COMPUTATIONAL DEVELOPABILITY ASSESSMENT OF ANTIBODY THERAPEUTICS

A thesis submitted to the University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Science and Engineering

2023

Rahul Khetan

Department of Chemical Engineering

LIST OF CONTENTS

Word Count: 86,556

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACC STAB - Accelerated Stability Assay

AC SINS - Affinity-Capture Self-Interaction Nanoparticle Spectroscopy

ADC - Antibody-drug conjugate

ADCC - Antibody dependent cellular cytotoxicity

ADR - Adverse drug reaction

APR - Aggregation-prone region

BCE - B-cell epitope

BCR - B-cell receptor

BLA - Biologics License Application

BVP - Baculovirus Particle assay

CDA - Computational developability assessment

CDR - Complementarity-determining region

CIC - Cross-Interaction Chromatography

CSI BLI - Clone Self-Interaction by Biolayer Interferometry

DSF - Differential Scanning Fluorimetry

ELISA - Enzyme-Linked Immunosorbent Assay

HEK - Expression Titer in HEK cells

HIC - Hydrophobic interaction chromatography

IMGT - International ImMunoGeneTics Information System

IND - Investigational New Drug

MHC - Major histocompatibility complex

M&A - Mergers and Acquisitions

MTD - Maximum tolerated dose

NGS - Next-generation sequencing

PD - Pharmacodynamic

PDB - Protein data bank

PFA - Position frequency analysis

PK – Pharmacokinetic

PNC - Patches of Negative Charge Metric

PPC - Patches of Positive Charge Metric

PPV - Positive predictive value

PSH - Patches of Surface Hydrophobicity Metric

PSR - Poly-Specificity Reagent

PTM - Post-translational modification

ROC - Receiver Operator Characteristic

SAP - Spatial aggregation propensity

ScFv - Single-chain variable fragment

SEC - Size-Exclusion Chromatography

SFvCSP - Structural Fv Charge Symmetry Parameter

SGAC - Salt-Gradient Affinity-Capture Spectroscopy

SMAC - Standup Monolayer Absorption Chromatography

SVM - Support Vector Machine

TAP - Therapeutic antibody profiler

TCR - T-cell receptor

TheraSabDab - Therapeutic Structural Antibody Database

Computational Developability Assessment of Antibody Therapeutics

Rahul Khetan, The University of Manchester, 2023

Submitted for the degree of Doctor of Philosophy

ABSTRACT

Therapeutic monoclonal antibodies and their associated biologic derivatives are a key component of the commercial clinical pipelines of the global pharmaceutical industry. The availability of large datasets of antibody biophysical properties enables the search for predictive models and computational tools for the "developability assessment" of drug candidates. This thesis work has evaluated the scope of using biopharmaceutical informatics approaches for the prediction of developability issues such as stability, aggregation, and immunogenicity. We firstly establish developability guidelines based on *in-silico* metrics used for the assessment of antibody properties and derived from clinical-stage antibodies. These new computational developability guidelines serve as benchmarks for acceptable biophysical properties desired in antibody therapeutics. We have also highlighted the developability potential of natural human immune repertoire.

Our developability criteria were then utilized to compare the developability profile of major antibody discovery platforms to guide the selection of platform technologies for next-generation biotherapeutics. Next, we have used machine learning algorithms to estimate clinical trial progression of antibody therapeutics. Finally, we have validated our developability criteria performance in flagging antibodies that have caused serious adverse events or failure in clinical trials with an overall model accuracy of 80.6%. Finally, a summary and conclusion of the work are provided with a future outlook towards biopharmaceutical informatics for antibody drug discovery and optimization.

DECLARATION

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Signature: *Rahul Khetan*

Name of author: Rahul Khetan

Date: 06-11-2023

COPYRIGHT STATEMENT

PREFACE TO THE ALTERNATIVE FORMAT THESIS

This thesis has been presented in the University of Manchester's alternative format for a PhD thesis. This style has been chosen as it most clearly demonstrates the different aspects of the research conducted. The thesis structure consists of an overarching abstract and introduction, two review chapters, three result chapters and a conclusion with suggestions for further work. To aid the reader, the references have been collected at the end of the chapters using EndNote referencing tool. All figures and tables have been renumbered to ensure consistency throughout the body of work. The alternative format allows the writing of the thesis to be conducted in parallel with producing papers for publication.

Two of the chapters presented in this thesis have been published in peer-reviewed journals. Chapter 2 was written, peer-reviewed, and published as Khetan Rahul, Robin Curtis, Charlotte M. Deane, Johannes Thorling Hadsund, Uddipan Kar, Konrad Krawczyk, Daisuke Kuroda et al. "Current advances in biopharmaceutical informatics: guidelines, impact, and challenges in the computational developability assessment of antibody therapeutics." in Mabs, vol. 14, no. 1, p. 2020082. Taylor & Francis, 2022. Another article provided in Appendix was written, peer-reviewed, and published as Khetan Rahul "Biopharma licensing and M&A trends in the 21st-century landscape." in Journal of Commercial Biotechnology, vol. 25, no.3, 2020.

It is anticipated that each of the chapters presented as part of this work provide individual contributions to the existing literature and that collectively they describe the journey taken from a concept to implementation.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my PhD supervisor Dr Robin Curtis and co-supervisor Dr Jim Warwicker who have been a constant source of inspiration and guidance. Their knowledge, expertise, skill, and patience has impressed me at every opportunity. I am very grateful for their support and help in helping me realize my career objectives. A very special thanks to Max Hebditch for his advice and support throughout my PhD work.

My time at the Manchester Institute for Biotechnology would have been incomplete and a lot less fun without the support of everyone in my group: Matja, John, Farah, Bosco, Sophia, Sonia, Jack, Aisling, Nikita, and Nicole. It would have been a lot harder without you all.

I am grateful to North-West Biotech Initiative, Chemical Engineering Society, Masood Enterprise Centre, and Alliance Manchester Business School for providing me the opportunity to be a part of your amazing teams. I also thank the International Society and Reslife flatmates for making my life so much more enjoyable in Manchester. Thanks to Talgar, Aayush, Em, Lily, Nima, Joan, Ross, Scott, Emily and Charlotte for your support and motivation. Finally, I wish to thank amazing people at University of Manchester who have shaped my dream career – Laura, Carla, Max, Julie, Dr. John, Elizabeth, Dan, Dr. Barbara, Dr. Reza. and Igor.

I would very much like to thank my family. Special gratitude goes to my parents, my cousins - Minku and Tumku, and my friends for their continuous love and support in all my decisions in life. I would also like to thank all my family members for their prayers and support throughout my program. Lastly a big thank you to my IIT friends for their optimism and support, especially towards the finishing of my thesis.

CHAPTER 1

1 Introduction

1.1 Background to the research

Antibodies, also known as Immunoglobulins, are the proteins found in extracellular fluids and on the surface of immune cells that are at the core of the immune response towards any foreign antigen. The biological activity of antibodies depends on binding to the specific receptors on immune cells such as those that express receptors for the Fc portion of antibodies (FcR). These FcRs play various roles such as modulation of the immune response by released cytokines or phagocytosis.[1-3] So, antibodies are major molecular effectors of adaptive immune responses.

The basic unit of each antibody is an immunoglobulin G (IgG) monomer with a molecular weight of approximately 150 kDa, that is comprised of four polypeptide chains: two identical light chains (L) and two identical heavy chains (H). The antibodies are characterised by their specific 'Y' shaped structure – two antigen-binding fragments (Fab) and one constant region (Fc). Also, the structure of antibodies includes a specified variable region (Fv) at the end of the light and heavy chains. Also, single-chain variable fragments (scFv) are noncovalent heterodimers comprised of the variable regions of the heavy (V_H) and light chains (V_L). Figure 1 shows the antibody structure and different molecular building blocks for antibody therapeutic formats.

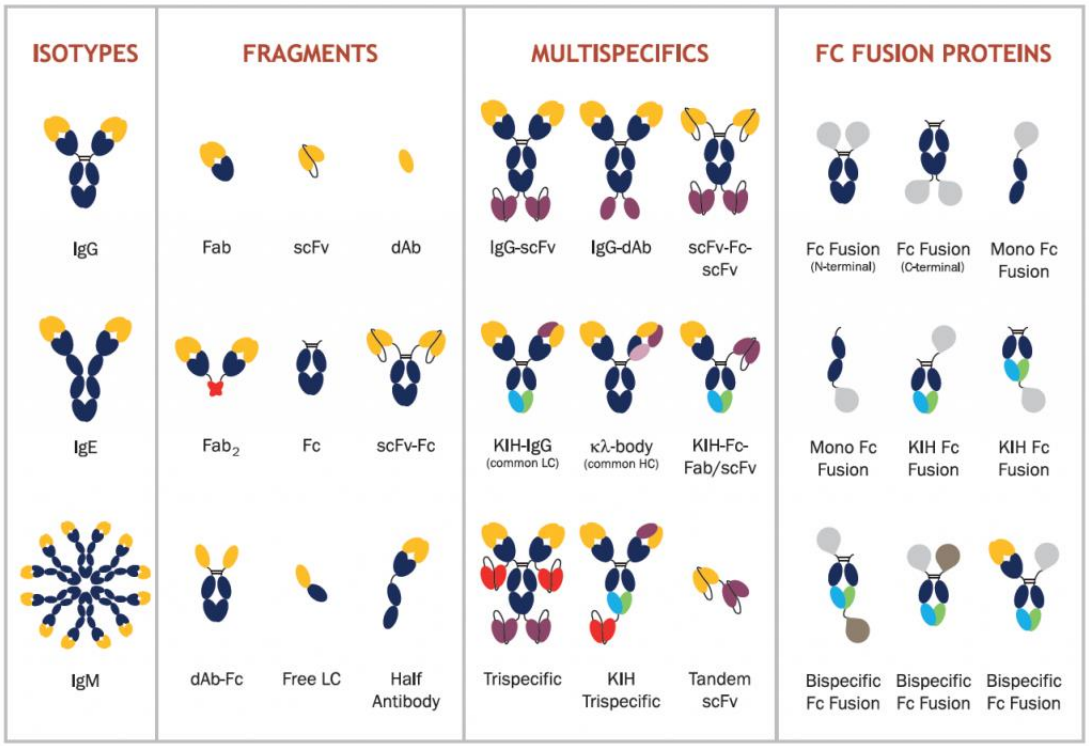


Figure 1: Overview of antibody structures. Wide range of antibody fragments created using IgGs, Fabs, and scFvs as building blocks. Image adapted from https://absoluteantibody.com/.

Fab as the name suggests contains the antigen-binding site (paratope) at the tips of the arms that determine the specificity of antibodies while the constant region (Fc domain) at the base of the antibody plays a role in modulating the effector function of an antibody, that may require prior binding of an antigen. Manipulations in the Fc regions can influence the pharmacokinetic properties of mAbs as well as improve the antibody dependent cellular cytotoxicity (ADCC).[4] IgG antibodies are known to mediate their effector functions through Fc gamma receptors (FcγR) on myeloid and Natural Killer (NK) cells. ADCC is an Fc-dependent effector function of IgG important for anti-viral immunity and anti-tumor therapies. In antibody dependent cellular cytotoxicity, FcγRs on the surface of effector cells (natural killer cells, macrophages, monocytes, and eosinophils) bind to the Fc region of an IgG which itself is bound to a target cell. An immune signalling pathway is triggered upon binding which results in the secretion of cytokines, lytic enzymes, perforin, granzymes and tumour necrosis factor (TNF), which mediate destruction of the target cell.[5] The level of ADCC effector function differs for various IgG subtypes in humans with high effector function for IgG1 and IgG3, and low for IgG2 and IgG4. Also, Complement-dependent cytotoxicity (CDC) is another major effector function of IgG and IgM antibodies. CDC is induced when the target-bound antibody is recognized by C1q protein, causing a cascade of events that result in the release of soluble C3a and C5a and the formation of the membrane attack complex (MAC) that lyses the target cell to achieve an antitumor effect.

The antigen-binding activity of mAbs is determined by the conformation of its amino acids in its complementary determining regions (CDRs) which are hypervariable loops of diverse lengths. Three CDRs are located in the variable region of both the light and the heavy chains of the antibody. The human immune system has the ability to create millions of different antibodies with high affinity to the target molecules because of the different unique combinations of CDRs. One of the greatest challenges in biomedical research on antibodies is to mimic the screening process of the human immune system as closely as possible in order to identify antibodies with the highest target/antigen specificity.

Therapeutic monoclonal antibodies (mAbs) have emerged as a reliable treatment option for serious clinical indications since the first monoclonal antibody - Orthoclone

OKT3, was approved in 1986.[6] Since the approval of OKT3, antibody therapeutics have evolved to include humanized and fully human monoclonal antibodies along with new formats of engineered antibodies such as bispecific antibody fragments, Fc-fusion proteins, and antibody-drug conjugates. Currently, over 166 monoclonal antibodies have been approved for the treatment of a variety of diseases that are summarized in Figure 2. The full details of the approved 166 mAbs are provided in Supplementary Information. Monoclonal antibody therapeutics have established themselves as a dominant and reliable biologic product class within the biopharmaceutical market. The global antibody therapeutics market has been valued at 217.3 billion USD and is projected to have an exorbitant 15% annual growth rate in the next decade.[7]

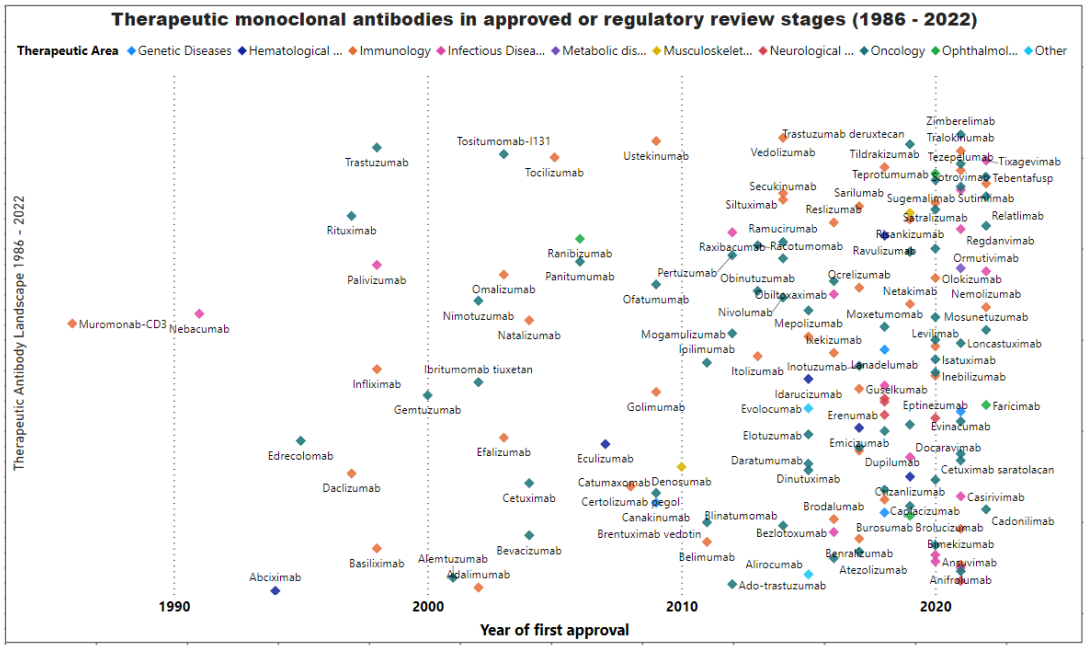


Figure 2: Therapeutic monoclonal antibodies in approved or review stages (1986 - 2022)

The therapeutic value of monoclonal antibodies (mAbs) originates from their intrinsic molecular properties that make them target-specific, biologically effective, stable, and manufacturable. Monoclonal antibodies offer exceptional antigen recognition and binding with a longer half-life. In addition, mAbs facilitate crucial effector functions such as antibody-dependent cellular cytotoxicity (ADCC) and antibody-dependent cell-mediated phagocytosis (ADCP). Some of the major factors contributing to the growth of the monoclonal antibodies market are high specificity towards molecular targets, excellent safety profiles, optimal pharmacokinetic properties, an easy route towards clinical proof-of-concept (PoC), and rapid commercialization for mAbs.

R&D advances continue to explore new therapeutic applications and expand the use of mAbs in various medical fields. Monoclonal antibodies are widely used in cancer therapy to specifically target cancer cells to inhibit their growth and promote immune activation for anti-tumour activity. Examples include trastuzumab (Herceptin) for HER2+ Breast Cancer and rituximab (Rituxan) for Non-Hodgkin's Lymphomas.[8] Several mAbs have been developed to directly neutralize pathogens such as viruses, bacteria, or toxins to combat various infectious diseases like COVID-19, Ebola, and HIV.[9] In transplantation medicine, mAbs can be used to prevent organ rejection by targeting and blocking specific immune cells responsible for rejection, allowing better acceptance of the transplanted organ. Examples include basiliximab (Simulect) and alemtuzumab (Campath) in organ transplantation.[10] Some mAbs have shown excellent therapeutic potential in major ophthalmic conditions. Ranibizumab (Lucentis) and bevacizumab (Avastin) are examples of mAbs used in ophthalmology. Finally, mAbs have been developed to treat allergic conditions, such as asthma and allergic rhinitis. These mAbs target and block specific immune molecules or checkpoint modulators involved in allergic responses and pathways, providing relief from symptoms.[11] These are just a few examples of the therapeutic value of monoclonal antibodies. Novel mAb applications and biology are being discovered each year with advances in new formats, manufacturing processes, design, and understanding of disease pathophysiology.

