary Data.

2.2.2 Binding affinity measurements

Surface plasmon resonance was used to screen the top 18 DARPin designs for binding to the biotinylated BCL-W using a Biacore T200 instrument (Cytiva Inc., Marlborough MA) at 25°C and with PBST running buffer (Teknova, Hollister CA) containing 0.05% Tween 20, 3.4 mM EDTA and an additional 350 mM NaCl. The strategy employed was to capture the biotinylated BCL-W onto the SPR surface with a CAP sensor chip (Cytiva Inc.) and flow a three-point concentration series of the DARPin scaffold using a 10-fold dilution series from 1 µM to cover a wide concentration range. From the resulting sensorgrams, the affinity constant of binding candidates can be determined. A CAP immobilization chip was prepared following the manufacturer's instructions. Each injection cycle consisted first of a 120-s injection at 5 µL/min of a 5-fold dilution of CAP reagent to indirectly immobilize streptavidin over flow-cells 1 and 2. This was followed by a 240-s capture of 5 µg/mL biotinylated BCL-W at 5 μL/min over flow cell 2 only to form the 60–62 RU BCL-W surface, and finally a three-point concentration injection of the DARPin scaffold or running buffer only using single-cycle kinetics was performed at 50 μL/min for 90 s with a 300-s dissociation phase. At the end of the dissociation phase, any BCL-W/DARPin complex was stripped from the SPR surface using a 60-s injection of 6 M GuCl/0.25 M NaOH taken from the CAP sensor chip reagent kit. The sensorgrams were double referenced and analyzed using the Biacore BiaEval software. Affinities of the DARPin scaffolds for BCL-W were determined using the steady state model, or the 1:1 binding model when kinetic rate constants could be evaluated.

2.2.3 Folding stability measurements

Differential scanning calorimetry (DSC) was used to determine the thermal transition midpoints (T_m) as previously performed (Schrag et al., 2019). DSC was carried out in a VP-Capillary DSC system instrument (Malvern Instruments Ltd., Malvern, United Kingdom). Samples were diluted in DPBS buffer to a final concentration of 0.4 mg/mL. DPBS blank and sample scans were carried out by increasing the temperature from 20°C to 100°C at a rate of 60°C/h, with feedback mode/gain set at “low”, filtering period of 8 s, pre-scan time of 3 min, and under 70 psi of nitrogen pressure. All data were analyzed with Origin 7.0 software (OriginLab Corporation, Northampton, MA). Thermograms were corrected by subtraction of corresponding DPBS blank scans and normalized to the protein molar concentration. The T_m values were determined using automated data processing with the rectangular peak finder algorithm for T_m. Melting temperatures are listed in Supplementary Table S1 and DSC thermograms are provided as Supplementary Data.

3 Results

3.1 Sequence-based and structure-based computational design

3.1.1 Overall design process

The flowchart in Figure 1 presents the overall computational design process devised and implemented for this rigid-docking based proof-of-concept engineering study based on the DARPin scaffold. It includes 6 steps: 1) definition of a single DARPin common framework sequence; 2) expansion of the common framework sequence into a DARPin sequence library with variable positions; 3) selection of a small DARPin sub-library consisting of diverse sequences; 4) grafting of the sequence sub-library onto DARPin structural templates; 5) docking of DARPin sub-library to target protein structures, the core component of the process; and 6) ranking docked DARPin variants for experimental testing. The first three steps operate in the sequence space, whereas the last three in the 3D structure space. All the steps are described in detail in the sub-sections of the Methods section. The following sub-sections focus more in-depth on results obtained in steps 3), 4), 5) and 6) of the process.

3.1.2 Selecting diverse DARPin sub-library sequences

Expanding a common framework sequence by varying 18 positions lining the concave face of the DARPin fold (Figure 2) resulted in 10^{23} theoretical library size. Millions of iterations were run to select a diverse sub-library fulfilling several design criteria (see Methods sub-Section 2.1.3). The resulting diverse sub-library comprised a total of 2,213 designs of which 1,429 were produced by mutations and 784 by permutations (Table 1). The closest designs in sequence are 9 amino-acid substitutions away from any of the two positive controls (Supplementary Figure S1), or 6 groups away when grouping amino acids by homology (see Methods section). The mutation-based designs have an even proportion of amino-acid groups at the variable positions (Supplementary Figure S2). In contrast, permutation-based designs have unevenly distributed amino-acid groups and lack Ala, His and Ser as inherited from the starting positive-control sequences (Supplementary Figure S2). In terms of net charge, mutation-based designs span a wide range from −16 to +3 with a mean net charge of −6.8, whereas permutation-based designs inherit the net charges of their respective parental positive control (Supplementary Figure S3).

In order to generate a sub-library that samples homogeneously the immense theoretical sequence space, designs were imposed to be orthogonal to each other. Clustering based on amino-acid properties indicated that most sequence space regions were covered by both mutation-based and permutation-based types of sequences, with a few areas only covered by the mutation-based set (Figure 3). While proximity in sequence might be perceivable between some of the designs and the two positive controls (Figure 3A), overall, the designed sequences were diverse and nearly equidistant from each other (Figure 3B).

3.1.3 Grafting sequence sub-library onto DARPin structural templates

Four crystal structures were used as templates in the modeling of the DARPin ligands (see Methods section). The variance in RMSD among these templates has a mean of 0.95 Å. The template 4drx [F] is more distant due in part to an opening of the last repeated motif of the scaffold. The magnitudes of backbone changes between each of these templates and any of the 2 known DARPin binders of BCL-W are larger than between the 2 known binders (0.42 Å). Thus, backbone RMSDs of 0.91, 0.75, 0.79 and 0.79Å were calculated to the 4k5a [B] known binder, and of 0.96, 0.75, 0.78 and 0.78Å to the 4k5b [A] known binder, for the template structures 4drx [F], 4j7w [A], 5le6 [A] and 5lw2 [A], respectively. More backbone variations could be observed in the unstructured region of the fourth ankyrin repeat, where the known BCL-W binders had a distinct conformational topology at the tip of this loop region. These variations in the templates relative to known binders were critical for testing the method in real-life application mode in which the bound backbone structure will be unknown *a priori*.

3.1.4 Docking DARPin sub-library structures to target

The entire set of sequence designs in the selected sub-library was grafted onto four template structures, then cross-docked against two target (BCL-W) structures, leading to 8 docking runs per DARPin sequence. The two backbone structures used for the target (4k5a [A] and 4k5b [C]) were relatively close from each other, with an RMSD of 0.77 Å. They also engaged their respective known DARPin binders (4k5a [B] and 4k5b [A]) via a well-preserved binding interface with backbone atoms deviating by an RMSD of 0.60 Å.

TABLE 1 Library design statistics.

Starting DARPin	PDB ID	Variable positions[a]	Set[b]	N_{seq}[c]	d_{seq}[d]	$d_{chemseq}$[e]	Q_{net}[f]
Common framework	4drx [F]	ASLTYIMSLITWDIMKFK	M	1,429	9	7	−6.8
Known binder	4k5a [B]	KYDMNFMRDNFWKQQKFK	P	284	12	7	−4.0
Known binder	4k5b [B]	RFWMEDLTMKIVYWEKFK	P	500	9	6	−6.0

[a] Position IDs, in the same order: 45, 46, 48, 56, 57, 78, 79, 81, 89, 90, 111, 112, 114, 122, 123, 144, 145 and 147.
[b] M: mutation; P: permutation.
[c] Number of sequences.
[d] Closest distance from a design to a known binder interface at 18 variable positions, expressed as number of substitutions.
[e] Closest distance from a design to a known binder interface at 18 variable positions, expressed as number of homology group changes.
[f] Mean net charge of designs within the set.

Hence, in this study, the docked poses for novel DARPins were required to bind around the same epitope that is targeted by these two known DARPin binders of BCL-W. In more technical terms, the predicted poses of designed DARPins were required to have an overlap of at least acceptable quality (according to CAPRI classification (Lensink et al., 2017) to either of these known binders. This was met by 1,033 designs (47% of the sub-library), and are referred to as "locus designs". (Increasing the stringency and imposing at least a medium quality of pose overlap with the known binders reduced the number of locus designs to 559.) We found no bias towards either of the two target BCL-W structures used for docking, as 811 designs docked to structure 4k5a [A] and 632 designs to structure 4k5b [C]. In terms of the template DARPin structures used for docking, 5lw2 [A] was the least successful template structure with 344 docked designs, followed by 380 designs docked on 5le6 [A], 533 on 4j7w [A] and 579 on 4drx [F]. The net charge distribution of the 1,033 locus designs is slightly different relative the entire docked sub-library of 2,213 designs, as it has sharper peaks at the −6 and −4 net charges (Supplementary Figure S3).

3.1.5 Ranking DARPin virtual hits

First, locus design DARPins were filtered based on the number of top-1 scored poses, N_{pose}, that were docked at the targeted epitope from the 8 docking runs for each DARPin. A total of 293 locus designs (13% of the sub-library) had at least 2 poses docked at the target epitope. These were retained for further ranking and were called "consensus designs". The net charge distribution among the consensus designs had even sharper peaks at the net charges −6 and −4, with the majority of consensus designs at charge −6 (Supplementary Figure S3).

For each of selected 293 consensus designs, a consensus score was derived as the arithmetic average over the docking scores of the poses binding to the targeted epitope. These consensus scores were normally distributed and ranged from −84.2 to −45.3, from strongest to weakest binder (Figure 4). The permutation-based designs were preferentially chosen according to the consensus scores with a median of −65 as opposed to a median of −61 for the mutation-based ones. In total, 152 (52%) and 71 (24%) designs that docked at the targeted epitope did so with values in N_{pose} of 2 and 3, respectively (inset in Figure 4). The lowest consensus score corresponded to a Z-score of −3.5. Consensus designs with Z-scores below −1.5 were selected for experimental validation, which formed a set consisting of 18 novel DARPins (Table 2). An overlay of all consensus poses for the selected designs is shown in Figure 5A.

The two positive-control interfaces, having the 18 variable positions imported from the two known DARPin binders of BCL-W and grafted onto the common framework sequence, were also docked in the same manner. These positive controls had consensus scores of −75.8 and −73.5, corresponding to Z-scores of −2.1 and −1.7, respectively. With the assumption that none of the designs are true binders, the separation of the positive controls from the designs had an AUC of 0.971 (Figure 4). The AUC dropped to 0.922 when best scores were used instead of consensus scores for designs with $N_{pose} > 1$. While working with rigid scaffolds, this observation suggested the need for structural ensembles to achieve better enrichments and thus motivated the use of consensus scores over best scores in the sections that follow. These positive controls were ranked within the range of the top-18 novel designs, and they were also subjected to experimental testing. It is important to note that to properly compare the scores for the two parental known binders of BCL-W with those of the mutants, we needed to base it on the modeled structures of the known binders rather than their crystal structures. Using the crystal structures would be a case of cognate backbone docking and perfect match in shape complementarity leading to out-of-range scores (DARPin/BCL-W docking scores of −144.5 and −123.3 were obtained for 4k5a [B]/4k5a [A] and 4k5b [B]/4k5b [C], respectively). The overlay of all locus docked poses for the two grafted positive control interfaces is shown in Figure 5B to have the same orientation with those of the selected designs (Figure 5A). These poses are further similarly oriented with those of the known binders, as exemplified in Figure 5C. A closer examination reveals that despite an excellent pose recovery for this cross-docking experiment, there are certain noticeable differences in the fine atomic details at the interface, which are likely due mainly to non-cognate backbone coordinates and to a lesser extent to changes of the framework sequence outside the 18 variable positions. Overall, cross-docking of positive controls predicted that they would retain similar binding relative to the corresponding known binders.

As presented in Table 2, most of these top consensus designs (13 of 18) were from the permutation set despite its smaller representation in the initial library. Also, 15 out of the 18 consensus designs had a net charge equal to that of a positive control (−6 or −4), despite the random sequence generation procedure employed. On average, the top-18 designs were 15 mutations away from the positive-control interfaces, with the closest design being 10 mutations away. These top consensus designs had between 2 and 7 top-1 poses bound at the target epitope. Interestingly, a strong bias towards an increased consensus was

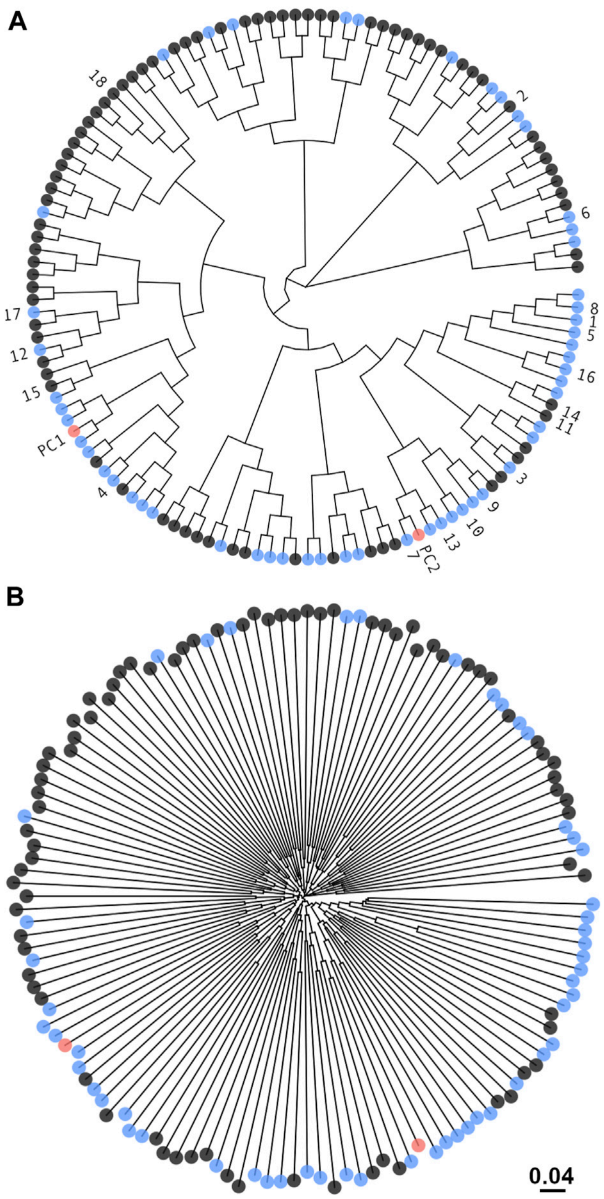


FIGURE 3
Diversity of the DARPin sequence sub-library. Unrooted phylogenetic tree from hierarchical clustering of sequences by the chemical properties of amino acids using defined amino-acid homology groups (see Methods section). Sequences marked in black are from the mutation-based set and in blue from the permutation-based set. The two positive-control sequences are shown in red. Only a 5% random sample of the sub-library consisting of 134 sequences is plotted. The top-18 consensus designs and positive controls were annotated. **(A)** For visual clarity, the terminal branches were equally trimmed down to a cladogram giving the illusion of sequence proximity (Yu, 2020). **(B)** The non-trimmed tree that preserves the ordering in **(A)** is shown to illustrate the true divergence in sequence between designs. For reference, the evolutionary distance is shown.

observed with 14 out of the 18 novel designs (78%) with at least 3 representative poses bound at the targeted epitope. This level has to be contrasted to only 24% of all consensus designs reaching an $N_{pose} > 2$. Hence, not only were the top designs predicted to bind stronger to the target, they also did so with a higher number of predicted consensus poses, with an average N_{pose} of 4.2. Comparably, the two positive controls had N_{pose} values of 4 and 7 (Figure 4).

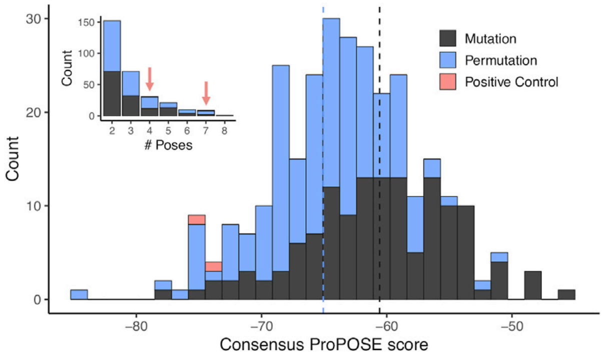


FIGURE 4
Distribution of scores from docking-based screening. Distribution of scores obtained from the docking experiments using ProPOSE on the entire set of designs in the library. The scores were obtained from a consensus of multiple predictions binding at the same locus while imposing an acceptable or better quality among the representatives of the cluster. The scores follow a normal distribution with the median marked as dashed lines. The underlying area-under-the-curve of the receiver operating characteristic (AUC-ROC) curve obtained from the separation of the two positive controls from the combined mutation and permutation design sets has a value 0.971. The inset shows the distribution in number of representatives used to calculate the consensus ProPOSE score. The two positive controls have 4 and 7 representatives.

3.2 Experimental testing of DARPin designs

The 18 top-ranked consensus designs, together with the 2 positive controls and the 2 parental known binder DARPins were produced in bacteria, purified by IMAC and screened for binding to BCL-W by SPR. The purity levels of the DARPins ranged from 45% to 99% with an average of 82% (Supplementary Table S1). While some of these levels could be considered as suboptimal for SPR experiments and might lead to non-specific binding, they were deemed sufficient for a first-pass screening. Tested DARPins were flowed at a fixed concentration over biotinylated target protein immobilized on the sensorchip. An overview of the SPR binding screen is given in Figure 6. Overall, binding in the nM range was detected for 3 designs, the 2 positive controls and the 2 known binders (Table 2). Additionally, 6 designs had weak binding in the μM range, with a caveat that some of the binding events detected in these cases could be non-specific. Among the top 10 designs, only 3 had no detected binding, while the 3 stronger binders and 4 of the weak binders were present in this group. All 7 binders in the top-10 group belonged to the permutation (P) set. In the group consisting of the 8 remaining tested designs, ranks 11–18, there were only 2 weak binders while the rest of designs had no detected binding. These 2 weak binders were both from the mutation (M) set. Overall, data in Table 2 indicate a certain level of enrichment in binding that follows the predicted docking scores within the set of 18 tested variants, with the caveat that SPR data is insufficient to confirm the predicted binding modes. We also measured the thermal stabilities of the

TABLE 2 Top-ranked consensus designs.

Rank	Variable positions[a]	Set[b]	N_{sub}[c]	N_{pose}[d]	Q_{net}[e]	Score[f]	Z-Score[g]	K_D (nM)[h]
1	RMTKEKFFWEILWYDMVK	P	14	7	−6	−84.2	−3.5	weak
2	RQIVHRHWFDVIKYWRHL	M	18 (17)	3	−1	−77.8	−2.4	n.d.b
3	KFWFETMDKMKRYEWVIL	P	14	7	−6	−77.2	−2.3	weak
4	KFWMEMLTDWIYEVRKKF	P	10	3	−6	−75.9	−2.1	44
5	KFMREEFWWLIKKTDYMV	P	15	6	−6	−75.3	−2.0	n.d.b
6	KFWYNDFQMDFQMRNKKK	P	13	4	−4	−75.2	−2.0	150
7	VWWEEDFKIKMMKFYTLR	P	14	3	−6	−75.1	−2.0	111
8	KYRKNKFWFNDQFKDQMM	P	14	3	−4	−74.8	−1.9	n.d.b
9	RKMDQKFKMNDYWNFQFK	P	15	2	−4	−74.7	−1.9	weak
10	KIMWFKWDYKELMVETFR	P	15	3	−6	−74.7	−1.9	weak
11	RAVNRTVFVYWAYNFRVV	M	18 (16)	2	−4	−74.6	−1.9	weak
12	KFWMQRFMQYKDFKDKNN	P	15	2	−4	−74.6	−1.9	n.d.b
13	KLMEYDFMVWITKFERWK	P	14	7	−6	−74.6	−1.7	n.d.b
14	KYWYRTTWYHAIWNFYKQ	M	18 (16)	5	−3	−73.5	−1.7	weak
15	KYFEWVQRVMFKVVLMNR	M	18 (14)	2	−4	−73.3	−1.7	n.d.b
16	FKMWEMLFWRVIYEDKKT	P	14	5	−6	−72.8	−1.6	n.d.b
17	KFFRNNKMDYWKKMDFQQ	P	14	3	−4	−72.8	−1.6	n.d.b
18	KKSQTSYHHQQMLRTHRV	M	18 (17)	5	0	−72.7	−1.6	n.d.b
	KYDMNFMRDNFWKQQKFK	PC1	0	4	−4	−75.8	−2.1	240
	RFWMEDLTMKIVYWEKFK	PC2	0	7	−6	−73.5	−1.7	0.9

[a] Position IDs, in the same order: 45, 46, 48, 56, 57, 78, 79, 81, 89, 90, 111, 112, 114, 122, 123, 144, 145 and 147.

[b] P: permutation; M: mutation; PC: positive control.

[c] Number of substitutions at 18 variable positions from the corresponding known binder for the P-set designs or from the initial sequence of the common framework-based library for the M-set designs. Number of substitutions from the closest known binder is also shown in parenthesis for the M-set designs.

[d] Number of poses predicted to bind at the target epitope.

[e] Net charge.

[f] Consensus docking score obtained from an arithmetic average of the docked poses at target epitope.

[g] Calculated from scores over the set of 293 "consensus designs" (see Results section).

[h] Determined by SPR measurements (see Methods section); weak: K_D > 1 μM; n.d.b.: no detected binding.

designed DARPins and obtained very high thermostabilities, with melting temperature (T_m) values typically in the 80–100°C range (Supplementary Table S1), comparable with those measured here for the positive controls and known binders, as well as previously for other DARPins (Schilling et al., 2014a). This stability data provides some level of confidence that the sequence perturbations introduced in the designed variants were able to maintain the folded structure of the archetypical DARPin scaffold.

For the two known DARPin binders of BCL-W, 4k5a [B] and 4k5b [B], we obtained dissociation constants, K_D, of 26 nM and 3.5 nM, respectively, which are in line with their previously published K_D data of 10 nM and 0.64 nM (Schilling et al., 2014a). The two corresponding positive control DARPins, which import only the 18 variable positions of the common framework scaffold from these known binders, bound with K_D values of 245 nM and 0.9 nM. These values represent comparable affinities to their respective parental known binders, although it seems that the framework change from the known binders to the common framework sequence impacted detrimentally the 4k5a [B] interface and beneficially the 4k5b [B] interface.

For the 3 novel DARPin designs exhibiting good binding, we obtained dissociation constants, K_D, in the 40–150 nM range, which are well within the range bracketed by the two positive controls (0.9–245 nM). These were ranked 4, 6 and 7 among the top-18 consensus designs (Table 2), with the 4th ranked design exhibiting the better K_D of 44 nM, which is similar to the affinity of one of the known binders (26 nM). Low binding, with K_D above 1 μM, could also be detected for designs with ranks 1, 3, 9, 10, 11 and 14. Further details about the binders on their amino-acid substitutions, net charges, sets and substitutions are listed in Table 2.

A retrospective analysis of ranking by best scores instead of consensus scores versus experiment indicated that this approach could also be suitable (Supplementary Table S2). By this ranking of the 18 tested designs, the top 4 gave binding signals and among them the 2nd and 4th ranked are the best designs with K_D values of 44 nM and 111 nM. Also, best scoring was able to correctly rank the two positive controls among themselves, i.e., the stronger binder has a more negative score. However, best scoring always

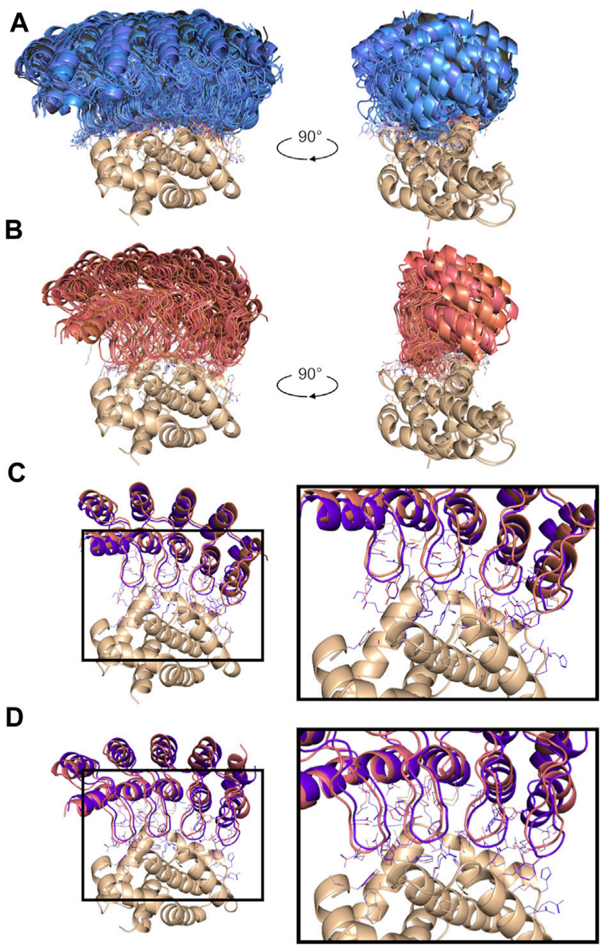


FIGURE 5
Non-cognate docking results for the top-ranked poses. **(A)** Overview of all poses at the target epitope for the top-18 consensus designs selected for testing. The novel designs are part of the permutation-based set (blue) and the mutation-based set (black). **(B)** Overview of all poses docked at the target epitope for the two positive controls (red). Comparisons of atomic details between the best-scored docked pose of a positive control and the crystal structure of the corresponding known binder sharing the same residues at the 18 variable positions are shown in panel **(C)** for the positive control PC1 and the known binder (4k5a [A]; in purple), and in panel **(D)** for the positive control PC2 and the known binder (4k5b [C]; in purple). All structure orientations are kept as in Figure 2C.

performed slightly worse than consensus ranking for the discrimination of binders against non-binders (Supplementary Figure S4).

While our computational strategy forced the designs to bind at a specified locus, our geometric criteria were loose enough to allow for some structural variability around the targeted epitope, that could lead to substantially different structural determinants required for binding. Despite the weak statistics due to the relatively low number of experimentally-validated designs, a close inspection of important structural determinants revealed that the non-binders bury more surface area on average than the validated strong binders (Supplementary Figure S5). Notably, a larger fraction in non-polar surface area on the BCL-W interface is predicted to be lost by the non-binders relative to binders (Supplementary Figure S5). This is an interesting finding to explore in future screening campaigns as docking algorithms are normally calibrated to attribute larger scores to burial of larger interfaces and would indirectly favor or enrich those designs achieving increased surface burial. For this set of binders, hydrophobic residues tend to be preferentially enriched only in the internal DARPin repeat 1 (Supplementary Figure S5).

4 Discussion

In this proof-of-concept study, we aimed at exploring if rigid-backbone docking can lead to meaningful biologics discovery. A first objective was to test, in a real-life scenario, the utility of our exhaustive protein-protein docking tool ProPOSE that incorporates side-chain flexibility (Hogues et al., 2018). ProPOSE performed very well in cognate-backbone docking, but returned a lower performance in unbound-backbone docking, thus hampering *de novo* antibody discovery efforts, mainly due to the hypervariable nature of the CDR-H3 loop. While work addressing the challenging problem of backbone sampling and scoring is highly relevant and remains to be pursued, here we explored the practical utility of ProPOSE in its current state by employing a more rigid scaffold, DARPin, which has already been used as an alternative scaffold in biologics discovery (Binz et al., 2003; Pluckthun, 2015). The overarching assumption is that ProPOSE can tolerate some minor level of backbone movements at the binding interface, but the extent of tolerated backbone movements has not been established yet.

From the technological perspective of rigid docking with unbound backbone conformation, employing four experimentally determined backbone conformations, each slightly different from bound backbone conformations, provided a test of the impact of backbone flexibility on biologics design. An initial measure of success was gleaned from so-called positive controls, in which 18 interfacial residues of known DARPin binders to a given target (BCL-W in this study) were transferred to a common DARPin framework sequence, assigned unbound backbone conformations, and cross-docked to the target. The predicted binding modes of these positive controls were similar to those of known binders, but docking scores were reduced almost in half relative to those obtained for the known binders in their bound backbone conformations. Yet, experimental testing of these positive controls showed retained binding affinities at comparable levels relative to the known binders, despite reduced scores. This established a new range of binding scores at a reduced magnitude which was adapted for cross-docking but remained predictive of true-positive binders. Consequently, novel DARPin designs cross-docked at that same target epitope were top-ranked and had scores within the re-established score level suitable for cross-docking. Upon their experimental testing, seven out of top-10 ranked designs demonstrated at least some level of binding to the target, with 3 of them exhibiting binding strengths similar to those of the positive controls as well as the previously known binders.

Despite this initial relative success, rigid-backbone docking remains challenging even for scaffolds with fairly rigid protein

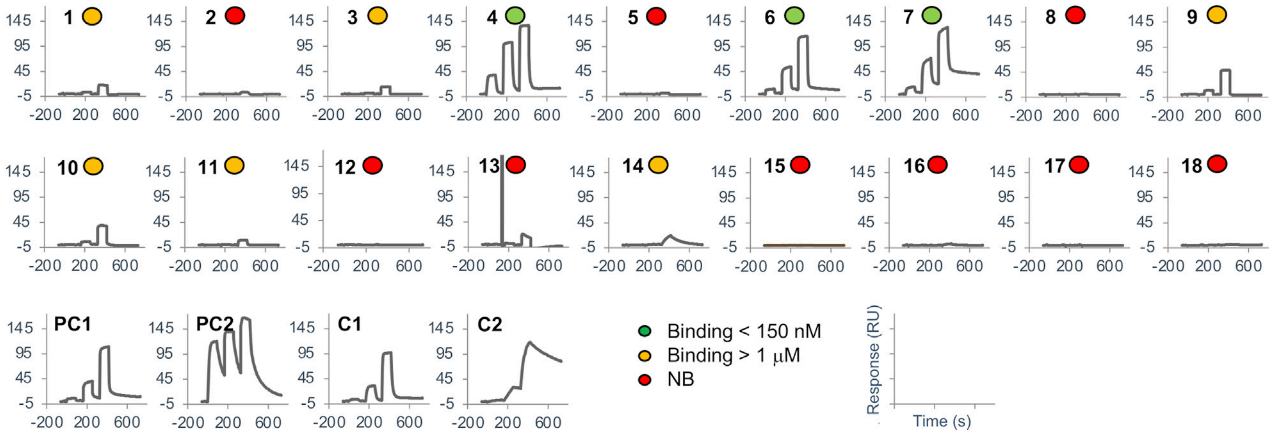


FIGURE 6
Surface plasmon resonance screening. SPR binding sensorgrams are shown for the 18 top-ranked designs, the positive controls and the known binders. Ranking of designs is based on the consensus score (see also Table 2). Sensorgrams are labeled according to 3 levels of binding affinity as shown in the legend.

backbone like DARPins. Several limitations of this approach and directions for possible improvements are noted below.

First, most novel binders belonged to the random permutation (P) set, which confines the library space with respect to certain global properties, for example, the net charge. These results thus point to the benefits of landing into the "right" regions of the library space after randomization at variable positions. While it was certainly harder for members from the random mutation (M) set to reach the top of the hit list, the finding of two weak binders belonging the M-set is extremely encouraging. In principle, real-life applications utilize mainly M-libraries. One way in this direction could be to enlarge the size of the docked diversified sub-library (only ~2,000 in this study). This could be feasible with access to large computing resources given the not overly prohibitive computational task involved in running ProPOSE. An alternative approach could be a focused expansion into P-subsets around initial M-set hits from a relatively sparse sub-library. This approach could set a preferred range for net charge, for example, and it would be especially beneficial as the number of randomized interfacial positions increases. Furthermore, the efficiency of the M-libraries at finding better hits could most likely be improved by applying structure-guided filters to search in more relevant regions of the sequence space. For instance, designs could be filtered based on their complementarity in charge or by their exposure of polar or non-polar surfaces at variable positions based on structural information of the selected binding epitope being targeted.

Secondly, while initial hits are often weak binders which are difficult to characterize, they should not be immediately discarded but rather treated as seeds for further optimization by affinity maturation, which can be done either experimentally (e.g., display methods) or computationally (e.g., ADAPT platform). This aspect has significant practical importance, given that by random sampling of the immense library space it is highly unlikely to obtain a very strong binder.

Thirdly, the unbound backbone conformations selected for cross-docking were from experimentally determined crystal structures. This is similar to the multiple protein structure approach used in small-molecule docking and virtual screening (Sheridan et al., 2008). Because the DARPin scaffold is not completely rigid and scoring functions used in docking are sensitive to atomic positions, including more than one backbone as templates in the cross-docking approach was felt to be beneficial. Carefully derived simulated structures obtained, for example, via backrub motions, molecular dynamics or Monte-Carlo simulations can be used as alternatives sources to experimentally-determined backbone conformations. The multiple template approach used here for docking was also extended to the stage of hit ranking, via consensus scoring. This seemed to provide a reasonable enrichment, although retrospectively we also found that the best-score approach might provide a similarly good, if not better ranking, among the small set of hits ranked by consensus scoring.

Despite some approximations in the underlying methodology adopted here, it is encouraging that cross-docking could identify binding sequences that differ substantially from known binders out of thousands of potential candidates. This relative success may be attributed to the foundational work underlying the methods used here to address the two intimately-related challenges of docking and scoring in computational drug discovery (Schneider et al., 2022). On one hand, for binding mode prediction, ProPOSE was used given its high accuracy in rigid-backbone docking when the bound-backbone conformation is provided. On the other hand, for ranking among different docked variants, ProPOSE employed a scoring function drawn from the solvated interaction energy (SIE) exhibiting high transferability from small-molecule to protein ligands (Purisima et al., 2023).

The data presented here support the notion that *de novo* biologics discovery *via* computational methods is a tractable problem that could complement the more traditional and matured wet-lab methods of library display screening and animal immunization. One main added benefit of the structure-based approach is directing the binding response towards desired target locations, e.g., functionally relevant, in a controlled manner. Further advances in several areas such as backbone sampling and depth of theoretical library screening, will be required for maturing *de novo* biologics discovery for routine applications in the not-so-distant future.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

FG: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing–original draft, Writing–review and editing. JB: Formal Analysis, Investigation, Methodology, Visualization, Writing–original draft, Writing–review and editing. YM: Formal Analysis, Investigation, Methodology, Writing–review and editing. AD: Formal Analysis, Investigation, Methodology, Writing–review and editing. HH. Conceptualization, Writing–review and editing. CC. Conceptualization, Writing–review and editing. EP: Conceptualization, Resources, Supervision, Writing–review and editing. MA: Formal Analysis, Methodology, Resources, Supervision, Writing–original draft, Writing–review and editing. TS: Conceptualization, Formal Analysis, Project administration, Resources, Supervision, Visualization, Writing–original draft, Writing–review and editing.

Funding

Some of the docking simulations were run on high-performance computing clusters from Digital Research Alliance of Canada under project number 4191 (TS).

Acknowledgments

We thank Emma Smith for performing DSC experiments.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2023.1253689/full#supplementary-material

References

Abanades, B., Wong, W. K., Boyles, F., Georges, G., Bujotzek, A., and Deane, C. M. (2022). ImmuneBuilder: deep-Learning models for predicting the structures of immune proteins, 2011.2004. bioRxiv, 514231. doi:10.1101/2022.11.04.514231

Adolf-Bryfogle, J., Kalyuzhniy, O., Kubitz, M., Weitzner, B. D., Hu, X., Adachi, Y., et al. (2018). RosettaAntibodyDesign (RAbD): a general framework for computational antibody design. *PLoS Comput. Biol.* 14, e1006112. doi:10.1371/journal.pcbi.1006112

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein Data Bank. *Nucl. Acids Res.* 28, 235–242. doi:10.1093/nar/28.1.235

Binz, H. K., Stumpp, M. T., Forrer, P., Amstutz, P., and Pluckthun, A. (2003). Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *J. Mol. Biol.* 332, 489–503. doi:10.1016/s0022-2836(03)00896-9

Chowdhury, R., Allan, M. F., and Maranas, C. D. (2018). OptMAVEn-2.0: de novo design of variable antibody regions against targeted antigen epitopes. *Antibodies (Basel)* 7, 23. doi:10.3390/antib7030023

Cohen, T., Halfon, M., and Schneidman-Duhovny, D. (2022). NanoNet: rapid and accurate end-to-end nanobody modeling by deep learning. *Front. Immunol.* 13, 958584. doi:10.3389/fimmu.2022.958584

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., et al. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* 117, 5179–5197. doi:10.1021/ja00124a002

DeFrancesco, L. (2019). Drug pipeline 1Q19. *Nat. Biotechnol.* 37, 579–580. doi:10.1038/s41587-019-0146-7

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., et al. (2022). Protein complex prediction with AlphaFold-Multimer, 2010.2004.463034. bioRxiv. doi:10.1101/2021.10.04.463034

Fischman, S., and Ofran, Y. (2018). Computational design of antibodies. *Curr. Opin. Struct. Biol.* 51, 156–162. doi:10.1016/j.sbi.2018.04.007

Forrer, P., and Jaussi, R. (1998). High-level expression of soluble heterologous proteins in the cytoplasm of *Escherichia coli* by fusion to the bacteriophage lambda head protein D. *Gene* 224, 45–52. doi:10.1016/s0378-1119(98)00538-1

Gao, M., Nakajima An, D., Parks, J. M., and Skolnick, J. (2022). AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat. Commun.* 13, 1744. doi:10.1038/s41467-022-29394-2

Hogues, H., Gaudreault, F., Corbeil, C. R., Deprez, C., Sulea, T., and Purisima, E. O. (2018). ProPOSE: direct exhaustive protein-protein docking with side chain flexibility. *J. Chem. Theory Comput.* 14, 4938–4947. doi:10.1021/acs.jctc.8b00225

Hogues, H., Sulea, T., and Purisima, E. O. (2014). Exhaustive docking and solvated interaction energy scoring: lessons learned from the SAMPL4 challenge. *J. Comput. Aided Mol. Des.* 28, 417–427. doi:10.1007/s10822-014-9715-5

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* 65, 712–725. doi:10.1002/prot.21123

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2

Kaplon, H., Crescioli, S., Chenoweth, A., Visweswaraiah, J., and Reichert, J. M. (2023). Antibodies to watch in 2023. *MAbs* 15, 2153410. doi:10.1080/19420862.2022.2153410

Kramer, M. A., Wetzel, S. K., Pluckthun, A., Mittl, P. R., and Grutter, M. G. (2010). Structural determinants for improved stability of designed ankyrin repeat proteins with a redesigned C-capping module. *J. Mol. Biol.* 404, 381–391. doi:10.1016/j.jmb.2010.09.023

Krivov, G. G., Shapovalov, M. V., and Dunbrack, R. L., Jr. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* 77, 778–795. doi:10.1002/prot.22488

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi:10.1093/bioinformatics/btm404

Lensink, M. F., Velankar, S., and Wodak, S. J. (2017). Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition. *Proteins* 85, 359–377. doi:10.1002/prot.25215

Lu, R. M., Hwang, Y. C., Liu, I. J., Lee, C. C., Tsai, H. Z., Li, H. J., et al. (2020). Development of therapeutic antibodies for the treatment of diseases. *J. Biomed. Sci.* 27, 1. doi:10.1186/s12929-019-0592-z

Pecqueur, L., Duellberg, C., Dreier, B., Jiang, Q., Wang, C., Pluckthun, A., et al. (2012). A designed ankyrin repeat protein selected to bind to tubulin caps the microtubule plus end. *Proc. Natl. Acad. Sci. U. S. A.* 109, 12011–12016. doi:10.1073/pnas.1204129109

Pluckthun, A. (2015). Designed ankyrin repeat proteins (DARPins): binding proteins for research, diagnostics, and therapy. *Annu. Rev. Pharmacol. Toxicol.* 55, 489–511. doi:10.1146/annurev-pharmtox-010611-134654

Purisima, E. O., Corbeil, C. R., Gaudreault, F., Wei, W., Deprez, C., and Sulea, T. (2023). Solvated interaction energy: from small-molecule to antibody drug design. *Front. Mol. Biosci.* 10, 1210576. doi:10.3389/fmolb.2023.1210576

R Development Core Team (2011). *R: a language and environment for statistical computing*. Vienna, Austria: The R Foundation for Statistical Computing.