1.2 Current advances in biopharmaceutical informatics: Guidelines, impact, and challenges in the computational developability assessment of antibody therapeutics

This chapter was written, peer-reviewed and published as: Khetan Rahul, Robin Curtis, Charlotte M. Deane, Johannes Thorling Hadsund, Uddipan Kar, Konrad Krawczyk, Daisuke Kuroda et al. ***"Current advances in biopharmaceutical informatics: guidelines, impact and challenges in the computational developability assessment of antibody therapeutics."*** In Mabs, vol. 14, no. 1, p. 2020082. Taylor & Francis, 2022. doi: https://doi.org/10.1080/19420862.2021.2020082

Keywords: - Developability guidelines; biopharmaceutical informatics; developability assessment; computational prediction; antibody engineering; therapeutic antibodies.

1.2.1 Abstract

Therapeutic monoclonal antibodies (mAbs) and their derivatives are key components of clinical pipelines in the global biopharmaceutical industry. The availability of large datasets of antibody sequences, structures, and biophysical properties are increasingly enabling the development of new predictive models and computational tools for the "developability assessment" of antibody drug candidates. In this review, we provide an overview of the available antibody informatics tools for the prediction of major developability issues such as stability, aggregation, immunogenicity, and chemical degradation. We further evaluate the key opportunities and challenges of using biopharmaceutical informatics for drug discovery and optimization. Finally, we discuss the potential of developability guidelines based on *in silico* metrics that can be used for the assessment of antibody stability and manufacturability.

1.2.2 Introduction

Monoclonal antibodies (mAbs) and antibody-based biotherapeutics represent a unique class of biologics that have greatly reshaped our modern biopharmaceutical industry since the first mAb drug, muromonab (Orthoclone®), was approved by the Food and Drug Administration in June 1986. The global mAb market is currently valued at 152.5 billion USD and is projected to exhibit an annual growth rate of 14.6% in the next decade.[12] Antibody therapeutics currently in late-stage clinical studies have more than tripled to 88 compared to 2010 and over 550 novel antibody therapeutics are currently in the early-stage commercial clinical pipeline.[12, 13] The antibody therapeutics are anticipated to be the key treatments in a broad range of disease areas, such as cancer, cardiovascular, inflammation, neurological, autoimmune, and infectious diseases.

Biopharmaceutical informatics is the application of computational methods and bioinformatics tools towards addressing challenges in biopharmaceutical drug development. It also includes development of databases containing biophysical data, molecular modelling and simulations, and statistical analysis of biopharmaceutical datasets. The term "Biopharmaceutical Informatics" was first introduced by Kumar *et al.*[14] as the umbrella term for applications of computational approaches in drug discovery and development. Here, we present different aspects of computational applications to antibody-based biopharmaceutical drug development by highlighting key scientific advances in the developability assessment of antibody-based biologic drug candidates.

One of the first practical applications of software relevant to antibody informatics was the antigenic index,[15] which was a program to generate surface contour profiles and predict antigenic sites from the linear amino acid sequence of proteins including antibodies. These techniques were the precursors of modern sequence-and structure-based bioinformatics tools used in biopharmaceutical discovery and development. The multitude of computational tools and algorithms now available have ushered in an era of high-throughput biopharmaceutical informatics.

This review is organized into four main sections. The first section outlines the databases and tools available for biopharmaceutical informatics relevant to antibody-based drugs. In the second section, we discuss the role of developability at early-stage development and computational developability assessment of antibody therapeutics. The third section describes the application of biopharmaceutical informatics to identify key developability issues in antibody-based drug discovery and design. The final section summarizes emerging trends in the use of biopharmaceutical informatics for antibody therapeutics. While we discussed antibody informatics tools and approaches for evaluating developability issues, a comprehensive review of every developability issue was not possible within this article. We have, however, cited previously published reviews that include more details for each developability issue in the respective sections below.

Creation of databases and data mining for comparison of biophysical attributes:

The availability of larger datasets with new high-throughput experimental methods has improved the predictions made by biopharmaceutical informatics tools. The challenge of data scarcity is now being resolved by open-source libraries and public databases of biopharmaceutical data. Data in biopharmaceutical informatics are highly heterogeneous and interrelated. Consequently, it is not possible to capture these broad ranges of properties in a single algorithm. Datasets currently used to assess the biophysical properties of antibodies are curated from internal releases by pharmaceutical companies or data points from scientific papers.[16-18] Experimental data sourced from scientific papers might not be comparable with one another because of differences in experimental setups, the plethora of developability assays, and different antibody formats tested. Additional data sources that potentially contain much antibody-engineering knowledge are patents, where one needs to scan the

documentation for primary sequence information.[19] Altogether, there is currently much yet-untapped data in the public domain, but these are often hard to curate and not immediately compatible and useful without much earlier pre-processing.

Further advantages from curating antibody databases to learn biophysical properties of antibodies can be obtained by linking information from heterogeneous sources. Current predictive approaches typically use either structural or sequence data that rarely link information from different sources (e.g., structural, and next-generation sequencing (NGS)). Collating information from different sources, however, can augment information available in heterogeneous sources. For instance, structural modelling can provide a conformational dimension to millions of sequences drawn from NGS,[20] whereas contrasting naturally sourced and therapeutically developed molecules can provide insights on commonalities and divergences between the two sources.[21] A good example of such an integrated approach is the INDI database,[22] which contains data for antibody-cognate nanobodies (single-domain antibodies VHH) collected from all major public sources, encompassing patents,[19] NCBI GenBank, Protein Data Bank (PDB), and NGS/AIRR[23] supplemented by manual curation from the scientific literature. The sequences and structures of antibodies from these heterogeneous sources are linked with textual information into an antibody-specific database. Integrating the heterogeneous sources in this manner facilitates searching and creation of custom datasets of nanobodies. Extrapolating such data integration approaches to antibodies should allow researchers to focus more on the machine learning/statistical approaches addressing the prediction of the biophysical properties of these molecules.

Norman *et al.*[24] have previously provided an overview of available databases and tools for computational antibody analysis. However, our specific focus here is on computational developability assessment tools and databases. Table 1 provides a list of relevant databases and datasets for antibody-based drugs that can be used for training, validation, and assessment of biopharmaceutical informatics tools.

Table 1. Relevant databases and datasets for biopharmaceutical informatics

S. No	Database Name	Application	Link
Sequence Databases			
1.	Observed Antibody Space (OAS)	Annotated immune repertoires of over a billion Ab sequences across diverse immune states and organisms.	http://opig.stats.ox.ac.uk/webapps/oas/

2.	International Immunogenetics Information System (IMGT)	IMGT® provides common access to sequence, genome, and structure Immunogenetics data.	http://www.imgt.org/
3.	Patented Antibody Database	The Patented Antibody Database contains sequence information found in patent documents for 267,722 antibody chains from 19,037 patent families.	https://www.naturalantibody.com/pad
4.	iReceptor	Antibody/B-cell and T-cell receptor repertoire data from multiple independent repositories.	https://gateway.ireceptor.org/login
5.	abYsis	Integrated antibody sequence and structure management, analysis, and prediction	http://www.abysis.org/
6.	EMBLIg	Antibody sequences automatically extracted from EMBL-ENA	http://www.abybank.org/emblig/
7.	Antibody Knowledge Graph	A framework for collecting antibody data from all major public sources.	https://www.naturalantibody.com/antibody-knowledge-graph/
8.	Integrated Nanobody Database for Immunoinformatics (INDI)	Database with structure data and sequence information of nanobodies created using an integrated curation approach from several sources.	http://research.naturalantibody.com/nanobodies [22]
Structure Databases			
9.	Protein Data Bank (PDB)	3D structure data for large biological molecules (proteins, DNA, and RNA).	https://www.rcsb.org/
10.	Structural Antibody Database (SAbDab)	An online resource containing all the publicly available antibody structures annotated with several properties.	http://opig.stats.ox.ac.uk/webapps/newsabdab/sabdab/
11.	Thera-SAbDab	Variable domain sequences and structural representations of all antibody therapeutics recognized by the WHO INN lists.	http://opig.stats.ox.ac.uk/webapps/newsabdab/therasabdab/search/ [25]
12.	SACS	Summary of antibody crystal structures in the PDB	http://www.abybank.org/sacs/
13.	AbDb	Information on redundancy and structures solved with and without antigens for Fv fragments extracted from PDB files.	http://www.abybank.org/abdb/
14.	PyIgClassify	A database of antibody CDR structural classifications	http://dunbrack2.fccc.edu/PyIgClassify/
15.	AAAAA	An automatic modelling and analysis tool for structural alignment of antibody and T cell receptor sequences.	https://plueckthun.bioc.uzh.ch/antibody/index.html
Immunogenicity			
16.	Immune Epitope Database (IEDB)	Experimental data on antibodies and T cell epitopes.	https://www.iedb.org/

17.	T Cell Epitope Database (TCED™)	Database of CD4+ T cell epitopes derived from T cell epitope mapping studies.	https://abzenaprod.wpengine.com/development-services/immunology/immunogenicity-assessment/itope-and-tced/
18.	MHCBN 4.0	A database of MHC/TAP binding peptides and T-cell epitopes.	http://crdd.osdd.net/raghava/mhcbn/
19.	Bcipep	Database of B-cell epitopes.	https://webs.iiitd.edu.in/raghava/bcipep/info.html
20.	Leadscope Toxicity Database	The Leadscope Toxicity Database contains over 180,000 chemical structures with over 400,000 toxicity study results.	https://www.leadscope.com/product_info.php?products_id=78
Antibody-antigen binding / Protein-Protein interactions			
21.	PCLICK	Antibody-Antigen Structures from a dataset of 403 antibody-antigen complexes using CLICK method.	http://mspc.bii.a-star.edu.sg/minhn/cluster_pclick.html
22.	AB-Bind: Antibody binding mutational database	Experimentally determined changes in binding free energies for 1101 mutants across 32 antibody-antigen structures.	https://github.com/sarahsirin/AB-Bind-Database [26]
23.	SKEMPI 2.0	Database of binding free energy changes upon mutation for structurally resolved protein-protein interactions.	https://life.bsc.es/pid/skempi2/
24.	AntigenDB	Database of antigens from several pathogenic species containing structural, sequence, and binding data	http://crdd.osdd.net/raghava/antigendb/
25.	AntiJen	Database containing quantitative binding data for peptides	http://www.ddg-pharmfac.net/antijen/AntiJen/antijenhomepage.htm
General Information, Regulatory			
26.	Tabs – Therapeutic Antibody Database (Commercial-use)	Data on 5,400+ antibodies, 1,350+ antigens, and 1,550+ companies, linked to clinical trials, patents, papers, news, and regulatory agencies.	https://tabs.craic.com/static_pages/4
27.	AbMiner	Database to match commercially available antibodies to their respective genomic identifiers.	https://discover.nci.nih.gov/abminer/

Table 1: Databases suggested for use in biopharmaceutical informatics that are relevant for antibody-based drugs. These databases have been selected by authors from several other available databases for general proteins to capture antibody-specific properties.

Relevance of biopharmaceutical informatics tools:

Biopharmaceutical informatics tools have the potential to be widely used for *in silico* screening of biophysical properties in an antibody library. These antibody informatics

approaches have been used to evaluate key biochemical and biophysical properties such as solubility, stability, viscosity, charge profiles, post-translational modifications (PTMs), immunogenicity, pharmacokinetic and pharmacodynamic (PK/PD) profiles, and hydrophobicity to rank the candidates. The prediction of protein tertiary structure is accomplished by either homology modelling approaches, fold recognition, or *ab initio* modeling approaches when similar sequences with known structures are absent. Several studies have implemented homology modeling to calculate the biochemical and biophysical properties of a mAb library.[27-29] Specific homology modeling algorithms for antibodies have been now developed for better accuracy and representation.[30-32] In general, antibody sequences and structures are well conserved except for the complementarity-determining regions (CDRs). The CDRs, except for CDR-H3, can be classified into a set of limited conformations called canonical structures[33-35] that can be predicted from sequence key residues, enabling sub-angstrom accuracy in structure prediction. However, predicting conformations of CDR-H3 is still challenging because it is the most diverse both in sequence and structure.[36] Sequence-structure correlations identified for CDR-H3 have been used as geometric constraints in simulations for structure prediction.[37, 38]

The antibody modeling tools provide an integrated computer-aided molecular design platform that can be used to access liabilities and optimize the affinity, solubility, and stability of antibody-based drug candidates. Several other biopharmaceutical informatics tools for various developability issues depend on protein sequence features that are based on amino acid physicochemical properties. There have been increasing efforts to compile these tools for integrated antibody sequence and structure management, analysis, and prediction. For instance, a large number of tools for antibody informatics are compiled under the abYsis database, abYmod antibody modeling program, and abYbank database. abYsis[39] incorporates a wide-ranging species-specific analysis of residue frequencies that can be combined with residue clustering to identify either hydrophobic or unusual patches that are likely to be important for the stability and immunogenicity of antibodies. The Scratch suite of predictors[40] also provides a set of comprehensive tools to evaluate the physicochemical properties of mAbs, such as solvent accessibility, secondary structure, tertiary structure, contact maps, protein antigenicity, and domain locations. The Oxford Protein Informatics Group (OPIG) also maintains several webservers and

databases relevant to antibody informatics. An up-to-date list of antibody-related resources is maintained at http://naturalantibody.com/tools. Table 2 provides a list of biopharmaceutical informatics tools for developability assessment of antibodies.

Table 2. Relevant Biopharmaceutical informatics tools

Software Name	Biophysical Property and Description	Link
Antibody modeling		
abYmod	Prediction of VH/VL packing and an extended loop database for modelling CDR-H3.	http://abymod.abysis.org
ABangle	A tool for calculating and analyzing the VH-VL orientation in antibodies.	http://opig.stats.ox.ac.uk/webapps/newsabdab/sabpred/abangle/
ABodyBuilder	Machine Learning based antibody Fv modelling using ABlooper.	http://opig.stats.ox.ac.uk/webapps/abodybuilder
PIGS	Modeling of immunoglobulin variable domains based on canonical structure method.	https://bio.tools/pigs
MODELLER	Comparative protein structure modeling by satisfaction of spatial restraints	https://salilab.org/modeller/
MOE	Integrated computer-aided molecular design platform for biologics	https://www.chemcomp.com/Products.htm
RosettaAntibody	Homology modeling program within the Rosetta suite for predicting high-resolution antibody FV structures.	https://new.rosettacommons.org/docs/latest/application_documentation/antibody/antibody-applications
LYRA	Lymphocyte Receptor Automated Modelling (LYRA) using homology modeling.	http://www.cbs.dtu.dk/services/LYRA/index.php
Repertoire Builder	Structural modeling of B cell / T cell receptors from their amino acid sequences	https://sysimm.org/rep_builder/
Solubility and Aggregation		
CamSol	CamSol method constitutes three algorithms to rationally design protein variants with enhanced solubility.	https://www-cohsoftware.ch.cam.ac.uk/index.php
Protein-Sol	A web tool for predicting protein solubility from 35 sequence-based features such as amino acid content, entropy, and disorder.	https://protein-sol.manchester.ac.uk/
SODA	Prediction of protein solubility from disorder and aggregation propensity.	http://old.protein.bio.unipd.it/soda/
SOLpro	Support vector machine (SVM) algorithm based protein solubility predictor	http://scratch.proteomics.ics.uci.edu/explanation.html#SOLpro
SOLart	A structure-based method to predict protein solubility and aggregation using solubility-dependent potentials.	http://babylone.ulb.ac.be/SOLART/
SAP	Aggregation Prediction Spatial aggregation propensity	[41]
Solubis	A webserver to reduce protein aggregation through mutation analysis.	http://solubis.switchlab.org/ [42]
GAP	Prediction of amyloid fibril-forming and amorphous ß-aggregating hexapeptides	https://www.iitm.ac.in/bioinfo/GAP/

AGGRESCAN 3D	Aggregation Prediction using structurally corrected aggregation value (A3D score)	http://bioinf.uab.es/aggrescan/ [43]
AggScore	Aggregation Prediction from distribution of hydrophobic and electrostatic patches	https://www.schrodinger.com/Aggscore
PASTA 2.0	Aggregation Prediction based on energy function of cross-beta pairings	http://old.protein.bio.unipd.it/pasta2/
TANGO	Aggregation Prediction from physico-chemical principles of secondary structure formation.	http://tango.crg.es/
Post-translational modifications/Stability		
MusiteDeep	A deep-learning based webserver for protein post-translational modification site prediction and visualization.	https://github.com/duolinwang/MusiteDeep_web
PTM prediction tools survey	Collection of publicly available PTM web resources, databases, and classification/prediction servers.	http://www.cbs.dtu.dk/databases/PTMpredictions/
MUpro	Prediction of protein stability changes for single-site mutations	http://mupro.proteomics.ics.uci.edu
FindMod	Tool to predict potential protein post-translational modifications	https://web.expasy.org/findmod/
SIDEpro	Prediction of protein side-chain conformations from rotamer probabilities for each residue	http://sidepro.proteomics.ics.uci.edu/
SCWRL4.0	Prediction of protein side-chain conformations using anisotropic hydrogen bonding function	http://dunbrack.fccc.edu/scwrl4/SCWRL4.php
PEARS	Prediction of protein side-chain conformations using the IMGT position-dependent distribution of rotamers.	http://opig.stats.ox.ac.uk/webapps/pears
Molecular docking		
DockThor	Web Server for Protein-ligand Docking	https://dockthor.lncc.br/v2/
SwissDock	Molecular docking based on the docking software EADock DSS.	http://www.swissdock.ch/
HADDOCK	High Ambiguity Driven protein docking with ambiguous interaction restraints	https://wenmr.science.uu.nl/haddock2.4/
MEGADOCK 4.0	FFT-grid-based protein-protein docking	https://www.bi.cs.titech.ac.jp/megadock/
RosettaDock	Monte Carlo (MC) based multi-scale docking algorithm that optimizes both rigid-body orientation and side-chain conformation.	https://new.rosettacommons.org/docs/latest/application_documentation/docking/docking-protocol
FTDock 2.0	Molecular docking based on the surface complementarity score between the two grids	http://www.sbg.bio.ic.ac.uk/docking/ftdock.html
AbAdapt	Antibody-specific epitope prediction	https://sysimm.org/abadapt/
Immunogenicity		
ANTIGENpro	Protein microarray data predictor to predict the likelihood that a protein is a protective antigen.	http://scratch.proteomics.ics.uci.edu/explanation.html#ANTIGENpro
COBEpro	Continuous B-cell epitope predictor using epitopic propensity scores on short peptide fragments.	http://scratch.proteomics.ics.uci.edu/explanation.html#COBEpro
BEpro (PEPITO)	Discontinuous B-cell epitope predictor.	http://pepito.proteomics.ics.uci.edu
DiscoTope	Prediction of discontinuous B cell epitopes from protein three-dimensional structures	http://www.cbs.dtu.dk/services/DiscoTope/

ElliPro	Antibody epitope prediction based on Protrusion Index averaged over residues	http://tools.iedb.org/ellipro/
SVMTriP	A tool to predict linear antigenic epitopes	http://sysbio.unl.edu/SVMTriP/
AbAdapt	Antibody-specific epitope prediction using antigen structural modeling with rigid docking	https://sysimm.org/abadapt/
EpiPred	Antibody-specific epitope prediction based on ranking of antigen structure patches	http://opig.stats.ox.ac.uk/webapps/newsabdab/sabpred/epipred/
RANKPEP	Immunogenicity risk assessment using Position Specific Scoring Matrices (PSSMs).	http://imed.med.ucm.es/Tools/rankpep.html
ProPred	Immunogenicity risk assessment using custom novel quantitative matrices.	http://crdd.osdd.net/raghava/propred/
NetMHCIIpan	Immunogenicity risk assessment using Artificial Neural Networks (ANNs).	http://www.cbs.dtu.dk/services/NetMHCIIpan/
MHCEpitopeEnergy	Rosetta-based biotherapeutic deimmunization platform with flexible scoring term.	https://new.rosettacommons.org/docs/latest/rosetta_basics/scoring/MHCEpitopeEnergy
Hu-mAb	Antibody humanization tool based on Random Forest models trained on sequence data	http://opig.stats.ox.ac.uk/webapps/newsabdab/sabpred/humab
TOPKAT	*in silico* toxicology assessments. TOxicity Prediction by Komputer Assisted Technology	https://www.toxit.it/en/services/software/topkat
MetaDrug	*in silico* toxicology assessments based on OMICs data analysis on pharmacogenomics and toxicogenomics datasets.	https://support.clarivate.com/LifeSciences/s/article/MetaDrug-Uses-and-benefits?language=en_US
Biophysical properties		
Abpred	Prediction of biophysical performance on 12 standard developability assays based on multiple machine-learning algorithms	https://protein-sol.manchester.ac.uk/abpred
QikProp	ADME prediction tool based on full 3D molecular structure.	https://www.schrodinger.com/products/qikprop
Delayed HIC retention time Prediction tool	Model for prediction of delayed HIC retention times directly from sequence.	[44]
General Developability		
Therapeutic Antibody Profiler (TAP)	Developability guidelines check and Identification of sequence liabilities.	http://opig.stats.ox.ac.uk/webapps/newsabdab/sabpred/tap
Developability Index	Developability Index is a function of an antibody's net charge and the spatial aggregation propensity, calculated on the complementarity-determining region structure.	[45]
abYsis	Integrated antibody sequence and structure management, analysis, and prediction.	http://www.abysis.org/
NaturalAntibody AbMapper	A data-driven suite of analytics to improve research decision support in screening and rational design of antibody therapeutics.	https://naturalantibody.com/antibody-analytics/

Table 2: Biopharmaceutical informatics tools for assessment of developability issues. Most of the tools listed are free for academic use or available on request. Some tools may have an upgraded commercial version for users. These tools have been selected by authors from several other available antibody informatics tools for general proteins.

1.2.3 Computational developability assessment of therapeutics using biopharmaceutical informatics

Novel criteria based on the biochemical and biophysical properties of mAbs are being increasingly used to select an antibody candidate from the early discovery to the development stage. Computational developability assessment approaches are now becoming a routine step in the drug discovery and development process. Developability assessments at the early stage of development can significantly de-risk development pipelines, thus saving valuable time and resources. Incorporating developability assessments in early-stage development provides an opportunity for us to re-engineer the molecule to mitigate any sequence or structural liabilities, or to select alternative molecules of similar potency, but with more favorable developability profiles. Previous studies have summarized various experimental platforms and computational tools to identify developability issues in therapeutic antibodies and antibody-like molecules.[46, 47] These key tools are summarized in Table 2.

In the past decade applications of techniques such as phage display, cell surface display, yeast display, hybridoma, and NGS have revolutionized biomedical research with the successful discovery of several therapeutic antibodies. Although most antibody libraries focus on maximizing library diversity, there are growing concerns regarding the developability of selected antibodies for successful commercialization.[17] Therefore, frameworks and procedures are being developed for the design of antibody libraries with improved developability and manufacturability.[48] *In silico* engineering and design of biologics using rational design principles has emerged as a faster and more economical alternative to traditional methods of lead generation such as hybridoma and phage display. Figure 3 provides a visual representation of the recommended biopharmaceutical informatics tools for computational developability assessment of antibody therapeutics and antibody-based drugs.

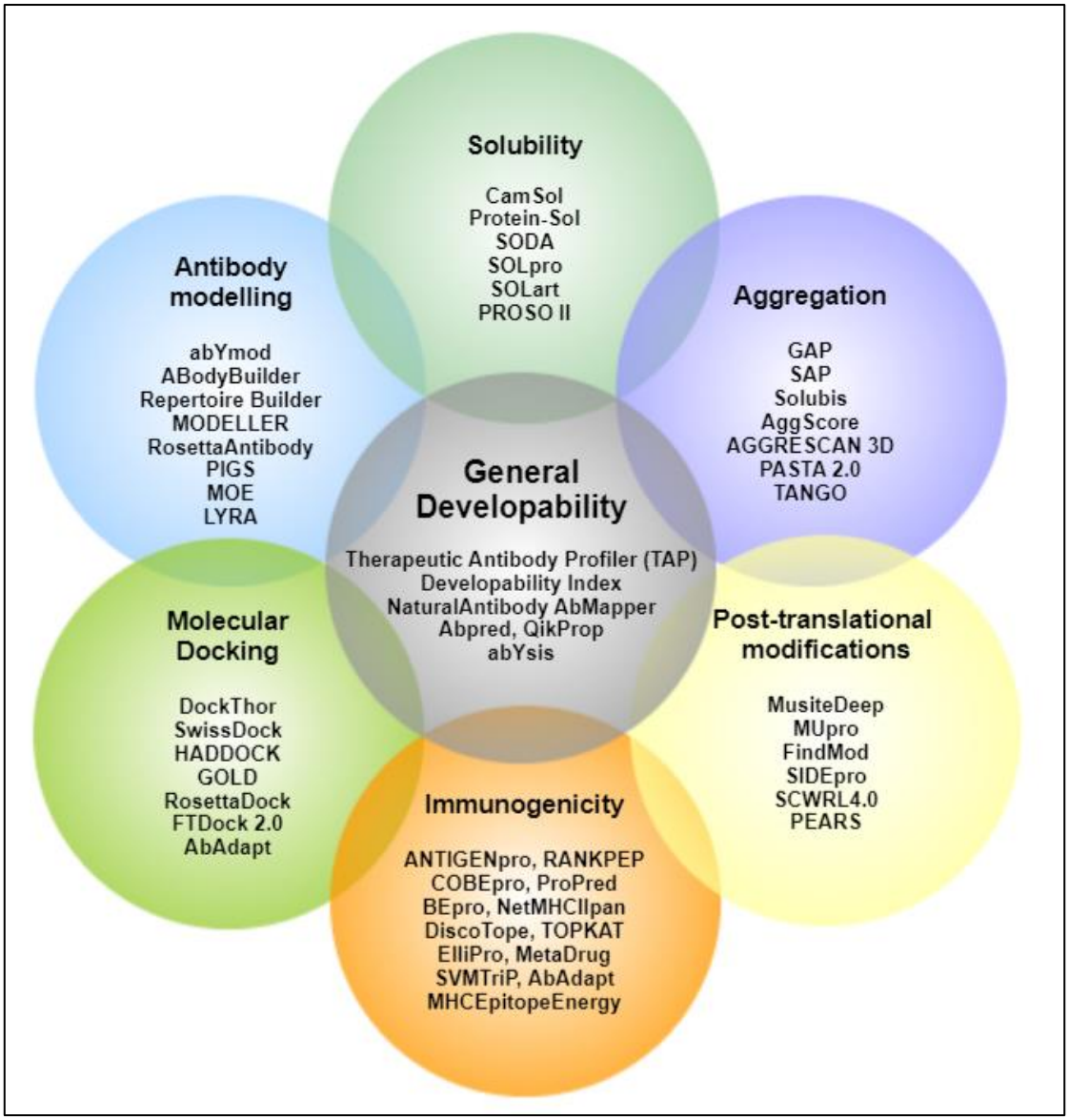


*Figure 3**:** Biopharmaceutical Informatics tools for computational developability assessment of antibody therapeutics. These tools have been selected by authors from several other available antibody informatics tools for general proteins.*

De-risking biopharmaceutical development using developability assessments:

Developability assessment is used to systematically evaluate mAb candidates that have the lowest risks for development to the final product. Previous studies have demonstrated the utility of the developability assessment of mAb lead candidates for screening out mAbs with low solubility and stability, low potency, high aggregation propensity, and high immunogenicity risk.[49] Any such predictions will inevitably

reject some antibodies that could have made excellent drugs, but not using such approaches comes with huge financial risk.

The general biophysical properties of approved mAbs can serve as a reference for the design of new mAb candidates. Several databases of biophysical properties of these approved mAb candidates have been reported such as the Jain dataset[17] and TheraSabDab.[25] The Jain dataset provides biophysical characterization across 12 different biophysical platforms for 137 clinical-stage and approved antibodies.[17] This benchmarking with approved mAbs provides an estimate of the acceptable ranges of the biophysical properties that can be considered in the developability assessments for new antibody candidates. Xu *et al.* have outlined some generally preferred quality attributes of a panel of approved and clinical-stage mAb products.[50] The general concept of examining the properties of successful antibody-based drugs has been exploited by Raybould *et al.*[16] resulting in Therapeutic Antibody Profiler (TAP) developability guidelines that are derived from the values of 377 clinical-stage antibody therapeutics. It relies on the hypothesis that antibodies that have deviating biophysical properties from clinically tested therapeutic mAbs are likely to have poor developability profiles. TAP can be used to analyze several properties linked to poor developability for any candidate mAb with known heavy and light chain variable domain sequences.

In addition, Abpred[51] tool can be used to predict the biophysical performance of commonly used developability assessment assays with just the amino acid sequence input. In Abpred, machine learning methods have been trained on heavy and light chain variable domain sequences from the Jain dataset using the amino acid composition and other fifteen sequence-derived features to represent physicochemical properties of antibodies. Other developability assessments using machine learning approaches have been used to predict and select the antibodies with optimal pH and thermal stabilities from 77 antibodies in development at Pfizer.[52] Lonza Biologics has also demonstrated the use of aggregation propensity screening along with other computational approaches during early drug development to select molecules with reduced risk of aggregation and optimal developability properties for screening several anti-interferon γ antibody variants.[18] Pfizer has implemented *in vitro* assays that correlate with *in vivo* human studies to differentiate mAbs at high risk for rapid clearance from those with favorable PK.[53] Finally, molecular dynamics simulation has

also implemented a high-throughput developability workflow on a panel of 152 human or humanized mAbs.[54] Here, physicochemical properties of these 152 mAbs were evaluated from multiple biophysical assays - size exclusion chromatography for aggregation, reverse phase chromatography and sodium dodecyl sulfate capillary electrophoresis for purity, differential scanning fluorimetry for thermostability, hydrophobic interaction chromatography (HIC) for hydrophobicity, affinity-capture self-interaction nanoparticle spectrometry for self-interaction and capillary isoelectric focusing for isoelectric point (pI) and charge variant analysis. These examined biophysical properties and key assay endpoints were also predictive of key downstream process parameters in development and clinical manufacturing.[54]

Design of antibody libraries with improved developability:

Screening libraries of antibodies is a commonly used strategy in antibody drug discovery. There are two main approaches to library design: creation of 1) a highly diverse library potentially containing binders to varied targets, or 2) a library focused on potential binders to a specific antigen or set of antigens. The ideal library contains genetically varied antibodies with the potential for high affinity and activity, but this can result in the generation of increasingly large libraries to achieve high diversity. With the huge amount of available sequence data and increased understanding of developability prediction, methods are being investigated for the optimal design of antibody libraries with high functionality and desired biophysical properties.

Natural:

Methods using B-cell receptor (BCR), i.e., antibody repertoires from antigen-exposed animals or humans ("immune libraries", to generate antigen-specific libraries) or from non-exposed humans ("naïve libraries", to generate functionally diverse libraries) try to capture the capabilities of the natural immune response in making functional, highly expressed, and low immunogenicity antibodies. However, not all naturally occurring antibodies are suitable drug candidates owing to other developability concerns, such as aggregation.[16] Libraries can aim to combat this by selecting the genes with known favorable characteristics using native heavy and light chains for improved specificity.[55-57] Limitations to natural-repertoire approaches also include the inherently biased nature, meaning diverse antibodies may be missed owing to sequence space restrictions. Nevertheless, available sequence space might not be as

constrained as previously expected, as multiple clinical-stage therapeutics have high sequence-identity matches in naturally sourced antibody repertoires.[21] Another way to select antibodies is by considering the "structural space". For example, a library of antibody structures identified in the repertoires of multiple individuals was found to contain structures highly similar to clinical-stage therapeutic antibodies[58] and may suggest antibodies with functionality and a low likelihood of immunogenicity.

Synthetic:

Synthetic libraries introduce diversity, often at defined regions of an antibody, to generate novel and varied sequences. Such methods can produce antibodies with higher affinity than natural repertoires,[59] but a proportion of the library may be non-folding or immunogenic. To reduce non-functionality, methods such as position frequency analysis (PFA) and deep learning have been applied. PFA introduces mutations based on the amino acid frequencies found at each CDR position in natural antibody repertoires, often using identical or only a small number of framework regions.[60, 61] Such methods do not account for correlations between residues at different sites. A different approach has used a database of antibodies with known functionality and interchanged CDR regions, assuming CDR regions are modular and can be interchanged without negative impact. In doing so they achieved high functionality.[62]

Deep learning:

Deep learning models aim to utilize the stability of natural repertoires and capture higher-order dependencies, missed by PFA, to avoid producing non-functional proteins. However, current limitations of deep learning approaches include a focus on only CDR regions or heavy chains, with a lack of experimental validation of predicted properties. For instance, 74% antigen binding was achieved in a mouse library designed by a variational autoencoder that generated novel CDR-H regions, but such an approach ignores non-CDR region contributions to the paratope, and the diversity of the sequences in this library is unknown.[63] Other generative approaches such as Generative Adversarial Networks can be trained on natural human antibody repertoires and biased via transfer learning (further training on antibodies with known properties such as solubility, stability, and predicted immunogenicity) to generate sequences predicted to have the desired biophysical properties.[64] However, more

information is needed to understand how such properties influence the overall developability of the antibody. Additionally, experimental validation of the predicted properties is necessary, as has been conducted for an enzymatically active protein library[65] and a nanobody library created by a generative deep neural network-powered autoregressive model trained on a native llama repertoire.[66]

Previous work has demonstrated the use of mammalian display libraries for the selection of antibody variants with optimal biophysical properties, reduced polyreactivity, and immunogenicity.[67] Here, they have described the use of a nuclease-directed integration system to generate antibody variants with differing biophysical properties based only on the display level achieved on the mammalian cell surface. Other studies have demonstrated the use of machine learning-guided directed evolution on the combinatorial sequence space.[68] Recently, a machine learning pipeline has been formulated to predict the developability of a library of 2400 antibodies from sequence.[69] These advances in bioinformatics and *in silico* methods have enabled the efficient development of commercially viable mAbs. Thus, antibody library variants are designed to exhibit better developability than the parent molecule.