Radom, F., Paci, E., and Pluckthun, A. (2019). Computational modeling of designed ankyrin repeat protein complexes with their targets. *J. Mol. Biol.* 431, 2852–2868. doi:10.1016/j.jmb.2019.05.005

Rothenberger, S., Hurdiss, D. L., Walser, M., Malvezzi, F., Mayor, J., Ryter, S., et al. (2022). The trispecific DARPin ensovibep inhibits diverse SARS-CoV-2 variants. *Nat. Biotechnol.* 40, 1845–1854. doi:10.1038/s41587-022-01382-3

Ruffolo, J. A., Chu, L.-S., Mahajan, S. P., and Gray, J. J. (2022). Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies, 2004.2020.488972. bioRxiv. doi:10.1101/2022.04.20.488972

Schilling, J., Jost, C., Ilie, I. M., Schnabl, J., Buechi, O., Eapen, R. S., et al. (2022). Thermostable designed ankyrin repeat proteins (DARPins) as building blocks for innovative drugs. *J. Biol. Chem.* 298, 101403. doi:10.1016/j.jbc.2021.101403

Schilling, J., Schoppe, J., and Pluckthun, A. (2014a). From DARPins to LoopDARPins: novel LoopDARPin design allows the selection of low picomolar binders in a single round of ribosome display. *J. Mol. Biol.* 426, 691–721. doi:10.1016/j.jmb.2013.10.026

Schilling, J., Schoppe, J., Sauer, E., and Pluckthun, A. (2014b). Co-crystallization with conformation-specific designed ankyrin repeat proteins explains the conformational flexibility of BCL-W. *J. Mol. Biol.* 426, 2346–2362. doi:10.1016/j.jmb.2014.04.010

Schneider, C., Buchanan, A., Taddese, B., and Deane, C. M. (2022). Dlab: deep learning methods for structure-based virtual screening of antibodies. *Bioinformatics* 38, 377–383. doi:10.1093/bioinformatics/btab660

Schrag, J. D., Picard, M. E., Gaudreault, F., Gagnon, L. P., Baardsnes, J., Manenda, M. S., et al. (2019). Binding symmetry and surface flexibility mediate antibody self-association. *MAbs* 11, 1300–1318. doi:10.1080/19420862.2019.1632114

Sheridan, R. P., McGaughey, G. B., and Cornell, W. D. (2008). Multiple protein structures and multiple ligands: effects on the apparent goodness of virtual screening results. *J. Comput. Aided Mol. Des.* 22, 257–265. doi:10.1007/s10822-008-9168-9

Strittmatter, T., Wang, Y., Bertschi, A., Scheller, L., Freitag, P. C., Ray, P. G., et al. (2022). Programmable DARPin-based receptors for the detection of thrombotic markers. *Nat. Chem. Biol.* 18, 1125–1134. doi:10.1038/s41589-022-01095-3

Warszawski, S., Borenstein Katz, A., Lipsh, R., Khmelnitsky, L., Ben Nissan, G., Javitt, G., et al. (2019). Optimizing antibody affinity and stability by the automated design of the variable light-heavy chain interfaces. *PLoS Comput. Biol.* 15, e1007207. doi:10.1371/journal.pcbi.1007207

Wood, T. K. (2021). Concerns with computational protein engineering programmes IPRO and OptMAVEn and metabolic pathway engineering programme optStoic. *Open Biol.* 11, 200173. doi:10.1098/rsob.200173

Yin, R., Feng, B. Y., Varshney, A., and Pierce, B. G. (2022). Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. *Protein Sci.* 31, e4379. doi:10.1002/pro.4379

Youn, S. J., Kwon, N. Y., Lee, J. H., Kim, J. H., Choi, J., Lee, H., et al. (2017). Construction of novel repeat proteins with rigid and predictable structures using a shared helix method. *Sci. Rep.* 7, 2595. doi:10.1038/s41598-017-02803-z

Yu, G. (2020). Using ggtree to visualize data on tree-like structures. *Curr. Protoc. Bioinforma.* 69, e96. doi:10.1002/cpbi.96

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 18 September 2023
DOI 10.3389/fmolb.2023.1237621

Check for updates

OPEN ACCESS

EDITED BY
Traian Sulea,
National Research Council Canada
(NRC), Canada

REVIEWED BY
Victor Greiff,
University of Oslo, Norway
Christopher Corbeil,
National Research Council Canada
(NRC), Canada

*CORRESPONDENCE
Charlotte M. Deane,
✉ deane@stats.ox.ac.uk

RECEIVED 09 June 2023
ACCEPTED 28 August 2023
PUBLISHED 18 September 2023

CITATION
Spoendlin FC, Abanades B, Raybould MIJ, Wong WK, Georges G and Deane CM (2023), Improved computational epitope profiling using structural models identifies a broader diversity of antibodies that bind to the same epitope.
Front. Mol. Biosci. 10:1237621.
doi: 10.3389/fmolb.2023.1237621

Improved computational epitope profiling using structural models identifies a broader diversity of antibodies that bind to the same epitope

Fabian C. Spoendlin[1], Brennan Abanades[1], Matthew I. J. Raybould[1], Wing Ki Wong[2], Guy Georges[2] and Charlotte M. Deane[1]*

[1]Oxford Protein Informatics Group, Department of Statistics, University of Oxford, Oxford, United Kingdom, [2]Large Molecule Research, Roche Pharma Research and Early Development, Roche Innovation Center Munich, Penzberg, Germany

The function of an antibody is intrinsically linked to the epitope it engages. Clonal clustering methods, based on sequence identity, are commonly used to group antibodies that will bind to the same epitope. However, such methods neglect the fact that antibodies with highly diverse sequences can exhibit similar binding site geometries and engage common epitopes. In a previous study, we described SPACE1, a method that structurally clustered antibodies in order to predict their epitopes. This methodology was limited by the inaccuracies and incomplete coverage of template-based modeling. In addition, it was only benchmarked at the level of domain-consistency on one virus class. Here, we present SPACE2, which uses the latest machine learning-based structure prediction technology combined with a novel clustering protocol, and benchmark it on binding data that have epitope-level resolution. On six diverse sets of antigen-specific antibodies, we demonstrate that SPACE2 accurately clusters antibodies that engage common epitopes and achieves far higher dataset coverage than clonal clustering and SPACE1. Furthermore, we show that the functionally consistent structural clusters identified by SPACE2 are even more diverse in sequence, genetic lineage, and species origin than those found by SPACE1. These results reiterate that structural data improve our ability to identify antibodies that bind to the same epitope, adding information to sequence-based methods, especially in datasets of antibodies from diverse sources. SPACE2 is openly available on GitHub (https://github.com/oxpig/SPACE2).

KEYWORDS
antibody, epitope profiling, structure prediction, clustering, epitope, clonotyping

1 Introduction

Antibodies are important components of the adaptive immune system. An antibody recognizes foreign particles by binding to a specific site—the epitope—on their surface. As antibody function is tightly linked to the epitope it engages, studying epitopes is essential to understand immunology. For example, determining epitope specificities of antibody repertoires can increase our understanding of the immune response to disease (Tsioris et al., 2015; Bashford-Rogers et al., 2019) or differences of the immune system between individuals (Briney et al., 2019). Furthermore, epitope profiling can be applied in antibody drug discovery to identify both new binders to a desired target (Reddy et al., 2010; Zhu et al., 2013; Tsioris et al., 2015) and binders with improved affinity (Hsiao et al., 2019).

Epitopes can be determined at high resolution by solving the structure of an antibody in complex with its antigen. However, structure determination methods are too resource-intensive to be used to explore large datasets (Nilvebrant and Rockberg, 2018). Experimental epitope binning methods, such as competition assays (Abdiche et al., 2009), scale better; however, it remains difficult to analyze very large datasets as costs grow at $O(n^2)$ with the number of antibodies (n) to be evaluated. Competition assays also only offer low resolution as they struggle to distinguish between antibodies that bind to the same site and those that bind to distinct sites but overlap sterically.

Prior computational clustering of antibodies into functional groups that engage the same epitope can reduce the number of experiments that needs to be run or even remove the need for experimental epitope determination entirely. Most computational epitope profiling methods group antibodies based on sequence similarity. Clonotyping, the most widely used method, attempts to link antibodies that originate from the same progenitor B-cell (Greiff et al., 2015; López-Santibáñez-Jácome et al., 2019). The exact definition of a clonotype varies across the literature. Commonly, antibodies that originate from the same heavy chain V and J genes, match in CDRH3 length, and exceed a threshold CDRH3 sequence identity are considered a clonotype. Threshold values between 80% and 100% have been reported. To introduce additional leniency, the requirement for matching J genes can be neglected (Greiff et al., 2015). Clonal clustering is usually highly accurate, and antibodies within a cluster tend to engage the same epitope.

Clonotyping was originally intended to trace lineages of antibodies within an individual. Its use in functional clustering thus makes the assumption that antibodies against a given epitope must originate from progenitor B cells with shared genetic origins. However, antibodies from different lineages and with highly dissimilar sequences can adopt a similar binding site geometry and engage the same epitope (Scheid et al., 2011; Joyce et al., 2016; Rijal et al., 2019; Robinson et al., 2021; Wong et al., 2021). The ability to determine functional convergence is especially important when comparing the immune response of individuals, as different individuals exhibit personalized immunoglobulin gene usages (Briney et al., 2019). As clonotyping is not able to link antibodies from distinct genetic lineages, it loses power when analyzing antibodies originating from different sources.

Alternative methods have been developed to try and identify functionally equivalent antibodies that are not similar in sequence. Clustering antibodies by sequence similarity across predicted paratope residues can link antibodies from different clonotypes (Richardson et al., 2021). However, methods that consider structural similarity to cluster antibodies are even better suited to detect less related sequences with functional convergence because the binding site structure provides more direct evidence of antibody function than its sequence. Several methods are available that attempt to functionally link antibodies based on a representation containing structural information in addition to physicochemical properties of paratope residues (Ripoll et al., 2021; Wong et al., 2021).

In a previous study, we described the SPACE1 method (Robinson et al., 2021), which clusters antibodies based on structural similarity of homology models. The algorithm accurately clusters antibodies that bind to the same epitope and is able to functionally link antibodies with diverse sequences. However, SPACE1 is limited by the coverage of homology modeling (in the original study, only 73% of the data could be modeled to a usable standard) and its inaccuracies. The method was also only benchmarked at the level of domain consistency on one virus class. Recent progress in machine learning-based antibody structure prediction has led to more accurate structural models than those obtained with homology-based approaches, especially in cases where no template with high-sequence similarity is available (Ruffolo et al., 2020; Baek et al., 2021; Jumper et al., 2021; Abanades et al., 2022a; Ruffolo et al., 2022a; Abanades et al., 2022b; Ruffolo et al., 2022b; Lin et al., 2022). Higher accuracy and higher confidence in structural models also allow increased coverage and have the potential to improve structure-based epitope profiling.

Here, we present the Structural Profiling of Antibodies to Cluster by Epitope 2 (SPACE2) algorithm. SPACE2 builds on recent progress in machine learning-based antibody structure prediction and uses a novel clustering protocol systematically optimized and extensively benchmarked on epitope-resolution binding data. We show that SPACE2 outperforms SPACE1 by improving data coverage and identifying clusters even more diverse in sequences, genetic lineages, and species origin. These results underline that structural data, which can now be rapidly and easily generated through structure prediction tools, contain orthogonal functional information to sequence and should be considered when investigating antibody function.

2 Materials and methods

2.1 Datasets

Six datasets of antigen-specific antibodies were used to analyze SPACE2 clustering performance.

The training set on which the clustering algorithm, thresholds, and antibody region were set consisted of 3,051 antibodies against the SARS-CoV-2 receptor-binding domain (RBD). Antibodies were annotated with groups of overlapping epitopes originating from mutation escape profiling (Cao et al., 2023). We refer to this dataset as the Cao et al. (2023) training set throughout the paper.

CoV-AbDab (Raybould et al., 2021a), a dataset of anti-lysozyme antibodies, a non-public dataset of antibodies against Ebola viruses (EVs), and two non-public dataset of antibodies against non-viral targets (NVA1 and NVA2) were used as additional datasets to evaluate SPACE2. CoV-AbDab is a database of antibodies against coronavirus antigens, such as those from SARS-CoV-2, SARS-CoV-1, and MERS-CoV. A version of CoV-AbDab timestamped 3 October 2022 was used containing 10,719 antibodies with sequence data. As CoV-AbDab is a collection of antibodies reported in the literature, it contains the Cao et al. (2023) training set. When using CoV-AbDab as a test set [denoted as

CoV-AbDab (test)], the training set was removed, and only the remaining 7,685 antibodies were included. Epitope data in CoV-AbDab are reported as in the original publications and range from the antigen to domain level.

A dataset of anti-lysozyme antibodies was created from all 53 lysozyme-specific antibodies in the structural antibody database (SAbDab) (Dunbar et al., 2014; Schneider et al., 2022), for which the antibody–antigen complex structure has been solved. Antibodies were grouped by their epitope using the Ab-ligity method (Wong et al., 2021) and annotated as binding to the same epitope if their Ab-ligity score was greater than a threshold of 0.1 (as in the original paper). Similarity of epitopes within an epitope group was confirmed by visual inspection.

The EV set contains 126 antibodies with epitope data ranging from antigen to domain level. The NVA1 set contains 31 antibodies with epitope data from competition assays. NVA2 contains 33 antibodies with epitope data from mutation escape profiling.

2.2 SPACE1

The original SPACE1 method clusters antibodies by the structural similarity of homology models. The algorithm was run as detailed in Robinson et al. (2021).

Homology models were produced using ABodyBuilder (Leem et al., 2016). ABodyBuilder uses structures from a database to build its models. In this study, we used quality-filtered SAbDab (Dunbar et al., 2014; Schneider et al., 2022) entries timestamped before 6 July 2022. Quality filtering restricts structures to those solved by X-ray crystallography and excludes structures with a resolution of >2.5 Å and structures containing residues with a B-factor >80. In a standard ABodyBuilder run, the method first attempts to model CDR loops with a template database search method (Choi and Deane, 2010). If no suitable template is found for CDRs, hybrid homology/*ab initio* modeling is performed (Leem et al., 2016). Only models for which homology templates for all six CDR loops were found are used for clustering in the SPACE1 method to keep the models as accurate as possible.

The remaining homology models are clustered by structural similarity of CDRs. The models are split into groups of antibodies with identical CDR lengths. Antibodies in each group are then clustered using a greedy clustering algorithm. The first antibody in the group is selected as the cluster center, and all antibodies with a CDR C_α RMSD smaller than a specified threshold after alignment of framework residues are added to the cluster. After all antibodies have been compared against the first cluster center, the algorithm selects the next unclustered antibody as a new cluster center, and cluster members are chosen as in the previous step. In addition to the RMSD threshold of 0.75 Å suggested by Robinson et al. (2021), we also assessed the performance at a 1.25 Å threshold.

2.3 SPACE2

Our novel SPACE2 algorithm clusters antibodies by the similarity of models obtained from an ML-based structure prediction tool. The method functions in four main steps. Initially, a structural model of the antibody Fv is produced using ABodyBuilder2 (Abanades et al., 2022b). ABodyBuilder2 is a deep-learning-based tool for antibody structure prediction and was trained on SAbDab structures timestamped up to 31 July 2021. The models are then split into groups of identical CDR lengths. The models in each group are then structurally aligned on the Cα of residues in framework regions, and a pairwise distance matrix is computed of the Cα RMSDs of CDR loop residues. The antibodies are then clustered based on these distances.

2.3.1 Clustering algorithms

Eight different clustering algorithms were explored (agglomerative clustering, affinity propagation, DBSCAN, OPTICS-xi, OPTICS-DBSCAN, K-means, Butina clustering, and greedy clustering). Agglomerative clustering (Murtagh and Contreras, 2012), affinity propagation (Frey and Dueck, 2007), DBSCAN (Schubert et al., 2017), OPTICS-xi, OPTICS-DBSCAN (Ankerst et al., 1999), and K-means (MacQueen, 1967) were implemented using the scikit-learn (Pedregosa et al., 2011). Butina clustering (Butina, 1999) was implemented using the RDKit (Landrum, 2006). A greedy clustering algorithm, grouping antibodies as the algorithm described in Section 2.2, was implemented.

Parameters and evaluated ranges for each algorithm are shown in Table 1. The K-means algorithm requires an additional parameter (K) that corresponds to the predetermined number of clusters. K was set to the number of clusters obtained from agglomerative clustering using the best performing parameters.

2.3.2 SPACE2-HC

A variation of the SPACE2 algorithm was implemented that clusters antibodies based on the structural similarity of heavy chains only (SPACE2-HC). The light chains were included for the modeling step, as ABodyBuilder2 (Abanades et al., 2022b) requires sequences of both chains as an input. After this step, light chains were ignored. Antibodies were grouped based on the length of the heavy chain CDRs, aligned on heavy chain framework regions, and the C_α RMSD of CDRs H1-3 calculated. Agglomerative clustering with a "complete" linkage criterion was used as the clustering algorithm of SPACE2-HC.

2.3.3 SPACE2-Paratope

A second variation of the SPACE2 algorithm was implemented that clusters antibodies based on the structural similarity of CDR loops, which are predicted to form part of the paratope (SPACE2-Paratope). Structural models were produced using ABodyBuilder2 (Abanades et al., 2022b). The Paragraph method (Chinery et al., 2023) with a classifier cut-off of 0.734, as suggested in the original paper, was then used to predict residues that are part of the paratope based on the models. All CDRs containing at least one paratope residue were then labeled as paratope CDRs. Antibodies were divided into groups containing the same set of paratope CDRs. Antibodies in each group were further grouped based on the length of the paratope CDRs, aligned on heavy chain framework regions, and clustered based on the C_α RMSD of paratope CDRs. Agglomerative clustering with a "complete" linkage criterion was used as the clustering algorithm of SPACE2-Paratope.

TABLE 1 Clustering algorithms and parameter ranges/values evaluated during optimization.

Algorithm	Parameter	Range/values	Optimal value
Greedy clustering	RMSD threshold	0.5–10 Å	1.25 Å
Agglomerative clustering	RMSD threshold	0.5–10 Å	1.25 Å
	Linkage criterion	Complete, average, and single	Complete
Affinity propagation	Preferences	−5 to 4	Median (RMSD matrix)
DBSCAN	RMSD threshold	0.5–5 Å	1 Å
	Minimum samples	2 and 5	2
OPTICS-xi	RMSD threshold	0.5–5 Å	1.5 Å
	Minimum samples	2	2
	xi	0.005–0.5	≤0.01
OPTICS-DBSCAN	RMSD threshold	0.5–2 Å	1 Å
	Minimum samples	2	2
K-means	Initialization	Random, K-means++	K-means++
Butina clustering	RMSD threshold	0.5–5 Å	1 Å
	Reordering	True and false	False

2.4 Numbering scheme and region definitions

IMGT numbering (Lefranc et al., 2003) and North CDR definitions (North et al., 2011) are used throughout.

2.5 Analysis of structural clusters

2.5.1 Domain/epitope-consistent clusters

Antibody clusters generated for the Cao et al. (2023) training set, NVA1 set, NVA2 set, and anti-lysozyme set were classified as “epitope-consistent” or “epitope-inconsistent.” “Epitope-consistent” clusters of the Cao et al. (2023) training, NVA1, and NVA2 sets only contain antibodies that bind to the same epitope group as determined by experimental epitope binning. “Epitope-consistent” clusters of the lysozyme dataset only contain antibodies that bind to the same residue-level epitope determined using crystal structures.

Owing to the lower resolution of epitopes reported in the EV set and CoV-AbDab, clusters of these datasets were classified as “domain-consistent” and “domain-inconsistent.” EV set clusters were labeled as “domain-consistent” if they only contain antibodies that engage the same antigen domain. CoV-AbDab clusters that satisfy the following rules, consistent with previous studies (Robinson et al., 2021), were determined to be “domain-consistent”:

1. Clusters that only contain antibodies that bind to the same antigen and domain.
2. Clusters that contain antibodies binding to the same domain and others that bind to the same antigen without domain-level resolution.
3. Clusters that only contain antibodies that bind to the same antigen but do not have domain-level resolution of the epitope data.
4. Clusters with internally consistent epitope data, e.g., a cluster of antibodies labeled to bind to the spike (S) protein N-terminal domain (NTD) and others labeled as S non-RBD binders, as S NTD is a subdomain of S non-RBD.

2.5.2 Performance metrics

Throughout this study, we used seven metrics to analyze functional clustering. Two accuracy metrics, the fraction of epitope-consistent clusters (number of epitope-consistent multiple-occupancy clusters/number of multiple-occupancy clusters) and the fraction of clustered antibodies in epitope-consistent clusters (number of antibodies in epitope-consistent multiple-occupancy clusters/number of antibodies in multiple-occupancy clusters), were used. Two coverage metrics, the number of multiple-occupancy clusters and the number of antibodies in multiple-occupancy clusters, were used. In order to examine accuracy and coverage with one measure, we also calculated the number of antibodies in consistent multiple-occupancy clusters. Two further metrics were used to assess the diversity of antibodies within clusters: the fraction of functionally consistent clusters containing antibodies from more than one clonotype and the mean CDRH3 sequence identity within functionally consistent clusters.

2.5.3 Random baseline

Random clustering was performed as a baseline. The distribution of cluster sizes obtained from the evaluated clustering algorithm with specific parameters was recorded. Clusters with an identical size distribution were then sampled randomly from the dataset, and performance metrics were calculated. Sampling was repeated 100 times, and the metrics averaged.

2.5.4 Clonotyping

Clonotyping was performed using an in-house script. Lenient VH-clonotyping and Fv-clonotyping threshold conditions based on community standards were used (Greiff et al., 2015; López-Santibáñez-Jácome et al., 2019). A VH-clonotype was defined as a match in *IGHV* genes, length-matched CDRH3, and >80% CDRH3 sequence similarity. Fv-clonotypes were defined as a match in VH-clonotype, matching of IG[K/L]V genes, length-matched CDRL3, and >80% sequence identity of CDRL3.

3 Results

The original SPACE1 algorithm was developed to cluster antibodies by structural similarity with the aim of better identifying functional convergence. It grouped antibodies based on the structural similarity of homology models. This method was not systematically optimized and only benchmarked on a single dataset of low-resolution epitope data. Newly available ML-based structure prediction tools produce more accurate models and have better coverage than homology modeling. Here, we introduce SPACE2, which uses a state-of-the-art antibody structure prediction method and a novel clustering protocol that has been extensively optimized and then benchmarked on several datasets of high-resolution epitope data.

SPACE2 clusters antibodies in four main steps. Initially, structural models are produced using ABodyBuilder2 (Abanades et al., 2022b). The models are then separated into groups of antibodies with identical lengths of the six CDRs, followed by the computation of a pairwise distance matrix of CDR C_α RMSDs. In the final step, a clustering algorithm divides the antibodies into structural clusters. Although some loops of different lengths can adopt similar structures, we have decided to restrict structural comparison to antibodies with identical CDR lengths for the SPACE2 method as evidence suggests length-independent structural similarities are infrequent (Nowak et al., 2016; Wong et al., 2019). Restricting structural comparison to CDRs of the same length also allows for more rapid computation as RMSDs do not have to be calculated between all pairs of antibodies within the set. Optimization of the clustering protocol was performed on a training set of 3,051 antibodies against the SARS-CoV-2 receptor-binding domain (RBD) (Cao et al., 2023).

3.1 Evaluating an optimal clustering algorithm

We tested eight widely used clustering algorithms, greedy clustering, affinity propagation (Frey and Dueck, 2007), Butina clustering (Butina, 1999), DBSCAN (Schubert et al., 2017), OPTICS-DBSCAN, OPTICS-xi (Ankerst et al., 1999), agglomerative clustering (Murtagh and Contreras, 2012), and K-means (MacQueen, 1967), for their ability to correctly group functionally consistent antibodies in the Cao et al. (2023) training set. To assess the methods, we used the number of antibodies in epitope-consistent multiple-occupancy clusters as our target performance metric as it provides a trade-off between clustering accuracy and dataset coverage. High accuracy or coverage metrics individually do not necessarily indicate a good epitope profiling method (Figure 1). High accuracy can be achieved by dividing the dataset into very small clusters that are highly likely to be epitope/domain-consistent but do not cover the full diversity of antibodies able to engage a given epitope. Maximal coverage can be achieved by putting all antibodies into a single cluster, which does not provide any useful epitope information.

A parameter scan was carried out to find the optimal setting for each clustering method. The ranges and optimal values of the evaluated parameters are shown in Table 1. As expected, lenient parameters increased dataset coverage, whereas stringent parameters improve accuracy, and the trade-off was maximized at intermediate values. The best performing algorithms, as defined by maximizing the number of antibodies in epitope-consistent multiple-occupancy clusters, were agglomerative clustering (optimal parameters: linkage criterion = complete; RMSD distance threshold = 1.25 Å), OPTICS-xi (optimal parameters: xi ≤ 0.01; RMSD distance threshold = 2 Å), and K-Means (optimal parameters: initialization method = K-means++), where K was set to the number of clusters obtained by agglomerative clustering with optimal parameters (Figure 2). As K-means does not lead to an improvement over agglomerative clustering, it was disregarded for further analysis. A visualization of the clustering obtained by the eight algorithms is shown in Supplementary Figure S1.

Agglomerative clustering and OPTICS-xi clustering were compared in more detail (Supplementary Table S1). Both algorithms achieve a similar clustering accuracy and dataset coverage. Agglomerative clustering produces larger clusters with a mean cluster size of 3.0 members and a maximum of 28 than OPTICS-xi clusters with mean 2.7 and maximum 11. When epitope-consistent clusters are larger, it suggests that they are better capturing the full diversity of the antibodies able to engage a given epitope. Therefore, agglomerative clustering was selected for use in SPACE2.

3.2 Examining the behavior of agglomerative clustering across different structural similarity thresholds

The RMSD threshold parameter of agglomerative clustering determines the leniency of the algorithm as it sets the maximum distance between any two antibodies in a cluster. Small thresholds restrict clustering to highly similar structures, whereas larger values allow clusters to contain more dissimilar antibodies. We evaluated agglomerative clustering for threshold values between 0.5 and 5 Å to assess how clustering results are affected.

Four metrics were monitored to assess the accuracy of clustering and dataset coverage. The fraction of epitope-consistent clusters (number of epitope-consistent multiple-occupancy clusters/number of multiple-occupancy clusters) and the fraction of clustered antibodies in epitope-consistent clusters (number of antibodies in epitope-consistent multiple-occupancy clusters/number of antibodies in multiple-occupancy clusters) were used as an accuracy measure. The number of multiple-occupancy clusters and the number of antibodies in multiple-occupancy clusters provide information on dataset coverage.

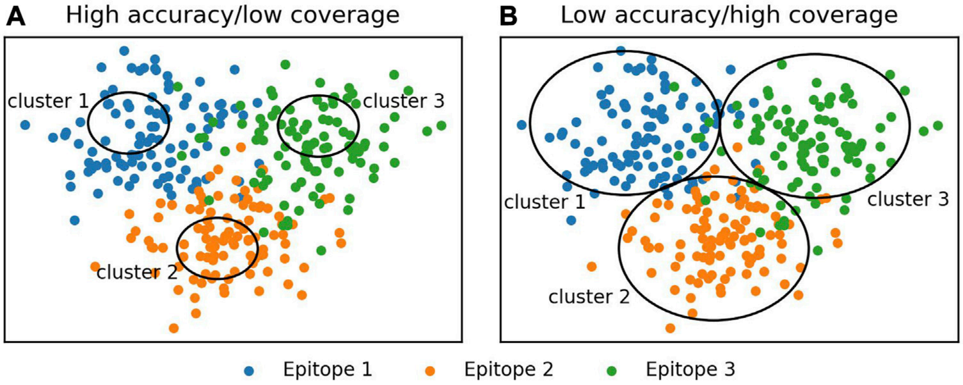


FIGURE 1
Illustration of the evaluation of clustering algorithms. Accuracy and coverage metrics were used to analyze clustering algorithms. Individually, these metrics do not necessarily indicate a good clustering algorithm, instead a trade-off between accuracy and coverage should be monitored. **(A)** High accuracy is achieved by making small clusters. These are likely to be epitope-specific; however, most antibodies are not contained in a cluster. **(B)** High coverage is achieved by the formation of large clusters. These contain most of the antibodies in the dataset but do not tend to be epitope-specific.

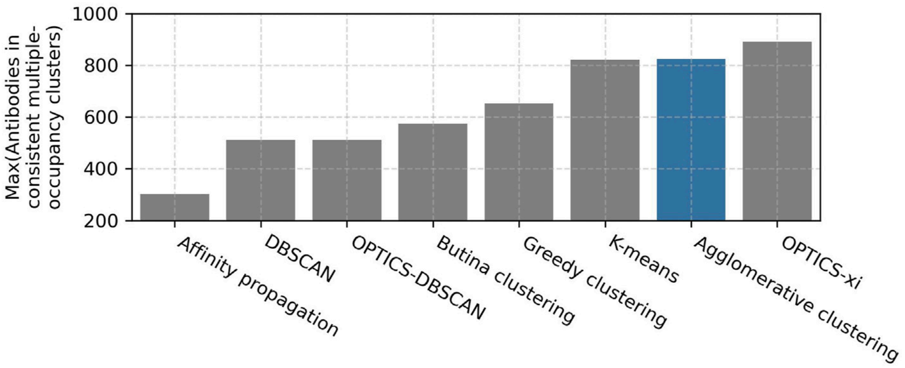


FIGURE 2
Examination of clustering algorithms. Parameter scans of eight clustering algorithms were performed using the Cao et al. (2023) training set. The performance of clustering was measured in terms of the number of antibodies in epitope-consistent multiple-occupancy clusters (*y*-axis). The maximum value of this metric achieved by a specific algorithm across all evaluated parameters when clustering the Cao et al. (2023) training set is shown. The ranges and optimal values of the evaluated parameter are shown in Table 1. The agglomerative clustering algorithm selected for SPACE2 is highlighted in blue.

Clustering accuracy and data coverage show a strong dependence on the RMSD threshold (Figure 3). At thresholds ≤0.75 Å, the clustering is highly accurate. More than 80% of clusters are epitope-consistent, and approximately 80% of clustered antibodies are in epitope-consistent clusters. Increasing the threshold leads to a rapid drop in accuracy but improves dataset coverage. The number of antibodies in multiple-occupancy clusters starts to plateau at approximately 3 Å. The large changes in accuracy and data coverage as a function of threshold suggest that the threshold should be adjusted depending on the aim of the epitope profiling task. Optimal clustering is achieved at a value of 1.25 Å, as defined by maximizing the number of antibodies in epitope-consistent multiple-occupancy clusters. However, the threshold can be set to any value between 0.75 and 3 Å to increase accuracy or coverage.

In all the analysis to this point, we have reported only on clusters that are 100% epitope-consistent (i.e., only contain antibodies against the same epitope). To measure the inconsistency of the remaining clusters, we analyzed the fraction of clusters in which at least 70% of the antibodies engage the same epitope. An additional 12% of clusters are >70% epitope-consistent, and these clusters contain an extra 26% of all antibodies contained in multiple-occupancy clusters (Supplementary Figure S2). This result indicates that even those clusters our standard performance metrics are marking as incorrect may contain large amounts of useful information.

3.3 Evaluating the optimal region for clustering

The SPACE2 method calculates structural similarity of antibodies across all six CDRs. However, not all CDRs are equally involved in binding and we expect the structure of some

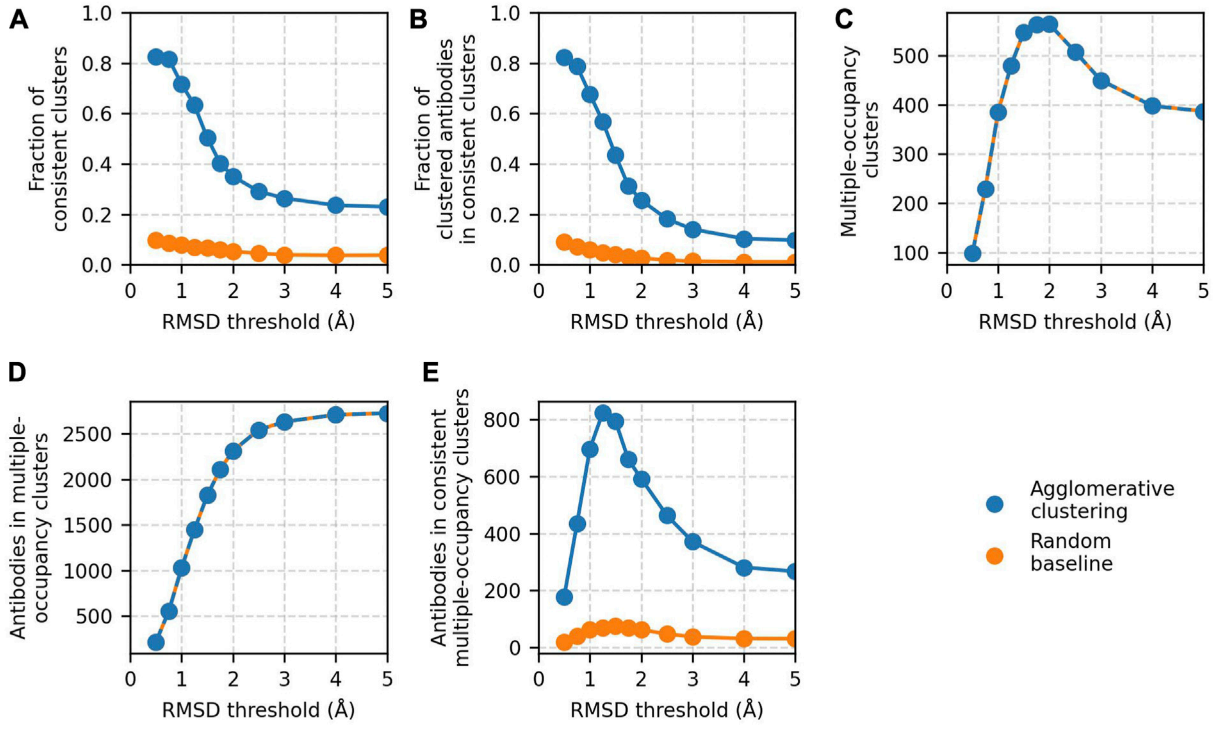


FIGURE 3
Results of agglomerative clustering as a function of RMSD threshold on the Cao et al. (2023) training set. Agglomerative clustering with a "complete" linkage criterion was performed for threshold values between 0.5 and 5 Å. The values of the five performance metrics are plotted against evaluated threshold values: **(A)** fraction of epitope-consistent clusters, **(B)** fraction of clustered antibodies in epitope-consistent clusters, **(C)** number of multiple-occupancy clusters, **(D)** number of antibodies in multiple-occupancy clusters, and **(E)** number of antibodies in epitope-consistent multiple-occupancy clusters. Results of a random clustering baseline (see Materials and Methods) are shown for comparison. Values for the number of multiple-occupancy clusters and antibodies in multiple-occupancy clusters for the random baseline are matched to agglomerative clustering.

CDRs to be more important in determining epitope specificity than the structure of others. Therefore, we investigated how the choice of CDRs over which RMSDs are calculated impacts clustering. We assessed two variations of SPACE2 that cluster based on subsets of CDRs.

In the first variation, the algorithm was adapted to consider structural similarity of heavy chain CDRs only (SPACE2-HC). This approach was motivated by sequence-based methods, such as clonotyping, which often achieve good performance considering only the heavy-chain sequence. In SPACE2-HC, antibodies were grouped based on the length of the heavy-chain CDRs, aligned on heavy-chain framework regions, C_α RMSD of CDRs H1-3 calculated, and clustered with an agglomerative clustering algorithm with a "complete" linkage criterion. An RMSD threshold of 1.25 Å was found to optimize SPACE2-HC (Supplementary Figure S3). SPACE2-HC performed worse than the standard SPACE2 algorithm as measured by a 33% drop in the trade-off metric of antibodies in epitope-consistent clusters (Supplementary Table S2). Although SPACE2-HC slightly increased dataset coverage, a substantial decrease in accuracy was observed.

A second variation of SPACE2 was implemented to cluster antibodies based only on the similarity of CDR loops that contain paratope residues (SPACE2-Paratope). Paratope residues were predicted using the Paragraph method (Chinery et al., 2023). Models were grouped by the combination of CDR loops that contain paratope residues (paratope CDRs). The models were then grouped again based on the length of paratope CDRs, aligned on framework regions, and the C_α RMSD of paratope CDRs was calculated. A RMSD threshold of 1.5 Å was found to optimize agglomerative clustering for SPACE2-Paratope (Supplementary Figure S3). Measured by the trade-off metric, SPACE2-Paratope performed worse than standard SPACE2 (Supplementary Table S2). A slight drop in both clustering accuracy and data coverage was observed.

The best clustering results were achieved by clustering based on the structural similarity of all six CDR loops. Therefore, the standard SPACE2 method was chosen as the clustering protocol for further analysis.

3.4 SPACE2 performs well on sets of antibodies against diverse targets

SPACE2 was tested on five datasets of antigen-specific antibodies using the clustering algorithm (agglomerative clustering) and parameter choices (complete linkage criterion, 1.25 Å RMSD threshold) defined on the Cao et al. (2023) training set. The test sets comprised a dataset of anti-lysozyme antibodies, a non-public dataset of anti-Ebola virus antibodies, two non-public datasets of antibodies against non-viral targets (NVA1 and NVA2), and CoV-AbDab (test), a version of CoV-AbDab with training set overlap removed (see Materials and

TABLE 2 Performance of SPACE2 on test datasets.

Dataset	Anti-lysozyme mAbs	CoV-AbDab (test)	EV	NVA1	NVA2
Antibodies in set	53	7,685	126	31	33
Fraction of antibodies modeled	1.0	1.0	0.87	1.0	1.0
Fraction of consistent clusters	1.0	0.85	0.78	0.83	1.0
Fraction of clustered antibodies in consistent clusters	1.0	0.80	0.74	0.86	1.0
Multiple-occupancy clusters	5	1,267	9	6	5
Antibodies in multiple-occupancy clusters	50 (94%)	4,188 (54%)	19 (15%)	14 (45%)	16 (48%)
Antibodies in consistent multiple-occupancy clusters	50 (94%)	3,353 (44%)	14 (11%)	12 (39%)	16 (48%)

Values of the five performance metrics and the fraction of antibodies successfully modeled using ABodyBuilder2 are shown for each dataset. For the two metrics of number of antibodies in multiple-occupancy clusters and number of antibodies in epitope-consistent multiple-occupancy clusters, a percentage is shown additionally, indicating the percentage of antibodies in the dataset. CoV-AbDab (test) denotes the subset of CoV-AbDab that is not contained in the Cao et al. (2023) training set (see Materials and Methods). CoV-AbDab (test) was used for this analysis to prevent testing on training set antibodies.

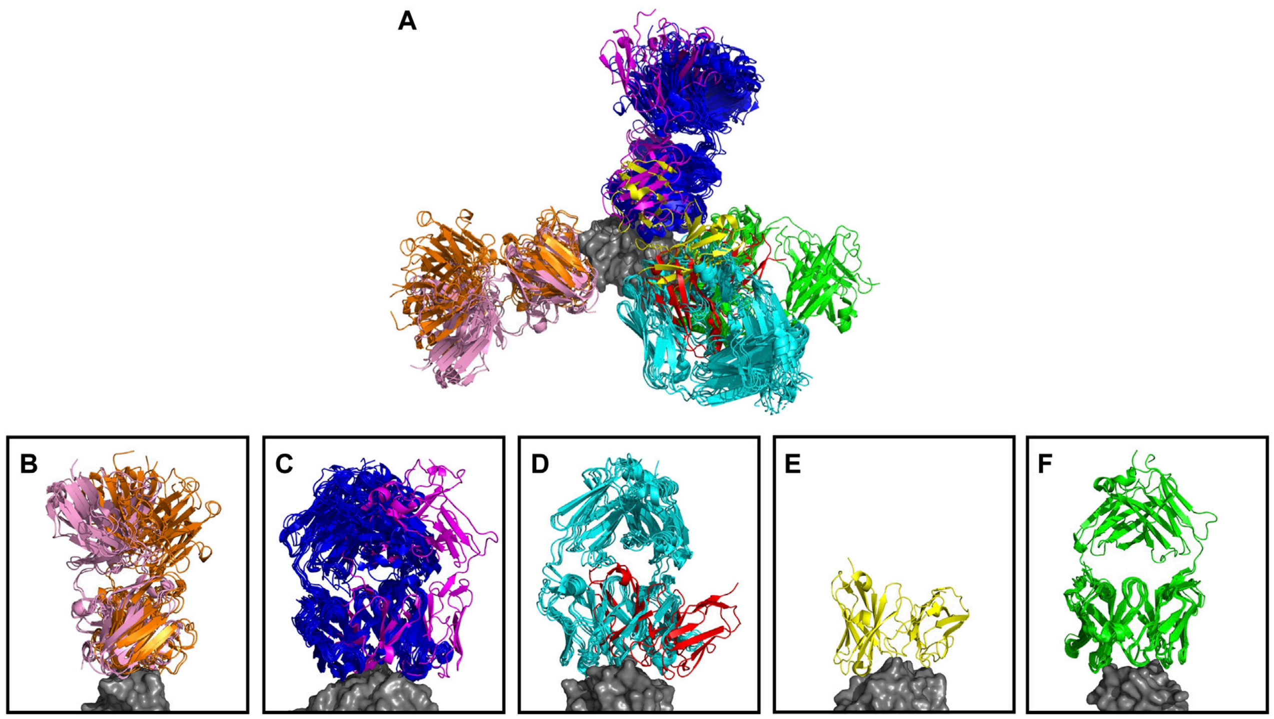


FIGURE 4
Anti-lysozyme antibodies. Crystal structures of 53 antibody–lysozyme complexes are shown aligned on the antigen structure (gray). Antibodies are colored according to the clusters assigned by SPACE2. **(A)** Overlay showing all 53 antibody–lysozyme complexes. **(B–F)** Each panel shows all antibodies that bind to one of the five lysozyme epitopes as defined by Ab-ligity (Wong et al., 2021). Panels **(B–D)** Each contain two sets of antibodies that do not overlay perfectly indicating a difference in binding pose. SPACE2 separates antibodies binding to the same epitope in a different binding pose into distinct clusters as indicated by coloring.