Mitigating aggregation and post-translational modifications:

Aggregation:

Aggregation of antibody-based drugs can lead to precipitation and decreased shelf-life of drugs before administration, while aggregation *in vivo* can increase the immunogenicity of the drug. The aggregation propensity is a critical attribute correlated with product failure.[70] Indeed aggregate levels in the final drug product are key quality indicators.[71, 72] Seeliger *et al.* have highlighted four key factors that must be avoided to minimize aggregation, many of which can be predicted computationally: 1) the number of "reactive sites", such as those susceptible to oxidation, deamidation, or proteolysis, should be minimized; 2) thermodynamic stability should be high to minimize protein unfolding; 3) the structure should not contain hydrophobic or charged surface patches; and 4) the sequence should not contain cross-beta-sheet aggregation hotspots.[73]

Van der Kant *et al.* showed that mutating residues in predicted aggregation hotspots could reduce aggregation and found that those hotspots having the largest impact on thermodynamic instability are frequently found in the CDRs.[74] The solubility can be

improved in mAbs having aggregation-prone regions (APRs) by inserting glycosylation sites near these APRs.[75, 76] Several other studies have used protein-engineering approaches to reduce self-association and aggregation to achieve high solubility and low viscosity.[77-80] A specific prediction of the tendency to aggregation is the AggScore,[81] which uses structural modeling to identify patches at risk of driving aggregation. Several methods have been developed to create so-called "developability indices" for antibodies and these tend to focus on aggregation propensity. For example, Lauer *et al.* used data from the storage of 12 IgG antibodies for periods of up to two years to examine aggregation. They then combined net charge (at a given pH using a calculated pKa) with a "spatial aggregation propensity" (SAP) score (derived from accessibility and residue hydrophobicity and calculated over a molecular dynamics simulation) to create their developability index and correlated this with the experimental aggregation propensity.[45] Developability Index[45] is a well-known tool for estimating the developability of a candidate antibody. However, a potential drawback of the Developability Index is that it is based only on the full-length antibody's net charge and the SAP of the CDR region, and, therefore, may ignore other indicators of developability.

The Therapeutic Antibody Profiler (TAP) has been demonstrated to be very useful in selectively highlighting antibodies with expression or aggregation issues.[16] Further, Lonza's aggregation prediction tool[18] has been instrumental in the selection of lead antibody candidates from combinatorial libraries with improved developability. abYsis[39] incorporates a wide-ranging species-specific analysis of residue frequencies that can be combined with residue clustering to identify either hydrophobic or unusual patches that are likely to be important for the stability and immunogenicity of biopharmaceuticals. Therefore, using these computational aggregation prediction tools can identify aggregation issues early in biopharmaceutical development and avoid expensive late-stage product failures.

Post-translational modifications:

PTMs can lead to several issues encountered with the development of antibodies. By their nature, PTMs lead to heterogeneity, something that generally concerns regulators since variants must be considered in risk assessments and during characterization to

assess the impact on product quality, safety, and efficacy. This includes potential effects on antigen binding, immunogenicity, and Fc-mediated effector functions.

In antibodies, the N-terminal glutamate or glutamine is frequently cyclized by nucleophilic attack of the lone pair of electrons from the backbone terminal NH2 onto the sidechain carboxy or amide, forming a 5-membered lactam ring known variously as pyroglutamic acid (pyroGlu), pyrrolidone carboxylic acid (PCA), 5-oxoproline, or pidolic acid, and this has been shown to occur *in vitro*.[82-84] The N-terminus is comparatively close to the antigen binding site, so the difference in charge could have an effect on antigen binding, particularly for large antigens that may approach close to this part of the antibody. In addition to N-terminal heterogeneity, "clipping" frequently occurs at the C-terminus of the heavy chain. The last three residues of the heavy chain are Pro-Gly-Lys; the proline is the last residue of the CH3 domain, and the glycine and lysine form the CHS region. The C-terminal lysine is mostly clipped post-translationally by endogenous carboxypeptidases during cell culture, or by endogenous serum carboxypeptidase B once the antibody is administered to a patient.[85] However, this PTM is unlikely to have any serious effect on the *in vivo* performance of antibody-based drugs since the C-terminus is remote from any functional sites. That said, C-terminal clipping has been shown to be required for optimal complement activation and the presence of the lysine can affect the blood circulation time.[86] The third major PTM in antibodies is the N-linked glycosylation present in the CH2 domain. While these are the three best-known PTMs present in the vast majority of antibodies, many other sequence-specific PTMs are also observed, all of which led to heterogeneity potentially affecting charge, pI, aggregation, and binding. Heterogeneity as a result of PTMs and their effects are reviewed by Liu *et al.*,[87] while a comprehensive analysis of charge heterogeneity in adalimumab (Humira®) was performed by Füssl *et al.*[88]

Asparagine and aspartate residues form hot spots susceptible to deamidation and isomerization.[50, 89] In addition to the effect of antibody deamidation, there have been reports of deamidation in protein antigens in severe diseases such as anthrax.[90] Oxidation of methionine and tryptophan residues is another sequence liability that can lead to low potency, decreased thermal stability, and high aggregation propensity.[91, 92] Disulfide scrambling due to cysteine residues is another phenomenon causing configurational changes in the hinge region of antibodies, thus impeding antigen

binding and mAb functionalities.[93, 94] The variable domains of mAbs may also contain N-glycosylation sites, which may cause variable domain glycosylation that results in the formation of Fab-associated oligosaccharides with α1,3-galactose that are known to cause immunogenicity.[95-97] These PTMs often lead to low potency, immunogenicity, and instability of circulating mAbs.[98] Consequently, suitable developability assessment protocols must be designed to capture these sequence liabilities.

abYsis[39] (http://www.abysis.org) provides screens for a number of these PTMs for the optimization of therapeutic antibodies. It also annotates residues as being exposed, buried, or intermediate based on averaged information from several hundred known structures and can be used in concert with abYmod (http://abymod.abysis.org) to build an antibody model from which more detailed exposure information can be obtained. As described, PTMs could seriously hamper the safety or efficacy of therapeutic antibodies and this safety concern calls for an immediate need for appropriate tools to relate a biophysical property to a single, or a set of, molecular sequence-structural motifs in biologic drugs. In summary, biopharmaceutical informatics tools are used to locate the amino acids critical for biophysical properties that are in undesirable ranges.

Biopharmaceutical informatics for drug safety and *in vivo* performance:

Drug safety:

A strategic framework for using computational tools for predicting chemical degradation sites in biologic drugs has been presented in a previous study by Sandeep *et al.*[14] Several computational tools for predicting the toxicity of antibody-based drugs are now available.[99] A critically important step in drug development for establishing clinical safety is the identification of adverse drug reactions (ADRs). Computer-aided prediction of ADRs provides an alternative to recognize ADRs before clinical trials. Kuang *et al.* have reviewed and compared the computational models available for predicting ADRs.[100] Here, among the topological features of drug-ADR association networks, the Jaccard coefficient (a measure of the relationship between the neighbourhood set of homology nodes) was the most important feature for the prediction of drug-ADR associations. Consequently, the Jaccard coefficient of drug-ADR association networks is an important topological feature that should be used in models designed for prediction of antibody drug safety.

Previous computational approaches have estimated *in vivo* performance descriptors such as the PK, PD, and immunogenicity of biologics.[53, 101-104] Avery *et al.* have demonstrated a combinatorial triage approach on *in vitro* assay parameters and categories for screening therapeutic mAb candidates with desirable PK properties and minimal non-target-related PK risk.[53] Here, threshold values of *in vitro* assays reflecting non-specific interactions and self-association were established to define criteria for avoiding the selection of mAbs with rapid *in vivo* clearance. Grinshpun *et al.* have also analyzed biophysical and sequence-based *in silico* properties that are predictive of PK properties such as clearance for a panel of 64 clinical-stage mAbs.[105] They have concluded that experimental poly-specificity assay results and *in silico* estimated pIs were the best predictors to estimate clearance in therapeutic antibodies.

Antigen-antibody interactions:

General protein-protein interaction prediction tools for proteins frequently do not work well for antigen-antibody interactions because antibody-antigen binding is a rather distinct mechanism. Unlike normal protein interfaces, the epitope on an antigen has evolved to be an exposed region rather than to be involved in a protein-protein interface. Consequently, other computational techniques such as epitope mapping are used to identify the regions of an antigen likely to form the epitope before docking. B-cell Epitope (BCE) mapping tools can broadly be divided into linear epitope predictors, which attempt to identify epitopes consisting of continuous amino acid primary sequences, and conformational epitope predictors, predicting discontinuous epitopes in three-dimensional (3D) space.[24] However, like other protein-protein interfaces, antibody-antigen interactions involve a combination of non-polar van der Waals interactions, hydrogen bonding, charge interactions, and the hydrophobic effect. Consequently, along with these epitope prediction tools, several docking algorithms such as Megadock, Haddock, RosettaDock, and Piper are being actively used to understand the binding between an antibody and the target. However, their performance is often poor compared with general protein-protein docking.

Immunogenicity:

The presence of T-cell and B-cell epitopes influences the immunogenicity of antibody therapeutics, and, therefore, bioinformatics approaches to avoid immunogenicity fall into two major categories: T-cell epitope prediction and B-cell epitope prediction.

Computational tools for immunogenicity risk assessment provide an alternative to *in vitro* or *in vivo* immunogenicity assays. The use of *in silico* tools to identify lead candidates with a reduced risk of immunogenicity is an important step in biologic drug development.

T-cell epitope prediction, which is relatively well-established, requires predicting linear peptides within a protein sequence that will bind to the Major Histocompatibility Complex (MHC). MHC molecules present peptides to T cells, which trigger T-cell immune responses. MHC molecules can be classified into class I and class II. MHC class I molecules present peptides derived from intracellular proteins, whereas MHC class II presents peptides from extracellular proteins. Since antibodies are extracellular, the focus is on the prediction of peptide binding to MHC class II molecules. These tools usually examine the primary sequences of candidate antibodies to identify binding motifs of MHC class II allotypes or for similarity to epitopes known to elicit an immune response. Several MHC class II binding predictors are available and the overall prediction performance is generally good.[106, 107]

For example, some tools such as RANKPEP,[108] Propred,[109] Tepitope,[110] and NetMHCII[111] make predictions based on algorithms trained on MHC class II binding assay data. Other tools such as NetMHCIIpan and IEDB (Consensus)[112] are based on sequence alignments with MHC class II binding peptide databases. Overall, studies have established that NetMHCIIpan, Propred, IEDB (Consensus), and MULTIPRED[113] were the best predictors of MHC class II binding and these are the most commonly used tools in the industry for the prediction of MHC class II binding. Other previous studies compared nine different MHC class II binding prediction tools and six different methods showing that NetMHCIIpan was the best method to predict peptide binding to MHC class II epitopes with an updated version, having improved predictions, now available.[114, 115] While less important for antibody-based drugs, computational tools for determining binding to MHC class I molecules require locating motifs that bind to the binding groove. Prediction methods for interrogating peptide binding to MHC class I alleles include NetMHC-3.0,[116] NetMHCpan-1.0, the Kernel-based Inter-allele peptide binding prediction system,[117] and Adaptive Double Threading.[118] Based on this predicted T-cell epitope information, Yachnin *et al.* recently developed a Rosetta-based platform to deimmunize therapeutic proteins.[119] They incorporated a new score term utilizing predicted or experimentally identified T-

cell epitope information into the scoring function so that computational protein design calculations can be guided based on the epitope information as well as the energetic stability.

In contrast to the prediction of T-cell epitopes, a much harder task is B-cell epitope (BCE) prediction – predicting sites where the patient antibodies will bind to the drug. Such approaches have not been very successful, mostly owing to the discontinuity of antigen binding sites. As mentioned above, the problem is made harder by the fact that B-cell epitopes are, by their nature, regions of a protein surface that have not evolved to be involved in protein-protein interactions. Consequently, they do not have clearly recognizable features that are bound by antibodies.[120] Nonetheless, some regions will be more likely to interact with an antibody than others, but making mutations to remove a dominant B-cell epitope can simply result in the immune response switching to a less dominant epitope.

Several predictors have been produced that work at either the sequence level or the level of 3D structure. The earliest BCE prediction methods attempted to predict linear epitopes (i.e., a continuous stretch of amino acid sequence) using sequence features such as hydrophilicity,[121] amino acid composition,[122] and predicted accessibility and mobility.[123] An early evaluation showed that no single sequence feature performed well, leading to attempts to combine features.[124] However, machine learning efforts[125] and additional features such as sequence conservation[126] have provided limited improvements to BCE prediction. In general, conformational epitope predictors such as CBTOPE, BETOPE, CEP, and DISCOTOPE are more accurate than linear epitope predictors such as LBTope, SYMTriP, and ABCored.[127-129]

The performance of computational epitope prediction tools and tools for predicting immunogenicity has been reviewed previously to establish guidelines for the deimmunization of protein therapeutics.[130] It is worth nothing that epitope databases are not exhaustive because of the heterogeneity of proteins involved in the immune response across the human population.[101] This variability of immune response for the same antigen limits the utility of *in silico* immunogenicity assessment methods as stand-alone tools. Therefore, this key limitation of immune response diversity needs to be captured by the forthcoming immunogenicity prediction tools.

Guidelines for the design of developability assessment protocols:

Assessment of developability by biopharmaceutical informatics protocols at an early stage in a development pipeline reduces the costs of development failures. Companies using transgenic mice to produce antibodies can generate as many as a million sequences a week (after cleaning the high-throughput sequence data) and it is impractical to take all these through to experimental validation. Even computational evaluation requires significant computing resources and optimization. If each sequence takes one second to analyze, a million sequences will require ~11.5 days of computer time. Consequently, it makes sense to use a triaging pipeline that performs evaluations that can be done quickly first and leave more computer-intensive evaluations to be performed only on those sequences that have survived the initial rapid triages.

A screening paradigm used in the industry for selecting mAbs with desirable PK properties during mAb discovery and lead selection has been demonstrated in a previous study.[53] This staged approach for developability assessment involves using the high-throughput assays first when hundreds of mAbs are available for screening. Here mAbs scoring above assay thresholds or having results outside the acceptable range are deprioritized because they have unfavorable physicochemical properties. Next, additional physicochemical properties such as thermal stability are evaluated for only the mAbs that have passed the previous stage. These additional screens include assays measuring properties such as biological activity, expression, and stability that are often low-throughput and need higher quantities of mAbs.

Finally, a combinatorial triage approach is used that ranks and classifies the mAbs based on the aggregate result of all the assays. It is very important to combine the results of multiple assays together since individual developability assays can have some false-positive results. This ensures that mAbs with desirable physicochemical properties advance to scale-up and costly preclinical and clinical development.

A Computational Developability Assessment (CDA) workflow should follow a similar strategy where a panel of high-throughput computationally undemanding tools is applied first to a mAbs library followed by specific computationally intensive antibody informatics tools as per the required objective, such as those for immunogenicity assessment. The final step in the CDA workflow as shown in Figure 4 is to use a combinatorial triage approach to combine scores and rankings from

multiple tools together and classify the mAbs based on the aggregate result of all the informatics tools.

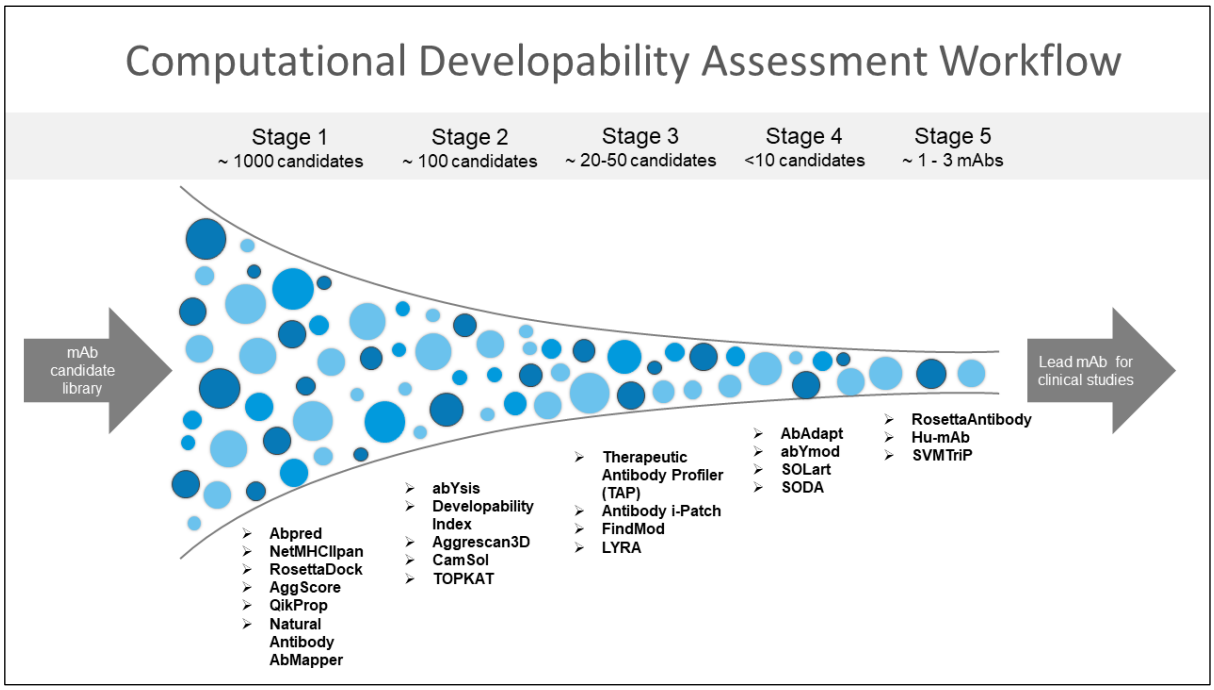


Figure 4: Computational Developability Assessment workflow for screening mAbs with optimal biophysical properties. An orthogonal combination of conceptually different algorithms is used to reduce method-specific biases. High-throughput antibody informatics tools are implemented first to an antibody library. mAbs scoring above assay thresholds or having results outside the acceptable range are deprioritized. Next, more computationally intensive antibody informatics tools are applied to evaluate additional developability issues. The final step in the CDA workflow is to use a combinatorial triage approach to combine scores and rankings from multiple tools together to classify the mAbs from aggregate result.