Methods) (Raybould et al., 2021a). An overview of results from clustering the test sets is shown in Table 2.

The anti-lysozyme dataset contains antibodies against five distinct epitopes. SPACE2 clusters antibodies in this set with high accuracy as 100% of clusters are epitope-consistent. Good data coverage is observed, and 50 of the 53 antibodies fall into multiple-occupancy clusters. SPACE2 divides the dataset into eight clusters (Figure 4). We observe three cases where antibodies binding to a common epitope are separated into two clusters. Looking at these cases in more detail shows that despite engaging the same epitope, the antibody structures do not overlay perfectly. In each case, we observe antibodies that bind to the epitope in two different binding poses, and these are separated into distinct clusters by SPACE2. These results show that SPACE2 groups antibodies with a high resolution.

SPACE2 also achieves a high clustering accuracy on CoVAbDab (test), the EV set, the NVA1 set, and the NVA2 set; 85%, 78%, 83%, and 100% of clusters in the four sets are domain/epitope-consistent, respectively. Domain/epitope-consistent clusters comprise 80%, 74%, 86%, and 100% of all antibodies grouped into multiple-occupancy clusters in the four sets, respectively.

Data coverage differs for the four datasets (see Materials and Methods for definition of coverage metrics). Coverage of CoVAbDab (test), the NVA1 set, and the NVA2 set is high, with 54%, 45%, and 48% of all antibodies contained within multiple-occupancy clusters, respectivelwhich balances both accuracy and cy. In comparison, only 19 of 126 EV set antibodies are grouped into multiple-occupancy clusters. The EV set is relatively small and contains antibodies against the Ebola virus glycoprotein, a large multi-domain protein with many potential epitopes (Lee et al., 2008). We do not expect to observe many antibodies engaging the same residue-level epitope in a small dataset of antibodies against a target with many epitopes, which is likely why we see low coverage.

Antibodies within the same epitope group in CoV-AbDab, the EV set, the NVA1 set, and the NVA2 set tend to be split across a large number of SPACE2 clusters. This is explained by the low resolution of epitope labels in these datasets and antibodies annotated with the same epitope label likely bind to a large number of different residue-level epitopes.

Overall, SPACE2 generalizes well to the test sets. The algorithm achieves a high clustering accuracy on all five datasets and a good coverage on CoV-AbDab (test), NVA1, NVA2, and anti-lysozyme datasets. Coverage of the EV set is comparably low, indicating a challenge in clustering smaller datasets of epitope-diverse antibodies.

3.5 Advances in structure prediction improve structure-based computational epitope profiling

We compared the performance of SPACE2 to SPACE1, our previous structural epitope profiling method. SPACE1 (Robinson et al., 2021) groups antibodies based on structural similarity of homology models produced using ABodyBuilder (Leem et al., 2016) followed by greedy structural clustering at an RMSD threshold of 0.75 Å.

We, once again, used the number of antibodies in epitope/domain-consistent multiple-occupancy clusters, which balances both accuracy and coverage, as our metric for comparing performance. SPACE2 outperforms SPACE1 using its suggested threshold (RMSD threshold 0.75 Å) (Supplementary Table S3). As SPACE2 uses an RMSD threshold of 1.25 Å, we also explored a range of RMSD values to see whether the difference in the RMSD threshold is the driver for the difference in performance. We found that a threshold of 1.25 Å improved SPACE1 clustering (Supplementary Table S3), but it was still significantly worse than SPACE2. SPACE1 with a 1.25 Å threshold results in an 18% and 9% decrease in antibodies in epitope/domain-consistent multiple-occupancy clusters on the two largest datasets, CoV-AbDab and the Cao et al. (2023) training set, respectively (Table 3). SPACE2's better performance is driven by better coverage while achieving a similar accuracy.

Modifications of the SPACE2 and SPACE1 methods reveal that the better performance of SPACE2 arises due to the larger number of antibodies modeled with ML-based structure prediction compared to homology modeling and better clustering with the agglomerative clustering protocol compared to greedy clustering. The higher quality of models obtained from ML-based structure prediction does not lead to clear improvements in clustering (Figure 5; Supplementary Table S4).

3.6 SPACE2 improves coverage compared to clonotyping

Clonotyping is the most commonly used epitope profiling method. It clusters antibodies based on sequence similarity. As clonotyping assumes that antibodies against a given epitope must originate from progenitor B cells with shared genetic origins, it cannot detect functional convergence. Thus, the method is limited when clustering datasets of antibodies from different individuals or species. Here, we compare SPACE2 to two lenient clonotyping protocols, VH- and Fv-clonotyping (see Materials and Methods), on the two largest datasets which contain antibodies from diverse sources. The Cao et al. (2023) training set consists of antibodies isolated from 165 human patients (Cao et al., 2023), and CoV-AbDab contains antibodies from a range of studies (~450) and several species (Raybould et al., 2021a).

The performance of SPACE2 and the two clonotyping protocols are shown in Table 3. SPACE2 outperforms both VH- and Fv-clonotyping in the key metric of antibodies in epitope/domain-consistent clusters on both datasets. Improvement in this metric is driven by increased dataset coverage by SPACE2. We observe 33% and 21% more antibodies in multiple-occupancy clusters for Fv-clonotyping of CoV-AbDab and the Cao et al. (2023) training set, respectively. Data coverage by VH-clonotyping is better but still substantially lower than SPACE2. On the other hand, SPACE2 is less accurate than both clonotyping protocols. However, the increase in coverage achieved by SPACE2 exceeds the drop in accuracy.

3.7 SPACE2 identifies functional convergence signals

We next analyzed the diversity of antibodies within the SPACE2 clusters of the Cao et al. (2023) training set and CoV-AbDab to see whether we were identifying functionally similar antibodies with very different sequences.

The majority of SPACE2 clusters contain antibodies belonging to several clonotypes, highlighting the ability to link antibodies from different genetic lineages. Specifically, 55% of epitope-consistent clusters from the Cao et al. (2023) training set and 81% of domain-consistent clusters from CoV-AbDab contain antibodies from more than one VH-clonotype (Table 3).

Moreover, we investigated the sequence similarity of antibodies within epitope/domain-consistent clusters. Clonotyping is limited to linking sequence-similar antibodies as the method uses a CDRH3 sequence identity cutoff to cluster antibodies. Here, we use a lenient cutoff of 80%. We observed a mean CDRH3 sequence identity of 86% within epitope-consistent VH-clonotypes of the Cao

TABLE 3 Comparison of SPACE2, SPACE1, and clonotyping.

Dataset	CoV-AbDab				Training set			
Method	SPACE2	SPACE1	Clonotyping		SPACE2	SPACE1	Clonotyping	
			VH	Fv			VH	Fv
Fraction of antibodies modeled	**1.0**	0.71	-	-	**1.0**	0.7	-	-
Fraction of consistent clusters	0.87	0.87	0.98	**0.99**	0.63	0.64	0.84	**0.83**
Fraction of clustered antibodies in consistent clusters	0.82	0.81	0.97	**0.99**	0.57	0.58	0.75	**0.79**
Multiple-occupancy clusters	**1,811**	1,165	1,191	970	**480**	314	361	303
Antibodies in multiple-occupancy clusters	**6,271 (59%)**	4,010 (37%)	4,045 (38%)	2,754 (26%)	**1,446 (47%)**	935 (31%)	1,060 (35%)	793 (26%)
Antibodies in consistent multiple-occupancy clusters	**5,126 (48%)**	3,255 (30%)	3,916 (37%)	2,733 (25%)	**823 (27%)**	538 (18%)	797 (26%)	628 (21%)
Fraction of clusters containing >1 VH-clonotypes	**0.81**	0.71	0	0	0.55	**0.58**	0	0
Mean CDRH3 sequence identity	**0.54**	0.57	0.88	0.88	**0.66**	0.67	0.86	0.87

The original SPACE1 algorithm was evaluated at an RMSD threshold of 1.25 Å. Two protocols were used for clonotyping (see Materials and Methods). VH-clonotyping is restricted to genes and sequence of the heavy chain. Fv-clonotyping considers both heavy and light chains. For the two metrics of number of antibodies in multiple-occupancy clusters and number antibodies in epitope-consistent multiple occupancy clusters, a percentage is shown additionally, indicating the percentage of antibodies in the dataset. The most important performance metric to consider when comparing different epitope profiling methods is the number of antibodies in epitope/domain-consistent multiple-occupancy clusters as high accuracy or coverage metrics individually may not indicate good performance. The fraction of epitope/domain-consistent clusters containing more than one VH-clonotype and the mean CDRH3 sequence identity observed within epitope/domain-consistent clusters are also given. The best result for each metric is highlighted in bold.

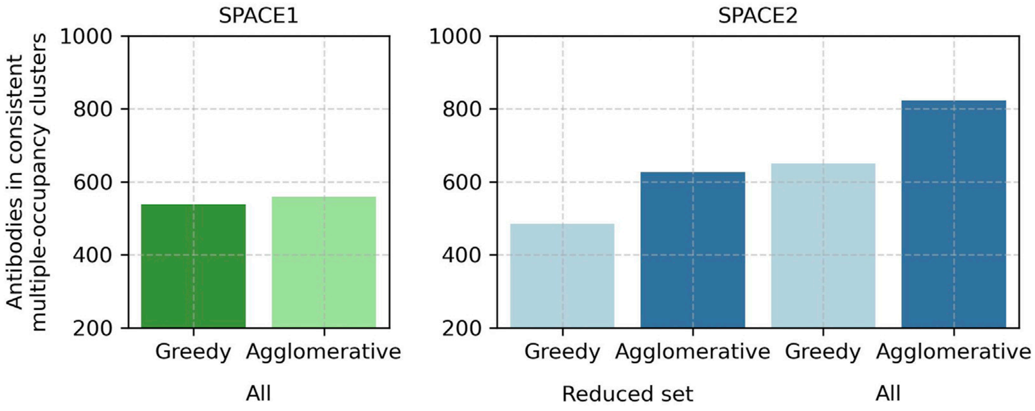


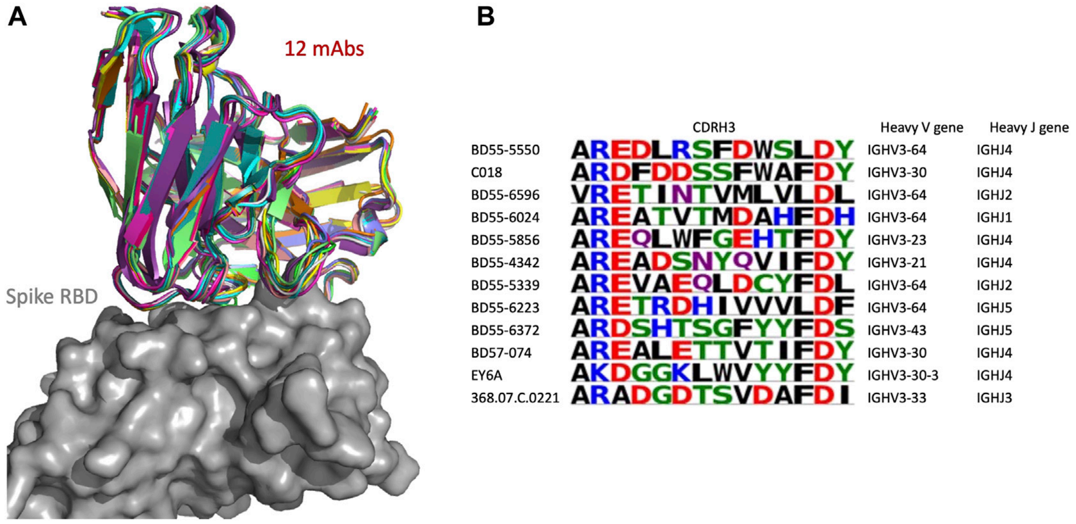
FIGURE 5
In-depth comparison of SPACE2 and SPACE1 performance on the Cao et al. (2023) training set. SPACE1 with a 1.25 Å RMSD threshold (green) and SPACE2 (blue) as well as an adaptation of SPACE1 (light green) using the default agglomerative clustering algorithm of SPACE2 (complete linkage criterion, 1.25 Å RMSD threshold) and an adaptation of SPACE2 (light blue) using the default greedy clustering algorithm of SPACE1 (1.25 Å RMSD threshold) were evaluated on the complete data set (all). Additionally, SPACE2 and its adaptation were evaluated on a reduced dataset which only included the 2,140 antibodies successfully modeled by homology modeling (reduced set).

et al. (2023) training set and 88% for domain-consistent VH-clonotypes of CoV-AbDab. In comparison, SPACE2 clusters tend to be highly diverse in sequence. Epitope/domain-consistent clusters have a mean CDRH3 sequence identity of 54% and 66% for CoV-AbDab and the Cao et al. (2023) training set, respectively (Table 3). A large number of CoV-AbDab clusters were observed with a mean sequence identity below 40% (Supplementary Figure S4), and some clusters even contain pairs of antibodies with no common CDRH3 residues.

Structural clustering is also able to functionally link antibodies from different organisms. For CoV-AbDab, SPACE2 produced 26 functionally consistent clusters containing antibodies from more than one species and was able to group antibodies from human, mouse, and rhesus macaque origins. In comparison, optimized SPACE1 was only able to detect 18 domain-consistent inter-species clusters, and clonotyping is unable to link antibodies from different species.

3.7.1 SPACE2 informs on functional convergence of sequence-dissimilar antibodies

We examined in more detail a SPACE2 cluster of the CoV-AbDab dataset with 12 members (368.07.C.0221, BD55-4342,



	CDRH3	Heavy V gene	Heavy J gene
BD55-5550	AREDLRSFDWSLDY	IGHV3-64	IGHJ4
C018	ARDFDDSSFWAFDY	IGHV3-30	IGHJ4
BD55-6596	VRETINTVMLVLDL	IGHV3-64	IGHJ2
BD55-6024	AREATVTMDAHFDH	IGHV3-64	IGHJ1
BD55-5856	AREQLWFGEHTFDY	IGHV3-23	IGHJ4
BD55-4342	AREADSNYQVIFDY	IGHV3-21	IGHJ4
BD55-5339	AREVAEQLDCYFDL	IGHV3-64	IGHJ2
BD55-6223	ARETRDHIVVVLDF	IGHV3-64	IGHJ5
BD55-6372	ARDSHTSGFYYFDS	IGHV3-43	IGHJ5
BD57-074	AREALETTVTIFDY	IGHV3-30	IGHJ4
EY6A	AKDGGKLWVYYFDY	IGHV3-30-3	IGHJ4
368.07.C.0221	ARADGDTSVDAFDI	IGHV3-33	IGHJ3

FIGURE 6
SPACE2 identifies functional convergence of antibodies. Twelve-membered CoV-AbDab cluster (368.07.C.0221, BD55-4342, BD55-5339, BD55-5550, BD55-5856, BD55-6024, BD55-6223, BD55-6372, BD55-6596, BD57-074, C018, and EY6A) with a mean CDRH3 sequence identity of 33%. The crystal structure of EY6A in complex with its antigen is available (PDB 6ZCZ). SPACE2 suggests that the 11 remaining antibodies bind to the same residue-level epitope. **(A)** Structural models of the 12 members are overlaid with the crystal structure of EY6A (not shown) in complex with the spike protein RBD (gray). **(B)** CDRH3 sequence alignment of all 12 cluster members colored by chemical properties of amino acid residues [produced with Logomaker (Tareen and Kinney, 2020)] and heavy-chain V and J genes.

BD55-5339, BD55-5550, BD55-5856, BD55-6024, BD55-6223, BD55-6372, BD55-6596, BD57-074, C018, and EY6A) (Figure 6). Eleven of the antibodies engage the spike protein RBD, and the final member is annotated as a spike protein binder with an unknown domain. Clustering by SPACE2 suggests that these antibodies, determined to bind to the same domain of the spike protein, engage the same residue-level epitope.

The 12 antibodies are highly diverse in sequence and genetic lineage. The cluster shows a mean CDRH3 sequence identity of 33%. The antibodies possess a CDRH3 of length 12, and eight of these residues differ on average. The most distant pair of antibodies in the cluster is BD55-6596 and EY6A, which differ in 11 of 12 CDRH3 residues. The 12 antibodies originate from seven different IgGH genes and fall into 12 separate VH-clonotypes.

The improvement of SPACE2 over SPACE1 can be seen when examining how these antibodies were clustered by SPACE1. Using SPACE1 with an optimized threshold, only six of these antibodies (BD55-6024, BD55-6223, BD55-6596, BD57-074, C018, and EY6A) were grouped together, and even these six were a part of a larger functionally inconsistent cluster with 44 members. BD55-5339 was in a separate functionally inconsistent SPACE1 cluster with four members, and the remaining five antibodies were not placed in multiple-occupancy clusters.

3.7.2 SPACE2 identifies epitopes targeted by multiple species

As the CoV-AbDab database contains antibodies from multiple species (human, mouse, and rhesus macaque), we examined whether SPACE2 can identify epitopes targeted by multiple species. There were 26 SPACE2 clusters of the CoV-AbDab database that contained antibodies from more than one species. We examined a SPACE2 cluster with seven members (368.02a.C.0049, B13, BD55-6574, BD57-092, DK15, Fab-160, and SW186) (Figure 7). The cluster contains six antibodies that engage the spike protein RBD and one spike-specific antibody without domain-level epitope data. Five of the antibodies have human genetics and originate from human patients, phage-display, and transgenic mice. The remaining antibodies have murine genetics and were raised by immunized mice. SPACE2 suggests that these genetically human and mouse antibodies engage the same residue-level epitope which highlights its ability to detect public epitope targeted by multiple species.

SPACE1 with an optimized threshold was only able to link one of the two murine antibodies in this cluster to human structures, while clonotyping is unable to link mouse and human antibodies due to different gene usage.

4 Discussion

Accurately identifying the epitope of antibodies is a key step in understanding immunology and in the design of new biological drugs. Such data are conventionally determined experimentally either by solving individual antibody–antigen crystal structures or by epitope binning methods, such as competition binding assays. Prior computational clustering of antibodies into functional groups could reduce the number of experiments that needs to be carried out or even remove the need for them entirely. Clonal clustering is most commonly used for this purpose, where antibodies are grouped by sequence identity and genetic lineage. However, these types of methods will miss antibodies with low sequence identity that have functionally converged and target common epitopes (Scheid et al., 2011; Joyce et al., 2016; Rijal et al., 2019; Robinson et al., 2021; Wong et al., 2021; Shrock et al., 2023).

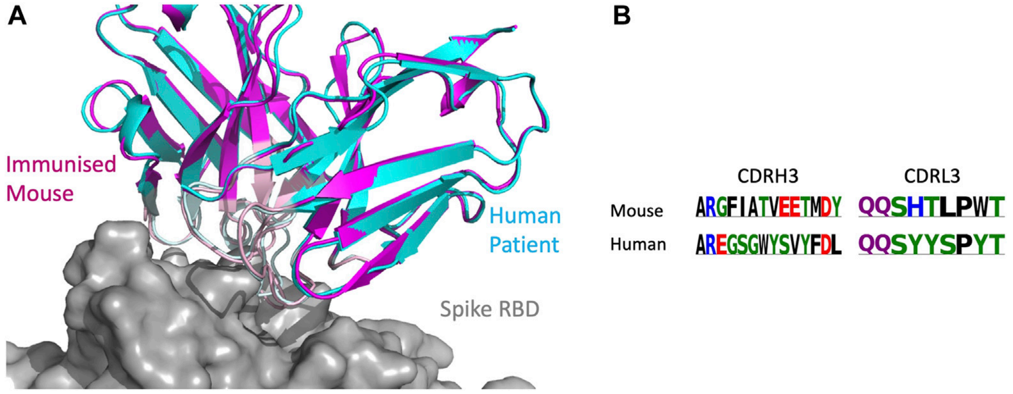


FIGURE 7
SPACE2 identifies epitopes targeted by multiple species. Two representatives from a SPACE2 CoV-AbDab cluster comprising murine (SW186) and human (BD57-092) antibodies are shown. The crystal structure of SW186 is available (PDB 8DT3). SPACE2 suggests that BD57-092 binds to the same residue-level epitope. **(A)** Structural models of SW186 (pink) and BD57-092 (cyan) were overlaid with the crystal structure of SW186 (not shown) in complex with the spike protein RBD (gray). CDR regions of both antibodies are highlighted by lighter coloring. **(B)** CDRH3 sequence alignment of the two antibodies colored by chemical properties of amino acid residues [produced with Logomaker (Tareen and Kinney, 2020)].

In a previous study, we reported the SPACE1 method (Robinson et al., 2021), which clusters antibodies by structural similarity of their homology models. This method showed that structure-based epitope profiling is better able to detect the full breadth of functional convergence. However, SPACE1 was limited by the coverage of template-based modeling and its inaccuracies. The method was also only benchmarked at the level of domain-consistency of antibodies against one virus class. Here, we introduce SPACE2, an updated method which uses the latest machine learning-based antibody structure prediction technology (Abanades et al., 2022b) and a novel clustering protocol systematically optimized on epitope-resolution data.

We show across six datasets that SPACE2 can accurately bin antibodies that engage the same epitope and achieve high data coverage. Available crystal structures of antibody–antigen complexes reveal that SPACE2 tends to group antibodies that bind to the same residue-level epitope in an identical binding pose. Epitope resolution of SPACE2 appears to be similar to that obtained from crystal structures and higher than data from epitope binning methods which struggle to distinguish between antibodies that bind to the same site and those that bind to distinct sites but overlap sterically.

SPACE2 outperforms our previous epitope profiling tool SPACE1 (Robinson et al., 2021) and clonotyping when considering the number of antibodies in epitope-consistent multiple-occupancy clusters. Clonotyping is more accurate than SPACE2 but has far lower coverage. The lower accuracy of SPACE2 is explained by the fact that antibodies with similar CDR structures may engage different epitopes if chemical properties of the CDR residues are significantly different.

We also highlight how our methodology allows the detection of functional convergence across populations of antibodies. Across functionally consistent clusters of our largest dataset, CoV-AbDab (Raybould et al., 2021a), we detect a mean CDRH3 sequence identity as low as 54%. Furthermore, we observe 26 functional clusters containing antibodies from multiple species including human, mouse, and rhesus macaque antibodies. In comparison, sequence-based epitope profiling such as clonotyping is severely restricted in grouping sequence-diverse antibodies and is not able to link antibodies from different genetic origins and species (Greiff et al., 2015; López-Santibáñez-Jácome et al., 2019; Raybould et al., 2021b). Although it is possible to cluster nanobodies with the SPACE2-HC implementation, we were unable to detect functional convergence to antibodies when testing on CoV-AbDab. No clusters were detected containing both antibodies and nanobodies suggesting that the two formats use different binding site structures to engage common epitopes.

SPACE2 clusters antibodies based on the length and structural similarity of all six CDRs. This approach may constrain the detection of functional convergence to some extent as it assumes that antibodies require the same length of all six CDRs to engage the same epitope. We tried to address this issue by evaluating two adaptations of SPACE2 that reduce the number of CDRs required to have the same length. An implementation clustering antibodies based on heavy-chain structural similarity (SPACE2-HC) caused a drastic decrease in clustering accuracy. This indicates that light chain structures are important for determining antibody binding specificity, which is in line with previous findings on the functional selection of light chains (Jaffe et al., 2022) and their structural importance (Guloglu and Deane, 2023). Similarly, combining SPACE2 with information from paratope prediction (SPACE2-Paratope) (Chinery et al., 2023) and computing structural similarity only across CDRs predicted to contain paratope residues currently led to fewer functionally consistent clusters. Furthermore, some loops of different lengths can adopt similar structures (Nowak et al., 2016; Wong et al., 2019). Although evidence suggests that this is infrequent, future work could focus on being able to detect functional convergence across different CDR lengths.

The ability to detect functional convergence of antibodies will provide valuable insights into the humoral immune response. SPACE2 is able to provide more complete information on public epitopes targeted by antibodies originating from different individuals and species. Although previous studies show several public epitopes are largely distinct between species (Shrock et al., 2023), here, we identify a number of inter-species clusters. Structural clustering of larger datasets of antibodies isolated from various species will further improve our understanding of differences in their immune responses.

Although SPACE2 is computationally more expensive than sequence-based epitope profiling, it is tractable for datasets of 10^4 antibodies, a typical number of sequences obtained from methods such as 10× sequencing (Supplementary Figure S5). The rate limiting step of SPACE2 is currently the prediction of antibody structures. Improvements in the speed of structure prediction tools as well as the release of antibody databases containing pre-modeled structures (Abanades et al., 2022b) will contribute to reducing the computational cost of structure-based epitope profiling.

Overall, SPACE2 efficiently detects functional convergence of antibodies with highly diverse sequences, genetic lineage, and species origins, further illustrating that predicted structures should be considered when investigating the function of antibodies. SPACE2 is openly available on GitHub (https://github.com/oxpig/SPACE2).

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

FS and CD contributed to the conception and design of the study. FS and BA wrote the code for the SPACE2 method. FS performed the data analysis and wrote the manuscript. MR, WW, GG, and CD contributed to critical revision of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by the Engineering and Physical Sciences Research Council (EPSRC), under grant number EP/S024093/1, and Roche (FS and BA) and a Postdoctoral Research grant funded by Boehringer Ingelheim (MR).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2023.1237621/full#supplementary-material

SUPPLEMENTARY FIGURES AND TABLES

Data Sheet 1.pdf.

TRAINING SET

Data Sheet 2.csv.

COV-ABDAB

Data Sheet 3.cvs.

ANTI-LYSOZYME ANTIBODIES

Data Sheet 4.csv

References

Abanades, B., Georges, G., Bujotzek, A., and Deane, C. M. (2022a). ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation. *Bioinformatics* 38, 1877–1880. doi:10.1093/bioinformatics/btac016

Abanades, B., Wong, W. K., Boyles, F., Georges, G., Bujotzek, A., and Deane, C. M. (2022b). ImmuneBuilder: deep-Learning models for predicting the structures of immune proteins. *bioRxiv*. doi:10.1101/2022.11.04.514231

Abdiche, Y. N., Malashock, D. S., Pinkerton, A., and Pons, J. (2009). Exploring blocking assays using Octet, ProteOn, and Biacore biosensors. *Anal. Biochem.* 386, 172–180. doi:10.1016/j.ab.2008.11.038

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). Optics: ordering points to identify the clustering structure. *ACM SIGMOD Rec.* 28, 49–60. doi:10.1145/304181.304187

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373, 871–876. doi:10.1126/science.abj8754

Bashford-Rogers, R. J. M., Bergamaschi, L., McKinney, E. F., Pombal, D. C., Mescia, F., Lee, J. C., et al. (2019). Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature* 574, 122–126. doi:10.1038/s41586-019-1595-3

Briney, B., Inderbitzin, A., Joyce, C., and Burton, D. R. (2019). Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* 566, 393–397. doi:10.1038/s41586-019-0879-y

Butina, D. (1999). Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. *J. Chem. Inf. Comput. Sci.* 39, 747–750. doi:10.1021/ci9803381

Cao, Y., Jian, F., Wang, J., Yu, Y., Song, W., Yisimayi, A., et al. (2023). Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution. *Nature* 614, 521–529. doi:10.1038/s41586-022-05644-7

Chinery, L., Wahome, N., Moal, I., and Deane, C. M. (2023). Paragraph—Antibody paratope prediction using graph neural networks with minimal feature vectors. *Bioinformatics* 39, btac732. doi:10.1093/bioinformatics/btac732

Choi, Y., and Deane, C. M. (2010). FREAD revisited: accurate loop structure prediction using a database search algorithm. *Proteins Struct. Funct. Bioinforma.* 78, 1431–1440. doi:10.1002/prot.22658

Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., et al. (2014). SAbDab: the structural antibody database. *Nucleic Acids Res.* 42, D1140–D1146. doi:10.1093/nar/gkt1043

Frey, B. J., and Dueck, D. (2007). Clustering by passing messages between data points. *Science* 315, 972–976. doi:10.1126/science.1136800

Greiff, V., Miho, E., Menzel, U., and Reddy, S. T. (2015). Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends Immunol.* 36, 738–749. doi:10.1016/j.it.2015.09.006

Guloglu, B., and Deane, C. M. (2023). Specific attributes of the VL domain influence both the structure and structural variability of CDR-H3 through steric effects. *bioRxiv*. doi:10.1101/2023.05.16.540974

Hsiao, Y.-C., Shang, Y., DiCara, D. M., Yee, A., Lai, J., Kim, S. H., et al. (2019). Immune repertoire mining for rapid affinity optimization of mouse monoclonal antibodies. *mAbs* 11, 735–746. doi:10.1080/19420862.2019.1584517

Jaffe, D. B., Shahi, P., Adams, B. A., Chrisman, A. M., Finnegan, P. M., Raman, N., et al. (2022). Functional antibodies exhibit light chain coherence. *Nature* 611, 352–357. doi:10.1038/s41586-022-05371-z

Joyce, M. G., Wheatley, A. K., Thomas, P. V., Chuang, G.-Y., Soto, C., Bailer, R. T., et al. (2016). Vaccine-induced antibodies that neutralize group 1 and group 2 influenza A viruses. *Cell* 166, 609–623. doi:10.1016/j.cell.2016.06.043

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2

Landrum, G. (2006). *RDKit: Open-source cheminformatics.*

Lee, J. E., Fusco, M. L., Hessell, A. J., Oswald, W. B., Burton, D. R., and Saphire, E. O. (2008). Structure of the Ebola virus glycoprotein bound to an antibody from a human survivor. *Nature* 454, 177–182. doi:10.1038/nature07082

Leem, J., Dunbar, J., Georges, G., Shi, J., and Deane, C. M. (2016). ABodyBuilder: automated antibody structure prediction with data-driven accuracy estimation. *mAbs* 8, 1259–1268. doi:10.1080/19420862.2016.1205773

Lefranc, M.-P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., et al. (2003). IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.* 27, 55–77. doi:10.1016/S0145-305X(02)00039-3

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., et al. (2022). Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*. doi:10.1101/2022.07.20.500902

López-Santibáñez-Jácome, L., Avendaño-Vázquez, S. E., and Flores-Jasso, C. F. (2019). The pipeline repertoire for ig-seq analysis. *Front. Immunol.* 10, 899. doi:10.3389/fimmu.2019.00899

MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth berkeley symposium on mathematical statistics and probability, volume 1: Statistics 5.1*, 281–298.

Murtagh, F., and Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *WIREs Data Min. Knowl. Discov.* 2, 86–97. doi:10.1002/widm.53

Nilvebrant, J., and Rockberg, J. (2018). "An introduction to epitope mapping," in *Epitope mapping protocols*. *Methods in molecular biology*. Editors J. Rockberg and J. Nilvebrant (New York, NY: Springer), 1–10. doi:10.1007/978-1-4939-7841-0_1

North, B., Lehmann, A., and Dunbrack, R. L. (2011). A new clustering of antibody CDR loop conformations. *J. Mol. Biol.* 406, 228–256. doi:10.1016/j.jmb.2010.10.030

Nowak, J., Baker, T., Georges, G., Kelm, S., Klostermann, S., Shi, J., et al. (2016). Length-independent structural similarities enrich the antibody CDR canonical class model. *mAbs* 8, 751–760. doi:10.1080/19420862.2016.1158370

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Raybould, M. I. J., Kovaltsuk, A., Marks, C., and Deane, C. M. (2021a). CoV-AbDab: the coronavirus antibody database. *Bioinformatics* 37, 734–735. doi:10.1093/bioinformatics/btaa739

Raybould, M. I. J., Rees, A. R., and Deane, C. M. (2021b). Current strategies for detecting functional convergence across B-cell receptor repertoires. *mAbs* 13, 1996732. doi:10.1080/19420862.2021.1996732

Reddy, S. T., Ge, X., Miklos, A. E., Hughes, R. A., Kang, S. H., Hoi, K. H., et al. (2010). Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat. Biotechnol.* 28, 965–969. doi:10.1038/nbt.1673

Richardson, E., Galson, J. D., Kellam, P., Kelly, D. F., Smith, S. E., Palser, A., et al. (2021). A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-pertussis toxoid antibodies. *mAbs* 13, 1869406. doi:10.1080/19420862.2020.1869406

Rijal, P., Elias, S. C., Machado, S. R., Xiao, J., Schimanski, L., O'Dowd, V., et al. (2019). Therapeutic monoclonal antibodies for Ebola virus infection derived from vaccinated humans. *Cell Rep.* 27, 172–186. doi:10.1016/j.celrep.2019.03.020

Ripoll, D. R., Chaudhury, S., and Wallqvist, A. (2021). Using the antibody-antigen binding interface to train image-based deep neural networks for antibody-epitope classification. *PLOS Comput. Biol.* 17, e1008864. doi:10.1371/journal.pcbi.1008864

Robinson, S. A., Raybould, M. I. J., Schneider, C., Wong, W. K., Marks, C., and Deane, C. M. (2021). Epitope profiling using computational structural modelling demonstrated on coronavirus-binding antibodies. *PLOS Comput. Biol.* 17, e1009675. doi:10.1371/journal.pcbi.1009675

Ruffolo, J. A., Chu, L.-S., Mahajan, S. P., and Gray, J. J. (2022a). Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nat. Commun.* 14 (1), 2389. doi:10.1038/s41467-023-38063-x

Ruffolo, J. A., Guerra, C., Mahajan, S. P., Sulam, J., and Gray, J. J. (2020). Geometric potentials from deep learning improve prediction of CDR H3 loop structures. *Bioinforma. Oxf. Engl.* 36, i268–i275. doi:10.1093/bioinformatics/btaa457

Ruffolo, J. A., Sulam, J., and Gray, J. J. (2022b). Antibody structure prediction using interpretable deep learning. *Patterns* 3, 100406. doi:10.1016/j.patter.2021.100406

Scheid, J. F., Mouquet, H., Ueberheide, B., Diskin, R., Klein, F., Oliveira, T. Y. K., et al. (2011). Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science* 333, 1633–1637. doi:10.1126/science.1207227

Schneider, C., Raybould, M. I. J., and Deane, C. M. (2022). SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker. *Nucleic Acids Res.* 50, D1368–D1372. doi:10.1093/nar/gkab1050

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., and Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.* 42, 1–21. doi:10.1145/3068335

Shrock, E. L., Timms, R. T., Kula, T., Mena, E. L., West, A. P., Guo, R., et al. (2023). Germline-encoded amino acid–binding motifs drive immunodominant public antibody responses. *Science* 380, eadc9498. doi:10.1126/science.adc9498

Tareen, A., and Kinney, J. B. (2020). Logomaker: beautiful sequence logos in Python. *Bioinformatics* 36, 2272–2274. doi:10.1093/bioinformatics/btz921

Tsioris, K., Gupta, N. T., Ogunniyi, A. O., Zimnisky, R. M., Qian, F., Yao, Y., et al. (2015). Neutralizing antibodies against West Nile virus identified directly from human B cells by single-cell analysis and next generation sequencing. *Integr. Biol.* 7, 1587–1597. doi:10.1039/C5IB00169B

Wong, W. K., Leem, J., and Deane, C. M. (2019). Comparative analysis of the CDR loops of antigen receptors. *Front. Immunol.* 10, 2454. doi:10.3389/fimmu.2019.02454

Wong, W. K., Robinson, S. A., Bujotzek, A., Georges, G., Lewis, A. P., Shi, J., et al. (2021). Ab-ligity: identifying sequence-dissimilar antibodies that bind to the same epitope. *mAbs* 13, 1873478. doi:10.1080/19420862.2021.1873478

Zhu, J., Ofek, G., Yang, Y., Zhang, B., Louder, M. K., Lu, G., et al. (2013). Mining the antibodyome for HIV-1–neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proc. Natl. Acad. Sci.* 110, 6470–6475. doi:10.1073/pnas.1219320110

Nomenclature

RBD	receptor-binding domain
S protein	spike protein
NTD	N-terminal domain
GP1	envelop glycoprotein 1
RMSD	root mean square deviation
CDR	complementarity-determining region

Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 28 September 2023
DOI 10.3389/fmolb.2023.1249247

OPEN ACCESS

EDITED BY
Traian Sulea,
National Research Council Canada (NRC), Canada

REVIEWED BY
M. Frank Erasmus,
Specifica Inc., United States
Hubert Kettenberger,
Roche Diagnostics GmbH, Germany

*CORRESPONDENCE
Andreas Evers,
Andreas.evers@merckgroup.com

RECEIVED 28 June 2023
ACCEPTED 31 August 2023
PUBLISHED 28 September 2023

CITATION
Arras P, Yoo HB, Pekar L, Clarke T, Friedrich L, Schröter C, Schanz J, Tonillo J, Siegmund V, Doerner A, Krah S, Guarnera E, Zielonka S and Evers A (2023), AI/ML combined with next-generation sequencing of VHH immune repertoires enables the rapid identification of *de novo* humanized and sequence-optimized single domain antibodies: a prospective case study.
Front. Mol. Biosci. 10:1249247.
doi: 10.3389/fmolb.2023.1249247

AI/ML combined with next-generation sequencing of VHH immune repertoires enables the rapid identification of *de novo* humanized and sequence-optimized single domain antibodies: a prospective case study

Paul Arras[1,2], Han Byul Yoo[1], Lukas Pekar[1], Thomas Clarke[3], Lukas Friedrich[4], Christian Schröter[5], Jennifer Schanz[5], Jason Tonillo[5], Vanessa Siegmund[6], Achim Doerner[1], Simon Krah[1], Enrico Guarnera[1], Stefan Zielonka[1,2] and Andreas Evers[1]*

[1]Antibody Discovery and Protein Engineering, Merck Healthcare KGaA, Darmstadt, Germany, [2]Institute for Organic Chemistry and Biochemistry, Technical University of Darmstadt, Darmstadt, Germany, [3]Bioinformatics, EMD Serono, Billerica, MA, United States, [4]Computational Chemistry and Biologics, Merck Healthcare KGaA, Darmstadt, Germany, [5]ADCs & Targeted NBE Therapeutics, Merck KGaA, Darmstadt, Germany, [6]Early Protein Supply and Characterization, Merck Healthcare KGaA, Darmstadt, Germany

Introduction: In this study, we demonstrate the feasibility of yeast surface display (YSD) and nextgeneration sequencing (NGS) in combination with artificial intelligence and machine learning methods (AI/ML) for the identification of de novo humanized single domain antibodies (sdAbs) with favorable early developability profiles.

Methods: The display library was derived from a novel approach, in which VHH-based CDR3 regions obtained from a llama (Lama glama), immunized against NKp46, were grafted onto a humanized VHH backbone library that was diversified in CDR1 and CDR2. Following NGS analysis of sequence pools from two rounds of fluorescence-activated cell sorting we focused on four sequence clusters based on NGS frequency and enrichment analysis as well as in silico developability assessment. For each cluster, long short-term memory (LSTM) based deep generative models were trained and used for the in silico sampling of new sequences. Sequences were subjected to sequence- and structure-based in silico developability assessment to select a set of less than 10 sequences per cluster for production.

Results: As demonstrated by binding kinetics and early developability assessment, this procedure represents a general strategy for the rapid and efficient design of potent and automatically humanized sdAb hits from screening selections with favorable early developability profiles.

KEYWORDS

artificial intelligence and machine learning (ML), deep learning, *in silico* developability, long short-term memory (LSTM), next-generation sequencing (NGS), single domain antibodies (VHH), yeast surface display (YSD), protein engineering

Introduction

VHHs (variable domain of the heavy chain of a heavy chain-only antibodies), commercially known as nanobodies, are single-domain antibody (sdAb) fragments derived from camelid heavy chain-only antibodies (HcAbs). VHHs exhibit small size, high stability, and exceptional binding specificity, making them valuable tools for therapeutics, diagnostics, and research applications (Krah et al., 2016; Könning et al., 2017; Wang et al., 2022; Jin et al., 2023). Owing to their simple molecular architecture, they offer a plethora of engineering options with respect to the generation of bi- and multispecific antibody designs involving different paratope valences and spatial orientations of individual domains within a given molecule (Bannas et al., 2017; Chanier and Chames, 2019; Pekar et al., 2020; Yanakieva et al., 2022; Lipinski et al., 2023a; Lipinski et al., 2023b). However, VHH domains usually have to be humanized and further sequence-optimized to be suitable for therapeutic applications.

A classical cascade for antibody and VHH discovery typically involves (camelid) immunization and antibody library construction after immunization followed by antibody selections or panning. Subsequently, Sanger sequencing of high prevalent clones can be applied (typically in the range of a couple of hundred clones) that are then profiled for the desired on-target effect, and functional or phenotypic assays. The best hits are then nominated for sequence optimization, usually including humanization (Vincke et al., 2009; Sulea et al., 2022), replacement of chemically labile and post-translational modification (PTM) motifs and ideally considering further developability-related aspects (Lauer et al., 2012; Sormanni et al., 2015; Raybould et al., 2019; Ahmed et al., 2021; Khetan et al., 2022; Negron et al., 2022; Evers et al., 2023a; Fernández-Quintero et al., 2023; Jain et al., 2023; Mieczkowski et al., 2023; Svilenov et al., 2023). Sometimes, the complexity of these different optimization parameters might require multiple design cycles and in some cases it might not be even possible to optimize such hits towards a favorable overall profile (Rabia et al., 2018). This process of iterative sequence optimization is generally on the critical path in early biologics drug discovery projects. Therefore, it is highly desirable to find new approaches that accelerate the discovery and design of humanized sequences with a favorable early developability profile, both in terms of project timelines and to reduce attrition in the downstream process.

In contrast to the traditional approach of Sanger sequencing, next-generation sequencing (NGS) of screening pools obtained from selection campaigns enables a rapid and cost-effective analysis of the vast sequence spaces of binders (Larman et al., 2012; Mathonet and Ullman, 2013; Hu et al., 2015; Barreto et al., 2019). Integration of Sequence-Activity-Relationship (SAR), frequency and enrichment analyses with *in silico* developability assessment on NGS data can furthermore provide a rational approach to identify potent sequences with improved developability profiles. Moreover, recent studies have shown the versatility of artificial intelligence/machine learning (AI/ML) techniques on antibody NGS data to design new sequences with potentially further improved potency or developability (Liu et al., 2020; Mason et al., 2021; Saka et al., 2021; Makowski et al., 2022; Hie et al., 2023; Parkinson et al., 2023). In these studies, regions of specific antibody candidates were diversified in combinatorial mutagenesis display libraries, followed by the generation of ML models from NGS data. Saka et al. (2021), for example, employed long short-term memory (LSTM) based on NGS derived sequences from different panning rounds of a library diversified in CDR-H1, -H2 and -H3 and FR1 of a kynurenine binding antibody. The affinities of newly designed sequences were over 1800-fold higher than for the parental clone. LSTM is a widely used deep learning architecture in natural language processing that is also particularly effective in predicting new protein sequences, as it is capable of modeling long-term dependencies and capturing the complex relationships between amino acids that determine structure and function. Such LSTMs have not only been successfully applied for the design of new antibodies (Saka et al., 2021), but also for peptides (Müller et al., 2018) and small molecules (Gupta et al., 2018; Merk et al., 2018; Segler et al., 2018; Z et al., 2022). While the above-mentioned studies used combinatorial synthetic display libraries in combination with NGS and AI/ML to optimize existing lead antibodies, this concept might also be employed to discover new and potent antibody sequences with favorable developability profiles from diverse antibody repertoires obtained from animal immunization.

As part of our integrated VHH hit discovery strategy, we have recently implemented a semi-immune/semi-synthetic library approach for the high-throughput *de novo* identification of humanized VHHs following camelid immunization (Arras et al., 2023). For this, VHH-derived CDR3 regions obtained from a llama, immunized against recombinant human (rh) Natural Cytotoxicity Receptor NKp46 (Barrow et al., 2019), were grafted onto a humanized VHH backbone library comprising sequence-diversified CDR1 and CDR2 regions that were tailored towards favorable *in silico* developability properties, by considering human-likeness and excluding potential sequence liabilities and predicted immunogenic motifs. NKp46 is an activating receptor on Natural Killer cells (NK cells) and was successfully harnessed for the generation of potent NK cell engagers (Gauthier et al., 2019; Gauthier et al., 2023; Lipinski et al., 2023). Target-specific humanized VHHs were readily obtained in our previous study by YSD (Arras et al., 2023). By exploiting this approach, high affinity VHHs with optimized developability profiles can principally be generated against any antigen of interest upon camelid immunization. The process of CDR3 engraftment onto our generic humanized and sequence-optimized VHH scaffold library is characterized by its low complexity and duration similar to the generation of wild-type VHH display libraries following immunization (Roth et al., 2020); thereby this procedure significantly accelerates VHH hit discovery by reducing or even eliminating the need for subsequent sequence optimization. Due to the setup of our library approach, all resulting VHHs have a fixed humanized framework sequence, e.g., any differences in antigen binding and developability properties are driven by sequence variations in the CDR regions. Providing NGS data from different rounds of YSD (Valldorf et al., 2022) based FACS screens from this library therefore represent ideal inputs to train AI/ML models for the design of new sequences with even further improved potency and developability.

Goal of the present study was to investigate the feasibility of our integrated approach of combining i) camelid immunization, ii) humanized VHH library generation, iii) YSD, iv) FACS screening, v) NGS analysis, vi) AI/ML based sequence sampling and vii) *in silico* developability assessment to identify potent and readily sequence optimized VHH hits in a single procedure. The display library was derived from our humanized VHH library that was directed against (rh) NKp46 (Arras et al., 2023). Based on NGS analysis, we selected four diverse CDR3 sequence clusters in the present study that showed high frequency or enrichment over two rounds of FACS screening. These repertoires were used to train LSTM deep generative models for the automated design of new sequences that were subsequently filtered based on *in silico* developability criteria using our recently described *Sequence Assessment Using Multiple Optimization criteria* (SUMO) approach (Evers et al., 2023a). We finally selected a set of only up to ten sequences per cluster for synthesis and experimental profiling. As demonstrated in binding measurements and early developability assays, the proposed methodology has the capability to generate diverse and potent VHH hits directly from screening collections upon camelid immunization that do ideally not require further humanization and sequence optimization. Furthermore, it provides sequence activity (SAR) and sequence-property (SPR) relationships for each of the investigated sequence clusters. Taken together, as exemplified and demonstrated on a typical early drug discovery project, this workflow has the potential to significantly accelerate hit discovery and optimization and reduce the risk for developability-related attrition.

Results

Previous work: humanized VHH library construction after camelid immunization, yeast surface display and cell sorting

As outlined in detail in our previous study (Arras et al., 2023) and schematically illustrated in Figure 1A, we have recently developed a semi-immune/semi-synthetic strategy that relies on grafting the PBMC-amplified CDR3 VHH repertoire of llamas following immunization onto two internally optimized humanized backbone libraries with a framework germline sequence derived from human IGHV3-23*1 (Arras et al., 2023). Both libraries were diversified in CDR1 and CDR2 towards favorable *in silico* developability properties, i) considering amino acid distributions observed in naïve and immunized llamas, eliminating residue combinations ii) that would result in potential N-glycosylation sites (Asn-X-Ser/Thr) or highly susceptible chemical liability motifs (Asn-Gly, Asp-Gly, Met, unpaired Cys) and iii) strong predicted MHC-II binding peptide motifs, while taking into account iv) diversity with respect to charge, size and hydrophobicity and v) occurrence in the equivalent positions in NGS data of human antibody repertoires. To identify novel binders against (rh) NKp46, we had opted for PBMC-derived total RNA of a (rh) NKp46 immunized llama for the generation of both CDR3-engrafted humanized libraries for YSD. As demonstrated in a head-to-head comparison, sequences from the CDR3-engrafted humanized library that were selected after two rounds of FACS showed similar activity against NKp46 compared to CDR3-analogues from immunized WT llama sequences with improved early developability profiles (Arras et al., 2023). In that study, 96 clones were selected after FACS by random picking and Sanger sequencing from each library. For the present study, we re-analyzed the sequence pools of the CDR3-engrafted humanized library from the different selection rounds by NGS (Figure 1B).

Identification of sequence clusters based on NGS analysis and *in silico* developability assessment

The application of NGS in combination with AI/ML approaches can represent a quick and cost-effective way to identify potent and developable binders that might not be picked with the traditional approach of random clone selection and Sanger sequencing. To exhaustively assess sequence diversity from our previous display campaign, NGS data for screening pools obtained from the different FACS rounds of the CDR3-engrafted humanized library were generated using the MiSeq system (Figure 1B). Table 1 summarizes the absolute number of NGS reads that were obtained after the different rounds of FACS for all sequences and for those CDR3 sequence clusters that were used for LSTM deep generative model generation as outlined below.

Sequences were annotated with Geneious Biologics (Antibody Discovery Software, 2023) using IMGT numbering and clustered based on 50% CDR3 sequence identity. We assumed that this cutoff assures that i) within each cluster most VHHs bind in a similar manner to the same epitope, and ii) at the same time provides sufficient sequence diversity within each cluster for ML model generation, SAR analysis and automated multi-parameter optimization towards improved potency and developability. All sequence clusters were ranked by either i) their absolute frequency (total number of reads), i.e., the number of clones observed after the second round of FACS or ii) their enrichments (as described in Materials and Methods) observed over FACS round 2 vs. round 0 (Figure 1B; Table 1). The ranking of clusters and sequences based on their absolute frequency should principally result in similar selections compared to the random selection and Sanger sequencing approach that is usually applied in the traditional screening cascade. Conversely, selection based on enrichment is potentially able to identify rare clones with superior affinity and specificity (Rouet et al., 2018; Barreto et al., 2019). In a first feasibility study, we selected the most occurring CDR1-3 amino acid sequence from the i) five most frequent and ii) five most enriched CDR3 clusters for production and binding affinity determination against NKp46. Since two CDR3 clusters occurred in both sets, a total of eight sequences were produced and tested (Table 2). Remarkably, seven sequences showed binding affinity in the 1-digit nanomolar range. Only the representative of the most frequent cluster exhibits a slightly lower binding affinity (KD = 19.8 nM). These results are in agreement with previous literature reports that enrichment-based selection based on NGS data can provide additional potent sequences (Rouet et al., 2018; Barreto et al., 2019).

As mentioned above, due to our library design strategy, all sequences are identical in their framework regions that were derived from a humanized germline sequence. In the next step, we analyzed the sequence and computed property space within each CDR3 sequence cluster. To visualize diversity (based on sequence identity) after each round of FACS enrichment, the respective

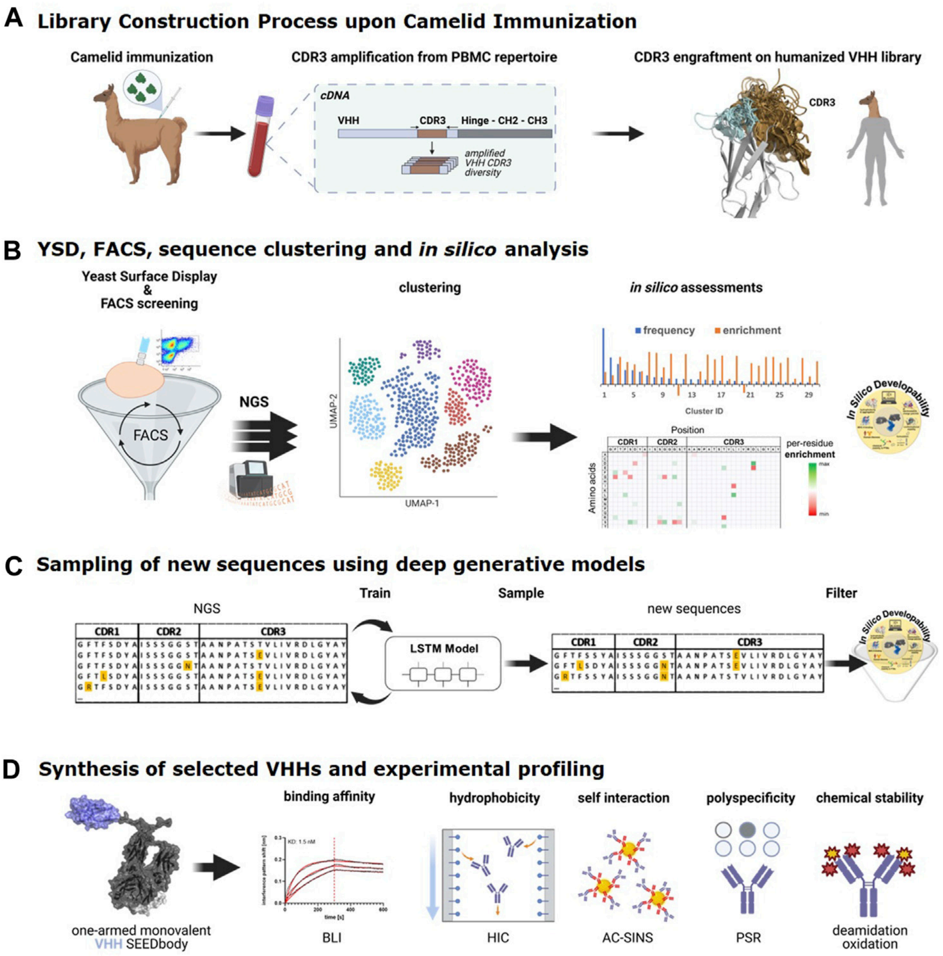


FIGURE 1
The end-to-end process consists of the following steps: **(A)**. Library construction process. VHH-derived CDR3 regions obtained from a llama, immunized against (rh) NKp46 are grafted onto a generic humanized and sequence-optimized VHH backbone library. **(B)**. Process of binder identification from Yeast Display Library based on multiple rounds of FACS and next-generation sequencing (NGS) analysis of sequence pools before and after FACS, followed by sequence clustering, per-cluster frequency and enrichment analyses in combination with *in silico* developability predictions to identify most interesting sequence clusters. **(C)**. Per-cluster LSTM deep generative model generation and sampling of new sequences that are subjected to *in silico* developability assessment to identify sequences for synthesis and experimental profiling. **(D)**. Selected VHH sequences are produced as one-armed monovalent SEEDbodies and experimentally characterized for binding against NKp46 and in early developability assays. (Figures partially created with BioRender.com).

TABLE 1 Summary of NGS data. VHH genes of screening samples were analyzed using MiSeq. Sequences were clustered based on 50% CDR3 sequence identity. Number of NGS reads are shown for all sequences and for those clusters that were selected for sampling of new sequences, antibody production and experimental profiling based on enrichment analysis and *in silico* developability assessment. Sequences obtained from FACS round 2 were used for LSTM deep generative model generation.

clusterID	NGS reads			Enrichment factor round 2 vs. round 0
	FACS round 0	FACS round 1	FACS round 2	
1	0	942	2,630	3,095
2	1	2,790	2,991	1760
3	36	4,964	4,147	132
4	888	8,573	11,954	16
ALL	887,881	1,138,880	754,669	

TABLE 2 Most occurring CDR1-3 sequences and binding affinity data against NKp46 of the five i) most enriched and five ii) most frequent clusters that were obtained from NGS of sequence pools obtained from YSD-FACS. Of note, two CDR3 clusters occurred in both sets. Hence a total of eight sequences were produced and tested. To visualize sequence diversity, amino acid differences to the most frequent residue in each position are shown in orange boxes.

cluster ranking		KD [nM]	CDR1-3 sequence																																			
most enriched	most frequent		CDR1								CDR2								CDR3																			
1		1.1	G	G	T	F	G	N	Y	A	I	S	R	G	G	G	S	T	A	A	A	G	G	M	G	S	T	T	V	V	V	S	T	I	P	Y	K	Y
2		8.6	G	F	T	F	S	S	Y	A	I	S	S	S	G	S	N	T	A	A	V	F	T	P	T	D	T	V	V	F	T	N	K	E	P	Y	N	Y
3		31.7	G	F	T	F	G	N	Y	A	I	S	S	S	G	D	S	T	A	A	V	E	A	D	S	S	E	V	V	F	L	S	P	H	I	Y	Q	Y
4	4	2.3	G	R	T	F	G	N	Y	A	I	S	R	S	G	G	S	T	A	A	N	P	A	T	S	T	V	-	L	I	V	R	D	L	G	Y	A	Y
5	5	1.7	G	F	T	F	S	S	Y	A	I	S	G	S	G	D	S	T	A	A	D	Q	Q	P	P	S	V	-	A	V	V	A	A	R	G	Y	R	Y
	3	3.1	G	F	T	F	S	D	Y	A	I	S	S	S	G	G	N	T	A	T	S	L	T	Y	D	Q	T	T	V	Y	V	S	P	L	A	Y	V	D
	2	4.0	G	R	T	L	S	S	Y	V	I	S	S	S	G	D	R	T	A	A	A	L	A	P	S	G	T	L	V	V	V	S	P	L	G	Y	T	Y
	1	19.8	G	R	T	F	S	S	Y	A	I	S	S	S	G	G	S	T	A	A	I	R	T	P	A	E	S	Q	V	I	V	T	L	D	W	Y	R	Y

sequences pools were projected into a two-dimensional space using UMAP (Becht et al., 2018) (Supplementary Figure S1). In addition, i) the per-residue frequency distributions of clones obtained after the second round of FACS and ii) the per-residue enrichment ratio through FACS enrichment rounds 1–2 were computed and analyzed, as shown in Figure 2 and Supplementary Figures S2–S4. Finally, for each cluster the 100 most frequent unique sequences obtained from FACS round 2 were subjected to *in silico* developability assessment using our previously described SUMO approach (Evers et al., 2023a). This method automatically generates structural VHH models from provided sequences, evaluates their human-likeness, and identifies potential surface-exposed chemical liabilities and post-translational modification motifs. Additionally, a small set of computed physico-chemical descriptors is reported, including the isoelectric point (pI), AggScore (Sankar et al., 2018) as predictor for hydrophobicity and aggregation tendency, and the positive patch energy of the CDRs. Analysis of sequence and predicted property data was used to assess the sequence spaces within each cluster regarding their potential to provide i) potent sequences, ii) broad sequence diversity and SAR information and iii) favorable *in silico* developability properties. We were particularly interested in selecting clusters with considerable sequence diversity to investigate how LSTM sampling could provide new sequence combinations to increase diversity and ideally improve affinity and/or developability properties. Based on these analyses, we picked four sequence clusters (termed cluster IDs 1–4 in the following) for LSTM based deep generative model generation and sampling of new sequences. The original data files used for sequence and *in silico* property analysis are provided in Supplementary Tables S1–S4 and illustrated for CDR cluster 3 in Supplementary Figures S5, S6.

LSTM model structure, training, sequence generation and scoring

As illustrated in Figure 1C, the LSTM model training and design was conducted using a recurrent network structure that has previously been successfully applied for the design of peptides [details in ref (Müller et al., 2018)]. LSTM models capture patterns in sequential data and generate new data instances from the learned distributions. Like their utility in peptide applications, the amino acid sequences of VHHs serve as appropriate inputs for these machine learning models. Since all sequences of the current study have identical framework regions, only the CDR1-3 sequences were concatenated and used for the training of LSTM models. For each of the four selected CDR3 sequence clusters, these CDR1-3 sequences (including all redundant sequences) from the second FACS round (Table 1) as determined by NGS were used for training. The best models were selected by evaluating the calculated validation losses on the left-out training datasets using a five-fold cross-validation approach (Supplementary Figure S7). Based on the learning distribution of the trained LSTM models, new sequences were sampled. We sampled 10,000 new sequences per cluster. These new sequences were combined with the original training sequences and ranked by their calculated *negative logarithm of likelihood* (NLL), a score that reflects the observed frequency of individual amino acids along the sequences of the training data sets (see Methods, Supplementary Figure

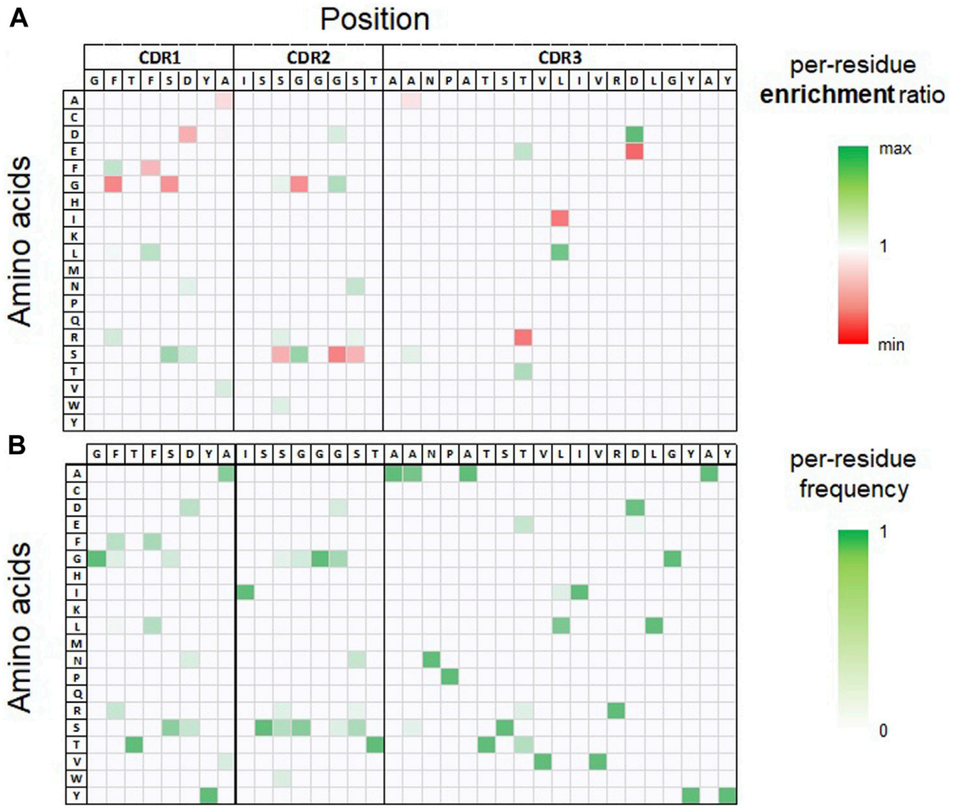


FIGURE 2
Per-residue enrichment and frequency analysis, both illustrated as heat-map for CDR3 sequence cluster 3. The table headers show the CDR1-3 sequence of the most frequent clone observed in the NGS data set after the second round of FACS selection within this cluster. **(A)**. Per-residue enrichment ratio over YSD-FACS rounds 1–2. Residues with a high enrichment (colored green) are observed with a higher relative frequency after FACS round 2 compared to round 1. **(B)**. Per-residue frequency distribution observed after FACS round 2.

S5 and Supplementary Tables S1–S4). The NLL score is not a predictor for binding affinity *per se*. However, since it reflects the sequence bias of amino acid distributions in the training data set sorted for favorable binding by FACS, it has been shown to represent a pragmatic score for selecting new sequences with an increased likelihood for high binding affinity (Saka et al., 2021).

In silico developability assessment to identify sequences for production and experimental profiling

Within each cluster, the top-ranked 100 non-redundant sequences obtained from LSTM sampling and NGS analysis were subjected to *in silico* developability assessment (see Supplementary Tables S1–S4) using our SUMO approach (Evers et al., 2023a). With the available sequences and their *in silico* profiles, the primary goal was to select a set of ≤10 sequences (for each cluster) for synthesis from which at least one sequence (per cluster) should be suited for further project progression after experimental profiling without the need for further iterative sequence optimization. For the nomination of these sequences, the following criteria were taken into account.

1. NLL scores: To assess the NLL's effectiveness in estimating binding affinities, we chose binders within each cluster with highly favorable scores, nominating at least three sequences from the top 100 scoring sequences. Additionally, we intentionally selected further sequences beyond the top 100 to cover a broad range of NLL scores, facilitating subsequent correlation analyses with experimental binding affinities.
2. *In silico* developability criteria: To minimize the risk of aggregation and non-specific binding, we selected sequences with computed aggregation propensity and positive charged CDR patch scores below defined cutoff scores. These cutoffs were set to the computed average scores plus one standard deviations over a data set of 79 marketed antibodies (see Table 4 legend). Additionally, as general de-risking approach, we intentionally picked sequence variants covering a certain pI range (Supplementary Table S5). The pI of an antibody/VHH can significantly impact various developability properties, such as solubility, aggregation during purification, virus inactivation (Jin et al., 2019), colloidal stability, viscosity in formulation (Kingsbury et al., 2020; Gupta et al., 2022), or non-specific binding or clearance (Ahmed et al., 2021; G et al., 2021). Small sequence modifications have been shown to improve colloidal stability and viscosity behavior (Kumar et al., 2018; Evers et al., 2019). Considering that the optimal pI for an

TABLE 3 CDR1-3 sequences of VHHs obtained from NGS analysis and AI/ML (LSTM) predictions. Sequences are grouped by their CDR3 cluster ID (50% SEQ-ID cutoff) with the most potent sequence at the top of each group. To visualize sequence and property relationships, amino acid differences to the most potent sequence within each group are shown in orange boxes. Residues that might theoretically be prone to chemical degradation are colored red (Asn deamidation, Asp isomerization, Met oxidation). In addition, the predicted NLL score and experimentally measured binding affinities (KD) as well as the k_{on} and k_{off} values are provided. NB: no binding.

ID	CDR3 cluster	source	KD [nM]	kon [1/Ms]	koff [1/s]	NLL	CDR1								CDR2								CDR3																			
1	1	AI/ML	5.3	3.2E + 05	1.7E – 03	4.9	G	R	T	F	S	N	Y	A	I	S	R	G	G	D	N	T	A	A	V	F	T	P	T	D	T	V	V	F	I	N	K	E	P	Y	N	Y
2	1	NGS	7.4	3.4E + 05	2.5E – 03	4.8	G	F	T	F	S	S	Y	A	I	S	S	S	G	S	N	T	A	A	V	F	T	P	T	D	T	V	V	F	T	N	K	G	P	Y	N	Y
3	1	AI/ML	8.1	6.9E + 04	5.6E – 04	6.2	G	R	T	F	S	S	Y	A	I	S	R	G	G	D	N	T	A	A	V	F	T	P	T	D	T	V	V	F	I	N	K	E	S	Y	N	Y
4	1	NGS	8.6	1.3E + 05	1.1E – 03	2.9	G	F	T	F	S	S	Y	A	I	S	S	S	G	S	N	T	A	A	V	F	T	P	T	D	T	V	V	F	T	N	K	E	P	Y	N	Y
5	1	NGS	9.4	1.9E + 05	1.8E – 03	2.5	G	F	T	F	S	S	Y	A	I	S	S	G	G	D	S	T	A	A	V	F	T	P	T	D	T	V	V	F	T	N	K	E	P	Y	N	Y
6	1	AI/ML	11.3	1.8E + 05	2.1E – 03	2.5	G	F	T	F	S	S	Y	A	I	S	S	S	G	G	S	T	A	A	V	F	T	P	T	D	T	V	V	F	T	N	K	E	P	Y	N	Y
7	1	NGS	11.7	2.3E + 05	2.7E – 03	2.8	G	R	T	F	S	S	Y	A	I	S	S	S	G	G	S	T	A	A	V	F	T	P	T	D	T	V	V	F	T	N	K	E	P	Y	N	Y
8	1	AI/ML	13.9	1.9E + 05	2.7E – 03	2.7	G	F	T	L	S	S	Y	A	I	S	S	G	G	G	S	T	A	A	V	F	T	P	T	D	T	V	V	F	T	N	K	E	P	Y	N	Y
9	1	AI/ML	21.9	5.4E + 04	1.2E – 03	20.8	G	G	T	F	S	I	Y	A	I	S	R	G	G	S	N	T	A	A	V	F	T	P	T	D	T	V	V	F	I	N	K	E	R	Y	N	Y
10	2	AI/ML	< 0.1	1.9E + 05	< 1.0E – 07	2.4	G	G	T	F	G	S	Y	A	I	S	R	S	G	G	S	T	A	A	A	G	G	M	G	S	T	T	V	V	V	S	T	I	P	Y	K	Y
11	2	NGS	0.8	2.3E + 05	2.0E – 04	2.3	G	G	T	F	S	S	Y	A	I	S	S	S	G	G	S	T	A	A	A	G	G	M	G	S	T	T	V	V	V	S	T	I	P	Y	K	Y
12	2	NGS	1.1	2.5E + 05	2.8E – 04	2.8	G	G	T	F	G	N	Y	A	I	S	R	G	G	G	S	T	A	A	A	G	G	M	G	S	T	T	V	V	V	S	T	I	P	Y	K	Y
13	2	NGS	1.3	2.0E + 05	2.7E – 04	2.3	G	G	T	F	S	S	Y	A	I	S	R	S	G	G	S	T	A	A	A	G	G	M	G	S	T	T	V	V	V	S	T	I	P	Y	K	Y
14	2	AI/ML	1.5	3.6E + 05	5.3E – 04	2.6	G	G	T	F	S	N	Y	A	I	S	S	S	G	G	S	T	A	A	A	G	G	M	G	S	T	T	V	V	V	S	T	I	P	Y	K	Y
15	2	NGS	2.3	1.1E + 05	2.6E – 04	7.1	G	R	T	F	G	S	Y	A	I	S	S	S	G	D	S	T	A	A	A	G	G	I	G	S	S	T	V	V	V	S	P	I	P	Y	A	Y
16	2	NGS	4.0	1.6E + 05	6.3E – 04	8.4	G	R	T	L	S	S	Y	V	I	S	S	S	G	D	R	T	A	A	A	L	A	P	S	G	T	L	V	V	V	S	P	L	G	Y	T	Y
17	2	AI/ML	4.4	1.0E + 05	4.5E – 04	2.8	G	G	T	F	G	N	Y	A	I	S	R	G	G	G	S	T	A	A	A	G	G	I	G	S	T	T	V	V	V	S	T	I	P	Y	K	Y
18	2	AI/ML	NB	NB	NB	17.8	G	G	T	F	G	N	Y	A	I	S	R	G	G	G	S	T	A	A	A	G	G	M	G	S	T	T	V	V	V	S	T	I	P	P	K	Y
19	2	AI/ML	NB	NB	NB	17.8	G	G	T	F	G	N	Y	A	I	S	R	G	G	G	S	T	A	A	A	G	G	M	G	S	T	T	E	V	V	S	T	I	P	Y	K	Y
20	3	AI/ML	2.1	1.3E + 05	2.8E – 04	3.6	G	G	T	F	S	D	A	A	I	S	R	S	G	D	S	T	A	A	N	P	A	T	S	E	V	L	I	V	R	D	L	G	Y	A	Y	
21	3	NGS	2.3	1.4E + 05	3.2E – 04	3.0	G	R	T	F	G	N	Y	A	I	S	R	S	G	G	S	T	A	A	N	P	A	T	S	T	V	L	I	V	R	D	L	G	Y	A	Y	
22	3	AI/ML	4.9	1.2E + 05	5.8E – 04	3.2	G	R	T	F	S	S	Y	A	I	S	S	S	G	G	N	T	A	A	N	P	A	T	S	T	V	L	I	V	R	D	L	G	Y	A	Y	
23	3	NGS	7.2	1.0E + 05	7.3E – 04	3.6	G	R	T	F	S	S	Y	A	I	S	S	G	G	G	N	T	A	A	N	P	A	T	S	T	V	L	I	V	R	D	L	G	Y	A	Y	
24	3	AI/ML	7.3	9.3E + 04	6.9E – 04	3.0	G	F	T	F	S	S	Y	A	I	S	S	S	G	G	S	T	A	A	N	P	A	T	S	E	V	L	I	V	R	D	L	G	Y	A	Y	
25	3	NGS	7.6	9.6E + 04	7.3E – 04	2.8	G	F	T	F	S	D	Y	A	I	S	S	S	G	G	S	T	A	A	N	P	A	T	S	T	V	L	I	V	R	D	L	G	Y	A	Y	
26	3	NGS	7.9	7.0E + 04	5.5E – 04	5.5	G	F	T	F	G	N	Y	A	I	S	R	S	G	S	S	T	A	A	N	P	A	T	S	R	V	I	I	V	R	D	L	G	Y	A	Y	
27	3	NGS	8.0	8.2E + 04	6.6E – 04	7.2	G	L	T	F	S	S	Y	A	I	S	G	S	G	D	N	T	A	A	N	P	A	T	S	R	V	I	I	V	R	E	L	G	Y	A	Y	
28	3	AI/ML	15.5	5.9E + 04	9.1E – 04	18.7	G	F	T	L	S	D	Y	V	I	S	S	S	G	G	N	T	A	A	N	E	A	T	S	E	V	L	I	V	R	D	L	G	Y	A	Y	
29	4	NGS	0.8	9.6E + 04	8.0E – 05	5.8	G	R	T	L	G	N	Y	A	I	S	W	G	G	S	R	T	A	T	S	L	T	Y	D	Q	T	T	V	Y	V	S	P	L	A	Y	V	D

(Continued on following page)

TABLE 3 (*Continued*) CDR1-3 sequences of VHHs obtained from NGS analysis and AI/ML (LSTM) predictions. Sequences are grouped by their CDR3 cluster ID (50% SEQ-ID cutoff) with the most potent sequence at the top of each group. To visualize sequence and property relationships, amino acid differences to the most potent sequence within each group are shown in orange boxes. Residues that might theoretically be prone to chemical degradation are colored red (Asn deamidation, Asp isomerization, Met oxidation). In addition, the predicted NLL score and experimentally measured binding affinities (KD) as well as the k_{on} and k_{off} values are provided. NB: no binding.

ID	CDR3 cluster	source	KD [nM]	kon [1/Ms]	koff [1/s]	NLL	CDR1								CDR2								CDR3																			
30	4	NGS	1.3	1.5E + 05	2.0E − 04	5.0	G	R	T	F	S	N	Y	A	I	S	G	G	G	G	N	T	A	T	S	L	T	Y	D	Q	T	T	V	Y	V	S	P	L	A	Y	G	D
31	4	AI/ML	2.0	5.4E + 05	1.1E − 03	3.5	G	R	T	L	S	N	Y	A	I	S	S	S	G	G	S	T	A	T	S	L	T	Y	D	Q	T	T	V	Y	V	S	P	L	A	Y	V	D
32	4	AI/ML	2.1	4.1E + 05	8.6E − 04	4.6	G	R	T	L	S	N	Y	A	I	S	S	G	G	S	N	T	A	T	S	L	T	Y	D	Q	T	T	V	Y	V	S	P	L	A	Y	V	D
33	4	AI/ML	2.6	3.6E + 05	9.2E − 04	3.6	G	R	T	F	S	S	Y	A	I	S	W	S	G	G	S	T	A	T	S	L	T	Y	D	Q	T	T	V	Y	V	S	P	L	A	Y	N	N
34	4	NGS	2.8	4.5E + 05	1.3E − 03	3.8	G	F	T	L	S	N	Y	A	I	S	S	S	G	D	S	T	A	T	S	L	T	Y	D	Q	T	T	V	Y	V	S	P	L	A	Y	V	D
35	4	NGS	3.1	2.9E + 05	9.2E − 04	3.9	G	F	T	F	S	D	Y	A	I	S	S	S	G	G	N	T	A	T	S	L	T	Y	D	Q	T	T	V	Y	V	S	P	L	A	Y	V	D
36	4	NGS	3.4	4.0E + 05	1.4E − 03	3.3	G	R	T	F	S	S	Y	A	I	S	S	S	G	G	S	T	A	T	S	L	T	Y	D	Q	T	T	V	Y	V	S	P	L	A	Y	V	D
37	4	NGS	4.7	6.0E + 04	2.8E − 04	4.3	G	F	T	F	S	S	Y	A	I	S	W	S	G	G	R	T	A	T	S	L	T	Y	D	Q	T	T	V	Y	V	S	P	L	A	Y	N	N

antibody drug product may vary depending on environmental factors, such as a solution or formulation pH, often not yet defined in the early project phase, selecting additional pI variants of a lead sequence provides potential backups for efficient project progression and de-risking.

3. Sequence diversity within each CDR3 cluster for SAR generation and chemical liability site elimination: Our humanized VHH library design strategy (Arras et al., 2023) omits N-glycosylation sites (Asn-X-Ser/Thr) and highly susceptible chemical liability sites (Asn-Gly, Asp-Gly, Met, Cys) in CDR1 or CDR2 (Table 3). However, such liabilities may still occur in CDR3, which is directly grafted from NKp46-immunized llama VHHs. Additionally, other theoretical chemical liability motifs (e.g., Asn-Ser, Asn-Asn, Asn-Thr; Asp-Ser, Asp-Asp, Asp-Thr, *etc.*) may be present in CDR1 or CDR2. These had not been excluded from library design, since degradation of these motifs occurs significantly less frequently based on internal and literature data (Lu et al., 2019) and are therefore assessed case-by-case, either by post-filtering based on more rigorous *in silico* liability assessments or by experimental profiling as exemplified below. As shown in Table 3, several selected sequences possess such "less severe" liability motifs. As part of our de-risking strategy, we intentionally selected sequence variants within each cluster where residues theoretically prone to chemical degradation (e.g., Asn, Asp, Met) are replaced by chemically non-reactive residues (e.g., sequences 15–17, where a Met residue in CDR3 is replaced by Ile).
4. Finally, we ensured that for all four clusters, sequences were selected from both the NGS output and LSTM sampled sequences to assess, through experimental profiling, the extend to which LSTM sampling provided additional or improved "chemical matter".

Table 3 and Supplementary Table S5 display the CDR1-3 sequences that were ultimately selected, along with their computed developability properties. For the specific rationale behind selecting each sequence for synthesis and experimental profiling, please refer to Supplementary Tables S1–S4. As shown in Supplementary Table S5 and Figure 3, due to our humanized VHH library design strategy all selected sequences show a high human-likeness in the framework region of 91.3%. Furthermore, due to our selection strategy, no sequence shows pronounced computed aggregation propensity or positive charged patches in the CDRs. However, as intended by the selection criteria, the sequences cover a certain diversity in NLL scores, pI, sequence diversity and chemical liability motifs.

NGS and AI/ML derived sequences display high-affinity antigen binding and favorable early developability properties

As illustrated in Figure 1D, the selected sequences (Table 3) were utilized to synthesize one-armed, monovalent paratope-Fc fusion constructs as described previously (Klausz et al., 2022; Lipinski et al., 2023) to exclude avidity-related interactions that might enhance apparent binding affinity (Vauquelin and Charlton, 2013). For this, we utilized the strand-exchanged engineered domain (SEED) technology for Fc heterodimerization (Davis et al., 2010).

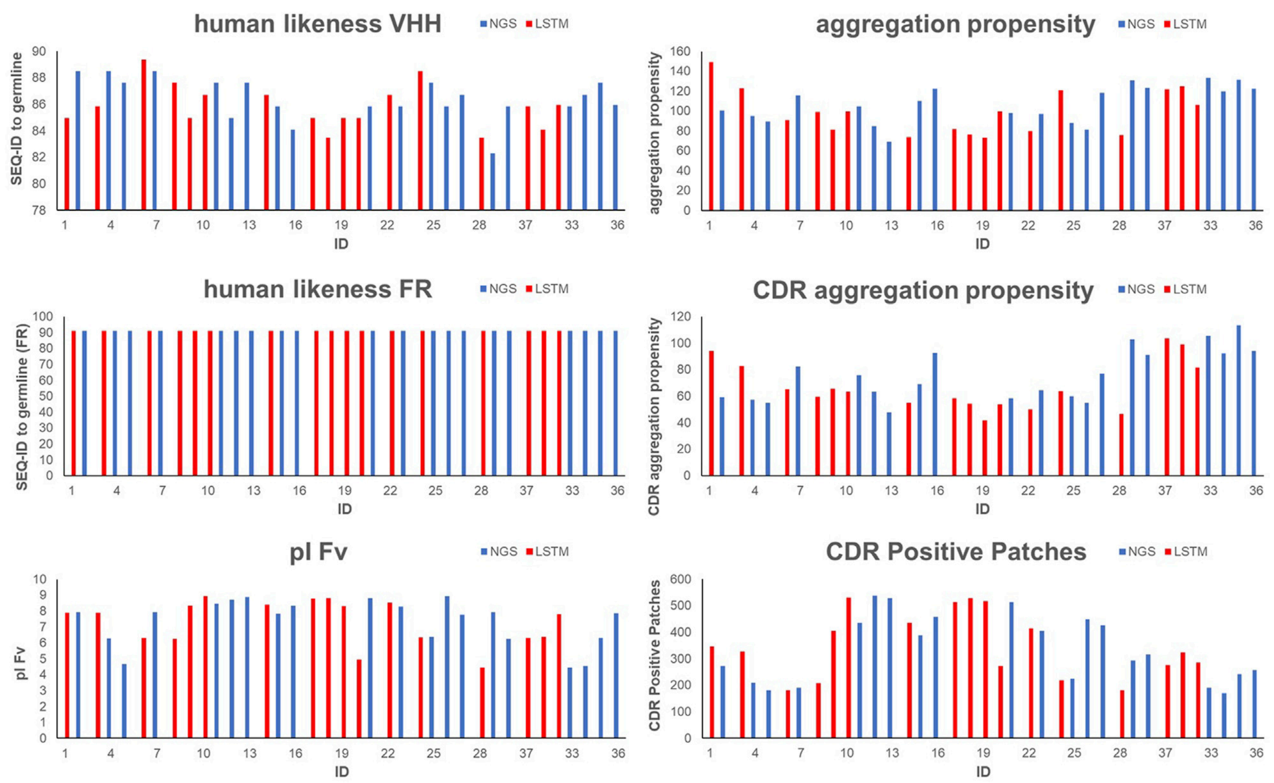


FIGURE 3
Graphical visualization of *in silico* properties for VHH domains that were selected for synthesis and experimental profiling. Blue bars indicate sequences obtained from NGS, red bars indicate sequences obtained from AI/ML (LSTM) sampling.