Together with previously discussed approaches to assessing developability, Raybould *et al.* have described five computational developability guidelines for therapeutic antibody profiling: 1) total CDR length, 2) patches of surface hydrophobicity (PSH) metric across the CDR vicinity, 3) patches of positive charge (PPC) metric across the CDR vicinity, 4) patches of negative charge (PNC) metric across the CDR vicinity, and 5) structural Fv charge symmetry parameter.[16] Overall, local charge and global charge asymmetry between the CDR and the framework have been correlated with higher aggregation and poor developability. Here, the approach was to look at the characteristics of clinically successful antibodies and rank candidate antibodies by ensuring they stay within these bounds. This is conceptually similar to Lipinski's rules used in small-molecule drug design.[131] An efficient high-throughput developability workflow was also demonstrated by Bailly *et al.* on a panel of 152 mAbs for rank

ordering of molecules during early-stage discovery screening.[54] Here, they have demonstrated that key physicochemical properties from multiple biophysical assays correlated well with major downstream process parameters.

As above, most types of analysis performed for developability assessment include identification of PTM sites, analysis of likely aggregation propensity (largely through examining surface hydrophobicity), pI, prediction of stability, and identification of T-cell epitopes/B-cell epitopes together with humanness scoring or unusual surface patches. Other considerations that can be included early in the pipeline include checking for the presence of the standard two cysteines present in antibody variable domains, the Trp-Gly motif present immediately after CDR-H3, and the length of CDR-H3 since unusually long CDR-H3 loops have been correlated with poor developability.[132] However, each tool relies on different interpretations and weighting of the essential features that determine developability. Therefore, an orthogonal combination of conceptually different algorithms should be used in computational developability assessment protocols to reduce method-specific biases.

1.2.4 Applications of biopharmaceutical informatics

Biopharmaceutical informatics for solubility predictions:

Solubility is one of the key biophysical properties that underpins developability potential, as high solubility typically translates into high expression yields, low aggregation, and provides the opportunity of formulating products at high concentrations while retaining a good shelf life.

Identifying antibodies with low solubility and high aggregation propensity from combinatorial libraries remains a hurdle for antibody development. Several *in silico* predictors have been reported that are now able to predict solubility or aggregation propensity accurately in many cases, a feature that makes them highly competitive with experiments.[48, 133-135] These solubility predictors include CamSol, Protein-Sol, SOLpro, SODA, Aggrescan, SAP, and Solubis. They have been effective at predicting the solubility and aggregation propensity of diverse antibody libraries.[133, 135]

As an example, the CamSol method of predicting solubility relies on a combination of physicochemical properties of amino acids. These include charge, hydrophobicity, and propensity to form secondary structure elements, which are first considered at the

individual residue level, then averaged locally across sequence regions, and finally considered globally to yield a solubility score.[135, 136] In particular, while a structural model is necessary to identify aggregation hotspots, the solubility prediction itself is performed using only the amino acid sequence. This aspect makes computational calculation significantly faster and makes the method readily applicable to the screening of antibody libraries without the need for structural modeling, thus it is fully independent of model accuracy.

For example, CamSol was used to rank the solubility of hits from a phage-display library from MedImmune.[136] The mAbs that were analyzed differed by up to 32 mutations in the Fv region, and the correlation between prediction and experiments of PEG-precipitation was R~0.97 after one outlier was removed ($p<10^{-4}$), which is fully consistent with the R~0.98 reported for a nanobody in the original report.[135] Similarly, a statistically significant correlation (R ~ 0.71 to 0.93) between CamSol predictions and solubility measurements was also reported for mutational variants of a troublesome mAb.[137] In a study on a library of 17 mAbs from Novo Nordisk, CamSol predictions were compared with a battery of commonly used developability assays and one measurement of relative solubility, and the correlations between CamSol and these experimental readouts were on a par with those seen between the assays.[138]

Notably, all these measurements were carried out with different experimental techniques, on widely different molecules, and in different laboratories. Taken together, these strong correlations suggest that CamSol predictions can greatly facilitate the screening of solubility and hence the developability potential. In particular, at the initial stages of antibody discovery campaigns, when the numbers of candidates can be very high while yield and purity are often low, such predictions may entirely replace experiments.

Kingsbury *et al.* have previously predicted the solution behaviour of a diverse dataset of 59 mAbs, including 43 approved antibodies, using a comprehensive array of 23 molecular descriptors categorized as colloidal, electrostatic, conformational, hydrodynamic, and hydrophobic.[139] They have shown that the diffusion interaction parameter (k_D), a measure of colloidal self-interaction is the key parameter that is most predictive of solution viscosity and opalescence for mAbs. So, they have postulated that computational developability assessment protocols should use a threshold value

of the diffusion interaction parameter, k_D (10 mM histidine-HCl buffer at pH 6.0) to screen antibodies with optimal antibody solution behaviour.

Biopharmaceutical informatics for predicting protein stability and interactions:

There can be opportunities to address the underlying balance of biophysical forces that drive interactions when developing models to predict the properties of biopharmaceutical candidates. Two such examples are discussed here, one relating to the measurement of hydrophobic interactions and the other to the protein structural basis of hydrophobic interaction between proteins. Several machine learning methods to predict the HIC retention time from antibody sequence input have been reported previously in the literature.[44, 51, 140] Assessment of aggregation propensity using HIC was the best-predicted biophysical property across 12 models produced using Abpred (www.protein-sol.manchester.ac.uk/abpred), one for each of the 12 biophysical properties measured across a set of antibodies.[51] Even so, there was a marked reduction in performance of the model for antibodies with higher retention times in HIC, leading to a model in which the salt gradient that is used to modulate hydrophobic interaction strength also affects interactions between charged proteins. A revised scheme was derived in which charge interactions play a role alongside hydrophobic effects in the HIC method. In this scheme, proteins with higher net charge repel more within the column when salt concentration (ionic strength) is lower, and are eluted faster, than proteins with lower net charge but the same hydrophobicity.

In this second example, another set of HIC data for 24 antibodies was used.[141] Here, the aromatic sidechain content of CDRs correlated well with the experimental data, but the equivalent correlation was much lower for the solvent-accessible surface area calculated for non-polar atoms in the CDRs[142] and it was concluded that hydrophobic interaction strength may be dependent on non-polar surface shape as well as surface area, consistent with thermodynamic measurements made for mutations in an antibody-antigen interface.[143] These examples demonstrate that models rooted in biophysical descriptions of protein stability and interactions, and benchmarked against experimental data, can both provide predictive insight for biopharmaceuticals and further the understanding of the underlying biophysical mechanisms.

Biopharmaceutical informatics for pre-clinical immunogenicity risk assessment:

A key concern with any biologic drug is immunogenicity, the effects of which range from simply having an immune response, meaning that the drug is rapidly cleared from the body when administered, through to the possibility of anaphylactic shock. As described above, methods can be applied to predict T-cell epitopes and (to some extent) B-cell epitopes, but a more practical approach has been to ensure that antibody-based drugs are as human as possible, and this has become one of the main aims in producing antibody-based drugs. As described earlier, the first monoclonal antibody-based drug to be approved was a mouse antibody (muromonab). However, since then, efforts have gone into making antibody-based drugs less immunogenic, first by producing chimerics (where the variable domains are from the donor species while constant domains are human) and then by "humanization" (where the CDR loops that form the antibody combining site are from the donor species and the rest of the variable domains is predominantly human, as well as human constant domains).[144]

A halfway step between chimerics and humanization to reduce immunogenicity has been "resurfacing" of chimeric antibodies in which surface residues of the variable domain, away from the CDRs, are mutated to human residues.[145] This is done to remove primarily B-cell epitopes on the antibody surface. Many antibody-based drugs are now "fully human", being produced by phage display, using transgenic mice, or by identifying antibodies from recovering patients. However, antibodies produced by such methods can still be immunogenic. For example, adalimumab (Humira®, the world's top-grossing drug), while "fully human" (produced by guided phage display), elicits an immune response in >25% of patients with only 4% of these patients having sustained remission, compared with 34% of patients who did not have antibodies against adalimumab.[146]

Thus, even with fully human antibodies, computational BCE and TCE predictors can be used to predict B-cell epitopes and T-cell epitopes, which can also be experimentally identified through proteomics assays.[147, 148] It is then desirable to remove these potentially immunogenic regions in advance of clinical trials. As well as the application of BCE and TCE predictors, various "humanness" scores have been proposed based on sequence information of human antibodies, enabling the *in silico* assessment of human-likeness given sequences of antibodies.[149-151]

Recently, Schmitz *et al.* developed a computational method that maps the sequence of a given antibody onto human B-cell repertoires comprising 326 million sequences of human antibodies.[152] Chin *et al.* built a machine learning-based predictive model that distinguishes human antibody sequences from non-human ones, which was trained on a large-scale repertoire dataset.[153] These human-likeness scoring approaches will be useful when assessing how much the given antibodies are close to human repertoires; the more human-like antibodies are, the less immunogenic they are expected to be. As described above, another approach is to identify patches of unusual residues on the protein surface that may lead to an immune response.

1.2.5 Future perspectives in biopharmaceutical informatics

Decoding human antibody gene repertoires and their role in target validation and drug discovery:

New high-throughput sequencing methods have generated a vast amount of antibody sequence data, with over one billion antibody sequences publicly accessible in repositories.[23, 39, 154-159] A sequenced human B-cell receptor (BCR, i.e., antibody) repertoire provides a snapshot of the BCRs present, typically those circulating in the blood, at a given time. BCR sequence and structure datasets can be used to investigate immune system mechanisms for improved library design, understand disease pathogenesis, and identify antibodies for potential therapeutic development.[160]

Immune system mechanisms:

The diversity of BCR repertoires can be used to develop an understanding of the mechanisms underlying the immune system. Typical BCR repertoire profiling includes sequence-based analysis, such as clonotyping. Clonotyping involves clustering sequences into clones, usually based on identical V and J genes and high CDR-H3 identity.[161] Such analysis can reveal dominant antibody sequences, potentially indicative of a response to an antigen, e.g., after vaccination. The availability of large datasets has been useful in characterizing the response to antigens and estimating true antibody genetic diversity,[162] though these are still far from fully understood. Sequence-based analysis has revealed that the immune systems of unrelated individuals have similarities; an estimated 0.02% of clones are “public” – shared across multiple individuals.[163] However, differences identified between

identical twins indicate the complexity of the immune response and the importance of epigenetics and environmental factors.[164] Understanding such mechanisms is useful for antibody drug development, for example, to design antibody libraries for drug discovery.

Understanding disease pathogenesis:

Immune responses to disease, and also therapies, can be profiled using BCR repertoires to investigate B cell subtype involvement and levels of antibody response. Using such analysis, we can distinguish between healthy and disease repertoires and learn about disease mechanisms, particularly those associated with B cells, such as autoimmune diseases, chronic lymphoid leukemia, and other cancers.[165, 166] In the future, such information will hopefully be used to improve patient outcomes by identifying the most at-risk patients, tracking disease progression and monitoring response to therapies. A better understanding of the immune system involvement in disease may also indicate targets for potential therapeutic intervention and even suggest antibody drug candidates present in the BCR repertoires of patients with the disease.

Therapeutic antibody candidate identification – using sequence information:

BCR sequence repertoires can be used to suggest suitable candidates for drug development. A previous study has contextualized the sequence and structural properties of clinical-stage antibodies with human immunoglobulin datasets (Ig-seq) to evaluate the extent of humanness/originality of antibodies in clinical investigation.[16] Whilst not all naturally occurring antibodies make good drug candidates, 29 clinical-stage therapeutic antibodies were found to share 100% CDR-H3 identity with a BCR sequence from a healthy human repertoire.[21, 163] By looking for antibody sequences frequently found after exposure to an antigen, we can identify those that might bind specifically to that particular antigen. When assessing individuals with the same disease or who have been exposed to the same antigen (either through infection or vaccination), these sequence-convergent responses can be a useful starting point for a potential therapeutic. Evidence to support this approach for drug discovery comes from vaccine studies[167] and more recently SARS-CoV-2-infected individuals, where convergent antibodies had sequence similarity with identified SARS-CoV-2-binding

antibodies.[168] In addition to being potential binders, public clones may also have low immunogenicity, making them attractive as drug candidates.[58]

If existing binders are already known, likely drug candidates can be identified from a BCR repertoire by comparing them with known antibodies binding to the desired antigen. Identification can be based on sequence identity, such as clonotyping,[169] or prediction of similar binding properties.[170] As such, sequence data from BCR repertoires can be useful starting points for suggested therapeutic antibody candidates, with or without knowledge of existing binding antibodies.

Therapeutic antibody candidate identification – incorporating structural information:

Whilst most examination of immune repertoires focuses on sequence analysis, utilizing available structural information may also be important when identifying potential therapeutic antibody candidates. Conventional antibody modeling tools are inefficient for building 3D models of entire repertoires of BCRs, with the fastest taking seconds per antibody model via homology modeling methods[31, 171], or ~285 CPU hours[172] with *ab initio* methods. Therefore, structural modeling methods have been developed specifically for large-scale BCR or TCR repertoire data analysis. Incorporating structural information from models can allow the prediction of antibody properties in a repertoire and we may be able to predict antibody domain binding by performing structural clustering of antibody models with known-function antibody datasets, such as CoV-AbDab.[173, 174]

A high-throughput alternative to modeling utilizes structural annotation to rapidly predict antibody CDR loop shapes, based on sequence identity matching to a template.[175] Repertoires can be evaluated based on predicted CDR structures, for example, to identify over-represented CDR-H3 templates or clusters of templates that may represent a response to an antigen, and therefore be a useful starting point in therapeutic antibody design. Using structural prediction tools with BCR repertoire sequence data can reveal antibody drug candidates not seen using sequence-only analysis.

Current limitations for utilizing BCR repertoire data in drug development include the major challenges of predicting antibody-antigen binding and affinity. In addition, existing BCR sequencing datasets often contain only heavy chain information and methods for obtaining BCR repertoires and binding affinities are varied and lack any

standardized protocols or analysis pipelines. With the development of high throughput methods for single-cell sequencing and antigen specificity mapping,[176] increased amounts of high-quality, antigen-labeled antibody data might enable new accurate and reliable computational methods for drug discovery.

Biopharmaceutical informatics for design and optimization of next-generation biotherapeutics:

The spectrum of biological activities accessible to antibody therapeutics is being expanded by exploring novel mechanisms of action. For example, bispecific antibodies can be created by engineering different specificities into each arm of the antibody, and multi-specific antibodies can be created by adding further VH/VL domains on the heavy and light chains or as a single-chain Fv (scFv) appended on the N- or C-terminus. In addition, novel binding functions can be created using scFvs or nanobodies (heavy-chain only), often combined in tandem for higher avidity or multi-specificity. Other technologies include antibody-drug conjugates (ADCs) created by conjugating cytotoxic drugs (payloads) for site-specific delivery. These novel antibody constructs are often collectively referred to as "next-generation antibodies"[177] and are emerging as potential therapeutics with unique properties.

The sequences of these antibody formats may differ substantially from those of immune-system-derived immunoglobulins, as extensive engineering is typically required to bring about the desired functionality. It is often the case that engineering additional functionality comes at the expense of other important properties that underpin developability, including conformational and colloidal stability, solubility, immunogenicity, and PK. Therefore, the successful development of next-generation biotherapeutics presents additional challenges, which are usually system-specific. For example, ADCs are complex molecules that require careful attention to various components, including the mAb, the engineered drug conjugation sites, the selected linker, the payload, and the drug load distribution.[178-180] Similarly, multi-specific antibodies require the selection of multiple binding domains that must be successfully combined to ultimately yield a homogeneous product with the desired functionality and suitable developability profile.[177]

In general, the computational prediction of the developability potential of these novel antibody-based formats presents two overarching challenges. The first is that there is

no guarantee that combining together components with suitable properties will translate into a final therapeutic that has desirable characteristics. For example, a bispecific antibody obtained by combining two Fvs with good developability profiles may present unexpected liabilities, such as increased oligomerization brought about by cross interactions between its components. Therefore, while the tools described here may be used to pre-select or engineer binding domains and mAbs with optimal characteristics when these are combined in a multi-specific format the resulting construct may not necessarily be well-behaved. The second challenge lies in the combinatorial nature of combining multiple constructs, which amplifies prediction errors and hence the risk of failure, even assuming that different components behave independently. As an example, consider a computational predictor of a "good" characteristic (such as having good solubility) with precision, or "positive predictive value" (PPV) of 0.9 that implies a false discovery rate (FDR = 1-PPV) of 0.1 (i.e., of the positive predictions, 90% of them are correct, or in other words, 1 in 10 antibodies that are predicted as good are actually poor). If we apply this method to select two distinct Fvs for a bispecific antibody, then the probability of introducing at least one liability in this construct is given by $1-(PPV)^2$, i.e., 0.19 or 19%. Similarly, for a tri-specific construct, such as nanobodies in tandem, the probability of introducing a "poor" binding domain becomes 27.1%. Therefore, even when neglecting the first challenge and considering the different components as fully independent from each other, the accurate prediction of the developability profile of next-generation biotherapeutics will require exceedingly precise methods.