Production was performed in ExpiCHO™ cells at a scale of 5 mL for experimental profiling. Expression yields were in the double-digit milligram-per-liter scale for most sequences, indicating adequate productivities for transient expression (Table 4). Furthermore, aggregation propensities as determined by analytical size-exclusion chromatography (SEC) post protein A purification indicated favorable biophysical properties for most sequences (Table 4). Binding experiments utilizing bio-layer interferometry (BLI) at varying (rh) NKp46 concentrations revealed specific antigen binding of the vast majority of tested VHHs from both approaches, NGS and AI/ML, respectively (Table 3; Figure 4). Encouragingly, within each sequence cluster, we obtained multiple sequences binding in the 1-digit nanomolar or even sub-nanomolar range to (rh) NKp46 (Table 3). Notably, although the affinity improvements are not significant, for three of the four sequence clusters, the most potent binder was obtained from the LSTM-predicted sequences, suggesting that the deep generative model approach can propose improved sequences in terms of binding affinities within the sequence space spanned by the NGS data set. Analysis of the NLL scores do not show a linear correlation to the experimentally observed binding affinities. However, within this specific dataset, high predicted (*i.e.*, unfavorable) NLL scores qualitatively translated to low or no detectable affinities, suggesting the use of more stringent NLL cutoff scores in future studies to eliminate true negatives from the list of candidates to be synthesized.

To experimentally assess early developability properties (Table 4; Figure 5), we exploited analytical size-exclusion chromatography (SEC) after protein A purification as a first filter. Generally, purities above 85% target peak are considered as adequate attributes for transient antibody expression, while purities of more than 90% indicate favorable properties. Overall, most sequences showed a high target purity above 90%. As additional early developability attribute we also scrutinized one-armed VHH SEEDbodies using analytical hydrophobic interaction chromatography (HIC) assuming that a low overall hydrophobicity would contribute to a good developability profile. For this, we utilized two marketed therapeutic antibodies as assay controls, cetuximab and avelumab, with HIC retention times of 5.8 min and 7.2 min, respectively. Overall, HIC retention times of the vast majority of VHH SEEDbodies were in the lower favorable range. In this respect most molecules displayed even shorter retention times compared to cetuximab, indicating a beneficial (low) relative hydrophobicity of the VHH domains. Only variants of CDR3 cluster 4 (IDs 30–37) showed retention times in the range of 6.0–6.7 min that are in between the ones from cetuximab and avelumab. Notably, although there is no ideal linear correlation between HIC retention times (Table 4) and computed aggregation propensities (Supplementary Table S5 and Figure 6), these *in silico* scores are (in agreement with their higher retention times) on average higher for IDs 30–37 (cluster 4) compared to the other sequences; supporting their usefulness for early *in silico* ranking and filtering of sequences. The observed degree of correlation between predicted and experimental hydrophobicity is in agreement with a recent systematic study on antibody structures

TABLE 4 Analytical and early developability data for selected one-armed VHH SEEDbodies and antibody controls, including amount of protein, SEC Purity, mean T_{onset}, HIC retention time, AC-SINS and PSR-BLI.

ID	source	amount of protein [mg/L]	SEC Purity [%]	Mean Tonset [°C]	HIC tR [min]	AC-SINS [Δλmax (nm)]	PSR/ BLI
1	AI/ML	49.1	92.7	59.1	4.9	-0.705	-0.011
2	NGS	51.8	90.9	59.2	5.0	-0.076	0.036
3	AI/ML	33.7	89.4	58.4	4.9	-1.189	0.004
4	NGS	28.0	96.9	58.1	4.9		
5	NGS	23.8	94.5	59.8	5.1	-0.550	0.016
6	AI/ML	25.2	96.5	58.1	5.1		
7	NGS	29.4	91.4	58.4	4.9	-0.570	0.030
8	AI/ML	32.2	95.0	59.8	5.3	-0.596	0.037
9	AI/ML	43.5	83.1	59.4	5.0	-0.550	0.005
10	AI/ML	26.5	97.1	59.1	4.8	-0.604	-0.031
11	NGS	23.7	97.3	58.9	4.8	-0.516	-0.012
12	NGS	18.1	100.0	58.0	4.8		
13	NGS	22.3	97.1	58.7	4.9	-0.578	0.021
14	AI/ML	25.1	99.2	57.4	4.9		
15	NGS	25.1	100.0	59.0	6.4	3.296	-0.011
16	NGS	19.5	98.8	57.4	5.4		
17	AI/ML	20.9	100.0	58.8	5.3	-0.497	0.007
18	AI/ML	50.9	98.8	58.7	5.2	-0.343	0.015
19	AI/ML	46.0	96.2	58.6	4.7	-0.548	0.015
20	AI/ML	36.8	99.1	56.3	5.0		
21	NGS	25.1	99.2	56.4	5.0		
22	AI/ML	37.7	98.3	58.2	5.2	-0.434	0.016
23	NGS	47.4	98.7	58.6	5.2	-0.504	0.035
24	AI/ML	34.9	97.3	58.5	5.1	-0.617	0.017
25	NGS	43.2	97.9	58.3	5.1	-0.511	0.024
26	NGS	30.7	97.4	58.0	5.0	-0.119	0.014
27	NGS	37.7	96.4	58.6	4.9	-0.310	0.039
28	AI/ML	22.3	82.2	58.5	5.0	-0.526	0.049
29	NGS	24.8	99.0	58.9	4.8	-0.395	0.019
30	NGS	37.2	100.0	58.4	6.0	-0.671	-0.020
31	AI/ML	133.2	99.1	58.7	6.4	-0.534	-0.011
32	AI/ML	71.6	100.0	59.0	6.4	-0.605	0.000
33	AI/ML	22.2	100.0	57.8	6.4		
34	NGS	33.1	100.0	58.1	6.3	-0.307	0.018
35	NGS	26.2	94.4	57.4	6.3		
36	NGS	48.2	97.8	58.6	6.7	-0.631	-0.009
37	NGS	22.2	97.1	58.7	6.5	-0.476	0.030

(Continued on following page)

TABLE 4 (*Continued*) Analytical and early developability data for selected one-armed VHH SEEDbodies and antibody controls, including amount of protein, SEC Purity, mean T_{onset}, HIC retention time, AC-SINS and PSR-BLI.

ID	source	amount of protein [mg/L]	SEC Purity [%]	Mean Tonset [°C]	HIC tR [min]	AC-SINS [Δλmax (nm)]	PSR/ BLI
Trastuzumab						-0.001	0.009
Briakinumab				63.6		25.961	0.115
Avelumab					7.2		
Cetuximab					5.8		

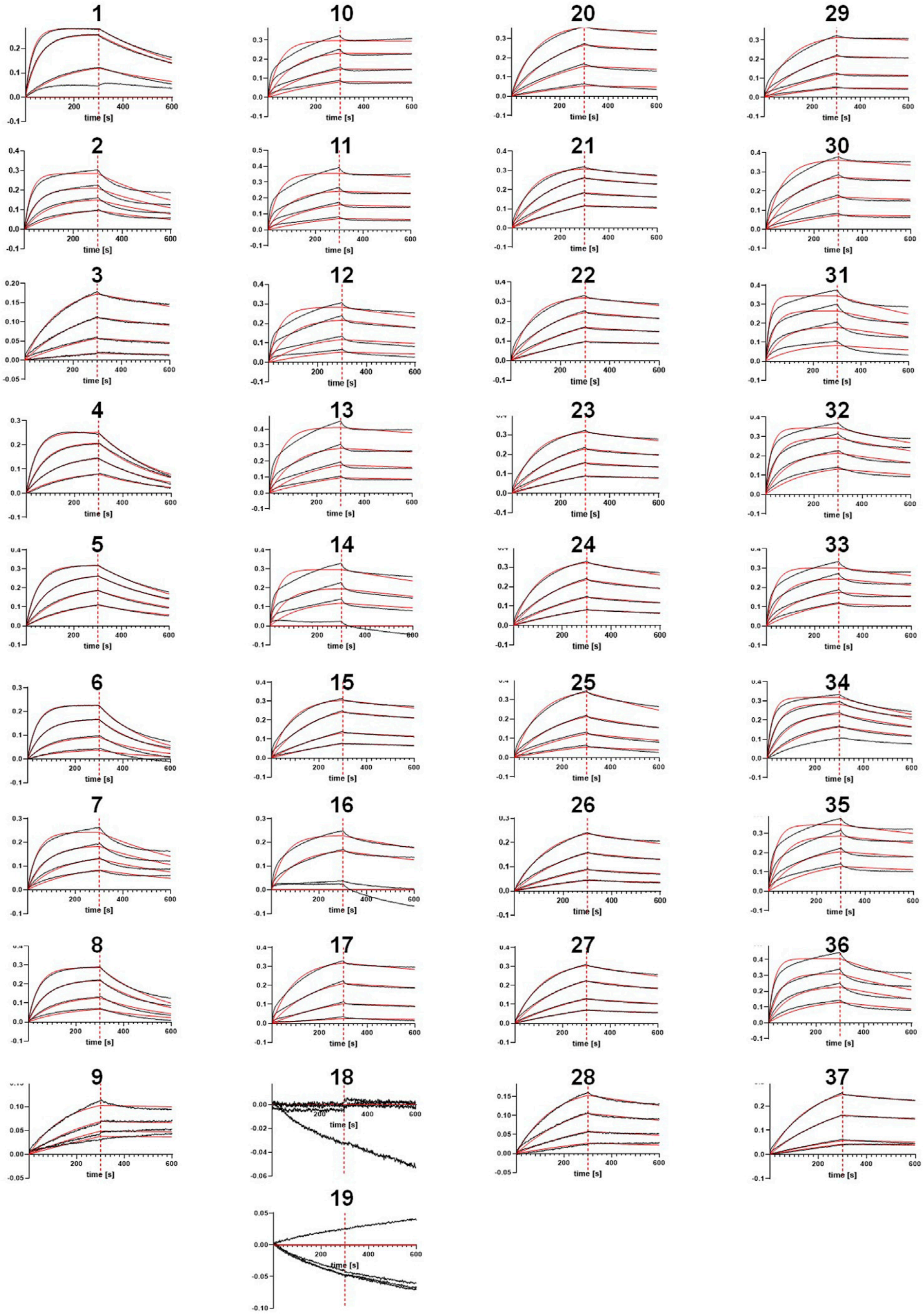


FIGURE 4
Bio-Layer Interferometry (BLI) curves (in black) and fitting curves (in red) obtained for all sequences.

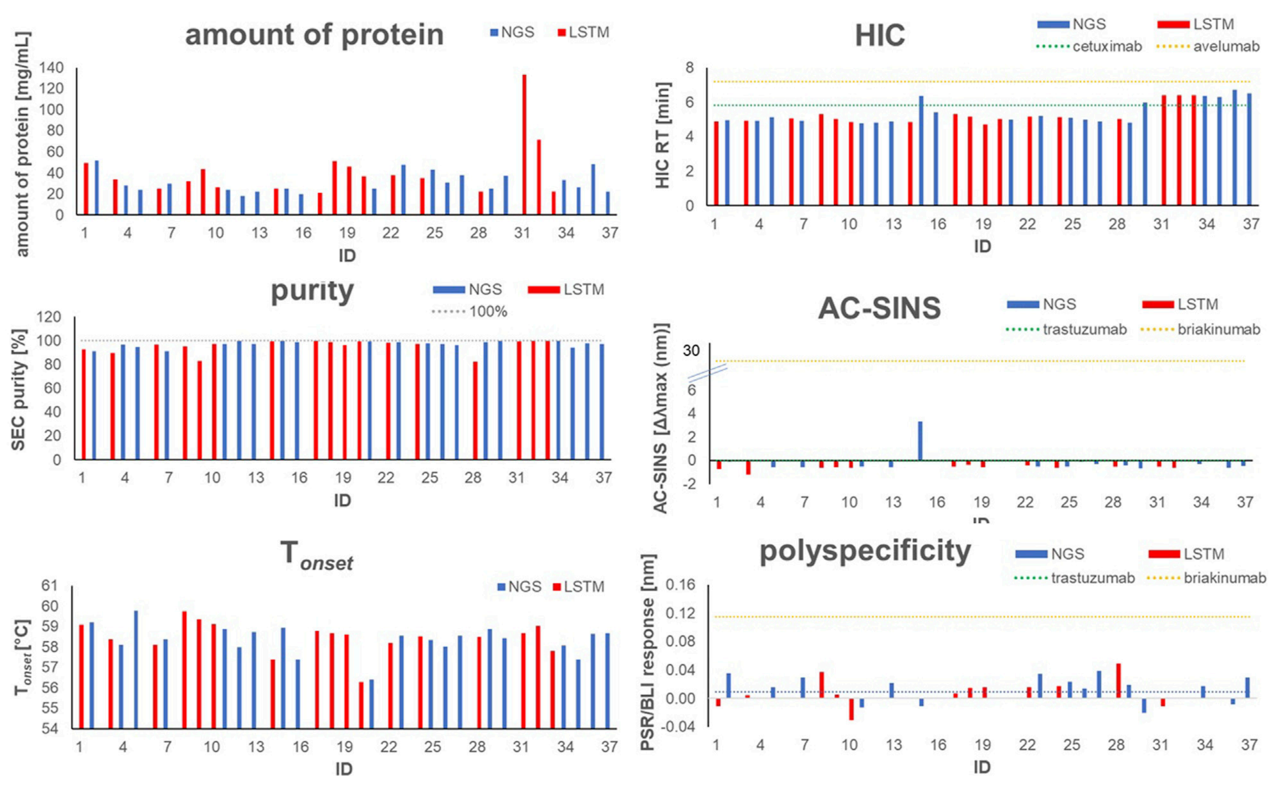


FIGURE 5
Graphical visualization of experimental analytical and early developability data for selected one-armed VHH SEEDbodies and antibody controls, including amount of protein, SEC Purity, mean T_{onset}, HIC retention time, AC-SINS and polyspecificity (PSR-BLI). Blue bars indicate sequences obtained from NGS, red bars indicate sequences obtained from AI/ML (LSTM) sampling.

(Waibl et al., 2022). Based on that study, prediction accuracy for HIC retention scales might be further improved by i) exploring alternative approaches for 3D model generation and by i) using hydrophobicity scales derived from experimental HIC data.

To further investigate the biophysical properties of the herein identified VHHs, we checked the thermostability of the molecules by nanoDSF. The T_{onset} of a dedicated molecule represents the temperature where the variable domain of a VHH construct starts to unfold while applying a temperature gradient and as such, is an indicator of its thermostability in a certain buffer and pH environment. The T_{onsets} we measured were in the range between 56˚C and 59 ˚C for all tested molecules, representing an overall adequate thermostability for further development (Mieczkowski et al., 2023). As obvious from Figure 5, no significant differences in T_{onset} are observed between the sequences obtained from NGS and LSTM sampling, supporting the claim that LSTM is capable of correctly modeling long-term dependencies and capturing relationships between amino acids that determine structure and function. Additionally, we evaluated available VHH SEEDbodies (that were selected based on remaining substance availability) in affinity-capture self-interaction nanoparticle spectroscopy (AC-SINS) as early experimental predictor for colloidal stability (Liu et al., 2014). Clinical antibody trastuzumab was used as assay control indicating favorable biophysical properties with mean Δλmax values of ~0.2 nm after subtraction of buffer blanks. Final AC-SINS scores for the tested VHH SEEDbodies were calculated via subtraction of blank and trastuzumab scores (Table 4). The calculated scores indicate favorable colloidal stability properties for all tested SEEDbodies, very similar to trastuzumab and significantly better compared to briakinumab, which was used as reference with a known propensity for self-interaction (Jain et al., 2017). As further early developability assessment, the selected SEEDbodies were evaluated in the polyspecificity reagent (PSR) assay which provides insights into the general off-target interactions/specificity and selectivity of the VHH domains, again using trastuzumab as indicator for reduced unspecific interactions and briakinumab reference indicating more pronounced polyspecificity (Table 4). Compared to these assay controls, no SEEDbody shows pronounced non-specific binding.

Although we have to keep in mind that the monospecific IgG1 control antibodies might not be ideal references for benchmarking our one-armed VHH SEEDbodies, the available experimental data indicate favorable intrinsic developability properties for the VHH domains.

To experimentally assess the risk for the formation of chemical degradation products along the drug development process, which might potentially affect its efficacy and safety, one potent sequence from each of the CDR3 clusters was subjected to forced oxidation and deamidation studies (Nowak et al., 2017) (Table 5; see Materials and Methods for experimental details and Supplementary Table S6 for detailed experimental results). Within the CDR regions of the four selected sequences, we could only observe significant

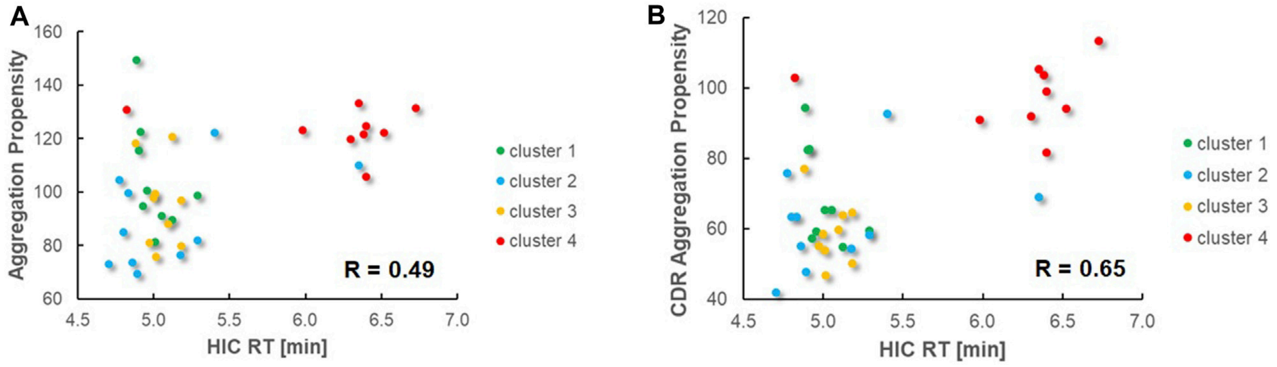


FIGURE 6
Comparison of predicted aggregation propensities vs. experimental HIC retention times and Pearson correlation values. Sequences from different clusters are shown in different colors. **(A)**. Predicted aggregation propensities based on the entire variable VHH regions. **(B)**. Predicted aggregation propensities based on the CDR regions only.

deamidation within CDR1 of sequence **1**, attributed to the Asn-Tyr sequence motif. This non-canonical motif is generally known as non-highly susceptible to deamidation (Lu et al., 2019), but in the present case this chemical liability is a potential critical quality attribute (CQA) that would require additional efforts for monitoring and control in the development process. The SAR data shown in Table 3 demonstrate that several alternative sequence variants with similar potency are available, which are devoid of this chemical liability motif and might be selected as alternative optimized hits. This example illustrates the benefit that the explicit selection of sequence variants within specific CDR3 clusters provide valuable SAR data that do not only point to mutations that finetune binding affinity, but also to optimize the physico-chemical property profiles (regarding chemical liabilities, PTMs, electrostatic and hydrophobic properties).

AI/ML derived sequences fill gaps within the sequence space spanned by NGS data

The experimental results demonstrate that several optimized hit sequences were obtained within each cluster, suitable for further project progression, including experimental characterization in functional assays, early formulation studies, and/or *in vivo* experiments. These sequences were derived from both the NGS data and the LSTM sampled sequences. To investigate the benefit of LSTM sampling, we analyzed the number and diversity of additional unique sequences designed by LSTM in comparison to the NGS sequences. Our analysis focused on the top-ranked 100 NLL scorers within each CDR3 cluster, since all tested variants from these lists showed favorable binding affinities (Supplementary Tables S1–S4). As illustrated in a UMAP dimension reduction based on sequence diversity, the LSTM approach generated a considerable number of new sequence combinations, effectively filling gaps within the sequence space spanned by the NGS dataset (see Figure 7 and the underlying sequences in Supplementary Tables S1–S4), thereby increasing not only the number of potent sequences but also the likelihood of including variants with lower risks of chemical degradation or post-translational modification motifs. For CDR3 cluster 1, 41 of the top 100 sequences were obtained from LSTM (cluster 2: 23, cluster 3: 45, cluster 4: 19). The predicted physical properties (pI, hydrophobicity/aggregation propensity, CDR Positive Patches) of the LSTM sampled sequences covered a similar range and diversity as those obtained from NGS (see Supplementary Figure S8). Moreover, a comparative inspection of production yield, melting temperatures, and other biophysical properties (Figure 5) between the LSTM and NGS-derived sequences that had been synthesized did not reveal any significant differences. This finding supports the claim that LSTM sampling can enrich the pool of NGS sequences with additional potent and developable binders, which increases the overall chance of discovering optimized hits with favorable developability profiles.

Discussion

In the past, the discovery and optimization of antibodies and VHHs were predominantly reactive in nature (Evers et al., 2023b): Traditional screening methods were used to obtain antibody or VHH sequences, which were subsequently sequence-optimized with regards to factors such as binding affinity, human-like characteristics, and chemical stability. Following the identification of the top-performing optimized hits, developability assessments were carried out. These assessments, since conducted after sequence optimization, aimed to identify any suboptimal developability characteristics, such as aggregation, low solubility, poor expression, non-specific binding, or unfavorable pharmacokinetic properties. Consequently, issues arising from these suboptimal properties were passed on to downstream functions, e.g., Drug Metabolism and Pharmacokinetics (DMPK), non-clinical safety, and Chemistry, Manufacturing and Controls (CMC) to adjust and optimize downstream process development and dosing regimens, thereby often imposing delays in development, increased costs and finally a considerable risk for the project to achieve approval for First in Human and further clinical studies (Evers et al., 2023b). To mitigate these risks, in this work we propose

TABLE 5 Deamidation and oxidation modifications observed within CDR1-3 in accelerated oxidation and deamidation studies, shown as % modified species after 24 h vs. 0 h. CDR residues that are typically prone to Asn-deamidation N) or Met-oxidation (M) are colored in red. ND: no degradation detected.

ID	CDR1								CDR2								CDR3																				CDR1% modified	CDR2% modified	CDR3% modified
1	G	R	T	F	S	N	Y	A	I	S	R	G	G	D	N	T	A	A	V	F	T	P	T	D	T	V	V	F	I	N	K	E	P	Y	N	Y	18.7	<1	ND
10	G	G	T	F	G	S	Y	A	I	S	R	S	G	G	S	T	A	A	A	G	G	M	G	S	T	T	V	V	V	S	T	I	P	Y	K	Y			ND
22	G	R	T	F	S	S	Y	A	I	S	S	S	G	G	N	T	A	A	N	P	A	T	S	T	V	L	I	V	R	D	L	G	Y	A	Y			<1	<1
30	G	R	T	F	S	N	Y	A	I	S	G	G	G	G	N	T	A	T	S	L	T	Y	D	Q	T	T	V	Y	V	S	P	L	A	Y	G	D	<1	<1	

an integrated and efficient *de novo* design strategy comprising camelid immunization, library generation, YSD, FACS, NGS analysis, AI/ML methods, *in silico* developability assessment as well as synthesis and early experimental characterization of the selected sequences. In an ideal scenario, these subsequent steps can be accomplished in less than 4 months without the need for subsequent time-consuming steps of iterative sequence optimization. This comprehensive approach was successfully applied for an early drug discovery project to generate automatically humanized and sequence optimized VHH binders against NKp46 with favorable early developability profiles.

The *in silico* steps described in this study are computationally inexpensive (<1 week in this study) and can be combined into a fully automated workflow. Furthermore, our process of CDR3 engraftment upon camelid immunization onto a generic humanized and sequence-optimized scaffold library is characterized by its low complexity and duration (<1 week). Besides camelid VHH library generation, we have established a similar CDR grafting approach for the generation of ultralong CDR-H3 antibodies following the immunization of cattle (Pekar et al., 2021). Since finally NGS is meanwhile quick and cost-effective, the herein described combination of experimental and *in silico* approaches represent a general strategy for a fast and efficient hit discovery and optimization upon camelid immunization. An alternative option that bypasses animal immunization and thereby can even further accelerate the *de novo* identification of developable antibodies or VHHs is the screen of diverse synthetic libraries that were tailored towards human-likeness and favorable physico-chemical properties (Teixeira et al., 2021; Khetan et al., 2022; Evers et al., 2023b). Binders obtained from antibody selections and NGS analysis of such diverse libraries might further be optimized towards improved binding and developability applying AI/ML approaches as described in the present study. As recently discussed (Gray et al., 2020; Gray et al., 2020; Custers and Steyaert, 2020; Laustsen et al., 2021), both animal immunization and synthetic library technologies have their own benefits and drawbacks for antibody discovery. For example, while synthetic libraries bypass the need of animal immunization, the immune system of animals has evolved over millions of years to efficiently produce highly specific antibodies against a diverse range of antigens. The semi-immune/semi-synthetic procedure presented in this study combines the advantages of both technologies and is coupled with the benefits of NGS and AI/ML approaches for rapid and efficient antibody discovery and optimization (Laustsen et al., 2021).

In this study, we opted for a LSTM, a recurrent neural network (RNN) architecture, as the basis of sequence prediction models based on NGS data. The selection of this approach was based on the fact that it has been successfully applied to diverse modalities (Saka et al., 2021; Müller et al., 2018; Gupta et al., 2018; Merk et al., 2018; Segler et al., 2018; Z et al., 2022) and that the code was already available (Müller et al., 2018). From a scientific perspective, LSTM models are known for their capability to learn complex patterns and dependencies within sequences. Therefore, by training on existing protein sequences from NGS data, the LSTM can capture essential structural and functional motifs present in the library, potentially generating new functional sequence combinations not observed in the NGS dataset. The experimental data from the present study confirmed that the LSTM sampled sequences did not exhibit

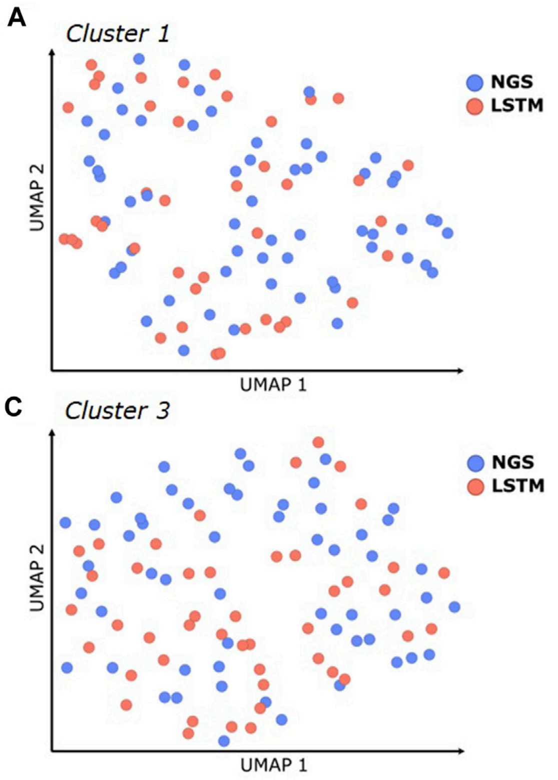


FIGURE 7
Similarity of CDR1-3 sequences within the best 100 scoring sequences (based on their NLL) for each CDR3 sequence cluster **(A–D)**, illustrated using UMAP dimensionality reduction. Blue dots represent sequences that were obtained from NGS, red dots represent new sequence combinations that were automatically designed with LSTM.

significant disadvantages compared to the NGS-derived sequences in terms of production yield, melting temperatures, or binding affinities. Various other ML approaches have also demonstrated effectiveness for the identification of complex patterns from sequence input data and were successfully employed for antibody design based on NGS data (Liu et al., 2020; Mason et al., 2021; Makowski et al., 2022; Hu et al., 2023; Li et al., 2023; Parkinson et al., 2023). Furthermore, additional deep generative modelling methods such as variational auto-encoders (VAEs) and generative adversarial networks (GANs) may also be explored to optimize sequence spaces obtained from NGS data (Akbar et al., 2022).

LSTM sampling efficiently filled diversity gaps in the sequence space beyond what is covered by the NGS training data (Figure 7). However, since the present LSTM approach uses one-hot amino acid encoding, it will generate new sequence combinations that only interpolate within the sequence space covered by the NGS data. Therefore, another aspect that might be further investigated is the representation of amino acids in the context of *in silico* sequence processing. Most approaches utilize one-hot encoding, which does not capture structural features, inherent relationships, or the physicochemical similarities between amino acids. Several alternative encoding schemes, such as amino acid embeddings, physicochemical descriptors or position-specific scoring matrices (PSSMs) might be suited to increase the model's ability to extrapolate into new sequence spaces.

Another crucial aspect for AI/ML based prediction and identification of improved binders is the scoring function used to rank the sequences based on their assumed binding affinity against the target. In this study, we utilized NLL that assumes a correlation of binding affinity with the observed amino acid distribution in the NGS set of sequences obtained after FACS. Notably, the majority of synthesized VHH constructs (>80%) exhibited binding affinities in the (sub-)1-digit nanomolar range. Therefore, based on the limited experimental data from this study, we consider the NLL ranking as the suited criterion for selecting sequences with a high likelihood of binding. For a more comprehensive conclusion, future systematic studies would be required to explore correlations with other scoring functions for identifying high-affinity binders. However, such analyses would necessitate a large dataset of sequences with experimental binding affinity data.

Recent studies have already shown the successful application of AI/ML techniques on antibody NGS data to design new sequences with even further improved potency or developability (Liu et al., 2020; Mason et al., 2021; Saka et al., 2021; Makowski et al., 2022; Hie et al., 2023; Parkinson et al., 2023). While these studies focused on optimizing previously identified antibody candidates through sequence diversification and library generation, the present study represents, to the best of our knowledge, the first prospective application of AI/ML for the *de novo* identification of diverse, potent, and developable VHHs. In contrast to these previous studies, our approach was applied on a humanized library that originated from a highly diverse camelid repertoire upon immunization.

To validate the efficacy of our approach, we conducted experimental profiling to assess binding affinity and developability for multiple sequences per cluster and gained valuable SAR and SPR information directly from the initial set of synthesized variants. This procedure mirrors the well-established "hit-triaging" approach for small molecules obtained from high-throughput screens, where multiple molecules within different chemical series are evaluated to identify the most promising candidates for further development (Kitchen and Decornez, 2015). As an advantage, this procedure can directly point to lead molecules and backups without the need for additional time-consuming sequence optimization cycles.

The present study represents a first successful application of our integrated VHH discovery approach on NKp46 as specific target. Further ongoing and future studies on internal projects will demonstrate the robustness of this process and certainly point to aspects that may be further optimized, e.g., regarding the design of a follow up humanized VHH scaffold library (Arras et al., 2023), *in silico* property predictors and further aspects as described above.

Finally, the findings and results of this study should also be considered in the light of some limitations and inspirations for further future studies (Jin et al., 2023). In the present study, we applied a CDR3 sequence identity cutoff of 50% for sequence clustering as a compromise to find i) sequences within one cluster that all bind in a similar mode to the same epitope and ii) at the same time provide sufficient sequence diversity for SAR analysis and automated multi-parameter sequence optimization. It is generally known that similar protein sequences have similar folds (Baker and Sali, 2001). However, if this is also true for CDR3 loops and whether the 50% cutoff is the most ideal cutoff for this purpose will require additional dedicated studies (Könning et al., 2017). One might question the general need for LSTM sampling if the sequences obtained from NGS analysis of the semi-immune/semi-synthetic strategy are already "good" enough. The present study demonstrates that i) high affinity binders with favorable early developability profiles can already be obtained from data mining of the available NGS data, but in addition ii) that LSTM sampling is able to fill sequence gaps with additional potent and developable sequences that have not obtained from the NGS data. The timeframe for LSTM model generation and sequence sampling (<1 day in the present study) is negligible in the context of a standard hit discovery campaign. Therefore, our general recommendation is to add the LSTM-based designs alongside NGS-derived sequences. Then, select the best binders from the combined pool based on their predicted likelihood of binding and relevant *in silico* developability parameters, aligned with the specific target product profile (TPP) of the project. This approach enhances the overall project success probability (Krah et al., 2016). To ensure proper assay controls for early experimental developability assessments, we used four well-characterized monospecific IgG1s (avelumab, cetuximab, trastuzumab, and briakinumab) as references. While these control sequences allow assay comparisons across different studies, they may not serve as ideal benchmarks for drawing final conclusions about the general developability of our VHHs, since we fused them to SEED Fc domains that show considerable sequence differences to IgG1 Fc domains. As a conclusion, the data presented in this study only indicate favorable intrinsic developability properties for the VHHs generated here. Further in-depth studies, including the identification and use of specific VHH-based controls for benchmarking, will be necessary to assess how these developability properties extend to different multi-specific architectures (Bannas et al., 2017; Chanier and Chames, 2019; Pekar et al., 2020; Yanakieva et al., 2022; Lipinski et al., 2023; Wang et al., 2022) Quality of NGS data is critical for any AI prediction tool, as it forms the basis for training. In this study, we used NGS data obtained from different round of FACS. As we learned through the course of the study, sample preparation, read depth, sequence complexity and sequencing error rates can significantly impact the results. The rate of enrichment over FACS round 2 vs. round 0 was used as an essential parameter for nominating sequence clusters, but this enrichment was biased due to the low number of reads in round 0, and the final selection might have varied based on variations in NGS data generation and analysis. Nevertheless, the reads used for LSTM sampling after FACS round 2 were sufficiently broad and frequent to discover potent binders with favorable early developability profiles.

In conclusion, the herein presented workflow comprising a combination of AI/ML methods, camelid immunization, library generation, NGS analysis, and *in silico* developability assessment can identify potent VHH binders with promising early developability profiles. This singular procedure mitigates the need for subsequent sequence optimization, thereby offering the potential to significantly accelerate hit discovery and optimization and at the same time to reduce the risk for developability-related attrition in the downstream process.

Materials and Methods

NGS, sequence clustering and ranking

To prepare RNA material for NGS analysis, two defined antisense primer sequences were used which specifically aligned with nucleotides in the upper hinge regions of camelid IgG2 and IgG3 antibody isotypes, facilitating directed cDNA synthesis. Within a subsequent PCR utilizing index primers for Illumina sequencing, the VHH sequences were amplified and tagged. For the samples derived from the VHH diversities embedded in the plasmid vector system, the sequences processed accordingly, but lacking the cDNA synthesis step. During the DNA amplification process, the AMPure system (Beckman Coulter) was used to purify the VHH amplicons, while for the purification of the final sequencing library a Pippin Prep (Sage Science) was used. For sequencing purposes, a MiSeq (Illumina) device with the v3 600 cycle kit according to the manufacturer's protocol was employed. Resulting FASTQ files were uploaded to Geneious Biologics (https://www.geneious.com/biopharma) for analysis and annotation. Reads were overlapped, filtered for length, and the VHH sequences were annotated using the *Lama glama* reference library. Normalized counts for each CDR3 were used to identify sequences that were enriched in the sorted samples relative to the baseline diversity.

Sequences were clustered based on 50% CDR3 sequence identity. All sequence clusters were assessed and ranked by their i) NGS counts after the second FACS round and their ii) enrichment ("Fold Change") over round 0 to 2. The enrichment factor EF ("Fold Change") was calculated according to the following formula:

$$EF = \frac{\left(\frac{N_{cluster}+1}{N_{total}+1}\right)_{S2}}{\left(\frac{N_{cluster}+1}{N_{total}+1}\right)_{S0}}$$

Where N represents the number of reads within the specific cluster and S0, S2 represent the FACS selection round.

LSTM model structure, training and sampling

The code from Müller et al. (Müller et al., 2018) (https://github.com/alexarnimueller/LSTM_peptides) has been used and slightly adapted to constrain the input training sequence length to the length of CDR1-CDR2-CDR3 output sequences of the individual clusters. The adapted code and the sequences used as input for training and sampling of new sequences are available from https://github.com/MCompChem/LSTM_CDRs. The input sequences had been exported from Geneious Biologics as csv file and used as input sequences without further preprocessing. Sequences are represented in one-hot encoding scheme, in which a one-hot residue represents a single amino acid (single letter code). The LSTM architecture was chosen based on hyperparameters described by Saka et al. (2021). The chosen network architecture for this study was a two-layer LSTM recurrent neural network consisting of 64 neurons and a 0.2 dropout rate and trained for 200 epochs. Remaining parameters were set to default values as described by Müller et al. were utilized for all other parameters in the network. Based on five-fold cross validation, the epoch with the best average performance were chosen for the given LSTM architecture for each cluster individually. For each cluster, 10,000 sequences were sampled from the selected best epoch model.

Likelihood for sequence ranking

The NLL (negative log-likelihood) is a statistical measure that describes the likelihood of observing each amino acid at each position within the set of sequences over a training data set. From a set of sequences, the NLL is computed for each sequence according to the following formula:

$$NLL = -\sum_{k=1}^{K} \ln p(x_k)$$

where $p(x_k)$ represents the generative probability of observing a residue x at the k-th position of the sequence and K is the sequence length.

In silico developability assessment

The *in silico* developability profiles were computed using an internal pipeline termed "Sequence Assessment Using Multiple Optimization Parameters (SUMO)" (22). This approach automatically generates VHH models based on the provided sequences of the variable regions, identifies the human-likeness by sequence comparison to the most similar human germline sequence, determines structure-based surface-exposed chemical liability motifs (unpaired cysteines, methionines, asparagine deamidation motifs and aspartate deamidation sites) as well as sites susceptible to post-translational modification (N-linked glycosylation). Moreover, a small set of orthogonal computed physico-chemical descriptors including the isoelectric point (pI) of the variable domain, Schrodingers AggScore as predictor for hydrophobicity and aggregation tendency calculated for the complete variable domain as well as the complementarity-determining regions (CDRs) only and the calculated positive patch energy of the CDRs were determined (Sankar et al., 2018). These scores were complemented with a green to yellow to red color coding, indicating scores within one standard deviation from the mean over a benchmarking dataset of multiple biotherapeutics approved for human application as green, scores above one standard deviation as yellow and those above two standard deviations as red (Ahmed et al., 2021). For the AggScore values, these cutoffs were slightly adjusted based on correlation analyses to internal experimental HIC data.

Protein expression and analysis

The sdAb variants were integrated into the pTT5 mammalian expression vector by fusing them at the hinge region of Fc immune effector-silenced (eff-) SEED AG chains (Thermo Fisher Scientific). This fusion allowed the generation of one-armed (oa) SEEDbodies, using a SEED-GA chain without paratope.

The proteins were produced using the ExpiCHO™ Expression System (Thermo Fisher Scientific) in either 5 or 25 mL scale, following the standard protocol provided by the manufacturer. The expression was carried out with a 2:1 ratio of AG to GA chain. After 7 days of expression, the supernatants containing the proteins were purified using MabSelect™ antibody purification chromatography resin (Cytiva) using 20 mM acetic acid followed by an neutralization (500 mM sodium phosphate buffer, 1.5 M NaCl, pH 8) to a final formulation pH of 6.8 in PBS. The purified proteins were then subjected to sterile filtration, and their concentrations were determined by measuring the absorbance at 280 nm (A_{280}).

To evaluate the monomer content of the protein samples, analytical size-exclusion chromatography (SEC) was performed. Each sample contained 7.5 µg of protein and was run on a TSKgel UP-SW3000 column (2 µm, 4.6 × 300 mm, Tosoh Bioscience) using an Agilent HPLC 1260 Infinity system. The mobile phase consisted of 50 mM sodium phosphate and 0.4 M NaClO4 at pH 6.3, with a flow rate of 0.35 mL/min. The signals were recorded at 214 nm.

For assessing the hydrophobicity of the different molecules, hydrophobic interaction chromatography (HIC) was employed. Each sample contained 20 µg of protein and was analyzed on a TSKgel Butyl-NPR column (2.5 µm, 4.6 × 100 mm, Tosoh Bioscience) using an Agilent HPLC 1260 Infinity system with a flow rate of 0.5 mL/min. Prior to injection, the samples were mixed with a 50% (v/v) solution of 2 M ammonium sulfate. A gradient was applied, running from mobile phase A (1.2 M ammonium sulfate in PBS) to mobile phase B (50% methanol in 0.1x PBS) over a period of 15 min at 25°C. Signals were recorded at 214 nm. The reference molecules, anti-PD-L1 Avelumab and anti-EGFR Cetuximab, were used for comparison.

To investigate the thermal unfolding properties of the antibodies, differential scanning fluorimetry (DSF) was performed using a Prometheus NT. PLEX nanoDSF instrument (NanoTemper). The samples were measured in duplicate using nanoDSF Standard Capillary Chips. A temperature gradient ranging from 20°C to 95°C at a slope of 1°C/min was applied. Fluorescence signals at 350 nm and 330 nm were recorded. The unfolding transition midpoints (Tm) and Tonset values were determined from the melting curves or the first derivative of the fluorescence ratio 350 nm/330 nm.

Bio-Layer Interferometry (BLI)

The biophysical properties of the sdAbs were evaluated using an Octet Red BLI system from Sartorius. The binding experiments were conducted in KB-buffer (PBS pH 7.4, 0.1% BSA, 0.02% Tween-20) using Protein G Biosensors. The biosensors were loaded with the one-armed antibody samples at a concentration of 3 µg/mL for 180 s. The samples were subjected to a 2-fold serial dilution of (rh) NKp46 (ACRO Biosystems), starting at a concentration of 100 nM using a measurement window of 300 s for association and dissociation each.

The obtained data was aligned to the association step, and inter-step correction was applied during the dissociation step. To reduce noise, Savitzky-Golay filtering was employed. The resulting data were analyzed using a 1:1 binding model to determine the binding kinetics and affinity between the binders and (rh) NKp46.

Forced oxidation and deamidation studies

Forced protein oxidation was introduced to the samples (30 µg, 1 mg/mL) by diluting with an equal volume of 0.1% H_2O_2 (Merck, 107,209) and incubation at room temperature. After 0, 6, and 24 h a 10 µL aliquot was taken and the oxidation reaction was stopped by buffer exchange to 25 mM NH_4HCO_3 (Merck, 101131), pH 7 with Amicon filter devices (Merck, UFC503096), respectively. To force protein deamidation 30 µg sample was buffer exchanged to 25 mM NH_4HCO_3 (Merck, 101131), pH 10 using Amicon filter devices (Merck, UFC503096). Subsequently, the sample volume was adjusted to 30 µL and incubated at 37°C. To stop the deamidation reaction the sample was buffer exchanged to NH_4HCO_3, pH 7 as described previously in the oxidation workflow.

Peptide mapping

Proteins were unfolded and reduced by addition of 5 µL 12 M Urea (Merck, 108487) and 1 µL 50 mM DTT (Merck, 111474) and subsequent incubation at 50 °C for 30 min. Reduced samples were then alkylated by addition of 2.5 µL 55 mM iodoacetamide (Merck, 804744) and incubation at room temperature for 30 min in the dark. Samples were then mixed with 30 µL 25 mM NH_4HCO_3 and 3 µL trypsin solution (0.1 mg/mL). After 6 h at 37°C, 0.5 µL 50% FA was added and the peptides were analyzed by LC-MS. LC-MS analysis was performed using an Exion HPLC system (Buffer A: 0.1% formic acid in water (Biosolve, 23244101), Buffer B: 0.1% formic acid in acetonitrile (Biosolve, 01934101)) coupled to a Sciex 6,600+ mass spectrometer by a Turbo V ESI source. 8 µg peptide solution was loaded onto an Aeris 1.7 µm PEPTIDE XB-C18 150 × 2.1 mm column (Phenomenex, 00B-4506-AN) and eluted with a linear gradient from 5% to 50% Buffer B within 49 min and 0.25 mL/min flow rate. Data were acquired in IDA mode with positive polarity, in a mass range from 230 to 1,600 m/z. Other instrument settings were as follows: source voltage 5.5 kV, declustering potential 80 V, accumulation time 0.25 s, source temperature 450°C, maximum number of candidate ions per cycle 10, gas1 45 L/h, and gas2 45 L/h. The mass spectrometer was calibrated with ESI positive calibration solution 5,600. Acquired data were processed with Genedata Expressionist 16.5. Chemical noise subtraction was applied to the data by clipping all data points below an intensity of 50. Furthermore, spectra were smoothed, and background subtracted. For peptide mapping the MS tolerance was 20 ppm and the MS/MS tolerance 0.1 Da. Trypsin was chosen as enzyme with maximum 2 missed cleavages and minimum 3 amino acid peptide length. Deamidation (NQ), glutamine to pyroglutamate conversion, c-terminal lysine loss, and oxidation (MW) were selected as variable modifications.

AC SINS

Molecules were captured onto particles via immobilized capture antibodies and self-association was judged in PBS buffer at pH 7.4 by shifts in the plasmon wavelengths (Makowski et al., 2021). Clinical antibody Trastuzumab was used as control indicating favorable biophysical properties with mean Δλmax values of ~0.2 nm after subtraction of buffer blanks. Final AC-SINS scores for molecules were calculated via subtraction of blank and Trastuzumab scores and the calculated scores of the molecules in the range of −0.46 and 0.06 indicate favorable developability properties very similar to Trastuzumab.

PSR-BLI

To assess non-specific antibody interactions to polyspecificity reagent (PSR), a published cytometric assay (Xu et al., 2013) was adapted for the application of fast and sensitive Bio-Layer Interferometry (BLI). PSR was derived from soluble membrane proteins (SMP) of CHO and HEK293-6E cells as described by Xu et al. (2013). Assays were performed at 25°C with orbital sensor agitation at 1,000 rpm in 200 µL volume with DPBS. Pre-hydrated AHC biosensors were loaded with antibody (10 µg/mL) for 300 s. Afterwards biosensors were blocked with 1% BSA for 200 s and a baseline was established by rinsing in DPBS for 60 s. Association with 20 µg/mL PSR (1:1 mixture of CHO and HEK293-6E SMP) was performed for 100 s. As reference, association was performed in DPBS. To calculate the PSR-BLI score, the binding response from the association step was normalized to the reference measurement by subtraction, followed by subsequent subtraction with non-loading control (DPBS).

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Materials. The adopted python code and the sequences that had been used for LSTM model generation are available from https://github.com/MCompChem/LSTM_CDRs. Further inquiries can be directed to the corresponding author.

Author contributions

AE designed and performed all *in silico* studies and took the lead in writing the manuscript with input from all authors. AE and SZ directed the project. PA, HY, LP, TC, JS, JT, VS, and CS performed experiments. All authors gave scientific advice, analyzed and interpreted the data. All authors contributed to the article and approved the submitted version.

Acknowledgments

We would like to thank Jakub Gunera, Maria Borisovska, Andreas Hildebrandt, Daniel Kuhn, Friedrich Rippmann, Satyendra Kumar, Sigrid Auth and Vanessa Lautenbach for their feedback and support for this research.

Conflict of interest

PA, HY, LP, LF, VS, AD, SK, EG, SZ, AE were employed by Merck Healthcare KGaA. CS, JS, JT were employed by Merck KGaA. TC was employed by EMD Serono.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2023.1249247/full#supplementary-material

References

Ahmed, L., Gupta, P., Martin, K. P., Scheer, J. M., Nixon, A. E., and Kumar, S. (2021). Intrinsic physicochemical profile of marketed antibody-based biotherapeutics. *Proc. Natl. Acad. Sci.* 118 (37), e2020577118. doi:10.1073/pnas.2020577118

Akbar, R., Bashour, H., Rawat, P., Robert, P. A., Smorodina, E., Cotet, T. S., et al. (2022). Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. *mAbs* 14 (1), 2008790. doi:10.1080/19420862.2021.2008790

Antibody Discovery Software (2023). Geneious biologics antibody discovery software. Available from: https://www.geneious.com/biopharma/.

Arras, P., Yoo, H. B., Pekar, L., Schröter, C., Clarke, T., Krah, S., et al. (2023). A library approach for the de novo high-throughput isolation of humanized VHH domains with favorable developability properties following camelid immunization. mAbs in press. doi:10.1080/19420862.2023.2261149

Baker, D., and Sali, A. (2001). Protein structure prediction and structural genomics. *Science* 294 (5540), 93–96. doi:10.1126/science.1065659

Bannas, P., Hambach, J., and Koch-Nolte, F. (2017). Nanobodies and nanobody-based human heavy chain antibodies as antitumor therapeutics. *Front. Immunol.* 8, 1603. doi:10.3389/fimmu.2017.01603

Barreto, K., Maruthachalam, B. V., Hill, W., Hogan, D., Sutherland, A. R., Kusalik, A., et al. (2019). Next-generation sequencing-guided identification and reconstruction of antibody CDR combinations from phage selection outputs. *Nucleic Acids Res.* 47 (9), e50. doi:10.1093/nar/gkz131

Barrow, A. D., Martin, C. J., and Colonna, M. (2019). The natural cytotoxicity receptors in Health and disease. *Front. Immunol.* 10, 909. doi:10.3389/fimmu.2019.00909

Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W. H., Ng, L. G., et al. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* 37, 38–44. doi:10.1038/nbt.4314

Chanier, T., and Chames, P. (2019). Nanobody engineering: toward next generation immunotherapies and immunoimaging of cancer. *Antibodies* 8 (1), 13. doi:10.3390/antib8010013

Custers, R., and Steyaert, J. (2020). Discussions on the quality of antibodies are no reason to ban animal immunization. *EMBO Rep.* 21 (12), e51761. doi:10.15252/embr.202051761

Davis, J. H., Aperlo, C., Li, Y., Kurosawa, E., Lan, Y., Lo, K. M., et al. (2010). SEEDbodies: fusion proteins based on strand-exchange engineered domain (SEED) CH3 heterodimers in an Fc analogue platform for asymmetric binders or immunofusions and bispecific antibodies. *Protein Eng. Des. Sel.* 23 (4), 195–202. doi:10.1093/protein/gzp094

Evers, A., Malhotra, S., and Sood, V. D. (2023b). Silico approaches to deliver better antibodies by design: the past, the present and the future. Available from: http://arxiv.org/abs/2305.07488.

Evers, A., Malhotra, S., Bolick, W. G., Najafian, A., Borisovska, M., Warszawski, S., et al. (2023a). "Sumo: in silico sequence assessment using multiple optimization parameters," in *Genotype phenotype coupling. Methods in molecular biology*. Editors S. Zielonka and S. Krah (New York, NY: Springer US), 383–398.

Evers, A., Pfeiffer-Marek, S., Bossart, M., Heubel, C., Stock, U., Tiwari, G., et al. (2019). Peptide optimization at the drug discovery-development interface: tailoring of physicochemical properties toward specific formulation requirements. *J. Pharm. Sci.* 108 (4), 1404–1414. doi:10.1016/j.xphs.2018.11.043

Fernández-Quintero, M. L., Ljungars, A., Waibl, F., Greiff, V., Andersen, J. T., Gjølberg, T. T., et al. (2023). Assessing developability early in the discovery process for novel biologics. *mAbs* 15 (1), 2171248. doi:10.1080/19420862.2023.2171248

Grinshpun, B., Thorsteinson, N., Pereira, J. N., Rippmann, F., Nannemann, D., Sood, V. D., et al. (2021). Identifying biophysical assays and *in silico* properties that enrich for slow clearance in clinical-stage therapeutic antibodies. *mAbs* 13 (1), 1932230. doi:10.1080/19420862.2021.1932230

Gauthier, L., Morel, A., Anceriz, N., Rossi, B., Blanchard-Alvarez, A., Grondin, G., et al. (2019). Multifunctional natural killer cell engagers targeting NKp46 trigger protective tumor immunity. *Cell* 177 (7), 1701–1713. doi:10.1016/j.cell.2019.04.041

Gauthier, L., Virone-Oddos, A., Beninga, J., Rossi, B., Nicolazzi, C., Amara, C., et al. (2023). Control of acute myeloid leukemia by a trifunctional NKp46-CD16a-NK cell engager targeting CD123. *Nat. Biotechnol.*, 1–11. doi:10.1038/s41587-022-01626-2

Gray, A., Bradbury, A. R. M., Knappik, A., Plückthun, A., Borrebaeck, C. A. K., and Dübel, S. (2020a). Animal-free alternatives and the antibody iceberg. *Nat. Biotechnol.* 38 (11), 1234–1239. doi:10.1038/s41587-020-0687-9

Gray, A. C., Bradbury, A., Dübel, S., Knappik, A., Plückthun, A., and Borrebaeck, C. A. K. (2020b). Reproducibility: bypass animals for antibody production. *Nature* 581 (7808), 262. doi:10.1038/d41586-020-01474-7

Gupta, A., Müller, A. T., Huisman, B. J. H., Fuchs, J. A., Schneider, P., and Schneider, G. (2018). Generative recurrent networks for de novo drug design. *Mol. Inf.* 37 (1–2), 1700111. doi:10.1002/minf.201700111

Gupta, P., Makowski, E. K., Kumar, S., Zhang, Y., Scheer, J. M., and Tessier, P. M. (2022). Antibodies with weakly basic isoelectric points minimize trade-offs between formulation and physiological colloidal properties. *Mol. Pharm.* 19 (3), 775–787. doi:10.1021/acs.molpharmaceut.1c00373

Hie, B. L., Shanker, V. R., Xu, D., Bruun, T. U. J., Weidenbacher, P. A., Tang, S., et al. (2023). Efficient evolution of human antibodies from general protein language models. *Nat. Biotechnol.*, 1–9. doi:10.1038/s41587-023-01763-2

Hu, D., Hu, S., Wan, W., Xu, M., Du, R., Zhao, W., et al. (2015). Effective optimization of antibody affinity by phage display integrated with high-throughput DNA synthesis and sequencing technologies. *PLOS ONE* 10 (6), e0129125. doi:10.1371/journal.pone.0129125

Hu, R., Fu, L., Chen, Y., Chen, J., Qiao, Y., and Si, T. (2023). Protein engineering via Bayesian optimization-guided evolutionary algorithm and robotic experiments. *Briefings Bioinforma.* 24 (1), bbac570. doi:10.1093/bib/bbac570

Jain, T., Boland, T., and Vásquez, M. (2023). Identifying developability risks for clinical progression of antibodies using high-throughput *in vitro* and *in silico* approaches. *mAbs* 15 (1), 2200540. doi:10.1080/19420862.2023.2200540

Jain, T., Sun, T., Durand, S., Hall, A., Houston, N. R., Nett, J. H., et al. (2017). Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl. Acad. Sci. U. S. A.* 114 (5), 944–949. doi:10.1073/pnas.1616408114

Jin, B. K., Odongo, S., Radwanska, M., and Magez, S. (2023). Nanobodies: A review of generation, diagnostics and therapeutics. *Int. J. Mol. Sci.* 24 (6), 5994. doi:10.3390/ijms24065994

Jin, W., Xing, Z., Song, Y., Huang, C., Xu, X., Ghose, S., et al. (2019). Protein aggregation and mitigation strategy in low pH viral inactivation for monoclonal antibody purification. *MAbs* 11 (8), 1479–1491. doi:10.1080/19420862.2019.1658493

Khetan, R., Curtis, R., Deane, C. M., Hadsund, J. T., Kar, U., Krawczyk, K., et al. (2022). Current advances in biopharmaceutical informatics: guidelines, impact and challenges in the computational developability assessment of antibody therapeutics. *mAbs* 14 (1), 2020082. doi:10.1080/19420862.2021.2020082

Kingsbury, J. S., Saini, A., Auclair, S. M., Fu, L., Lantz, M. M., Halloran, K. T., et al. (2020). A single molecular descriptor to predict solution behavior of therapeutic antibodies. *Sci. Adv.* 6 (32), eabb0372. doi:10.1126/sciadv.abb0372

Kitchen, D. B., and Decornez, H. Y. (2015). "Computational techniques to support hit triage," in *Small molecule medicinal Chemistry* (Hoboken, New Jerse: John Wiley & Sons, Ltd).

Klausz, K., Pekar, L., Boje, A. S., Gehlert, C. L., Krohn, S., Gupta, T., et al. (2022). Multifunctional NK cell-engaging antibodies targeting EGFR and NKp30 elicit efficient tumor cell killing and proinflammatory cytokine release. *J. Immunol.* 209 (9), 1724–1735. doi:10.4049/jimmunol.2100970

Könning, D., Zielonka, S., Grzeschik, J., Empting, M., Valldorf, B., Krah, S., et al. (2017). Camelid and shark single domain antibodies: structural features and therapeutic potential. *Curr. Opin. Struct. Biol.* 45, 10–16. doi:10.1016/j.sbi.2016.10.019

Krah, S., Schröter, C., Zielonka, S., Empting, M., Valldorf, B., and Kolmar, H. (2016). Single-domain antibodies for biomedical applications. *Immunopharmacol. Immunotoxicol.* 38 (1), 21–28. doi:10.3109/08923973.2015.1102934

Kumar, S., Roffi, K., Tomar, D. S., Cirelli, D., Luksha, N., Meyer, D., et al. (2018). Rational optimization of a monoclonal antibody for simultaneous improvements in its solution properties and biological activity. *Protein Eng. Des. Sel.* 31 (7–8), 313–325. doi:10.1093/protein/gzy020

Larman, H. B., Jing Xu, G., Pavlova, N. N., and Elledge, S. J. (2012). Construction of a rationally designed antibody platform for sequencing-assisted selection. *Proc. Natl. Acad. Sci.* 109 (45), 18523–18528. doi:10.1073/pnas.1215549109

Lauer, T. M., Agrawal, N. J., Chennamsetty, N., Egodage, K., Helk, B., and Trout, B. L. (2012). Developability index: A rapid in silico tool for the screening of antibody aggregation propensity. *J. Pharm. Sci.* 101 (1), 102–115. doi:10.1002/jps.22758

Laustsen, A. H., Greiff, V., Karatt-Vellatt, A., Muyldermans, S., and Jenkins, T. P. (2021). Animal immunization, *in vitro* display technologies, and machine learning for antibody discovery. *Trends Biotechnol.* 39 (12), 1263–1273. doi:10.1016/j.tibtech.2021.03.003

Li, L., Gupta, E., Spaeth, J., Shing, L., Jaimes, R., Engelhart, E., et al. (2023). Machine learning optimization of candidate antibody yields highly diverse sub-nanomolar affinity antibody libraries. *Nat. Commun.* 14 (1), 3454. doi:10.1038/s41467-023-39022-2

Lipinski, B., Arras, P., Pekar, L., Klewinghaus, D., Boje, A. S., Krah, S., et al. (2023a). NKp46-specific single domain antibodies enable facile engineering of various potent NK cell engager formats. *Protein Sci.* 32 (3), e4593. doi:10.1002/pro.4593

Lipinski, B., Unmuth, L., Arras, P., Becker, S., Bauer, C., Toleikis, L., et al. (2023b). Generation and engineering of potent single domain antibody-based bispecific IL-18 mimetics resistant to IL-18BP decoy receptor inhibition. *mAbs* 15 (1), 2236265. doi:10.1080/19420862.2023.2236265

Liu, G., Zeng, H., Mueller, J., Carter, B., Wang, Z., Schilz, J., et al. (2020). Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics* 36 (7), 2126–2133. doi:10.1093/bioinformatics/btz895

Liu, Y., Caffry, I., Wu, J., Geng, S. B., Jain, T., Sun, T., et al. (2014). High-throughput screening for developability during early-stage antibody discovery using self-interaction nanoparticle spectroscopy. *MAbs* 6 (2), 483–492. doi:10.4161/mabs.27431

Lu, X., Nobrega, R. P., Lynaugh, H., Jain, T., Barlow, K., Boland, T., et al. (2019). Deamidation and isomerization liability analysis of 131 clinical-stage antibodies. *mAbs* 11 (1), 45–57. doi:10.1080/19420862.2018.1548233

Makowski, E. K., Kinnunen, P. C., Huang, J., Wu, L., Smith, M. D., Wang, T., et al. (2022). Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. *Nat. Commun.* 13 (1), 3788. doi:10.1038/s41467-022-31457-3

Makowski, E. K., Wu, L., Gupta, P., and Tessier, P. M. (2021). Discovery-stage identification of drug-like antibodies using emerging experimental and computational methods. *mAbs* 13 (1), 1895540. doi:10.1080/19420862.2021.1895540

Mason, D. M., Friedensohn, S., Weber, C. R., Jordi, C., Wagner, B., Meng, S. M., et al. (2021). Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat. Biomed. Eng.* 5 (6), 600–612. doi:10.1038/s41551-021-00699-9

Mathonet, P., and Ullman, C. G. (2013). The application of next generation sequencing to the understanding of antibody repertoires. *Front. Immunol.* 4, 265. doi:10.3389/fimmu.2013.00265

Merk, D., Friedrich, L., Grisoni, F., and Schneider, G. (2018). De novo design of bioactive small molecules by artificial intelligence. *Mol. Inf.* 37 (1–2), 1700153. doi:10.1002/minf.201700153

Mieczkowski, C., Zhang, X., Lee, D., Nguyen, K., Lv, W., Wang, Y., et al. (2023). Blueprint for antibody biologics developability. *mAbs* 15 (1), 2185924. doi:10.1080/19420862.2023.2185924

Müller, A. T., Hiss, J. A., and Schneider, G. (2018). Recurrent neural network model for constructive peptide design. *J. Chem. Inf. Model* 58 (2), 472–479. doi:10.1021/acs.jcim.7b00414

Negron, C., Fang, J., McPherson, M. J., Stine, W. B., and McCluskey, A. J. (2022). Separating clinical antibodies from repertoire antibodies, a path to *in silico* developability assessment. *mAbs* 14 (1), 2080628. doi:10.1080/19420862.2022.2080628

Nowak, C. K., Cheung, J., Gu, M., Dellatore, S., Katiyar, A., Bhat, R., et al. (2017). Forced degradation of recombinant monoclonal antibodies: A practical guide. *MAbs* 9 (8), 1217–1230. doi:10.1080/19420862.2017.1368602

Parkinson, J., Hard, R., and Wang, W. (2023). The RESP AI model accelerates the identification of tight-binding antibodies. *Nat. Commun.* 14 (1), 454. doi:10.1038/s41467-023-36028-8

Pekar, L., Busch, M., Valldorf, B., Hinz, S. C., Toleikis, L., Krah, S., et al. (2020). Biophysical and biochemical characterization of a VHH-based IgG-like bi- and trispecific antibody platform. *mAbs* 12 (1), 1812210. doi:10.1080/19420862.2020.1812210

Pekar, L., Klewinghaus, D., Arras, P., Carrara, S. C., Harwardt, J., Krah, S., et al. (2021). Milking the cow: cattle-derived chimeric ultralong CDR-H3 antibodies and their engineered CDR-H3-only knobbody counterparts targeting epidermal growth factor receptor elicit potent NK cell-mediated cytotoxicity. *Front. Immunol.* 12, 742418. doi:10.3389/fimmu.2021.742418

Rabia, L. A., Desai, A. A., Jhajj, H. S., and Tessier, P. M. (2018). Understanding and overcoming trade-offs between antibody affinity, specificity, stability and solubility. *Biochem. Eng. J.* 137, 365–374. doi:10.1016/j.bej.2018.06.003

Raybould, M. I. J., Marks, C., Krawczyk, K., Taddese, B., Nowak, J., Lewis, A. P., et al. (2019). Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl. Acad. Sci.* 116 (10), 4025–4030. doi:10.1073/pnas.1810576116

Roth, L., Krah, S., Klemm, J., Günther, R., Toleikis, L., Busch, M., et al. (2020). "Isolation of antigen-specific VHH single-domain antibodies by combining animal immunization with yeast surface display," in *Genotype phenotype coupling: Methods and protocols. Methods in molecular biology*. Editors S. Zielonka and S. Krah (New York, NY: Springer US), 173–189. doi:10.1007/978-1-4939-9853-1_10

Rouet, R., Jackson, K. J. L., Langley, D. B., and Christ, D. (2018). Next-generation sequencing of antibody display repertoires. *Front. Immunol.* 9, 118. doi:10.3389/fimmu.2018.00118

Saka, K., Kakuzaki, T., Metsugi, S., Kashiwagi, D., Yoshida, K., Wada, M., et al. (2021). Antibody design using LSTM based deep generative model from phage display library for affinity maturation. *Sci. Rep.* 11 (1), 5852. doi:10.1038/s41598-021-85274-7

Sankar, K., Krystek, S. R., Jr, Carl, S. M., Day, T., and Maier, J. K. X. (2018). AggScore: prediction of aggregation-prone regions in proteins based on the distribution of surface patches. *Proteins Struct. Funct. Bioinforma.* 86 (11), 1147–1156. doi:10.1002/prot.25594

Segler, M. H. S., Kogej, T., Tyrchan, C., and Waller, M. P. (2018). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* 4 (1), 120–131. doi:10.1021/acscentsci.7b00512

Sormanni, P., Aprile, F. A., and Vendruscolo, M. (2015). The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.* 427 (2), 478–490. doi:10.1016/j.jmb.2014.09.026

Sulea, T. (2022). "Humanization of camelid single-domain antibodies," in *Single-domain antibodies: Methods and protocols*. Editors G. Hussack and K. A. Henry (New York, NY: Springer US). doi:10.1007/978-1-0716-2075-5_14

Svilenov, H. L., Arosio, P., Menzen, T., Tessier, P., and Sormanni, P. (2023). Approaches to expand the conventional toolbox for discovery and selection of antibodies with drug-like physicochemical properties. *mAbs* 15 (1), 2164459. doi:10.1080/19420862.2022.2164459

Teixeira, A. A. R., Erasmus, M. F., D'Angelo, S., Naranjo, L., Ferrara, F., Leal-Lopes, C., et al. (2021). Drug-like antibodies with high affinity, diversity and developability directly from next-generation antibody libraries. *mAbs* 13 (1), 1980942. doi:10.1080/19420862.2021.1980942

Valldorf, B., Hinz, S. C., Russo, G., Pekar, L., Mohr, L., Klemm, J., et al. (2022). Antibody display technologies: selecting the cream of the crop. *Biol. Chem.* 403 (5–6), 455–477. doi:10.1515/hsz-2020-0377

Vauquelin, G., and Charlton, S. J. (2013). Exploring avidity: understanding the potential gains in functional affinity and target residence time of bivalent and heterobivalent ligands. *Br. J. Pharmacol.* 168 (8), 1771–1785. doi:10.1111/bph.12106

Vincke, C., Loris, R., Saerens, D., Martinez-Rodriguez, S., Muyldermans, S., and Conrath, K. (2009). General strategy to humanize a camelid single-domain antibody and identification of a universal humanized nanobody scaffold. *J. Biol. Chem.* 284 (5), 3273–3284. doi:10.1074/jbc.M806889200

Waibl, F., Fernández-Quintero, M. L., Wedl, F. S., Kettenberger, H., Georges, G., and Liedl, K. R. (2022). Comparison of hydrophobicity scales for predicting biophysical properties of antibodies. *Front. Mol. Biosci.* 9, 960194. doi:10.3389/fmolb.2022.960194

Wang, J., Kang, G., Yuan, H., Cao, X., Huang, H., and de Marco, A. (2022). Research progress and applications of multivalent, multispecific and modified nanobodies for disease treatment. *Front. Immunol.* 12. doi:10.3389/fimmu.2021.838082

Xu, Y., Roach, W., Sun, T., Jain, T., Prinz, B., Yu, T. Y., et al. (2013). Addressing polyspecificity of antibodies selected from an *in vitro* yeast presentation system: A FACS-based, high-throughput selection and analytical tool. *Protein Eng. Des. Sel.* 26 (10), 663–670. doi:10.1093/protein/gzt047

Yanakieva, D., Pekar, L., Evers, A., Fleischer, M., Keller, S., Mueller-Pompalla, D., et al. (2022). Beyond bispecificity: controlled fab arm exchange for the generation of antibodies with multiple specificities. *mAbs* 14, 2018960. doi:10.1080/19420862.2021.2018960

Zeng, X., Wang, F., Luo, Y., Kang, S., Tang, J., Lightstone, F. C., et al. (2022). Deep generative molecular design reshapes drug discovery. *Cell Rep. Med.* 3 (12), 100794. doi:10.1016/j.xcrm.2022.100794

TYPE Original Research
PUBLISHED 01 March 2024
DOI 10.3389/fmolb.2023.1302737

Check for updates

OPEN ACCESS

EDITED BY
Didier Devaurs,
University of Edinburgh, United Kingdom

REVIEWED BY
Andrew Buchanan,
AstraZeneca, United Kingdom
Klaus R. Liedl,
University of Innsbruck, Austria

*CORRESPONDENCE
Daisuke Kuroda,
✉ dkuroda@niid.go.jp
Takao Hashiguchi,
✉ hashiguchi.takao.1a@kyoto-u.ac.jp
Kouhei Tsumoto,
✉ tsumoto@bioeng.t.u-tokyo.ac.jp

RECEIVED 27 September 2023
ACCEPTED 11 December 2023
PUBLISHED 01 March 2024

CITATION
Paul R, Kasahara K, Sasaki J, Pérez JF, Matsunaga R, Hashiguchi T, Kuroda D and Tsumoto K (2024), Unveiling the affinity–stability relationship in anti-measles virus antibodies: a computational approach for hotspots prediction.
Front. Mol. Biosci. 10:1302737.
doi: 10.3389/fmolb.2023.1302737

Unveiling the affinity–stability relationship in anti-measles virus antibodies: a computational approach for hotspots prediction

Rimpa Paul[1,2], Keisuke Kasahara[1], Jiei Sasaki[3], Jorge Fernández Pérez[1], Ryo Matsunaga[1,4], Takao Hashiguchi[3]*, Daisuke Kuroda[1,2,4]* and Kouhei Tsumoto[1,4,5]*

[1]Department of Bioengineering, School of Engineering, The University of Tokyo, Tokyo, Japan, [2]Research Center of Drug and Vaccine Development, National Institute of Infectious Diseases, Tokyo, Japan, [3]Institute for Life and Medical Sciences, Kyoto University, Sakyo-ku, Kyoto, Japan, [4]Department of Chemistry and Biotechnology, School of Engineering, The University of Tokyo, Tokyo, Japan, [5]The Institute of Medical Science, The University of Tokyo, Tokyo, Japan

Recent years have seen an uptick in the use of computational applications in antibody engineering. These tools have enhanced our ability to predict interactions with antigens and immunogenicity, facilitate humanization, and serve other critical functions. However, several studies highlight the concern of potential trade-offs between antibody affinity and stability in antibody engineering. In this study, we analyzed anti-measles virus antibodies as a case study, to examine the relationship between binding affinity and stability, upon identifying the binding hotspots. We leverage *in silico* tools like Rosetta and FoldX, along with molecular dynamics (MD) simulations, offering a cost-effective alternative to traditional *in vitro* mutagenesis. We introduced a pattern in identifying key residues in pairs, shedding light on hotspots identification. Experimental physicochemical analysis validated the predicted key residues by confirming significant decrease in binding affinity for the high-affinity antibodies to measles virus hemagglutinin. Through the nature of the identified pairs, which represented the relative hydropathy of amino acid side chain, a connection was proposed between affinity and stability. The findings of the study enhance our understanding of the interactions between antibody and measles virus hemagglutinin. Moreover, the implications of the observed correlation between binding affinity and stability extend beyond the field of anti-measles virus antibodies, thereby opening doors for advancements in antibody research.

KEYWORDS

antibody engineering, computer-aided design, measles virus hemagglutinin, hotspots, relative hydropathy, molecular dynamics

1 Introduction

In recent years, the application of computational methods has expanded significantly in the field of antibody engineering (Kuroda et al., 2012; Fischman and Ofran, 2018; Kuroda and Tsumoto, 2018; 2020; Akbar et al., 2022a; Wilman et al., 2022). The potential applications are vast; however, predicting biophysical properties can be still challenging when crystal structures of neither the antibody itself nor the antigen-antibody complex are available. This lack of binding information further complicates the task of guiding *in silico* antibody engineering. Numerous computational protocols have been developed to facilitate tasks such as affinity maturation, protein aggregation prediction, and stability enhancement. These aim to create biologically superior antibodies and often rely on initial structure predictions through techniques like homology modelling and molecular docking (Weitzner et al., 2017; Cannon et al., 2019; Liang et al., 2021).

Artificial intelligence (AI) technologies have made significant strides in tackling challenges within protein engineering. The advancements in machine learning (ML) and deep learning (DL) have revolutionized antibody research, particularly in areas such as structure prediction, antibody design, and epitope mapping (Jumper et al., 2021; Ripoll et al., 2021; Akbar et al., 2022b; Prihoda et al., 2022; Ruffolo et al., 2022; 2023). The integration of data-driven AI approaches holds immense promise for drug discovery. However, the accuracy and reliability of these AI predictions heavily rely on the quality of the training data. One significant advancement in developing more robust prediction models is the availability of comprehensive antibody libraries, such as the Observed Antibody Space (OAS) (Marks et al., 2021). OAS has played a crucial role in addressing challenges in antibody engineering, such as humanization and immunogenicity prediction (Olsen et al., 2022; Prihoda et al., 2022). Despite these advancements, certain problems, like trade-offs between antibody affinity and stability remains a challenge as it necessitates large-scale experimental data.

Several studies highlight the same concern of potential trade-offs between antibody affinity and stability in antibody engineering (Rabia et al., 2018). However, current approaches have not specifically addressed the exploration of this relationship. Seizing this opportunity, we followed a knowledge-based computational approach that can identify key residues, thereby revealing the intricate interplay between the affinity and stability of an antibody. This approach utilizes standard *in silico* protein engineering tools and focuses on the importance of residues in the complementarity determining regions (CDRs). CDR3 in the heavy and light chains is widely recognized for its critical role in antigen recognition and binding (Kuroda et al., 2008; Kuroda et al., 2009; Weitzner et al., 2015; D'Angelo et al., 2018). In general, other regions such as framework regions (FRs) in variable domain (Fv) and constant domains primarily contribute to antibody stability (Ionescu et al., 2008; Zabetakis et al., 2013). Nevertheless, we hypothesize that CDR3 residues also contribute to stability and could impact both affinity and stability. To substantiate this, we identified hotspots as sequential pair located within CDRs (particularly focusing on CDR3), by integrating MD simulations to *in silico* alanine scanning. These hotspots are capable of modulating both affinity and stability based on their local or relative hydropathy (Di Rienzo et al., 2021). Relative hydropathy is based on the surroundings of an amino acid side chain, which plays a crucial role in antigen binding and stability.

As a model system, we choose antibodies against the measles virus hemagglutinin (MVH). Measles is an infectious and highly contagious disease that continues to thrive in developing countries, despite the availability of an effective vaccine for decades (Suvvari et al., 2023). To fully eradicate the disease, there is an urgent need for advanced measles therapy. Although researchers have been developing antibodies against measles virus for epitope identification and other research purposes, none of these antibodies have yet entered clinical trials. Remarkably, no crystal structures for anti-measles virus antibodies or antibody-antigen complexes are available in the Protein Data Bank (PDB) (Berman et al., 2007). This lack of structural data is a significant hurdle to the development of antibody-based treatments against the measles virus. On the other hand, the crystal structures of the MVH (Hashiguchi et al., 2007) and a fusion protein, two glycoproteins present in the virus's envelope, are available in PDB in both apo and holo forms with cellular receptors such as signaling lymphocytic activation molecule (SLAM) (PDB ID: 3ALW, 3ALZ, 3ALX), Nectin-4 (PDB ID: 4GJT), and CD46 (PDB ID: 3INB) (Santiago et al., 2010; Hashiguchi et al., 2011; Zhang et al., 2013). This disparity makes antibodies against measles virus an intriguing subject for further research. In this context, Tadokoro and colleagues (Tadokoro et al., 2020) have extensively analyzed biophysical parameters such as equilibrium dissociation constant (K_D) or binding affinity, melting point T_m or thermal stability, and thermodynamic parameters for an anti-MVH antibody 2F4. The reported binding affinity for antibody 2F4 Fab at 25 °C was 18 nM, which is about 10 and 37-fold higher affinity than SLAM (K_D = 170 nM) and Nectin-4 (K_D = 670 nM), respectively. Neutralization of the virus by the antibody 2F4 has also been reported, along with three other antibodies, namely, 7C6, 8F6, and 10B5 (Sato et al., 2018). All the antibodies obtained from mouse immunization can neutralize the antigen MVH, differing to some extent in the neutralizing capability. These four antibodies have different germline origins (Supplementary Table S1).

In this study, based on homology modeling, docking simulations, MD simulations, and *in silico* alanine scanning, we computationally predicted residues that potentially coupled both stability and binding affinity, and experimentally analyzed physicochemical properties of anti-MVH antibodies. The antibodies we employed demonstrated high binding affinities less than 1 nM to MVH, but they differed in stability. Pairwise point mutational analysis offered insights into these differences and suggested a potential relationship between affinity and stability of anti-MVH antibodies.

2 Results

2.1 Experimental characterization of anti-measles virus neutralizing antibodies

We first performed physicochemical analysis of the four wild type (WT) antibodies: 2F4, 7C6, 8F6, and 10B5. These antibodies were previously obtained through mouse immunization (Sato et al., 2018) and, except for 2F4 (Tadokoro et al., 2020), they had not been biophysically characterized until this study. Ideally, antibodies should demonstrate a rapid association and a slow dissociation with antigens. Our SPR measurements confirmed that antibodies 7C6 and 8F6 exhibited these characteristics, resulting in an affinity of 0.4 ± 0.2 and 0.9 ± 0.2 nM, respectively, toward MVH (Table 1). On the other hand, 2F4 and 10B5 demonstrated a slower association and a faster dissociation, resulting in lower binding affinity of 54.1 ± 0.1 and 60.3 ± 19.4 nM, respectively.

The K_D of 2F4 antibody reported in a previous study (Tadokoro et al., 2020) was lower than our observed value

TABLE 1 Physicochemical analysis of the wild type anti-MVH antibodies. Kinetic parameters[a] and melting temperature (T_m) are shown.

Physicochemical analysis (wild type)		< 1 nM affinity group		> 50 nM affinity group	
		7C6	8F6	2F4	10B5
Binding affinity	k_{on} ($\times 10^5$ $M^{-1}s^{-1}$)	11.4 ± 4.8	3.4 ± 2.9	1.2 ± 0.5	0.1 ± 0.1
	k_{off} ($\times 10^{-4}$ s^{-1})	4.1 ± 1.7	2.8 ± 1.7	65.8 ± 29.1	8.6 ± 0.4
	K_D at 25°C (nM)	0.4 ± 0.2	0.9 ± 0.2	54.1 ± 0.1	60.3 ± 19.4
Thermal stability	T_m (°C)	73.9 ± 0.3	68.0 ± 0.1	72.7 ± 0.1	73.9 ± 0.1

[a]The simple 1:1 Langmuir binding model was used to fit and calculate the kinetic parameters of the binding.

(Table 1; Supplementary Figure S1). Despite this discrepancy, all four antibodies exhibited better binding affinity than the receptors, particularly 7C6 and 8F6. Although the reported thermal stability of the 2F4 Fab was 76°C (Tadokoro et al., 2020), our DSC measurements revealed a decrease in melting temperature (T_m = 72.7°C ± 0.1°C). Antibodies 7C6 and 10B5 demonstrated higher stability with melting temperatures of 73.9°C ± 0.3°C and 73.9°C ± 0.1°C, respectively, while 8F6 exhibited lower thermal stability of 68.0°C ± 0.1°C (Table 1; Supplementary Figure S2).

Based on these observations, we classified the antibodies into two affinity groups (Table 1). Subsequently, we focused on the high binding affinity (<1 nM) antibodies 7C6 and 8F6, which showed a significant difference in thermal stability (ΔT_m, ~6°C). Analyzing these characteristics may provide insights into the relationship between binding affinity and thermal stability in anti-MVH antibodies.

2.2 Homology modeling and antibody-antigen local docking

As the crystal structure of the antibodies are unavailable at the time of this writing, we performed antibody structure modeling with the RosettaAntibody protocol (Weitzner et al., 2017). The variable fragment of the antibody was modeled from the amino acid sequences (Figures 1A, B), and the best scored model was selected for docking with the MVH crystal structure (PDB ID: 2ZB6). While there was no prior binding information available for the high-affinity antibodies (7C6 and 8F6), it was available for the receptors. The head domain of the MVH has 6-bladed β-propeller folds (β1–6). It is the main target of neutralizing antibodies (Tahara et al., 2016). Among them, the receptor binding epitope, which is a group of amino acids in the receptor binding site, stands out because, as the name suggests, it is also recognized by the three receptors to MVH, as well as by antibody 2F4. It is worth noting that several other antibodies, which were not included in this study, have also been reported to target this epitope (Tahara et al., 2016). The receptor binding epitope is located primarily within β5 with some extension in β4 and β6. Since 2F4 is reported to interact with the receptor binding epitope (Tahara et al., 2016), we first constructed a putative structure of the 2F4 with MVH by placing the antibody within 7 Å of the MVH near the receptor binding epitope, so that the CDRs and the receptor binding epitope roughly face each other. Next, we performed a Monte Carlo-based rigid body docking using RosettaDock (Chaudhury et al., 2011), that predicted favorable binding modes of 2F4 with MVH. The best docking score obtained was −26.9 Rosetta Energy Unit (REU). The visual inspection of this docked model showed that amino acids 190, 533 and 541, which reported to recognize 2F4 is within 5 Å, in agreement with the reported experimental data (Tahara et al., 2013; 2016). The 2F4 docked model helped in our knowledge-based docking approach and we used it as a reference to construct the putative model for 7C6 and 8F6 followed by flexible antibody-antigen docking (Weitzner et al., 2017). The "core epitope" utilized in this study encompasses the following amino acids in the receptor binding site of MVH: 187, 190–200 and 571–579 in β6, 483 in β4, 505–552 in β5 (Figure 1C). Binding of antibodies to this core epitope could identify key interacting residues.