Some of the databases that can be used for the analysis of nanobody-derived therapeutics are the Single Domain Antibody Database[181] (sdAb-DB), Integrated Nanobody Database for Immunoinformatics[22] (INDI Nanobodies DB), Non-redundant Nanobody database[182] and database of Institute Collection and Analysis of Nanobodies[183] (iCAN). These databases host large collections of natural and synthetic camelid single-domain antibody sequences from literature sources and other online repositories. Each of these databases further provides unified annotation and integrative analysis tools for describing various single domain antibodies.

Overall, computational predictions of developability potential can already be used to aid the development of next-generation biotherapeutics. However, further developments are required before these methods will become highly competitive with

experimental readouts in terms of accuracy and reliability. To accelerate innovation in this area, it will be essential that experimental data of developability are published together with the antibody sequences used in the experiments, including any engineered modifications. We anticipate that, just as the Jain *et al.* study[17] and others[54, 184] spurred the development of several computational predictors,[16, 51] similar investigations using next-generation biotherapeutics will enable such methods to be refined, or new approaches to be developed, to yield accurate predictions of the developability profiles of these constructs.

Applications of artificial intelligence and machine learning towards antibody discovery, development, and manufacturing:

Machine learning algorithms have been used for the classification, regression, or clustering of biopharmaceutical experimental datasets. Machine learning models have been used for the prediction of protein secondary structure,[185, 186] relative solvent accessibility,[187-190] protein folding,[191-194] protein-protein interactions,[195-199] and PTMs.[200-203] Machine learning methods have also been applied to the prediction of aggregation using a classification tree ensemble with sequence-derived physicochemical properties.[18, 204] Other machine learning approaches such as gradient-boosting machines have been used for the prediction of CDR structure from protein sequence, particularly CDR-H3.[205, 206] The most common strategy used by these algorithms is the use of biophysical propensity scales as input features for machine learning methods to characterize the structural and functional properties of proteins.[207]

Narayanan *et al.* have reviewed the application of machine learning approaches in predicting the developability of antibody-based biologics.[208] A machine learning algorithm has been shown to predict antibody developability solely by sequence using a dataset of 2400 antibodies.[69] Here, a support vector machine model trained on physicochemical features with multiple sequence alignment emerged as the best machine learning pipeline combination to capture antibody developability from the sequence.

Deep learning approaches for antibody design and engineering are also becoming popular.[209] Several deep learning models have been described for predicting paratope regions in antibody sequences,[210] epitope-specific paratope identification,[211] predicting antibody/antigen binding,[212] CDR-H3 region optimization,[213] and virtual

screening for therapeutic antibody optimization.[214] Deep learning algorithms offer the ability to capture key biophysical features and properties for any developability objective without the need to create complex theoretical functions. Consequently, deep learning approaches are ideal for cases where mechanistic understanding of the underlying developability issue is not fully understood. However, deep learning algorithms generally require large amounts of data, and so can be unsuitable for smaller datasets.

The choice of the machine learning algorithm is decided by the dataset availability and the objectives of the application. Supervised machine learning methods such as support vector machines, random forests, and conditional random fields are usually more appropriate for balanced datasets.[202] Although machine learning-based methods lack the physical transparency of other approaches, their practical application is remarkably successful. Therefore, given that the amount of available training data across biological and structural databases is rapidly increasing, and that machine-learning algorithms are constantly improving, these methods are destined to play key roles in shaping the future of biopharmaceutical informatics.

1.2.6 Conclusion:

The past two decades have seen transformational advances in the biomedical sciences. In particular, the Human Genome Project has triggered the development of NGS technologies, which are enriching biological databases with millions of sequences of proteins, including antibodies from myriad different sources. Furthermore, improvements in the pace and accuracy of protein structure determination techniques are contributing unprecedented amounts of high-quality structural data, comprising large numbers of antibody-antigen complexes.[215-217] The increasing use of quantitative methods in biology has gradually transformed the way biological observations are made, and it is now possible to assemble large datasets of highly accurate measurements of antibody biophysical properties. Finally, computers able to perform complex calculations quickly are available, and extremely powerful algorithms for data mining and machine learning are constantly being developed. Taken together, these advances are enabling the antibody community to address questions that were essentially intractable a decade ago, including the development of highly accurate computational methods to streamline the development of biotherapeutics.

Here, we have described numerous metrics for computational developability assessment and established that no single tool or biophysical parameter can be used for predicting the developability potential of a biotherapeutic. The orthogonal combination of conceptually different algorithms should be used in developability assessment protocols to reduce method-specific biases. However, as stated by Narayaram *et al.*,[208] "one common disadvantage of such *in silico* tools is that they use only protein sequences or structure-based information as input and usually do not consider the impact of formulation conditions". The biophysical solution behaviour is also influenced by the excipients and solution conditions of the formulated product. Therefore, the developability assessment algorithms will have more real-life practical applications if they also consider the solution conditions and formulation parameters in the algorithms. In addition, minimal information has been provided in the available literature on the validation of these tools in the industrial setting. Therefore, it is important that biopharmaceutical informatics approaches are uniformly applied across the industry to expand and accelerate their potential for biotherapeutics development.

Biopharmaceutical informatics can also be a valuable guide for the commercialization and licensing of antibody-based drugs. The insights from computational developability assessments can aid the due-diligence activities performed during licensing and acquisition transactions.[218] The application of biopharmaceutical informatics tools is likely to increase in the future as new accurate and faster software are becoming available for generating antibody structure from the sequence for mAbs. Recently, AlphaFold,[219] a neural-network-based algorithm that was recognized as the optimal solution to the protein-folding problem at the Critical Assessment of protein Structure Prediction (CASP) competition, has received wide media attention, but its efficacy in modeling antibodies remains unproven. The recent success of AlphaFold at predicting protein structures demonstrates the power of bioinformatics applications. With increasing efforts devoted to data curation and method development as described here, biopharmaceutical informatics holds the potential to play a leading role in selection and engineering of safe therapeutics.

Acknowledgments:

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) via UK EPSRC grant EP/N024796/1. PS is a Royal Society University

Research Fellow (URF\R1\201461). We gratefully acknowledge Max Hebditch for his scientific discussions and critical reading of the manuscript.

Author contributions:

Conceptualization – RK, RC, JW; Original draft Preparation – RK, RC, JW, SR, KK, JH, AM, DK, UK, PS; Review and editing – RK, KK, RC, JW, AM, CD, PS; Figures and visualizations – RK, KK. All authors have read and agreed to the published version of the manuscript.

1.3 **Biopharmaceutical informatics**

Biopharmaceutical informatics refers to the application of computational methods and bioinformatics tools towards biopharmaceutical drug development. It incorporates a full strategic framework of computational tools and informatics applications such as database curation, big data analytics, molecular modelling, simulation, sequence and structure-based bioinformatics analyses, machine learning, and artificial intelligence.

The traditional biologic drug discovery and development driven solely by the use of experimental methods has been expensive and time-consuming. Moreover, from the cost perspective as well, the creation and screening of a large number of candidates in multiple biophysical characterization assays have been challenging. This has motivated the development of computational approaches for assessing biological product properties to minimize sample requirements and accelerate the biologic drug development process.

Computational tools are particularly useful in the early stages of biotherapeutic drug discovery where usually little or no experimental data is available. Biopharmaceutical informatics is, therefore, very useful for selecting antibodies to be prioritized and flagging antibodies to be deprioritized to shortlist antibodies for experimental testing.

1.3.1 Molecular modelling and simulations for biologic characterization

Computational modelling and simulation approaches have also been actively used for structure prediction and dynamic characterization of antibodies. Molecular dynamics (MD) simulations, in particular, provide a dynamic view of the conformational ensembles by numerically solving the Newton's equations of motion for a system of interacting particles to determine the trajectories of atoms and molecules.[220] So, these

simulations capture the fluctuations in the three-dimensional structure of proteins that may have timescale ranges from low nanoseconds up to seconds. Some of the major applications of MD simulations include antigen recognition, receptor binding, and identification of aggregation propensity and chemical modifications.[221]

There have been substantial advances in antibody structure prediction with more accurate biophysical surface property predictions. Deep learning methods like IgFold and AlphaFold are leading this revolution of reliable and accurate estimations for each residue and six hypervariable loops.[222-224] These can be used to build paratope ensembles to elucidate the function and properties of antibodies. Other specific molecular modelling tools have emerged for describing CDR loop movements, side chain orientations, interdomain rearrangements, and angle rearrangements of the antigen-binding site.

This conformational diversity of antibodies captured by modelling and simulation tools has been utilized to engineer developability liabilities and optimize biophysical properties. For instance, a previous study has used molecular dynamics simulations to estimate the thermal stabilities of antibodies by correlating the fraction of native atomic contacts (Q) to the melting temperature.[225] Also, conformational shifts toward more hydrophobic low-population states have been shown to accelerate aggregation of antibodies in a previous study.[226] Constant pH molecular dynamics simulations have been also used to reflect the sampling protonation changes to evaluate charge-dependent biophysical properties that facilitate antibody therapeutics development.

Multi-scale molecular simulations can help us to identify early formulation process challenges such as aggregation, viscosity, solubility, diffusion interaction, degradation and stress tolerance. Specifically, the use of explicit molecular dynamics simulations can potentially provide a molecular-level understanding of response to thermal and other stresses. Expanding the scope of MD simulations to include formulation parameters such as pH, buffers, salt, and excipients will pave the way towards *in silico* formulation development for antibody therapeutics.

1.3.2 Protein sequence-structural contexts for biologic product stability

Protein sequence-structural contexts play a crucial role in determining the stability of biologic products. The stability of a protein refers to its ability to maintain its folded

and functional state over time, under various environmental stress conditions. This is particularly important for biologic products, such as therapeutic proteins, antibodies, or enzymes, as their stability directly impacts their efficacy, safety, and shelf life.

The relationship between protein sequence and structure is well established.[227] The amino acid sequence of a protein determines its folding pattern and three-dimensional structure, which in turn influences its stability. The protein stability is an interplay between the presence and distribution of hydrophobic and hydrophilic patches on the protein surface, the ratio of secondary structure elements like α-helices and β-sheets, flexibility and conformational changes of the loop regions, linkers and disulfide bonds, and post-translational modifications.[228] This has paved the way for sequence-guided design approaches in biopharmaceutical informatics which provides insights into position-specific mutations that are tolerated and acceptable in therapeutic design.

We have witnessed, in particular, the emergence of *in silico* candidate screening tools which are based on *de novo* prediction of properties based on sequence and structure information. These *in silico* tools have been very successful in identifying high-value leads and new engineering targets for improving biological activity and stability. The era of machine learning has also arrived in biologic drug discovery and development expanding beyond image recognition and language translation. For example, machine learning enabled sequence analysis has been used to train a deep contextual language model on 86 billion amino acids across 250 million protein sequences spanning evolutionary diversity.[229] This merger between biopharmaceutical informatics and protein sequence-structural relationship insights has accelerated new asset discovery, optimized mAb designs, and de-risked development programs.

Some other applications of computational tools and informatics approaches in later drug development stages are in clinical trials workflow management actions such as patient selection and recruitment, study design optimization, outcome prediction and monitoring, data analysis and interpretation, and real-world evidence studies. Overall, these biopharmaceutical informatics approaches analyze historical clinical trial data to optimize future human clinical trial protocols. Figure 5 provides an overview of the application of biopharmaceutical informatics across the entire biologic drug discovery pipeline to accelerate new asset discovery, optimize design, and de-risk programs. We have focused on the *in silico* developability assessments part in this thesis work.

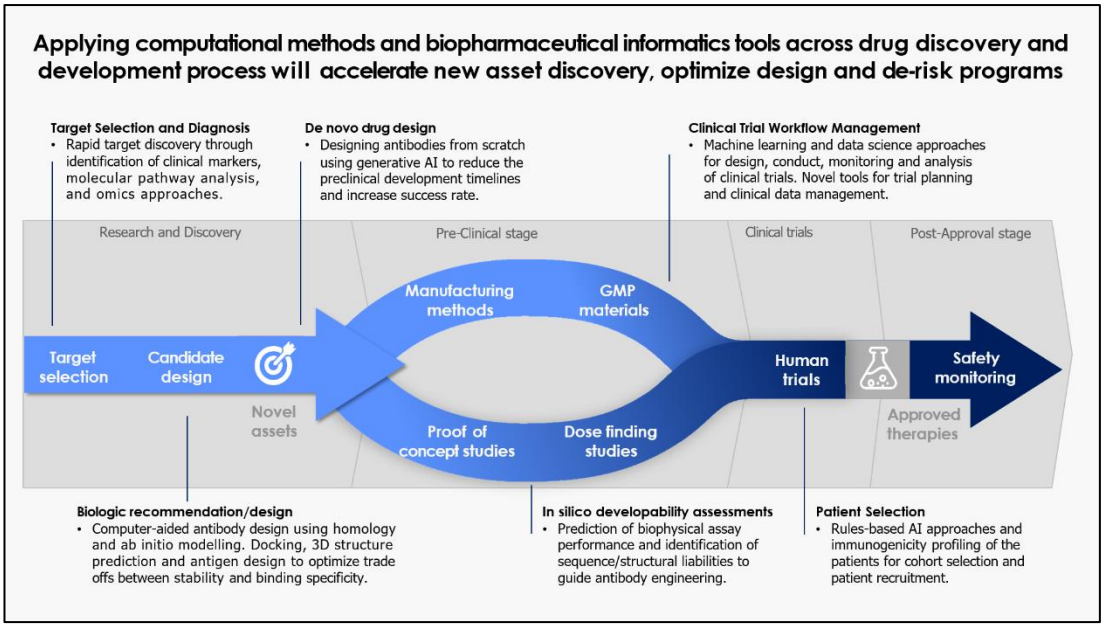


Figure 5: Application of biopharmaceutical informatics across the drug discovery pipeline

1.4 **Computational developability assessment**

Successful development of monoclonal antibodies into safe and effective commercial therapeutics can often be impeded by known developability liabilities such as poor expression, low solubility, high viscosity, and aggregation. A good developability profile is therefore a key attribute for a biological drug. Poor biophysical properties of antibody-based drugs can lead to decreased shelf-life of drugs before administration, while aggregation *in vivo* can lead to adverse reaction in the body. Clinical trial failures can arise due to poor developability factors which lead to several efficacy or safety concerns. Clinical trial failures are costly and result in setbacks for both the pharma company developing the drug and patients who are eagerly awaiting new treatment options. The risk of late-stage clinical failure of antibody candidates is highly valuable as the average cost of biologics development is USD 559 million per drug.[230]

1.4.1 Developability assessments at early-stage development

'Developability' refers to the likelihood of mAbs to be successfully developed as safe and effective drugs. Overall, good developability of monoclonal antibodies relies on various factors such as specificity, stability, safety, formulation, delivery, scalability pharmacokinetics, pharmacodynamics, manufacturability, and regulatory compliance.

To select the optimal candidate, an increasing amount of attention is being paid to the developability characterization of therapeutic antibodies. The insights gained from developability assessments help to devise better end-to-end antibody therapeutic drug discovery and development strategies. Various experimental platforms and antibody informatics tools have been created and employed to identify antibodies with bad or non-optimal biophysical properties, also referred to as developability issues.[46, 47, 138] The experimental platforms to evaluate non-specific binding are usually based on binding to a panel of antigen mixtures, defined protein reagents, or immobilized antibodies in immunosorbent chromatography assays.[231-233] Other key developability features especially important for concentrated mAb formulations are the propensity of self-association and aggregation which are evaluated using dynamic light scattering, nanoparticle spectroscopy, or self-interaction chromatography-based assays.[234-236] Several additional experimental assays have also been proposed in early-stage discovery such as surface-mediated stress assay and differential scanning fluorimetry.

High-throughput developability assessments are set to be actively used in biologic drug discovery and design to identify potential downstream risks that occur during manufacturing, undesired modifications, and aggregation during long-term storage, poor colloidal stability, solubility, and high viscosity precluding subcutaneous administration, poor pharmacokinetics and *in vivo* off-target interactions affecting the therapeutic objective. High-throughput formulation screening techniques like robotic formulation preparation, microfluidics, and automated formulation stability testing can assess the compatibility of a candidate molecule with different formulation approaches and excipients. Finally, high-throughput ADME assays like microsomal stability assays can provide rapid predictions of drug metabolism and clearance.