Subsequently, with the SnugDock algorithm (Sircar and Gray, 2010), we obtained the best docking scores of −41 REU and −39.5 REU for 7C6 and 8F6 antibodies, respectively. The order of these docking scores aligns with the experimental K_D values (0.4 ± 0.2 and 0.9 ± 0.2 nM for 7C6 and 8F6, respectively). We also employed docking local refinement in Rosetta to compute the docking score for the available crystal structure of the receptor-antigen complex as a positive control. The best docking scores for receptors SLAM (PDB ID: 3ALZ) (Hashiguchi et al., 2011), CD46 (PDB ID: 3INB) (Santiago et al., 2010) and Necin-4 (PDB ID: 4GJT) (Zhang et al., 2013) were −38.9, −33.7 and −30.9 REU, respectively. These docking scores are aligned well with the reported experimental binding affinity (K_D 170, 200 and 670 nM for SLAM, CD46 and Nectin-4, respectively) (Hashiguchi et al., 2007; Santiago et al., 2010). The resulting models for antibodies, representing the predicted holo form, were then evaluated through *in silico* and *in vitro* assessments. The workflow for the *in silico* assessments is depicted in Figure 2.

2.3 Visual inspection and MD simulations to identify interacting residues in predicted complex structures

In line with our proposed workflow for hotspot prediction (Figure 2), our initial step involves identifying the interface residues contributing to binding between the antibody and

A

```
Chothia No.  1                         25          35              49    52A
             |                         |           |               |     |
7C6 WT       EVQLQQSGTVLARPGASVKMSCKAS GYSFTSYYIH WVKQRPGQGLEWIG GIYPGNSDTAYNHNFKG
8F6 WT       QVQLQQSAAELARPGASVKMSCKAS GYTLNSYTMH WVKQRPGQGLEWFG YINPSSGYTDYNQKFKD
             |---------FR-H1---------| |-CDR-H1-| |---FR-H2----| |----CDR-H2-----|

Chothia No.   65                82A-C            94 95     100A-E   103      112
              |                 | |              |  |      |   |    |        |
7C6 WT        KAKLTAITSASTAYMEFSSLTNEDSAVYYCTR  SDWYF------DV  WGAGTTVTV
8F6 WT        KTTLTADTSSSTAYMQLSSLTSEDSAVYYCAS  QDAIYYGNYRFAY  WGQGTLVTV
              |------------FR-H3-------------|  |--CDR-H3---|  |-FR-H4-|
```

B

```
Chothia No.  1                      24          35               49
             |                      |           |                |
7C6 WT       DIKMTQSPSSIYASLGERVTITC KASQDINRYLS WFQQKPGKSPKTLIY RAHRLVD
8F6 WT       DIQMTQTTSSLSASLGDRVTISC RASQDISNYLN WYQQKPDGTVKLLIY YTSRLQS
             |--------FR-L1--------| |-CDR-L1--| |----FR-L2----| |CDR-L2|

Chothia No.   57                               88 89        98          109
              |                                |  |         |           |
7C6 WT        GVPSRFSGSGSGQDYSLTISSLEYEDMGIYYC LQYDAFPYT  FGGGTKLEIKRA
8F6 WT        GVPSRFSGSGSGTDYYITIANLEQEDFATYFC QQTNTLPWT  FGGGTKLEIKRA
              |------------FR-L3-------------| |-CDR-L3-| |-- FR-L4 -|
```

C

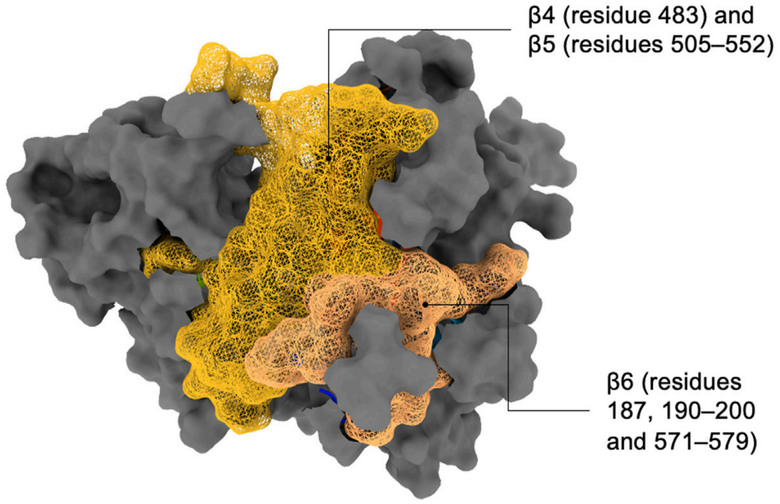


FIGURE 1
Antibody sequence and epitope of MVH. **(A)** and **(B)** display the antibody sequence of the heavy and light chains, respectively. **(C)** illustrates the head domain of the measles hemagglutinin, showcasing the epitope used in this study represented as a mesh-like surface. The non-epitope region is colored gray.

the core epitope. To achieve this, we performed interface analysis of the predicted holo form using UCSF Chimera (Pettersen et al., 2004). We considered residues within a 5 Å distance from both the core epitope and the antibody as interface residues. Among the interface residues identified for 7C6, 46 residues were found in MVH, with 62.8% of them belonging to the core epitope region. In contrast, a total of 31 residues were identified in the antibody as interacting residues (13 and 18 residues in the heavy and light chains, respectively). Notably, all of the identified interface residues

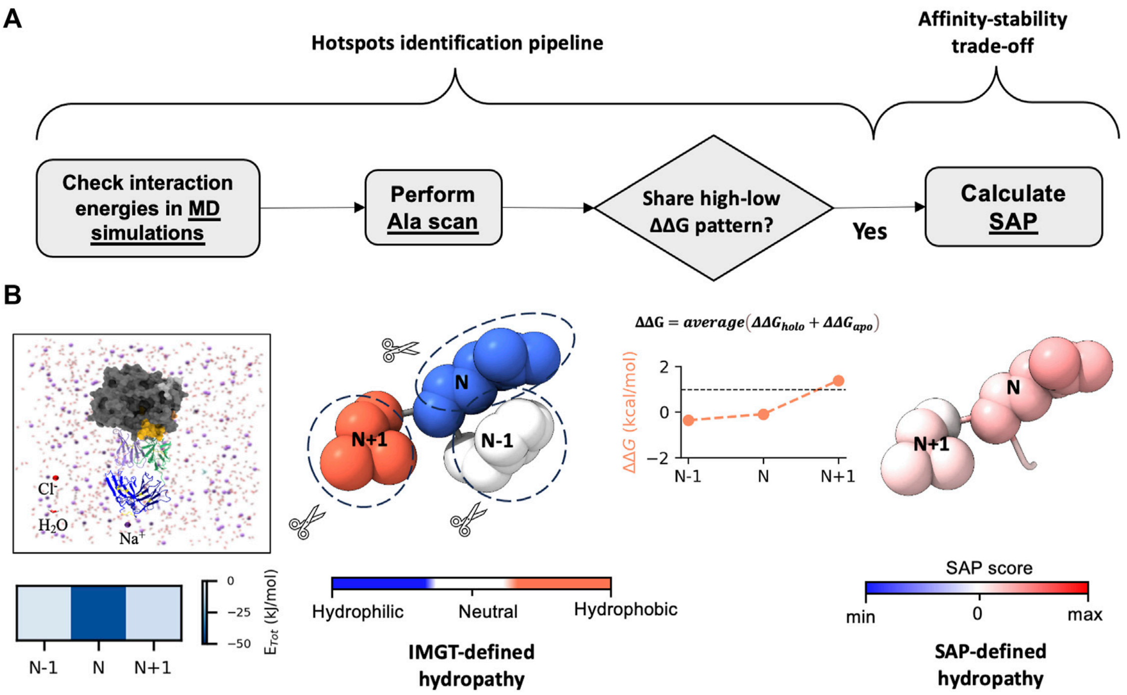


FIGURE 2
Workflow of hotspots identification pipeline. **(A, B)** display the steps for hotspots identification and their link to affinity-stability trade-offs. **(A)** presents the workflow of the *in silico* experiments, which involves checking the interaction energies of the complex structure by using MD simulations. We identify the residue N exerting strongest interaction energy (total non-bonded interaction energy (E_{Tot})). We then perform *in silico* alanine scan (Ala scan) on the structures (both apo and holo forms). After averaging the ΔΔG from both apo and holo structures, we check whether residue N shares a high-low ΔΔ*G* pattern with its neighbor residues N-1 and N+1. If it does, we pair the residues and predict the pair as hotspots. Finally, we check the relative hydropathy of the pair by calculating spatial aggregation propensity (SAP). This step helps in understanding the interplay between affinity-stability trade-offs. **(B)** illustrates the workflow shown in **(A)**. The residues are displayed as atoms in sphere style, with color coding based on the IMGT-defined hydropathy and SAP score.

in the heavy chain and 77.8% in the light chain were located within the CDRs. Throughout this study, we followed Chothia numbering scheme (Chothia and Lesk, 1987; Al-Lazikani et al., 1997) to define CDRs (Figures 1A, B). Moving on to the 8F6 antibody, we identified 36 interacting residues in MVH, with 74.4% of them belonging to the core epitope region. Additionally, we found 27 interacting residues in the 8F6 antibody, out of which 18 were in the heavy chain, and all of them situated within the CDRs. From this analysis, we deduced that the CDRs of the heavy chain exhibited a reasonable number of interacting residues in the antibodies, particularly in the case of 8F6. Notably, the light chain of 7C6 exhibited a higher presence of interfacial residues than the heavy chain, emphasizing its importance in the interactions.

To computationally assess the validity of the predicted interacting residues of the antibody-antigen complexes, we employed MD simulations. In MD simulations, model structures are refined as they interact with surrounding explicit water molecules. This makes MD simulations a common tool for refining model structures (Heo et al., 2021). To confirm the quality of the simulations we first checked convergence of the three independent MD simulations for each antibody-antigen complex. The convergence of the predicted complex is difficult to achieve since the crystal structure of the MVH (PDB ID: 2ZB6) we used in our docking simulations has missing residues (167–183 and 240–246) in the non-epitope region (Figure 1C). Therefore, we trimmed the terminals of MVH and repaired the missing residues 240–246 through Modeller (Fiser et al., 2000; Webb and Sali, 2016) before the MD simulations. In addition, we modeled the constant regions of the antibody to mimic the Fab format used in experiments. The contribution of the modeled regions was evident in the simulation runs which caused the higher structural deviations in the trajectories. Given that our above interface analysis of the docked models indicated that the interacting residues were primarily located in the core epitope, we focused our attention on verifying the potential interactions within the core epitope and Fv of antibody. Therefore, we checked the convergence using the root mean square deviation (RMSD) of the Cα atoms for these regions, which remained quite stable after 170 ns (Supplementary Figure S3). We used the last 70 ns of the trajectories after achieving convergence in the analyses below.

To identify the residue-wise contributions of interactions between antibody CDRs and the core epitope more quantitatively, we computed the interaction energies (comprising van der Waals and coulomb energy) based on the MD trajectories (Figure 3). The probability distribution function of the non-bonded energy components for both antibodies showed strong interaction energies toward the core

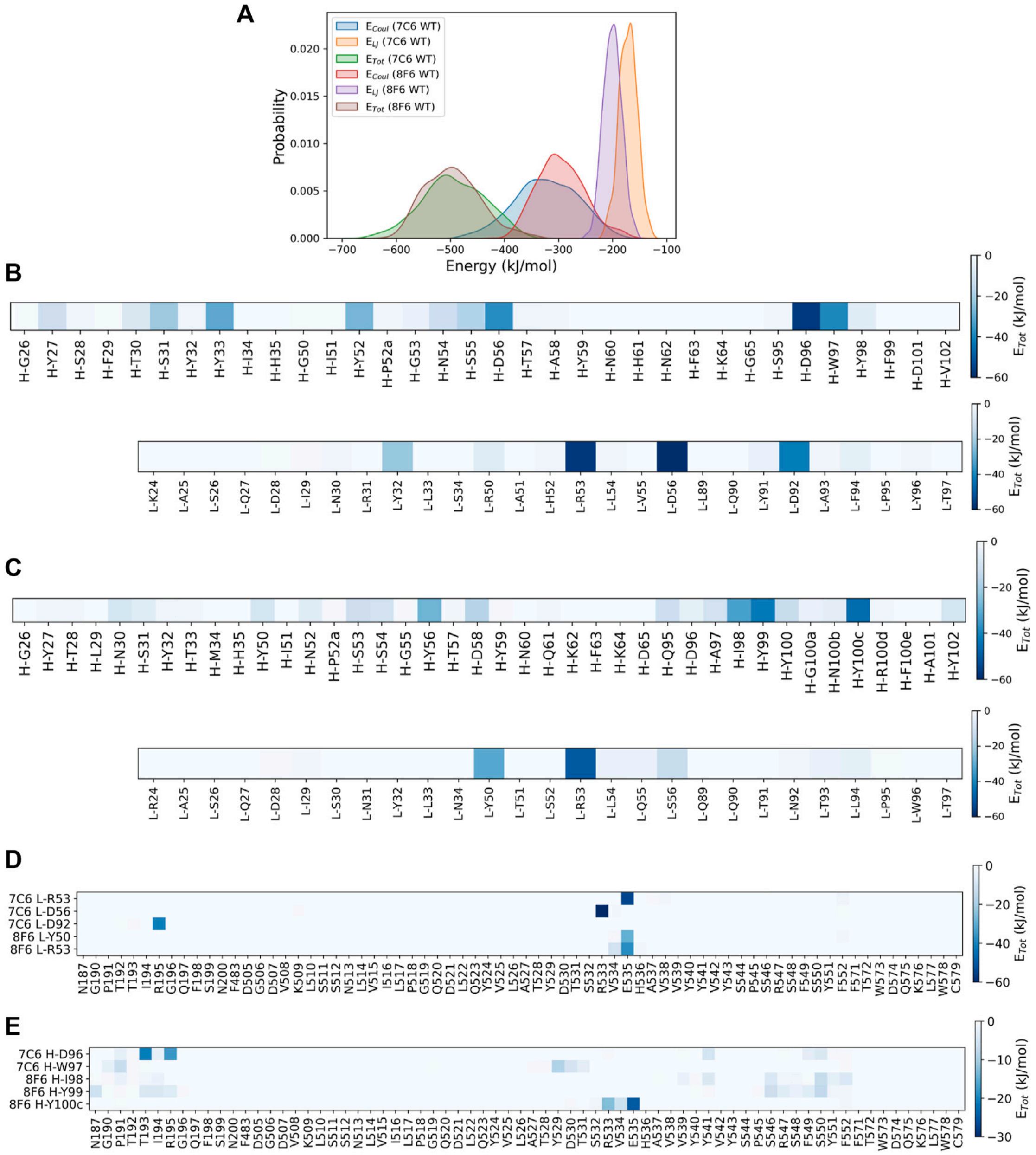


FIGURE 3
Identifying interacting residues by MD simulations. **(A)** probability distribution functions for interaction energies between antibody and epitope. The data shown for the average of three independent MD simulations. The total non-bonded interaction energy (E_{Tot}) shown in kJ/mol, E_{Tot} = Coulombic energy (Coul) + Lennard-Jones (LJ) energy. **(B, C)** display the heatmaps of residue-wise E_{Tot} between the epitope and CDRs of 7C6 and 8F6, respectively. Interaction energies for heavy chain and light chain CDR residues are shown. **(D, E)** present heatmaps of residue-wise E_{Tot} between the CDR and the epitope. These figures illustrate a quasi-epitope mapping for CDR-L and CDR-H residues with largest interaction energies shown in Figure 3BC, respectively.

epitope. The well-defined peaks observed in Figure 3A suggest the system was in stable configurations during the interactions. We also calculated the energy contribution from residues in all six CDRs (Figure 3BC). The total interaction energy observed for CDRs of 7C6 was −488.5 kJ/mol (H-CDRs: −285.4 kJ/mol and L-CDRs: −203.1 kJ/mol), which was stronger than the interaction energy of 8F6 CDRs at −409.6 kJ/mol (H-CDRs: −294.7 kJ/mol and L-CDRs: −114.9 kJ/mol), in agreement with our experimental results of SPR (Table 1). On a residue-wise basis, a few L-CDR residues contributed significantly to the interaction energy (Figure 3B), whereas multiple heavy chain residues made notable contributions. For 8F6, a similar energy contribution profile was observed for its H-CDR residues (Figure 3C). It is worth noting that all six CDRs contributed to the interaction energies observed in 7C6. In contrast, for 8F6, L-CDRs appeared to make no discernible contribution to the interaction energies except for CDR-L2. We further calculated the interaction energies between the core epitope residues and

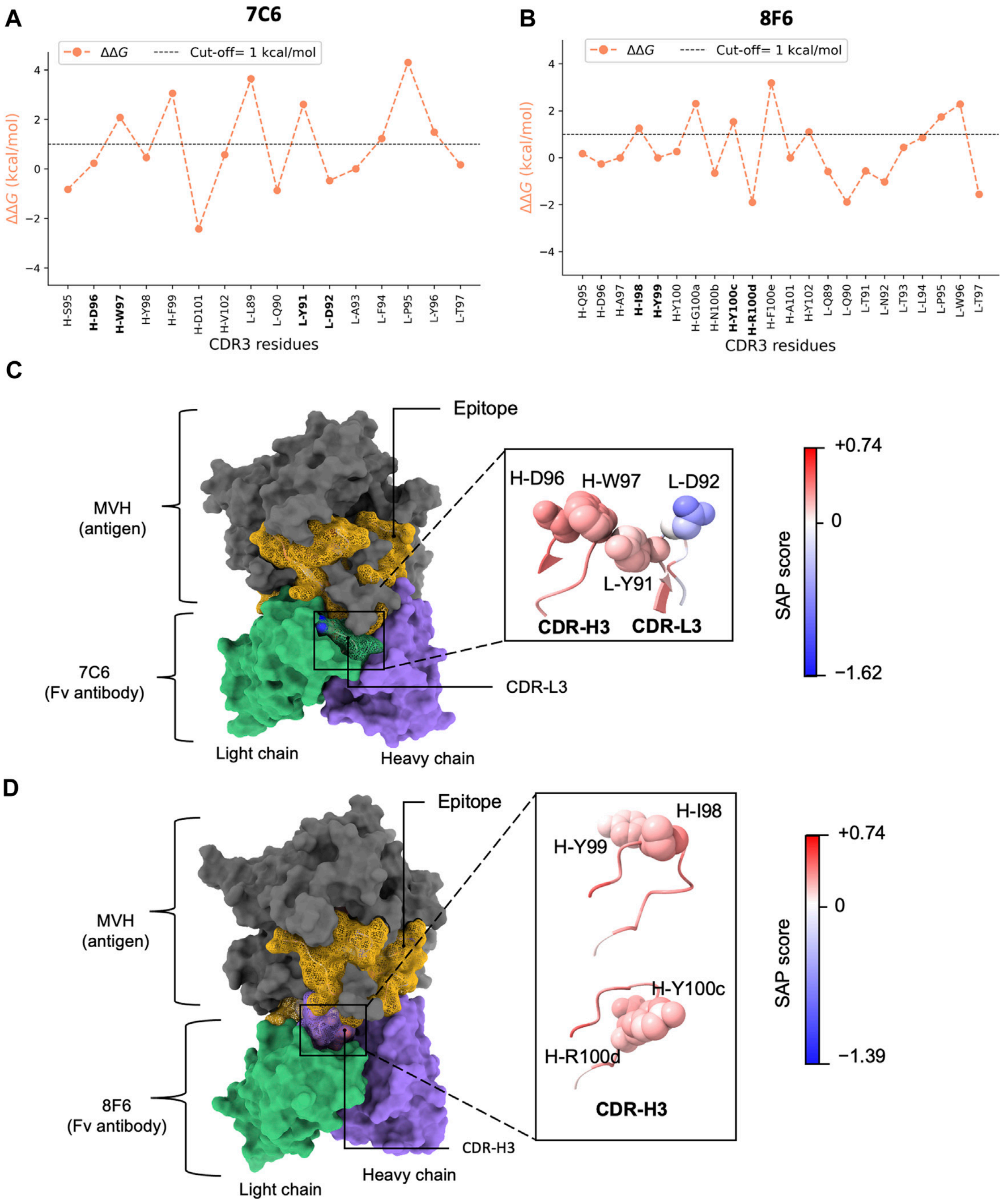


FIGURE 4
In silico alanine scanning and relative hydropathy analysis. **(A, B)** show the results of *in silico* alanine scanning using the FoldX AlaScan command. The results are depicted as an orange line. The ΔΔG cut-off = 1 kcal/mol is represented by dashed line. These plots highlight the four identified residue pairs for antibodies 7C6 (illustrated in **(C)** and 8F6 (illustrated in **(D)**. **(C)**, **D)** display the holo forms of the 7C6 and 8F6 antibodies, respectively. The epitope is represented as a golden mesh-like surface, the non-epitope region is colored in gray, and the heavy and light chains are shown in purple and green, respectively. The CDR3 region is highlighted with a mesh-like surface. The identified residue pairs are displayed as atoms in sphere style, with color coding based on the SAP score. The corresponding SAP scale used for both antibodies is also depicted in the image. The molecular representations were visualized using UCSF ChimeraX.

the selected CDR residues that exhibited significant interaction energies, as seen in Figures 3A, B. Residue L-R53 in the L-CDR2 of both antibodies demonstrated a pronounced interaction energy with residue E535 of the core epitope (Figure 3D). More core epitope residues interacted with H-CDR residues (Figure 3E) than with L-CDR residues (Figure 3D). Figures 3D, E illustrate a quasi-epitope mapping of the MVH for 7C6 and 8F6 antibodies. The possible binding site of 7C6 and 8F6 could be within β6 (187–195) and β5 (529–535, 541, and 546–552).

2.4 *In silico* alanine scanning to identify hotspots for thermal stability and binding affinity

The next step in our proposed workflow (Figure 2) entails confirming the key residues for binding. To achieve this, we performed *in silico* Ala scanning (hereafter Ala scan) using FoldX (Schymkowitz et al., 2005). We employed Ala scan on both apo (antibodies only) and holo (antibody-antigen complexes) forms. We included apo forms in this analysis because the loss of binding may originate from the collapse of the antibody structure itself. Hereafter, we referred to $\Delta\Delta G$ as the average value estimated from the $\Delta\Delta G$ of both the apo and holo forms. We utilized the standard cut-off of $\Delta\Delta G \geq 1$ kcal/mol for hotspot prediction in protein engineering (Liu et al., 2011; Peng et al., 2014). Positions with $\Delta\Delta G$ above the cut-off are identified as predicted hotspots. From the $\Delta\Delta G$ profile, we first noticed that hydrophobic residues tend to exhibit higher $\Delta\Delta G$ (Supplementary Figure S4, 5). This is likely because they were buried in the antibody structures or at the antibody-antigen interfaces and mutating such a buried residue to Ala would lead to an unstable structure in the apo and holo forms, respectively.

Second, we also observed a distinct visualization of the high-low $\Delta\Delta G$ pattern (Figure 2; Supplementary Figure S4, 5), which prompted us to further focus on a subset of 2 residues or "pair". Together with the MD results (Figure 3), we inferred that certain residues paired with its sequential adjacent residues. The sequential pairs for 7C6 were L-R53/L-L54, L-V55/L-D56, L-Y91/L-D92 and H-D96/H-W97 (Supplementary Figure S4). For 8F6, the sequential pairs were L-R53/L-L54, H-I98/H-Y99 and H-Y100c/H-R100d (Supplementary Figure S5). Since our focus of this study is to understand the intricate interplay between binding affinity and stability, we decided to focus on the sequential pairs found in CDR3: L-Y91/L-D92 in 7C6 CDR-L3, H-D96/H-W97 in 7C6 CDR-H3 and H-I98/H-Y99 and H-Y100c/H-R100d in 8F6 CDR-H3 (Figure 4AB). We hypothesized that focusing on the CDR3 region would provide insights into affinity-related trade-offs since, among the CDRs, CDR3 contributes primarily to the binding affinity.

Interestingly, considering the amino acid types, Tyr exhibited a duality nature in the Ala scan depending on the partner residues. When the partner residue is hydrophilic, i.e., Asp (7C6 L-D92) or Arg (8F6 H-R100d), Tyr showed high $\Delta\Delta G$. On the other hand, when the partner residue is hydrophobic (8F6 H-I98), Tyr showed low $\Delta\Delta G$. Despite being an aromatic residue, Tyr falls under the "neutral" class of IMGT-defined hydropathy (Pommié et al., 2004), which may explain this duality in the $\Delta\Delta G$ profile.

Thus, from the above analysis, it was suggested that a pattern of high-low $\Delta\Delta G$ observed in this study (Figure 4AB) may be utilized to identify residues in subset or pair that potentially contribute both thermal stability and binding affinity. MD simulation helped in drawing our attention to the residues in CDRs where the pattern is distinct. Even though more favorable interaction energies were observed for L-R53, we chose to focus on the residues in CDR3 that matched our criteria of selection. A high-low $\Delta\Delta G$ pattern shared by the pairs suggested that the hydrophobic partner residues likely aid in interactions by stabilizing the conformation of the partner residues tailored for binding.

2.5 Relative hydropathy analysis

The dual hydropathic nature of Tyr prompts questions about its relative hydropathy and its contribution to affinity and stability. To explore this, we investigated the factors that influence change in amino acid hydropathy. We observed non-bonded interactions (van der Waals and coulomb) between antibody and antigen with Coulombic interactions playing a dominant role (Figure 3A). Antibody 7C6 exhibited stronger attractive forces compared to 8F6. The surrounding environment, including water molecules (hydration) in a biological system, influences these interactions. Changes in the environment can alter the chemical nature of an amino acid, affecting the hydrophobic or hydrophilic nature of the amino acid side chain. Recently, Rienzo et al. characterized the hydropathy profiles of amino acid side chains at the protein-solvent interface (Di Rienzo et al., 2021). Inspired by their work, we were prompted to calculate the relative hydropathy of the identified pairs based on their surroundings.

We computed the relative hydropathy on the holo form (Figure 4CD) using spatial aggregation propensity (SAP) (Chennamsetty et al., 2009). SAP identifies hydrophobic patches on a protein's surface based on a defined radius (R) called SAP radius. Chennamsetty and colleagues (Chennamsetty et al., 2009) reported that hydrophobic interaction plays a key role in protein aggregation, thus impacting stability. A SAP radius of 5 Å could identify the aggregation-prone patches with detailed view. Conversely, a SAP radius of 15 or 20 Å tends to eliminate the hydrophobic patches and favor the hydrophilic patches (Chennamsetty et al., 2010). Thus, to identify the true nature of the amino acid pairs, we employed a SAP radius of 10 Å that could favor both hydrophobic and hydrophilic patches, maintaining a balance between them. We provided a schematic illustration of the alterations in hydropathy in Figure 2B. Upon analyzing the residue pairs in CDR3, we observed pair L-Y91/L-D92 in CDR-L3 of 7C6 (Figure 4C), have a balanced hydrophobic and hydrophilic nature respectively, while the other pairs H-D96/H-W97 in CDR-H3 of 7C6, and H-I98/H-Y99 and H-Y100c/H-R100d in CDR-H3 of 8F6 contributed to the hydrophobic gradient (Figure 4CD). The observation that an IMGT-defined hydrophilic Asp and Arg experiences a distinct change in its hydropathic nature (such as 7C6 H-D96 and 8F6 H-R100d becoming hydrophobic, while 7C6 L-D92 remains hydrophilic) may provide valuable insights into their connection with stability. This is particularly relevant since charged residues are typically not buried without neutralizing their charge, often by forming salt bridges with other residues. Without such compensation, buried charged residues could lead to unstable protein structures. This emphasizes the critical role of the protein environment in considerations of residue hydropathy and its impact on the trade-off between stability and binding affinity.

To further explore the relationship between affinity and stability, and to validate our computational predictions, we subjected the identified paired residues to *in vitro* alanine scanning experiments. This *in vitro* validation is particularly critical given the limited scope of our dataset, comprising only four pairs. Drawing broad conclusions from such a small dataset can be precarious. With this in mind, our experimental validations were designed to assess whether mutations at these positions could alter the characteristics of these pairs, thereby affecting both binding affinity and thermal

TABLE 2 Kinetic and thermal stability parameters of the 7C6 and 8F6 mutants.

	k_{on} (×10^5 $M^{-1}s^{-1}$)	k_{off} (×10^{-4} s^{-1})	K_D at 25°C (nM)	T_m (°C)	ΔT_m (°C)
7C6 WT	11.4 ± 4.8	4.1 ± 1.7	0.4 ± 0.2	73.9 ± 0.9	
L-Y91A	1.8 ± 0.5	168.8 ± 35.0	97.5 ± 8.2	71.3 ± 0.7	−2.6
L-D92A	7 ± 0.4	4.6 ± 0.4	0.7 ± 0	73.7 ± 1.7	−0.3
H-D96A	14.2 ± 0.3	90.8 ± 0.2	6.4 ± 0.1	75 ± 1.7	1.0
H-W97A	43.6 ± 5.1	37.6 ± 1.6	0.9 ± 0.1	72.9 ± 0.3	−1.0
L-Y91F	27.0 ± 1.0	13.2 ± 0.3	0.5 ± 0	71.7 ± 1.1	−2.2
L-D92F	14.4 ± 0.2	4.4 ± 0.2	0.3 ± 0	72.7[a]	−1.2
H-D96F	14.4 ± 1.3	363.1 ± 27.3	25.2 ± 0.4	72.8 ± 0.8	−1.1
8F6 WT	3.4 ± 2.9	2.8 ± 1.7	0.9 ± 0.2	68.4 ± 0.9	
H-I98A	2.2 ± 0.8	153.3 ± 48.8	70.5 ± 2.4	69.1 ± 1.0	0.7
H-Y99A	0.1 ± 0	8.9 ± 0.2	99.5 ± 2.8	68.6 ± 0.4	0.2
H-Y100cA	- -	- -	N.D.	69.7 ± 0.7	1.3
H-R100dA	0.1 ± 0	40.5 ± 0.3	284.0 ± 0.9	70.2 ± 0.4	1.8

N.D., not determined as kinetic fitting was not applicable.

[a] T_m measurements for 7C6 L-D92F were conducted only once due to insufficient protein quantity.

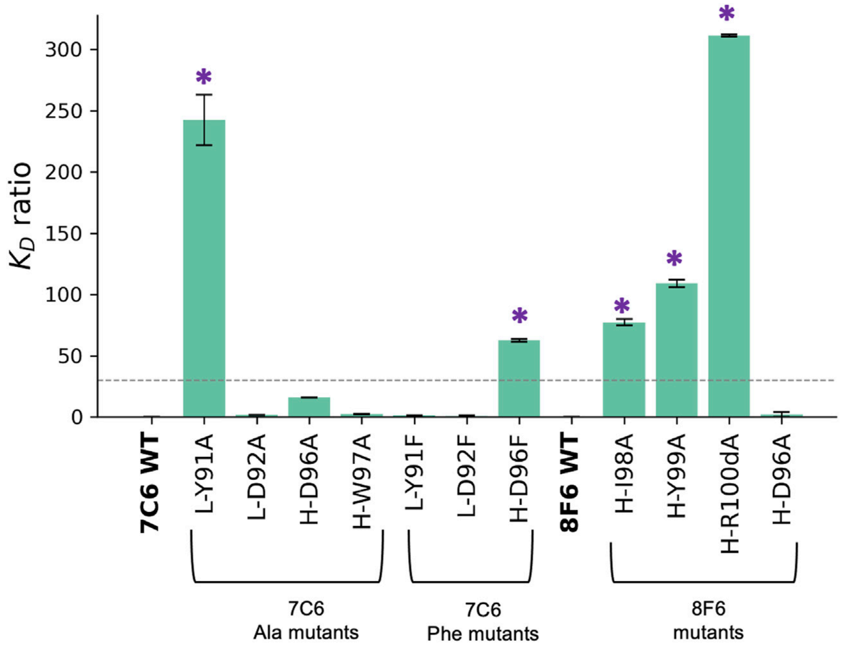


FIGURE 5
Effect of mutations on binding affinity. Binding affinity is measured by SPR. Effect on binding affinity was measured in terms of K_D ratio = K_D of mutant/K_D of wild type. The wild type (WT) 7C6 and 8F6 antibodies served as the baseline (i.e 0), indicating no change in binding affinity. Error bars were calculated from three independent measurements, and asterisks denote mutants that exhibit a significant change in binding affinity, which corresponds with the > 30-fold decrease in binding affinity (Akiba and Tsumoto, 2015).

stability. For the pairs identified in 7C6, which has two types of pairs within CDR3 (Figure 4C), in addition to introducing alanine, we also predicted other amino acid substitutions at the same positions using standard *in silico* tools.

We employed two methods to predict new mutations based on the high-low $\Delta\Delta G$ pattern derived from Ala scan analysis of FoldX. For residues with high $\Delta\Delta G$ values (such as 7C6 L-Y91 and H-W97), which we hypothesized have an impact on stability, we utilized

Rosetta's Cartesian_ddg application on the apo form to predict potential mutations. We chose to use two different methods for $\Delta\Delta G$ calculations–FoldX and Rosetta–because they are orthogonal methods. They utilize distinct rotamer libraries and scoring functions, capturing different aspects of the underlying physics. On the other hand, residues with low $\Delta\Delta G$ values (7C6 L-D92 and H-D96A) suggested that the effects of mutations at these positions are minimal. Therefore, we continued to use FoldX to predict mutations for these residues in both apo and holo forms. Mutations with values below the cut-off (−1 kcal/mol) from the *in silico* mutational analysis were chosen for the *in vitro* mutagenesis study (Supplementary Figure S6). The only exception was for 7C6 H-W97, which did not meet the cut-off. The amino acid Phe was predicted for residues L-Y91, L-D92 and H-D96.

2.6 Experimental physicochemical analysis of the antibody mutants

We expressed the mutants (Table 2) and purified them using size-exclusion chromatography (SEC). Similar to the WT antibodies, we conducted SPR analysis for the mutants to measure the binding affinity and compared the change in binding affinity or K_D ratio (Figure 5). For the Ala mutants of the predicted hydrophobic-hydrophobic pairs identified in 8F6 (H-I98/H-Y99 and H-Y100c/H-R100d), significant loss of binding affinity was observed. Ala mutation to H-Y100c exhibited a weak binding to the extent that kinetic fitting was not applicable (Supplementary Figure S7B), revealing that this position is also a hotspot for binding. This suggests that all residues involved in the hydrophobic-hydrophobic pairs of 8F6 were critical for binding. As a control, we chose 8F6 H-D96, which is spatially near the hotspot pair H-Y100c/H-R100d in 8F6 (Supplementary Figure S8). Although H-D96 was not predicted as a hotspot in our approach, its proximity and the charged nature of aspartic acid suggested its potential importance for binding. However, despite its location within the CDR-H3, H-D96 showed a negligible change in binding affinity (Figure 5 and S8). This outcome serves as validation for our hotspot identification pipeline (Figure 2), confirming the accuracy of not identifying this residue as a hotspot.

For the Ala mutants of 7C6, we identified L-Y91 from the hydrophobic-hydrophilic pair as a key residue with a loss in binding affinity of about 242-fold. On the other hand, its partner residue, L-D92, had no significant effect on binding affinity (Figure 5). In contrast, within the hydrophobic-hydrophobic pair, the H-D96A and H-W97A mutants in CDR-H3 of 7C6 showed a 16-fold loss and a negligible change in binding affinity, respectively. Additionally, the differences in the K_D ratio between key residues found in CDR-L3 and CDR-H3 suggested that light chain accommodated the primary hotspot. The Ala mutants resulting in reduced binding affinity of the high-affinity antibodies to MVH echoed one common cause of loss of binding, that is faster k_{off} (Table 2).

The Phe mutants to 7C6 showed tolerance for Phe mutation at the primary hotspot pair (L-Y91/L-D92), which is consistent with the docking scores (Supplementary Figure S9). Furthermore, the hydropathy of these Phe mutations aligned with the hydropathy of the pair in the WT, suggesting an explanation for the pair's ability to tolerate the mutations. In contrast, the mutant H-D96F in CDR-H3 showed a 63-fold loss in binding affinity. This suggests that the secondary hotspot is also contributing to the overall binding affinity of 7C6 and did not tolerate a mutation to a bulky residue like Phe.

We performed circular dichroism (CD) to observe any structural changes that may have occurred due to the point mutations causing these changes in binding affinity (Figure 6). The CD spectrum for all the mutants retained the beta-sheet like folding that generally observed for Fab antibodies (Cathou et al., 1968). In addition, some changes in molar ellipticity were observed for the mutants, but the results were not conclusive to provide sufficient information about the type of structural changes. Thus, we next performed thermal stability measurements to observe the effect of mutations on the melting temperature (T_m) of the mutant antibodies.

Due to insufficient yield, we employed CD measurements instead of DSC to determine the T_m of the mutants. The T_m of WT 7C6 remained consistent in both DSC and CD measurements (T_m in CD: 73.9°C ± 0.9°C and DSC: 73.9°C ± 0.3°C), while a negligible difference was observed for the 8F6 WT antibody (ΔT_m ~0.5°C). Therefore, we used the T_m obtained from CD measurements to compare the ΔT_m upon mutation (Table 2; Supplementary Figure S10). In the CD measurements, we observed that some mutants, such as L-D92A and H-D96A in 7C6, displayed larger error bars (±1.7°C). While differences in T_m values might seem insignificant, the slopes of the CD profiles in Supplementary Figure S10 could offer biophysical insights. For instance, although the ΔT_m value of L-D92A is only 0.3°C, a seemingly negligible difference from the WT, its slope increases more rapidly than the WT. This implies that the mutant unfolds faster than the WT upon exposure to increasing temperatures. Therefore, despite the need for caution, the subtle variations in T_m observed in this study could provide valuable insights into the affinity-stability trade-offs of the antibodies.

The thermal stability results offer revealing insights when correlated with the nature of the amino acid pairs, specifically their relative hydropathy. Figure 7 illustrates the relationship between molar Gibbs free energy (ΔG) and stability, highlighting the intricate interplay between affinity-stability trade-offs. For hydrophobic-hydrophobic pairs found in CDR-H3 of both antibodies, residues H-D96 in 7C6, as well as H-I98/H-Y99 and H-Y100c/H-R100d in 8F6, exhibited an increase in T_m, with a less favorable ΔG. This implies that the mutations have improved the thermal stability of the antibodies; however, this enhancement comes at the expense of an energetically less favorable binding reaction, resulting in a decrease in affinity. An exception among the hydrophobic-hydrophobic pairs was observed with H-W97 in 7C6. An alanine mutation in this residue led to a decrease in T_m (ΔT_m = −1.0°C), but did not significantly affect binding affinity (Table 2). Similar to a Tyr residue, a Trp residue seems to have a unique function; it contributes to aromatic interactions, acts as a hydrogen bond donor, possesses a large hydrophobic surface, and can shield delicate hydrogen bonds from water (Samanta et al., 2000). In contrast, in the case of the hydrophobic-hydrophilic pair within 7C6's CDR-L3 (L-Y91/L-D92), a less favorable ΔG was observed alongside a decrease in T_m. This suggests that the mutation has resulted in an energetically less favorable binding interaction, consequently leading to diminished binding affinity and a decrease in thermal stability. Notably, the negative ΔG associated

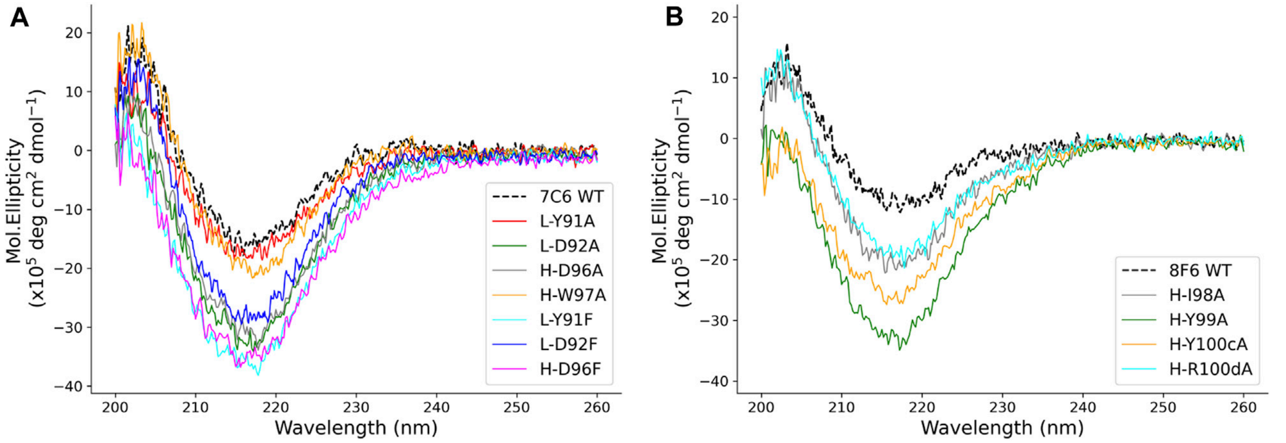


FIGURE 6
CD profile of the mutants. **(A, B)** show the CD profile of the mutants for 7C6 and 8F6, respectively.

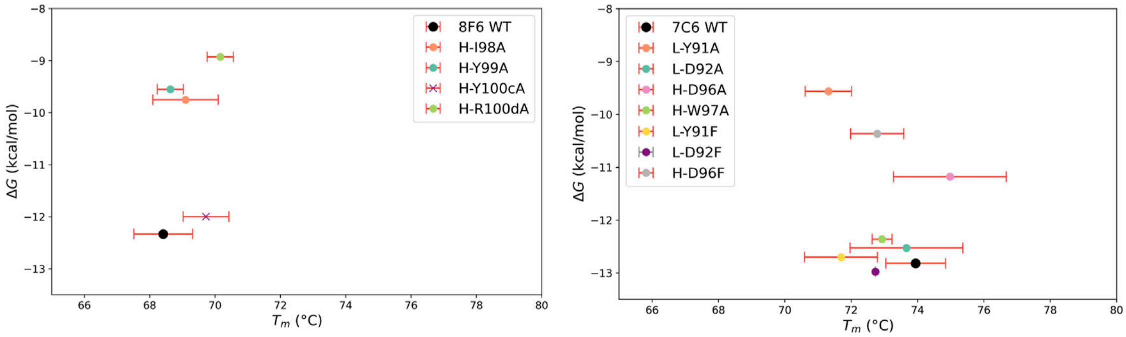


FIGURE 7
Relationship between the molar Gibbs free energy (ΔG) and thermal stability (T_m). ΔG is obtained from the equilibrium dissociation constant, K_D using the following equation: $\Delta G = -RT\ln(1/K_D)$, where R is gas constant (0.0019872 kcal/mol·K) and T is temperature in Kelvin (298.15 K). Thermal stability is represented by melting temperature (T_m). Error bars are shown for T_m calculated from three independent measurements for all antibodies except 7C6 L-D92F due to insufficient protein quantity. This is shown by gray error bar. Since kinetic parameters were not obtained for 8F6 H-Y100cA, the corresponding ΔG value does not directly reflect the effect on free energy for this mutant. To point out this discrepancy we marked this mutant with X.

with our predicted Phe mutant of 7C6 L-D92 suggests that this hydrophilic position is well-suited to accommodate the mutation and promotes an energetically favorable binding reaction. Among the identified hotspot pairs, the residues 7C6 L-Y91 and H-D96, along with 8F6 H-I98, H-Y100c and H-R100d, had a notable impact on T_m. The marginal effect of 8F6 H-Y99A on T_m corroborates our hypothesis about the dual role of Tyr, as evidenced by our pattern analysis.

Our computational analysis and experimental measurements suggest that relative hydropathy influences the trend in thermal stability, whether increasing or decreasing (Figure 4; Table 2), while the IMGT-defined hydropathy highlights the importance of a residue's contribution to stability (Figures 2, 4). This was particularly observed with the dual nature of Tyr (8F6 H-Y99). Recognizing the importance of both definitions provides a better understanding of the factors determining stability. Therefore, this study contributes to laying the groundwork for further exploration into the dual nature of Tyr in antibody and protein research.

3 Discussion

This study aims to investigate the binding affinity and stability of anti-MVH neutralizing antibodies, with the objective of exploring a potential correlation between binding affinity and stability. For this purpose, we proposed a hypothesis that high-affinity antibodies with differences in stability could provide valuable insights for our research objective. The physicochemical analysis revealed that antibodies 7C6 and 8F6 exhibited rapid association and slow dissociation with MVH, indicating high binding affinity (<1 nM). We focused on these two antibodies, which showed a significant difference in thermal stability (ΔT_m, ~6°C). Since no antibody crystal structure was available at the time of writing, homology modelling and knowledge-based local docking were performed to generate the apo (antibody) and holo (complex) forms, respectively.

Modeling antibody structures remains challenging, especially when the CDR-H3 extends beyond the average length (i.e., > 13–14 residues). While the modeling accuracy for non-CDR-H3 sections of antibodies is often satisfactory, even the state-of-the-art deep learning methods still struggle with CDR-H3 conformation predictions. On average, these predictions often deviate by more than 2.0 Å in backbone RMSD from crystal structures (Ruffolo et al., 2023). Such a 2 Å variance in backbone conformations is significant; even minor discrepancies (<1.0 Å) in backbone configurations can substantially alter the energy landscape of protein-protein interactions (Kuroda and Gray, 2016). Consequently, computer-guided affinity maturation studies without antibody crystal structures are scarce. A standout example is the work by Cannon et al. They integrated experiments with computational modeling to guide the affinity maturation of an antibody targeting an antigen (Cannon et al., 2019). Mutagenesis experiments were used to validate docking models and pinpoint the potential binding modes of the antibody-antigen complex. This was succeeded by re-docking of the complex and further design calculations based on the predicted model complex.

In our study, we sought to improve computational modeling accuracy by performing MD simulations immediately after modeling the antibody and docking it with the antigen. Within MD simulations, model structures undergo adjustments by interacting with the surrounding environment, including explicit water molecules. Based on these wholly computational outcomes, we were able to identify hotspots in the antibody-antigen interactions, a finding that our *in vitro* mutagenesis experiments subsequently validated. While the accuracy of $\Delta\Delta G$ calculations by FoldX may be influenced by the quality of the input structures (Buß et al., 2018), our study's strength lies in the experimental validations that corroborate our computational predictions. Although crystal structures of the complexes between MVH and the anti-MVH antibodies would offer valuable insights into molecular-level interactions, our study suggests that knowledge-based rigid-body docking simulations, followed by explicit solvent MD simulations, could serve as an effective alternative for exploring these interactions.

In protein engineering, the defined hotspots are a subset of residues composed of high affinity residues surrounded by low affinity residues as O-ring structure (Bogan and Thorn, 1998; Soga et al., 2010; Akiba and Tsumoto, 2015). We proposed a novel high and low $\Delta\Delta G$ pattern that appears to effectively recognize these hotspots as a subset of two partner residues or pair. This pattern aided in identifying the hotspots responsible for significant loss in binding affinity for both the 7C6 and 8F6 antibodies. Through our investigation of high-affinity anti-MVH antibodies, we suggested a potential relationship between affinity and stability, which may offer insights into their trade-offs. We noted two distinct types of pairs based on their relative hydropathy: a) hydrophobic-hydrophilic and b) hydrophobic-hydrophobic. While the former type tended to show a decrease in stability along with a loss in binding affinity, the latter type seemed to maintain or increase stability despite a decrease in affinity.

In general, CDR-H3 is primarily responsible for antigen recognition and binding. However, it is intriguing to note that the highest affinity antibody, 7C6, possesses a shorter CDR-H3 (consisting of only 7 residues) compared to the other anti-MVH antibodies (2F4 and 10B5 with CDR-H3 of 12 residues, and 8F6 with CDR-H3 of 13 residues). This disparity in CDR-H3 length may explain why the CDR-H3 of 7C6 acts as a secondary hotspot.

A comparison of MVH binding to its receptors and antibodies in Supplementary Figure S11 shows that the binding site of 7C6 is predicted to be located within the region composed of amino acid residues 190–200, which is part of the immunodominant epitope (amino acids: 190–200 and 571–579). This epitope has been identified as a recognition site for mAb BH26, which inhibits the binding of approximately 60% of human serum antibodies in vaccinees and individuals recovering from measles (Ertl et al., 2003; Tahara et al., 2016). On the other hand, the predicted binding site of 8F6 lies within amino acids 505–552, which corresponds to the receptor binding epitope (residues 187, 190, 483, and 505–552). Additionally, residues within the CDR-H3 of 8F6 were found to interact with R533, which is part of a conserved neutralizing epitope (residues F483, D505, R533, Y541, and Y543). While there was a clear correlation between docking score and binding affinity, we also observed that the docking pose correlates with the inhibition capabilities of the anti-MVH antibodies. Supplementary Figure S11B illustrates the footprints of both receptors and the antibody on MVH. Although the overall binding sites seem similar, the CDR-H3 of 8F6, containing a hotspot (H-R100), is located near the hydrophobic pocket within the β4-β5 groove, a region implicated in receptor binding (Zhang et al., 2013). In contrast, a hotspot of 7C6 (L-Y91), experimentally identified in this study, is positioned in a region more distal from the hydrophobic pocket (Supplementary Figure S11B). This difference in the location of hotspots may account for the lower neutralizing capability of 7C6 compared to 8F6, as reported by Sato and colleagues (Sato et al., 2018), despite having higher affinity among the anti-MVH antibodies (Supplementary Figure S11A). Mutations at these conserved neutralizing epitope residues have been shown to facilitate immune escape from neutralization by the monoclonal antibody 2F4 (Santiago et al., 2010; Tahara et al., 2013). Although a co-crystal structure is currently unavailable, this study suggested the plausible binding mode of high-affinity antibodies to MVH (Supplementary Figure S11). Consequently, these findings open up avenues for further research on anti-MVH antibodies, providing valuable insights into their development.

Thus, this study highlighted the importance of a balance between hydrophobic and hydrophilic residues for achieving high affinity and stability in anti-MVH antibodies (Supplementary Figure S9). These findings pave the way for computational design strategies aimed at enhancing the affinity and stability of low-affinity anti-MVH antibodies, such as 2F4 and 10B5, in future research endeavors. When applying our approach to analyze the low-affinity anti-MVH antibodies, we identified residues that form pairs with Gly (Supplementary Figure S12). The absence of a side chain in one of the paired residues in 2F4 and 10B5 may contribute to their low affinity (Table 1). While Gly is known to play important roles in conformational flexibility, its specific influence on affinity, stability, and neutralization requires further investigation. Additionally, the pairs identified for 2F4 and 10B5 exhibit a hydrophobic gradient, suggesting that hydrophobic-hydrophilic combinations are relatively uncommon.

In this study, the combination of high-low $\Delta\Delta G$ pattern and relative hydropathy analysis exhibit computational promise for addressing challenges related to the trade-offs between affinity and stability in antibody research. By training AI models with this pattern-driven analysis of antibodies, it may be possible to mitigate the need for large-scale experimental data. Therefore, it is essential to validate this pattern on additional antibodies targeting a range of antigens, in order to drive advancements in the field of antibody research facilitated by computational methods.

4 Methods

4.1 Antibody homology modeling and docking to antigen

The RosettaAntibody protocol (Weitzner et al., 2017) in Rosetta (Leman et al., 2020) was used to generate three-dimensional structure of the antibody Fv. To ensure comprehensive analysis, we generated 2000 structures for the top scored grafted model and 200 structures for the other grafted models. This enabled us to select the top-scoring model as a representative of the Fv structure among a wide variation of models.

At the time of our analysis, seven crystal structures of the MVH antigen were available in PDB, two of which were in the apo form, and the remaining structures were in complex with receptors. To identify the most suitable structure for docking, we selected the best resolution structure available (PDB: 2ZB6, 2.6 Å). Using Chimera v1.16 (Pettersen et al., 2004), we manually constructed a putative antigen-antibody complex. Subsequently, we employed the SnugDock protocol to perform a flexible backbone local docking, generating 1,000 poses of the anticipated antigen-antibody complex (Sircar and Gray, 2010).

4.2 Molecular dynamics simulations

The input structure for MD simulation were first modeled using Modeller 10.0 (Webb and Sali, 2016) for repairing the missing residues of MVH and constructing the constant regions of Fab. Then MD simulations were conducted using GROMACS 2022.4 (Berendsen et al., 1995; Lindahl et al., 2001; Abraham et al., 2015) with the CHARMM36m force field (Huang et al., 2017) to explore the behavior of the docked models. To solvate the system, TIP3P water (Madura et al., 1983) was used to fill a cubic box, and the protein was placed at the center with a 10 Å minimum distance to the box edge, while periodic boundary conditions were applied. Additional Na^+ or Cl^- ions were introduced to neutralize the protein charge and simulate a salt solution with a concentration of 0.15 M. Each system was energy-minimized for 5,000 steps with the steepest descent algorithm and equilibrated with position restraints of protein heavy atoms and NVT ensemble, where the temperature was increased from 50 to 298 K during 200 ps. Further non-restrained simulations were performed with the NPT ensemble at 298 K for 240 ns. The time step was set to 2 fs throughout the simulations. A cutoff distance of 12 A was used for Coulomb and van der Waals interactions. Long-range electrostatic interactions were evaluated by means of the particle mesh Ewald method (Darden et al., 1993). Covalent bonds involving hydrogen atoms were constrained by the LINCS algorithm (Hess et al., 1997). A snapshot was saved every 100 ps. We performed three independent production runs with distinct initial velocities. All subsequent analyses were conducted using the GROMACS package.

4.3 *In silico* alanine scanning and mutational design

FoldX (v4) *AlaScan* command was utilized to identify potential hotspots on the antibody (Schymkowitz et al., 2005). Both apo and holo models underwent alanine scanning to predict the effect of mutations on binding with the antigen and antibody. We obtained difference in the free energy, or $\Delta\Delta G$ values for both apo ($\Delta\Delta G_{\mathrm{apo}}$) and holo ($\Delta\Delta G_{\mathrm{holo}}$) forms in kcal/mol from each analysis and averaged them for each position ($\Delta\Delta G$).

$$\Delta\Delta \boldsymbol{G}_{\mathrm{holo}} = \Delta \boldsymbol{G}_{\mathbf{Mut\_}\boldsymbol{holo}} - \Delta \boldsymbol{G}_{\mathbf{WT\_holo}}$$
$$\Delta\Delta \boldsymbol{G}_{\mathrm{apo}} = \Delta \boldsymbol{G}_{\mathbf{Mut\_apo}} - \Delta \boldsymbol{G}_{\mathbf{WT\_apo}}$$
$$\Delta\Delta \boldsymbol{G} = \boldsymbol{average}\left(\Delta\Delta \boldsymbol{G}_{\mathbf{holo}} + \Delta\Delta \boldsymbol{G}_{\mathbf{apo}}\right)$$

Using $\Delta\Delta G$ from Ala scan as a reference, we performed mutational design. Mutations for positions with low $\Delta\Delta G$ were predicted using FoldX *BuildModel* command (van Durme et al., 2011), while positions with high $\Delta\Delta G$ were predicted using the Rosetta's Cartesian_ddg application (Kellogg et al., 2011; Park et al., 2016). A cut-off value of −1 kcal/mol was used for selecting mutants for *in vitro* mutagenesis study.

4.4 Spatial aggregation propensity (SAP)

The SAP (Chennamsetty et al., 2009) algorithm was used to predict relative hydropathy with an in-house CHARMM-based script (Brooks et al., 2009). The SAP was calculated on the holo form and score for each atom within a 10 Å radius was calculated by this algorithm. As a result, a residue wise score was obtained in an output file. The maximum (positive) and minimum (negative) values on the SAP scale indicate hydrophobicity and hydrophilicity of the scale.

4.5 Cloning, expression, and purification of antibodies

The DNA sequences encoding the heavy and light chains of the Fab antibodies were codon-optimized and synthesized by Integrated DNA Technologies, Inc. They were subcloned into separate pcDNA3.4 vectors (Thermo Fisher Scientific), with a His_6 tag fused to the C-terminus of the heavy chains by HiFi DNA assembly (NEB). The DNA of the mutants was prepared by site-directed mutagenesis PCR using the KOD -Plus- Mutagenesis Kit (TOYOBO). The protocol was slightly modified, as we used KOD One PCR Master Mix (TOYOBO) instead of KOD -Plus-. The Fab antibodies were expressed in ExpiCHO cells (Thermo Fisher

Scientific) following the max titer protocol for 8F6 antibody, and in Expi293 cells (Thermo Fisher Scientific) following the manufacturer's standard protocol for rest of the antibodies (Fang et al., 2017; Jain et al., 2017). The cells were cultured by rotating at 125 rpm at 37°C and 8% CO_2 for 5 days for Expi293 cells, and at 32°C and 5% CO_2 for 13 days for ExpiCHO cells after co-transfecting the cells with 13 µg of the heavy and light chain encoding plasmids. The culture supernatant was collected by centrifugation for 10 min at 5,000 g, dialyzed with a solution of 20 mM Tris-HCl (pH 8), 500 mM NaCl, 5 mM imidazole (binding buffer), and filtered with 0.8 µm filters (Advantec). It was loaded onto a Ni-NTA Agarose resin (Qiagen) equilibrated with binding buffer for immobilized metal affinity chromatography. After washing the resin with 10 mL of binding buffer, the protein was eluted with the buffers containing increasing concentrations of imidazole. The antibodies were obtained after further purification by size-exclusion chromatography (SEC) using HiLoad 26/600 Superdex 75 pg column (Cytiva) at 4°C equilibrated with phosphate-buffered saline (PBS) pH 7.4. The concentration of the proteins was calculated from the molecular weights and molar extinction coefficients ($cm^{-1}M^{-1}$) calculated from the amino acid sequences using ProtParam Tool (ExPASy) (Gasteiger et al., 2005) and the absorbance at 280 nm obtained on NanodropOne (Thermo Fisher Scientific).

4.6 Cloning, expression, and purification of antigen hemagglutinin

The pHLsec-vector plasmid with the MVH head domain (amino acid residues 149–617) was transiently transfected into 293S GnTI (−) cells (Hashiguchi et al., 2007). The cells were cultured for 4 days after transfection at 37°C and 5% CO_2. The culture supernatant was collected by centrifugation at 7,000 rpm for 20 min at 4°C and filtration. The collected supernatant was purified with a complete His-Tag Purification Resin (Roche, Cat# 5893682001) affinity column after equilibration with 50 mM $NaH_2PO_4 \cdot 2H_2O$, 150 mM NaCl, and 10 mM imidazole. The resin capturing the head domain of MVH was washed with 25 mM $NaH_2PO_4 \cdot 2H_2O$, 75 mM NaCl, and 5 mM imidazole, and subsequently, the protein was eluted with the buffers containing increasing concentrations of imidazole. The head domain of MVH was obtained after further purification by SEC using Superdex 200 Increase 10/300 GL column (Cytiva) equilibrated with PBS. The concentration of the head domain of MVH was also confirmed following the same protocol as above.

4.7 Surface plasmon resonance (SPR)

The kinetic parameters of the antigen-antibody binding were determined by SPR using Biacore T200 instrument (Cytiva). The antigen hemagglutinin was immobilized on a CM5 sensor chip (Cytiva) at around 500 resonance units following the manufacturer's amine coupling protocol. The Fabs were injected into the sensor chip at a flow rate of 30 µL/min at 25°C. The binding response at the following concentrations 62.5, 125, 250, 500, and 1,000 nM for 2F4 and 10B5, and 1.25, 2.5, 5, 10, and 20 nM for 7C6 and 8F6 wild type antibodies were used for the experiment. The concentrations used for 7C6 mutants except L-Y91A, were 1.25, 2.5, 5, 10, and 20 nM. For 7C6 L-Y91A mutation we used the following dilutions 6.25, 12.5, 25, 50, and 100 nM. For 8F6, two mutants H-D96A and H-I98A, used the following concentrations 1.25, 2.5, 5, 10, and 20 nM, like wild type antibody. For, 8F6 mutants H-Y99A, H-Y100 cA and H-R100dA, the following concentrations 190, 380, 750, 1,500, 3,000 nM; 250, 500, 1,000, 2000 and 4,000 nM; and 62.5, 125, 250, 500, and 1,000 nM, were used respectively. The association and dissociation time for wild 2F4, 10B5 and 8F6 mutant H-Y100 cA were 120 s and 600 s, respectively. For the rest of the Fabs including wild type and mutants for 7C6 and 8F6, a 120 s of association and 1,200 s of dissociation time were used in the experiment. The assays were carried out in HBS-T buffer (10 mM HEPES pH 7.5, 150 mM NaCl and 0.005% [v/v] Tween 20 surfactant). Biacore Insight Evaluation Software (Cytiva) was used to calculate the binding parameters.

4.8 Differential scanning calorimetry (DSC)

The thermal stability of the wild type antibodies was measured by DSC using MicroCal PEAQ-DSC (Malvern; Worcestershire, UK). The Fab samples (1 mg/mL) were prepared in PBS. At a scanning rate of 1°C/min the samples were heated from 20°C to 110°C. The data was fitted by non-two-state model using MicroCal PEAQ-DSC software (Malvern).

4.9 Circular dichroism (CD) measurements

The Fab's CD profile and thermal stability were measured using a JASCO J-820 spectropolarimeter. The CD spectra were obtained from 260 to 200 nm using a 1 mm quartz cuvette with a protein sample of 0.1 mg/mL in PBS. Each sample was measured five times with a 1 nm bandwidth. To analyze the protein denaturation profile, the thermal stability was measured at lower concentrations under the same buffer conditions and with three repetitions, at 1°C intervals from 30°C to 90°C and at a speed of 0.1°C/min, at 218 nm and 215 nm ellipticity for 7C6 and 8F6 wild type and mutants, respectively. The T_m was determined by fitting the ellipticity data against temperature using nonlinear least squares curve fitting that followed the below logistic function equation, followed by sigmoid curve fitting in Python 3.0 (Rossant, 2018) to obtain the fitted molar ellipticity and temperature values.

$$f_{L,m,k,x0}(x) = \frac{L}{1 + exp(-k(x - x0))} + m$$

Where, L, m, k and x0 are the vector parameters for optimization of the fitting. For better visualization of the T_m measurements, we represent the derivative of the fitted data.

Data availability statement

The MD trajectories and the docking model structures have been submitted to the Biological Structure Model Archive (BSM-Arc) under BSM-ID BSM000047 [https://bsma.pdbj.org/entry/47] (Bekker et al., 2020).

Author contributions

RP: Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing–original draft. KK: Data curation, Investigation, Validation, Writing–review and editing. JS: Resources, Writing–review and editing. JP: Data curation, Investigation, Validation, Writing–review and editing. RM: Data curation, Investigation, Validation, Writing–review and editing. TH: Conceptualization, Data curation, Funding acquisition, Resources, Supervision, Writing–review and editing. DK: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Visualization, Writing–original draft, Writing–review and editing. KT: Data curation, Funding acquisition, Project administration, Resources, Supervision, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was funded in part by the Japan Society for the Promotion of Science (JP 20K20596 to TH, JP19H04202 and JP21K18310 to DK, JP19H05766 and JP20H02531 to KT) and the Japan Agency for Medical Research and Development (JP22wm0325047 and JP23wm0325069 to TH and DK, JP22ama121033 to KT). We also thanks to Okawa Foundation for Information and Telecommunications (grant number 20-10 to DK), the Sumitomo foundation (to TH) and the Naito foundation (to TH).

Acknowledgments

The supercomputing resources in this study were provided in part by the Human Genome Center at the Institute of Medical Science, The University of Tokyo, Japan and by Research Center for Computational Science, Okazaki, Japan (Project: 23-IMS-C078).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2023.1302737/full#supplementary-material

References

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., et al. (2015). GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1 (2), 19–25. doi:10.1016/J.SOFTX.2015.06.001

Akbar, R., Bashour, H., Rawat, P., Robert, P. A., Smorodina, E., Cotet, T.-S., et al. (2022a). Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. *MAbs* 14, 2008790. doi:10.1080/19420862.2021.2008790

Akbar, R., Robert, P. A., Weber, C. R., Widrich, M., Frank, R., Pavlović, M., et al. (2022b). *In silico* proof of principle of machine learning-based antibody design at unconstrained scale. *MAbs* 14, 2031482. doi:10.1080/19420862.2022.2031482

Akiba, H., and Tsumoto, K. (2015). Thermodynamics of antibody–antigen interaction revealed by mutation analysis of antibody variable regions. *J. Biochem.* 158, 1–13. doi:10.1093/JB/MVV049

Al-Lazikani, B., Lesk, A. M., and Chothia, C. (1997). Standard conformations for the canonical structures of immunoglobulins. *J. Mol. Biol.* 273, 927–948. doi:10.1006/JMBI.1997.1354

Bekker, G. J., Kawabata, T., and Kurisu, G. (2020). The biological structure model archive (BSM-Arc): an archive for *in silico* models and simulations. *Biophys. Rev.* 12, 371–375. doi:10.1007/S12551-020-00632-5

Berendsen, H. J. C., van der Spoel, D., and van Drunen, R. (1995). GROMACS: a message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* 91, 43–56. doi:10.1016/0010-4655(95)00042-E

Berman, H., Henrick, K., Nakamura, H., and Markley, J. L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* 35, D301–D303. doi:10.1093/NAR/GKL971

Bogan, A. A., and Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* 280, 1–9. doi:10.1006/JMBI.1998.1843

Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., et al. (2009). CHARMM: the biomolecular simulation program. *J. Comput. Chem.* 30, 1545–1614. doi:10.1002/JCC.21287

Buß, O., Rudat, J., and Ochsenreither, K. (2018). FoldX as protein engineering tool: better than random based approaches? *Comput. Struct. Biotechnol. J.* 16, 25–33. doi:10.1016/J.CSBJ.2018.01.002

Cannon, D. A., Shan, L., Du, Q., Shirinian, L., Rickert, K. W., Rosenthal, K. L., et al. (2019). Experimentally guided computational antibody affinity maturation with *de novo* docking, modelling and rational design. *PLoS Comput. Biol.* 15, e1006980. doi:10.1371/JOURNAL.PCBI.1006980

Cathou, R. E., Kulczycki, A., and Haber, E. (1968). Structural features of γ-immunoglobulin, antibody, and their fragments. Circular dichroism studies. *Biochemistry* 7, 3958–3964. doi:10.1021/BI00851A024

Chaudhury, S., Berrondo, M., Weitzner, B. D., Muthu, P., Bergman, H., and Gray, J. J. (2011). Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PLoS One* 6, e22477. doi:10.1371/JOURNAL.PONE.0022477

Chennamsetty, N., Voynov, V., Kayser, V., Helk, B., and Trout, B. L. (2009). Design of therapeutic proteins with enhanced stability. *Proc. Natl. Acad. Sci. U. S. A.* 106, 11937–11942. doi:10.1073/pnas.0904191106

Chennamsetty, N., Voynov, V., Kayser, V., Helk, B., and Trout, B. L. (2010). Prediction of aggregation prone regions of therapeutic proteins. *J. Phys. Chem. B* 114, 6614–6624. doi:10.1021/jp911706q

Chothia, C., and Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* 196, 901–917. doi:10.1016/0022-2836(87)90412-8

D'Angelo, S., Ferrara, F., Naranjo, L., Erasmus, M. F., Hraber, P., and Bradbury, A. R. M. (2018). Many routes to an antibody heavy-chain CDR3: necessary, yet insufficient, for specific binding. *Front. Immunol.* 9, 395. doi:10.3389/fimmu.2018.00395

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98, 10089–10092. doi:10.1063/1.464397

Di Rienzo, L., Miotto, M., Bò, L., Ruocco, G., Raimondo, D., and Milanetti, E. (2021). Characterizing hydropathy of amino acid side chain in a protein environment by investigating the structural changes of water molecules network. *Front. Mol. Biosci.* 8, 2. doi:10.3389/fmolb.2021.626837

Ertl, O. T., Wenz, D. C., Bouche, F. B., Berbers, G. A. M., and Muller, C. P. (2003). Immunodominant domains of the Measles virus hemagglutinin protein eliciting a neutralizing human B cell response. *Archives Virology* 11 (148), 2195–2206. doi:10.1007/S00705-003-0159-9

Fang, X. T., Sehlin, D., Lannfelt, L., Syvänen, S., and Hultqvist, G. (2017). Efficient and inexpensive transient expression of multispecific multivalent antibodies in Expi293 cells. *Biol. Proced. Online* 19, 11. doi:10.1186/S12575-017-0060-7

Fischman, S., and Ofran, Y. (2018). Computational design of antibodies. *Curr. Opin. Struct. Biol.* 51, 156–162. doi:10.1016/j.sbi.2018.04.007

Fiser, A., Kinh, R., Do, G., and Ali, A. S. (2000). Modeling of loops in protein structures. *Protein Sci.* 9, 1753–1773. doi:10.1110/PS.9.9.1753

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., et al. (2005). Protein identification and analysis tools on the ExPASy server. *Proteomics Protoc. Handb.* 112, 571–607. doi:10.1385/1-59259-890-0:571

Hashiguchi, T., Kajikawa, M., Maita, N., Takeda, M., Kuroki, K., Sasaki, K., et al. (2007). Crystal structure of measles virus hemagglutinin provides insight into effective vaccines. *Proc. Natl. Acad. Sci. U. S. A.* 104, 19535–19540. doi:10.1073/PNAS.0707830104

Hashiguchi, T., Ose, T., Kubota, M., Maita, N., Kamishikiryo, J., Maenaka, K., et al. (2011). Structure of the measles virus hemagglutinin bound to its cellular receptor SLAM. *Nat. Struct. Mol. Biol.* 18, 135–141. doi:10.1038/NSMB.1969

Heo, L., Arbour, C. F., Janson, G., and Feig, M. (2021). Improved sampling strategies for protein model refinement based on molecular dynamics simulation. *J. Chem. Theory Comput.* 17, 1931–1943. doi:10.1021/ACS.JCTC.0C01238

Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997). LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* 18, 1463–1472. doi:10.1002/(sici)1096-987x(199709)18:12<1463::aid-jcc4>3.0.co;2-h

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., De Groot, B. L., et al. (2017). CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* 14, 71–73. doi:10.1038/NMETH.4067

Ionescu, R. M., Vlasak, J., Price, C., and Kirchmeier, M. (2008). Contribution of variable domains to the stability of humanized IgG1 monoclonal antibodies. *J. Pharm. Sci.* 97, 1414–1426. doi:10.1002/JPS.21104

Jain, N. K., Barkowski-Clark, S., Altman, R., Johnson, K., Sun, F., Zmuda, J., et al. (2017). A high density CHO-S transient transfection system: comparison of ExpiCHO and Expi293. *Protein Expr. Purif.* 134, 38–46. doi:10.1016/J.PEP.2017.03.018

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2

Kellogg, E. H., Leaver-Fay, A., and Baker, D. (2011). Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins Struct. Funct. Bioinforma.* 79, 830–838. doi:10.1002/PROT.22921

Kuroda, D., and Gray, J. J. (2016). Pushing the backbone in protein-protein docking. *Structure* 24, 1821–1829. doi:10.1016/J.STR.2016.06.025

Kuroda, D., Shirai, H., Jacobson, M. P., and Nakamura, H. (2012). Computer-aided antibody design. *Protein Eng. Des. Sel.* 25, 507–521. doi:10.1093/protein/gzs024

Kuroda, D., Shirai, H., Kobori, M., and Nakamura, H. (2008). Structural classification of CDR-H3 revisited: a lesson in antibody modeling. *Proteins Struct. Funct. Bioinforma.* 73, 608–620. doi:10.1002/prot.22087

Kuroda, D., Shirai, H., Kobori, M., and Nakamura, H. (2009). Systematic classification of CDR-L3 in antibodies: implications of the light chain subtypes and the V $_L$ -V $_H$ interface. *Proteins Struct. Funct. Bioinforma.* 75, 139–146. doi:10.1002/prot.22230

Kuroda, D., and Tsumoto, K. (2018). Antibody affinity maturation by computational design. *Methods Mol. Biol.* 1827, 15–34. doi:10.1007/978-1-4939-8648-4_2

Kuroda, D., and Tsumoto, K. (2020). Engineering stability, viscosity, and immunogenicity of antibodies by computational design. *J. Pharm. Sci.* 109, 1631–1651. doi:10.1016/j.xphs.2020.01.011

Leman, J. K., Weitzner, B. D., Lewis, S. M., Adolf-Bryfogle, J., Alam, N., Alford, R. F., et al. (2020). Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat. Methods* 17, 665–680. doi:10.1038/s41592-020-0848-2

Liang, T., Chen, H., Yuan, J., Jiang, C., Hao, Y., Wang, Y., et al. (2021). IsAb: a computational protocol for antibody design. *Brief. Bioinform* 22, bbab143. doi:10.1093/bib/bbab143

Lindahl, E., Hess, B., and van der Spoel, D. (2001). GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model* 7, 306–317. doi:10.1007/s008940100045

Liu, Q., Hoi, S. C. H., Su, C. T. T., Li, Z., Kwoh, C. K., Wong, L., et al. (2011). Structural analysis of the hot spots in the binding between H1N1 HA and the 2D1 antibody: do mutations of H1N1 from 1918 to 2009 affect much on this binding? *Bioinformatics* 27, 2529–2536. doi:10.1093/BIOINFORMATICS/BTR437

Madura, J. D., Jorgensen, W. L., Chandrasekhar, J., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *aip.scitation.Org.* 79, 926–935. doi:10.1063/1.445869

Marks, C., Hummer, A. M., Chin, M., and Deane, C. M. (2021). Humanization of antibodies using a machine learning approach on large-scale repertoire data. *Bioinformatics* 37, 4041–4047. doi:10.1093/BIOINFORMATICS/BTAB434

Olsen, T. H., Boyles, F., and Deane, C. M. (2022). Observed Antibody Space: a diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Sci.* 31, 141–146. doi:10.1002/PRO.4205

Park, H., Bradley, P., Greisen, P., Liu, Y., Mulligan, V. K., Kim, D. E., et al. (2016). Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J. Chem. Theory Comput.* 12, 6201–6212. doi:10.1021/ACS.JCTC.6B00819

Peng, H. P., Lee, K. H., Jian, J. W., and Yang, A. S. (2014). Origins of specificity and affinity in antibody-protein interactions. *Proc. Natl. Acad. Sci. U. S. A.* 111, E2656–E2665. doi:10.1073/pnas.1401131111

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi:10.1002/JCC.20084

Pommié, C., Levadoux, S., Sabatier, R., Lefranc, G., and Lefranc, M. P. (2004). IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J. Mol. Recognit.* 17, 17–32. doi:10.1002/JMR.647

Prihoda, D., Maamary, J., Waight, A., Juan, V., Fayadat-Dilman, L., Svozil, D., et al. (2022). BioPhi: a platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *MAbs* 14, 2020203. doi:10.1080/19420862.2021.2020203

Rabia, L. A., Desai, A. A., Jhajj, H. S., and Tessier, P. M. (2018). Understanding and overcoming trade-offs between antibody affinity, specificity, stability and solubility. *Biochem. Eng. J.* 137, 365–374. doi:10.1016/J.BEJ.2018.06.003

Ripoll, D. R., Chaudhury, S., and Wallqvist, A. (2021). Using the antibody-antigen binding interface to train image-based deep neural networks for antibody-epitope classification. *PLoS Comput. Biol.* 17, e1008864. doi:10.1371/JOURNAL.PCBI.1008864

Rossant, C. (2018). IPython Interactive Computing and Visualization Cookbook: over 100 hands-on recipes to sharpen your skills in high-performance numerical computing and. Available at: https://books.google.com/books?hl=en&lr=&id=GyBKDwAAQBAJ&oi=fnd&pg=PP1&dq=IPython+Cookbook+-+IPython+Cookbook,+Second+Edition+(2018)&ots=3YGtX01nzu&sig=TCChjkd0FRZtews3zimVdvPTmjI (Accessed July 6, 2023).

Ruffolo, J. A., Chu, L. S., Mahajan, S. P., and Gray, J. J. (2023). Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nat. Commun.* 14, 2389–2413. doi:10.1038/s41467-023-38063-x

Ruffolo, J. A., Sulam, J., and Gray, J. J. (2022). Antibody structure prediction using interpretable deep learning. *Patterns* 3, 100406. doi:10.1016/J.PATTER.2021.100406

Samanta, U., Pal, D., and Chakrabarti, P. (2000). Environment of tryptophan side chains in proteins. *Proteins Struct. Funct. Bioinforma.* 38, 288–300. doi:10.1002/(SICI)1097-0134(20000215)38:3<288::AID-PROT5>3.0.CO;2-7

Santiago, C., Celma, M. L., Stehle, T., and Casasnovas, J. M. (2010). Structure of the measles virus hemagglutinin bound to the CD46 receptor. *Nat. Struct. Mol. Biol.* 17, 124–129. doi:10.1038/NSMB.1726

Sato, Y., Watanabe, S., Fukuda, Y., Hashiguchi, T., Yanagi, Y., and Ohno, S. (2018). Cell-to-Cell measles virus spread between human neurons is dependent on hemagglutinin and hyperfusogenic fusion protein. *J. Virol.* 92, e02166-17. doi:10.1128/JVI.02166-17

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Res.* 33, W382–W388. doi:10.1093/NAR/GKI387

Sircar, A., and Gray, J. J. (2010). SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Comput. Biol.* 6, e1000644. doi:10.1371/JOURNAL.PCBI.1000644

Soga, S., Kuroda, D., Shirai, H., Kobori, M., and Hirayama, N. (2010). Use of amino acid composition to predict epitope residues of individual antibodies. *Protein Eng. Des. Sel.* 23, 441–448. doi:10.1093/PROTEIN/GZQ014

Suvvari, T. K., Kandi, V., Mohapatra, R. K., Chopra, H., Islam, M. A., and Dhama, K. (2023). The re-emergence of measles is posing an imminent global threat owing to decline in its vaccination rates amid COVID-19 pandemic: a special focus on recent outbreak in India - a call for massive vaccination drive to be enhanced at global level. *Int. J. Surg.* 109, 198–200. doi:10.1097/JS9.0000000000000228

Tadokoro, T., Jahan, M. L., Ito, Y., Tahara, M., Chen, S., Imai, A., et al. (2020). Biophysical characterization and single-chain Fv construction of a neutralizing antibody to measles virus. *FEBS J.* 287, 145–159. doi:10.1111/febs.14991

Tahara, M., Bürckert, J. P., Kanou, K., Maenaka, K., Muller, C. P., and Takeda, M. (2016). Erratum: Tahara, M., et al. Measles Virus Hemagglutinin Protein Epitopes: the Basis of Antigenic Stability. Viruses 2016, 8, 216. *Viruses* 8, 313. doi:10.3390/v8110313

Tahara, M., Ohno, S., Sakai, K., Ito, Y., Fukuhara, H., Komase, K., et al. (2013). The receptor-binding site of the measles virus hemagglutinin protein itself constitutes a conserved neutralizing epitope. *J. Virol.* 87, 3583–3586. doi:10.1128/JVI.03029-12

van Durme, J., Delgado, J., Stricher, F., Serrano, L., Schymkowitz, J., and Rousseau, F. (2011). A graphical interface for the FoldX forcefield. *Bioinformatics* 27, 1711–1712. doi:10.1093/BIOINFORMATICS/BTR254

Webb, B., and Sali, A. (2016). Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinforma.* 5, 5.6.1–5.6.37. doi:10.1002/CPBI.3

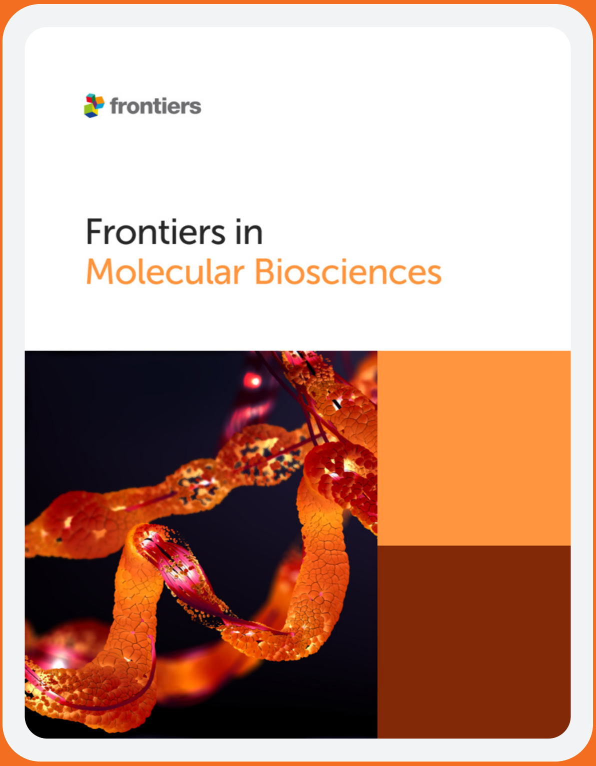
Weitzner, B. D., Dunbrack, R. L., and Gray, J. J. (2015). The origin of CDR H3 structural diversity. *Structure* 23, 302–311. doi:10.1016/j.str.2014.11.010

Weitzner, B. D., Jeliazkov, J. R., Lyskov, S., Marze, N., Kuroda, D., Frick, R., et al. (2017). Modeling and docking of antibody structures with Rosetta. *Nat. Protoc.* 2 (12), 401–416. doi:10.1038/nprot.2016.180

Wilman, W., Wróbel, S., Bielska, W., Deszynski, P., Dudzic, P., Jaszczyszyn, I., et al. (2022). Machine-designed biotherapeutics: opportunities, feasibility and advantages of deep learning in computational antibody discovery. *Brief. Bioinform* 23, bbac267. doi:10.1093/bib/bbac267

Zabetakis, D., Anderson, G. P., Bayya, N., and Goldman, E. R. (2013). Contributions of the complementarity determining regions to the thermal stability of a single-domain antibody. *PLoS One* 8, e77678. doi:10.1371/JOURNAL.PONE.0077678

Zhang, X., Lu, G., Qi, J., Li, Y., He, Y., Xu, X., et al. (2013). Structure of measles virus hemagglutinin bound to its epithelial receptor nectin-4. *Nat. Struct. Mol. Biol.* 20, 67–72. doi:10.1038/NSMB.2432



frontiers

Frontiers in
Molecular Biosciences

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us
+41 (0)21 510 17 00
frontiersin.org/about/contact