1.4.2 Computational approaches towards developability characterization

Several predictive tools have been developed to complement the cumbersome and expensive high-throughput experimentation by providing *in silico* quantitative estimates of developability. Antibody informatics tools are used for the prediction of developability issues such as stability, aggregation, and immunogenicity. The core concept behind computational developability assessments is to employ computational methods to discriminate good drug-like lead antibodies from the candidate library. Incorporating developability assessments in early-stage therapeutic development also

provides an opportunity to re-engineer the antibody molecules to mitigate any sequence or structural liabilities using available antibody engineering techniques.[237]

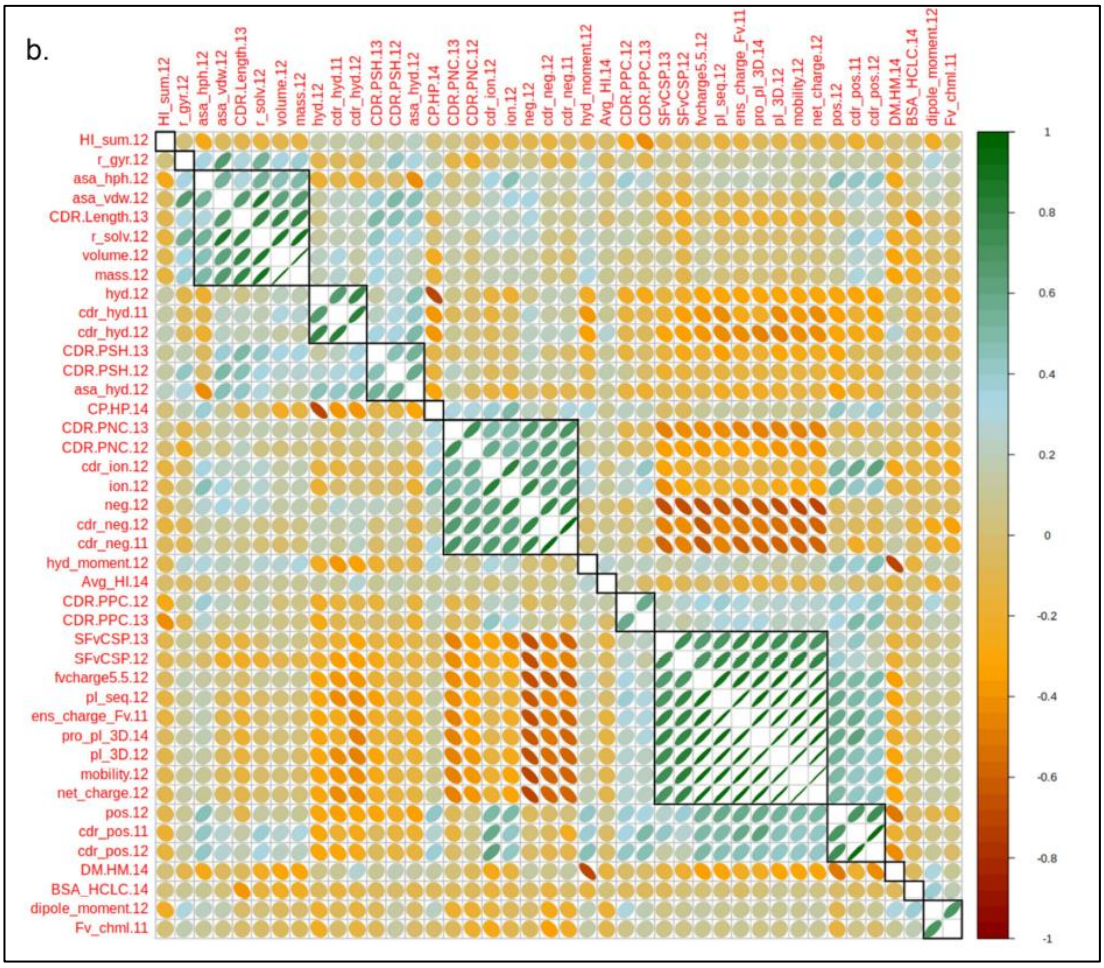


Figure 6: Spearman rank correlation matrix for in silico descriptors. (Adapted from Jain[238])

Typical features evaluated using developability assessments include MHC class II binding immune epitopes, aggregation-prone regions, motifs for glycosylation, motifs for chemical degradation, and presence of non-canonical cysteine residues. Advances in structural prediction of antibody variable regions have also encouraged the use of new physicochemical descriptors based on Fv models such as V_H-V_L interface surface patches, hydrophobic imbalance, charge symmetry, and other electrostatic properties.

Quintero et al. have reviewed the spectrum of features currently reviewed in assessing developability of biologics - cellular assays, preclinical pharmacokinetic assessment, cell line development, manufacturing, and biophysical characterization.[239] A recent study by Jain *et al.* has described major *in vitro* experimental descriptors and *in silico* descriptors to identify developability risks for the clinical progression of antibodies.[238]

They have used hierarchical clustering dendrogram and correlation matrix diagrams to group these descriptors for developability assessment. These are shown in Figure 6.

The general principle in computational developability assessments is to determine assay values for clinical or approved mAbs as reference sets and flag those candidates that lie at the extreme tail regions of these distributions. A key requirement, therefore, is to create reference datasets for approved antibody therapeutics. Several recent efforts have been carried out to address this gap. For instance, Martin *et al.* have previously studied the major product characteristics such as types and molecular formats, formulation, routes of administration, pharmacokinetics (PK), and product presentation of 89 marketed antibody-based biotherapeutics.[240] They demonstrate that a successful antibody drug candidate is most likely to be a humanized IgG1 kappa mAb in a mildly acidic formulation consisting of histidine buffer, sucrose as a stabilizer, and polysorbate 80 as a surfactant. Ahmed *et al.* have also previously studied 24 physicochemical descriptors for the variable regions (Fv) found in 77 marketed antibody-based biotherapeutics.[241] Such datasets serve as good references for optimal properties and especially are very useful to train machine learning models.

1.5 **Research hypothesis and objectives**

This thesis aims to establish a framework for computational developability assessment of therapeutic monoclonal antibodies. The availability of large datasets of antibody biophysical properties enables the search for new predictive models and computational tools for the developability assessment of biologic candidates. In this work, we apply antibody informatics tools for the prediction of developability issues such as stability, aggregation, and immunogenicity for several antibody candidates and platforms.

We believe clinical-stage antibodies serve as a good benchmark of acceptable biophysical properties and developability. So, our research hypothesis is that - **"Computational developability assessment criteria derived from clinical-stage antibodies can be successfully employed to estimate the clinical trial success or failure of antibody therapeutics"**. We hypothesize that a target therapeutic mAb with assay scores exceeding multiple threshold criteria is likely to fail due to adverse events or high immunogenicity caused by the underlying developability liabilities. We also have explored alternate research hypotheses such as: Which developability features can best predict the clinical trial success or development attrition of monoclonal

antibodies? How do different antibody discovery platforms compare in terms of developability profile? What are the possible scope and challenges in the application of biopharmaceutical informatics tools for the prediction of biophysical properties?

The primary objectives of the thesis have been identified as:

- Research and review the current scope and application of biopharmaceutical informatics for antibody therapeutics in industry and academia.
- Establish the developability guidelines based on clinical-stage antibodies and analyze biophysical property distributions of successful antibody therapeutics.
- Compare the biophysical features and developability profile of major antibody discovery platforms. Validate developability assessments with case studies.
- Train a machine learning classification algorithm to predict the clinical trial success outcome based on computational developability assessment results.

The secondary objectives of the thesis to be considered are:

- Evaluate the relevance and prediction accuracy of multiple biopharmaceutical informatics tools. Propose a framework for the sequential use of these available antibody informatics tools for the developability assessment in the triage stage.
- Apply antibody informatics tools to evaluate the dealmaking trends to explore additional applications in the industry beyond developability assessment.

1.6 Introduction to the thesis

Accurate developability assessment of antibody therapeutics remains challenging because many biophysical features influence antibody developability that are not captured by the existing datasets, benchmarks, and experimental methods. Instead, computational approaches to developability assessment are important in probing these several biophysical features in a fast and reliable manner. Successful implementation of computational developability assessment for estimating clinical trial progression remains a non-trivial task. We, therefore, approach this problem by creating a new computational developability assessment framework for antibody therapeutics covering the full spectrum of all mAbs. This thesis presents several approaches to predict and quantify several key developability features using biopharmaceutical informatics tools for traditional mAbs to novel engineered antibody formats, currently

in pre-clinical to advanced clinical or approved stages, and for mAbs originating from various antibody discovery platforms.

Firstly, in **Chapter 1**, biopharmaceutical informatics resources have been reviewed. The databases and tools available for the prediction or assessment of biophysical features relevant to antibody-based drugs have been tabulated and discussed. Applications of currently available biopharmaceutical informatics tools for the assessment of solubility, aggregation, degradation, immunogenicity, post-translational modifications, and *in vivo* safety are discussed in detail. Most importantly, in this chapter, based on this review we have introduced and proposed new guidelines for the design of developability assessment protocols.

A combinatorial triage approach based on the orthogonal combination of conceptually different algorithms to combine scores and rankings from multiple tools is presented that aims to combine scores and rankings from multiple tools together and classify the mAbs based on the aggregate result of all developability prediction tools. The final section summarizes the emerging trends in the use of biopharmaceutical informatics approaches for antibody therapeutics with a future outlook towards new techniques such as machine learning and artificial intelligence for antibody gene repertoires.

In **Chapter 2**, we have shown the detailed methodology and procedures used in this thesis work. This chapter provides insights on the datasets used, data-processing and calculations behind the tools used for developability assessments and data analysis.

In **Chapter 3**, we present developability guidelines based on clinical-stage antibodies that serve as benchmarks for acceptable biophysical properties desired in antibody therapeutics and represent the practical guidelines for antibody drug candidates in preclinical development. We discuss computational developability assessment results for two datasets of clinical-stage antibodies. Their biophysical property distributions are used to decide the assay thresholds to establish the developability criteria.

Next, we have compared the clinical-stage therapeutics and human immune repertoire dataset. A case study on True Human™ antibody therapeutics is also presented to demonstrate the unique developability considerations for natural human antibodies. Overall, the developability benchmarks based on clinical-stage mAbs established in this chapter will inform future developability predictions of preclinical antibodies and other therapeutic proteins. These criteria have been used throughout the next chapters.

In **Chapter 4**, we apply the developability assessment criteria on novel engineered antibodies and other next-generation antibody formats. We compare different antibody discovery platforms and discuss the trends among the several categories of engineered antibodies to capture their differences in biophysical performance. Firstly, we create and curate datasets for three unique antibody discovery platforms namely – bispecific antibodies, phage display antibodies, and transgenic mice antibodies.

Next, we compare the developability profiles of these antibody platform technologies and provide insights on which is the suitable platform technology in each category for a desired application. Overall, this chapter can be a valuable resource to select or filter appropriate platform technology for creating the required engineered antibodies with optimal biophysical properties during antibody drug discovery. This chapter serves as an example of how computational developability assessment can be used to guide the selection of platform technologies for creating next-generation biotherapeutics.

In **Chapter 5**, we have created and implemented various machine learning algorithms to train a prediction algorithm in MATLAB for estimating the clinical trial progression of preclinical antibody candidates. The resulting ROC curves and model performance for machine learning classification algorithms concluded that some other key factors beyond biophysical properties influence the outcome of clinical trials like therapeutic efficacy, adverse events, and clinical trial design.

Chapter 6 is a conclusion and summary of our findings with outlook towards possible future directions from the methods developed and insights gained in this work. We have, in particular, identified the limitations of current computational developability assessments and highlighted the key gaps that need to be addressed for the successful implementation of antibody informatics tools in the development of biologics.

1.7 References

1. Heyman B. Regulation of antibody responses via antibodies, complement, and Fc receptors. Annual review of immunology. 2000;18(1):709-37.
2. Perussia B, Trinchieri G, Jackson A, Warner NL, Faust J, Rumpold H, Kraft D, Lanier LL. The Fc receptor for IgG on human natural killer cells: phenotypic, functional, and comparative studies with monoclonal antibodies. Journal of immunology (Baltimore, Md: 1950). 1984;133(1):180-9.
3. Clynes R, Ravetch JV. Cytotoxic antibodies trigger inflammation through Fc receptors. Immunity. 1995;3(1):21-6.
4. Sarmay G, Lund J, Rozsnyay Z, Gergely J, Jefferis R. Mapping and comparison of the interaction sites on the Fc region of IgG responsible for triggering antibody dependent

cellular cytotoxicity (ADCC) through different types of human Fcγ receptor. Molecular immunology. 1992;29(5):633-9.
5. Dixon KJ, Wu J, Walcheck B. Engineering anti-tumor monoclonal antibodies and Fc receptors to enhance ADCC by human NK cells. Cancers. 2021;13(2):312.
6. Rodgers KR, Chou RC. Therapeutic monoclonal antibodies and derivatives: Historical perspectives and future directions. Biotechnology advances. 2016;34(6):1149-58.
7. Farid SS, Baron M, Stamatis C, Nie W, Coffman J. Benchmarking biopharmaceutical process development and manufacturing cost contributions to R&D. MAbs; 2020.
8. Zahavi D, Weiner L. Monoclonal antibodies in cancer therapy. Antibodies. 2020;9(3):34.
9. Sparrow E, Friede M, Sheikh M, Torvaldsen S. Therapeutic antibodies for infectious diseases. Bulletin of the World Health Organization. 2017;95(3):235.
10. Masroor S, Schroeder TJ, Michler RE, Alexander JW, First MR. Monoclonal antibodies in organ transplantation: an overview. Transplant Immunology. 1994;2(3):176-89.
11. Ozdemir C. Monoclonal antibodies in allergy; updated applications and promising trials. Recent Patents on Inflammation & Allergy Drug Discovery. 2015;9(1):54-65.
12. Kaplon H, Reichert JM. Antibodies to watch in 2021. Mabs; 2021. doi:10.1080/19420862.2020.1860476.
13. Kaplon H, Muralidharan M, Schneider Z, Reichert JM. Antibodies to watch in 2020. MAbs; 2020. doi:10.1080/19420862.2019.1703531.
14. Kumar S, Plotnikov NV, Rouse JC, Singh SK. Biopharmaceutical informatics: supporting biologic drug development via molecular modelling and informatics. Journal of Pharmacy and Pharmacology. 2018;70(5):595-608. doi:10.1111/jphp.12700.
15. Jameson B, Wolf H. The antigenic index: a novel algorithm for predicting antigenic determinants. Bioinformatics. 1988;4(1):181-6. doi:10.1093/bioinformatics/4.1.181.
16. Raybould MI, Marks C, Krawczyk K, Taddese B, Nowak J, Lewis AP, Bujotzek A, Shi J, Deane CM. Five computational developability guidelines for therapeutic antibody profiling. Proceedings of the National Academy of Sciences. 2019;116(10):4025-30. doi:10.1073/pnas.1810576116.
17. Jain T, Sun T, Durand S, Hall A, Houston NR, Nett JH, Sharkey B, Bobrowicz B, Caffry I, Yu Y. Biophysical properties of the clinical-stage antibody landscape. Proceedings of the National Academy of Sciences. 2017;114(5):944-9. doi:10.1073/pnas.1616408114.
18. Obrezanova O, Arnell A, de la Cuesta RG, Berthelot ME, Gallagher TR, Zurdo J, Stallwood Y. Aggregation risk prediction for antibodies and its application to biotherapeutic development. MAbs; 2015. doi:10.1080/19420862.2015.1007828.
19. Krawczyk K, Buchanan A, Marcatili P. Data mining patented antibody sequences. Mabs; 2021. doi:10.1080/19420862.2021.1892366.
20. Krawczyk K, Kelm S, Kovaltsuk A, Galson JD, Kelly D, Trück J, Regep C, Leem J, Wong WK, Nowak J. Structurally mapping antibody repertoires. Frontiers in immunology. 2018;9:1698. doi:10.3389/fimmu.2018.01698.
21. Krawczyk K, Raybould MI, Kovaltsuk A, Deane CM. Looking for therapeutic antibodies in next-generation sequencing repositories. MAbs; 2019. doi:10.1080/19420862.2019.1633884.
22. Deszynski P, Mlokosiewicz J, Volanakis A, Jaszczyszyn I, Castellana N, Bonissone S, Ganesan R, Krawczyk K. INDI-Integrated Nanobody Database for Immunoinformatics. medRxiv. 2021. doi:10.1101/2021.08.04.21261581.
23. Kovaltsuk A, Leem J, Kelm S, Snowden J, Deane CM, Krawczyk K. Observed antibody space: a resource for data mining next-generation sequencing of antibody repertoires. The Journal of Immunology. 2018;201(8):2502-9. doi:10.4049/jimmunol.1800708.

24. Norman RA, Ambrosetti F, Bonvin AM, Colwell LJ, Kelm S, Kumar S, Krawczyk K. Computational approaches to therapeutic antibody design: established methods and emerging trends. Briefings in bioinformatics. 2020;21(5):1549-67. doi:10.1093/bib/bbz095.
25. Raybould MI, Marks C, Lewis AP, Shi J, Bujotzek A, Taddese B, Deane CM. Thera-SAbDab: the therapeutic structural antibody database. Nucleic acids research. 2020;48(D1):D383-D8. doi:10.1093/nar/gkz827.
26. Sirin S, Apgar JR, Bennett EM, Keating AE. AB-Bind: antibody binding mutational database for computational affinity predictions. Protein Science. 2016;25(2):393-409. doi:10.1002/pro.2829.
27. Yadav S, Laue TM, Kalonia DS, Singh SN, Shire SJ. The influence of charge distribution on self-association and viscosity behavior of monoclonal antibody solutions. Molecular pharmaceutics. 2012;9(4):791-802. doi:10.1021/mp200566k.
28. Li L, Kumar S, Buck PM, Burns C, Lavoie J, Singh SK, Warne NW, Nichols P, Luksha N, Boardman D. Concentration dependent viscosity of monoclonal antibody solutions: explaining experimental behavior in terms of molecular properties. Pharmaceutical research. 2014;31(11):3161-78. doi:10.1007/s11095-014-1409-0.
29. Schoch A, Kettenberger H, Mundigl O, Winter G, Engert J, Heinrich J, Emrich T. Charge-mediated influence of the antibody variable domain on FcRn-dependent pharmacokinetics. Proceedings of the National Academy of Sciences. 2015;112(19):5997-6002. doi:10.1073/pnas.1408766112.
30. Marcatili P, Olimpieri PP, Chailyan A, Tramontano A. Antibody modeling using the Prediction of ImmunoGlobulin Structure (PIGS) web server. Nature protocols. 2014;9(12):2771. doi:10.1038/nprot.2014.189.
31. Leem J, Dunbar J, Georges G, Shi J, Deane CM. ABodyBuilder: Automated antibody structure prediction with data–driven accuracy estimation. MAbs; 2016. doi:10.1080/19420862.2016.1205773.
32. Weitzner BD, Jeliazkov JR, Lyskov S, Marze N, Kuroda D, Frick R, Adolf-Bryfogle J, Biswas N, Dunbrack Jr RL, Gray JJ. Modeling and docking of antibody structures with Rosetta. Nature protocols. 2017;12(2):401. doi:10.1038/nprot.2016.180.
33. North B, Lehmann A, Dunbrack Jr RL. A new clustering of antibody CDR loop conformations. Journal of molecular biology. 2011;406(2):228-56. doi:10.1016/j.jmb.2010.10.030.
34. Al-Lazikani B, Lesk AM, Chothia C. Standard conformations for the canonical structures of immunoglobulins. Journal of molecular biology. 1997;273(4):927-48. doi:10.1006/jmbi.1997.1354.
35. Martin AC, Thornton JM. Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. Journal of molecular biology. 1996;263(5):800-15. doi:10.1006/jmbi.1996.0617.
36. Weitzner BD, Dunbrack Jr RL, Gray JJ. The origin of CDR H3 structural diversity. Structure. 2015;23(2):302-11. doi:10.1016/j.str.2014.11.010.
37. Kuroda D, Shirai H, Kobori M, Nakamura H. Structural classification of CDR-H3 revisited: a lesson in antibody modeling. Proteins: Structure, Function, and Bioinformatics. 2008;73(3):608-20. doi:10.1002/prot.22087.
38. Weitzner BD, Kuroda D, Marze N, Xu J, Gray JJ. Blind prediction performance of RosettaAntibody 3.0: grafting, relaxation, kinematic loop modeling, and full CDR optimization. Proteins: Structure, Function, and Bioinformatics. 2014;82(8):1611-23. doi:10.1002/prot.24534.
39. Swindells MB, Porter CT, Couch M, Hurst J, Abhinandan K, Nielsen JH, Macindoe G, Hetherington J, Martin AC. abYsis: integrated antibody sequence and structure—management, analysis, and prediction. Journal of molecular biology. 2017;429(3):356-64. doi:10.1016/j.jmb.2016.08.019.

40. Cheng J, Randall AZ, Sweredoski MJ, Baldi P. SCRATCH: a protein structure and structural feature prediction server. Nucleic acids research. 2005;33(suppl_2):W72-W6. doi:10.1093/nar/gki396.
41. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL. Design of therapeutic proteins with enhanced stability. Proceedings of the National Academy of Sciences. 2009;106(29):11937-42. doi:10.1073/pnas.0904191106.
42. Van Durme J, De Baets G, Van Der Kant R, Ramakers M, Ganesan A, Wilkinson H, Gallardo R, Rousseau F, Schymkowitz J. Solubis: a webserver to reduce protein aggregation through mutation. Protein Engineering, Design and Selection. 2016;29(8):285-9. doi:10.1093/protein/gzw019.
43. Gil-Garcia M, Bano-Polo M, Varejao N, Jamroz M, Kuriata A, Díaz-Caballero M, Lascorz J, Morel B, Navarro S, Reverter D. Combining structural aggregation propensity and stability predictions to redesign protein solubility. Molecular pharmaceutics. 2018;15(9):3846-59. doi:10.1021/acs.molpharmaceut.8b00341.
44. Jain T, Boland T, Lilov A, Burnina I, Brown M, Xu Y, Vásquez M. Prediction of delayed retention of antibodies in hydrophobic interaction chromatography from sequence using machine learning. Bioinformatics. 2017;33(23):3758-66. doi:10.1093/bioinformatics/btx519.
45. Lauer TM, Agrawal NJ, Chennamsetty N, Egodage K, Helk B, Trout BL. Developability index: a rapid in silico tool for the screening of antibody aggregation propensity. Journal of pharmaceutical sciences. 2012;101(1):102-15. doi:10.1002/jps.22758.
46. Jarasch A, Koll H, Regula JT, Bader M, Papadimitriou A, Kettenberger H. Developability assessment during the selection of novel therapeutic antibodies. Journal of pharmaceutical sciences. 2015;104(6):1885-98. doi:10.1002/jps.24430.
47. Kohli N, Jain N, Geddie ML, Razlog M, Xu L, Lugovskoy AA. A novel screening method to assess developability of antibody-like molecules. MAbs; 2015. doi:10.1080/19420862.2015.1048410.
48. Sormanni P, Aprile FA, Vendruscolo M. Third generation antibody discovery methods: in silico rational design. Chemical Society Reviews. 2018;47(24):9137-57. doi:10.1039/C8CS00523K.
49. Kumar S, Singh SK. Developability of Biotherapeutics: computational approaches: CRC Press; 2015.
50. Xu Y, Wang D, Mason B, Rossomando T, Li N, Liu D, Cheung JK, Xu W, Raghava S, Katiyar A. Structure, heterogeneity and developability assessment of therapeutic antibodies. MAbs; 2019. doi:10.1080/19420862.2018.1553476.
51. Hebditch M, Warwicker J. Charge and hydrophobicity are key features in sequence-trained machine learning models for predicting the biophysical properties of clinical-stage antibodies. PeerJ. 2019;7:e8199. doi:10.7717/peerj.8199.
52. King AC, Woods M, Liu W, Lu Z, Gill D, Krebs MR. High-throughput measurement, correlation analysis, and machine-learning predictions for pH and thermal stabilities of Pfizer-generated antibodies. Protein Science. 2011;20(9):1546-57. doi:10.1002/pro.680.
53. Avery LB, Wade J, Wang M, Tam A, King A, Piche-Nicholas N, Kavosi MS, Penn S, Cirelli D, Kurz JC. Establishing in vitro in vivo correlations to screen monoclonal antibodies for physicochemical properties related to favorable human pharmacokinetics. MAbs; 2018. doi:10.1080/19420862.2017.1417718.
54. Bailly M, Mieczkowski C, Juan V, Metwally E, Tomazela D, Baker J, Uchida M, Kofman E, Raoufi F, Motlagh S. Predicting antibody developability profiles through early stage discovery screening. MAbs; 2020. doi:10.1080/19420862.2020.1743053.
55. Erasmus MF, D'Angelo S, Ferrara F, Naranjo L, Teixeira AA, Buonpane R, Stewart SM, Nastri HG, Bradbury AR. A single donor is sufficient to produce a highly functional in

vitro antibody library. Communications biology. 2021;4(1):1-16. doi:10.1038/s42003-021-01881-0.
56. Tiller T, Schuster I, Deppe D, Siegers K, Strohner R, Herrmann T, Berenguer M, Poujol D, Stehle J, Stark Y. A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. MAbs; 2013. doi:10.4161/mabs.24218.
57. Adler AS, Bedinger D, Adams MS, Asensio MA, Edgar RC, Leong R, Leong J, Mizrahi RA, Spindler MJ, Bandi SR. A natively paired antibody library yields drug leads with higher sensitivity and specificity than a randomly paired antibody library. MAbs; 2018. doi:10.1080/19420862.2018.1426422.
58. Raybould MI, Marks C, Kovaltsuk A, Lewis AP, Shi J, Deane CM. Public Baseline and shared response structures support the theory of antibody repertoire functional commonality. PLoS computational biology. 2021;17(3):e1008781. doi:10.1371/journal.pcbi.1008781.
59. Adams JJ, Sidhu SS. Synthetic antibody technologies. Current opinion in structural biology. 2014;24:1-9. doi:10.1016/j.sbi.2013.11.003.
60. Prassler J, Thiel S, Pracht C, Polzer A, Peters S, Bauer M, Nörenberg S, Stark Y, Kölln J, Popp A. HuCAL PLATINUM, a synthetic Fab library optimized for sequence diversity and superior performance in mammalian expression systems. Journal of molecular biology. 2011;413(1):261-78. doi:10.1016/j.jmb.2011.08.012.
61. Zhai W, Glanville J, Fuhrmann M, Mei L, Ni I, Sundar PD, Van Blarcom T, Abdiche Y, Lindquist K, Strohner R. Synthetic antibodies designed on natural sequence landscapes. Journal of molecular biology. 2011;412(1):55-71. doi:10.1016/j.jmb.2011.07.018.
62. Zhao Q, Buhr D, Gunter C, Frenette J, Ferguson M, Sanford E, Holland E, Rajagopal C, Batonick M, Kiss MM. Rational library design by functional CDR resampling. New biotechnology. 2018;45:89-97. doi:10.1016/j.nbt.2017.12.005.
63. Friedensohn S, Neumeier D, Khan TA, Csepregi L, Parola C, de Vries ARG, Erlach L, Mason DM, Reddy ST. Convergent selection in antibody repertoires is revealed by deep learning. bioRxiv. 2020. doi:10.1101/2020.02.25.965673.
64. Amimeur T, Shaver JM, Ketchem RR, Taylor JA, Clark RH, Smith J, Van Citters D, Siska CC, Smidt P, Sprague M. Designing feature-controlled humanoid antibody discovery libraries using generative adversarial networks. bioRxiv. 2020. doi:10.1101/2020.04.12.024844.
65. Repecka D, Jauniskis V, Karpus L, Rembeza E, Rokaitis I, Zrimec J, Poviloniene S, Laurynenas A, Viknander S, Abuajwa W. Expanding functional protein sequence spaces using generative adversarial networks. Nature Machine Intelligence. 2021;3(4):324-33. doi:10.1038/s42256-021-00310-5.
66. Shin J-E, Riesselman AJ, Kollasch AW, McMahon C, Simon E, Sander C, Manglik A, Kruse AC, Marks DS. Protein design and variant prediction using autoregressive generative models. Nature communications. 2021;12(1):1-11. doi:10.1038/s41467-021-22732-w.
67. Dyson MR, Masters E, Pazeraitis D, Perera RL, Syrjanen JL, Surade S, Thorsteinson N, Parthiban K, Jones PC, Sattar M. Beyond affinity: selection of antibody variants with optimal biophysical properties and reduced immunogenicity from mammalian display libraries. Mabs; 2020. doi:10.1080/19420862.2020.1829335.
68. Wu Z, Kan SJ, Lewis RD, Wittmann BJ, Arnold FH. Machine learning-assisted directed protein evolution with combinatorial libraries. Proceedings of the National Academy of Sciences. 2019;116(18):8852-8. doi:10.1073/pnas.1901979116.
69. Chen X, Dougherty T, Hong C, Schibler R, Zhao YC, Sadeghi R, Matasci N, Wu Y-C, Kerman I. Predicting antibody developability from sequence using machine learning. bioRxiv. 2020. doi:10.1101/2020.06.18.159798.

70. Starr CG, Tessier PM. Selecting and engineering monoclonal antibodies with drug-like specificity. Current opinion in biotechnology. 2019;60:119-27. doi:10.1016/j.copbio.2019.01.008.
71. Vázquez-Rey M, Lang DA. Aggregates in monoclonal antibody manufacturing processes. Biotechnology and bioengineering. 2011;108(7):1494-508. doi:10.1002/bit.23155.
72. Ratanji KD, Derrick JP, Dearman RJ, Kimber I. Immunogenicity of therapeutic proteins: influence of aggregation. Journal of immunotoxicology. 2014;11(2):99-109. doi:10.3109/1547691X.2013.821564.
73. Seeliger D, Schulz P, Litzenburger T, Spitz J, Hoerer S, Blech M, Enenkel B, Studts JM, Garidel P, Karow AR. Boosting antibody developability through rational sequence optimization. MAbs; 2015. doi:10.1080/19420862.2015.1017695.
74. van der Kant R, Karow-Zwick AR, Van Durme J, Blech M, Gallardo R, Seeliger D, Aßfalg K, Baatsen P, Compernolle G, Gils A. Prediction and reduction of the aggregation of monoclonal antibodies. Journal of molecular biology. 2017;429(8):1244-61. doi:10.1016/j.jmb.2017.03.014.
75. Kuhn AB, Kube S, Karow-Zwick AR, Seeliger D, Garidel P, Blech M, Schäfer LV. Improved solution-state properties of monoclonal antibodies by targeted mutations. The Journal of Physical Chemistry B. 2017;121(48):10818-27. doi:10.1021/acs.jpcb.7b09126.
76. Casaz P, Boucher E, Wollacott R, Pierce BG, Rivera R, Sedic M, Ozturk S, Thomas Jr WD, Wang Y. Resolving self-association of a therapeutic antibody by formulation optimization and molecular approaches. MAbs; 2014. doi:10.4161/19420862.2014.975658.
77. Wu S-J, Luo J, O'Neil KT, Kang J, Lacy ER, Canziani G, Baker A, Huang M, Tang QM, Raju TS. Structure-based engineering of a monoclonal antibody for improved solubility. Protein Engineering, Design & Selection. 2010;23(8):643-51. doi:10.1093/protein/gzq037.
78. Nichols P, Li L, Kumar S, Buck PM, Singh SK, Goswami S, Balthazor B, Conley TR, Sek D, Allen MJ. Rational design of viscosity reducing mutants of a monoclonal antibody: hydrophobic versus electrostatic inter-molecular interactions. MAbs; 2015. doi:10.4161/19420862.2014.985504.
79. Pindrus M, Shire SJ, Kelley RF, Demeule Bl, Wong R, Xu Y, Yadav S. Solubility challenges in high concentration monoclonal antibody formulations: relationship with amino acid sequence and intermolecular interactions. Molecular pharmaceutics. 2015;12(11):3896-907. doi:10.1021/acs.molpharmaceut.5b00336.
80. Yadav S, Sreedhara A, Kanai S, Liu J, Lien S, Lowman H, Kalonia DS, Shire SJ. Establishing a link between amino acid sequences and self-associating and viscoelastic behavior of two closely related monoclonal antibodies. Pharmaceutical research. 2011;28(7):1750-64. doi:10.1007/s11095-011-0410-0.
81. Sankar K, Krystek Jr SR, Carl SM, Day T, Maier JK. AggScore: Prediction of aggregation-prone regions in proteins based on the distribution of surface patches. Proteins: Structure, Function, and Bioinformatics. 2018;86(11):1147-56. doi:10.1002/prot.25594.
82. Liu YD, Goetze AM, Bass RB, Flynn GC. N-terminal glutamate to pyroglutamate conversion in vivo for human IgG2 antibodies. Journal of Biological Chemistry. 2011;286(13):11211-7. doi:10.1074/jbc.M110.185041.
83. Chelius D, Jing K, Lueras A, Rehder DS, Dillon TM, Vizel A, Rajan RS, Li T, Treuheit MJ, Bondarenko PV. Formation of pyroglutamic acid from N-terminal glutamic acid in immunoglobulin gamma antibodies. Analytical chemistry. 2006;78(7):2370-6. doi:10.1021/ac051827k.
84. Yu L, Vizel A, Huff MB, Young M, Remmele Jr RL, He B. Investigation of N-terminal glutamate cyclization of recombinant monoclonal antibody in formulation development.

Journal of pharmaceutical and biomedical analysis. 2006;42(4):455-63. doi:10.1016/j.jpba.2006.05.008.
85. Tang L, Sundaram S, Zhang J, Carlson P, Matathia A, Parekh B, Zhou Q, Hsieh M-C. Conformational characterization of the charge variants of a human IgG1 monoclonal antibody using H/D exchange mass spectrometry. MAbs; 2013. doi:10.4161/mabs.22695.
86. van den Bremer ET, Beurskens FJ, Voorhorst M, Engelberts PJ, de Jong RN, van der Boom BG, Cook EM, Lindorfer MA, Taylor RP, van Berkel PH. Human IgG is produced in a pro-form that requires clipping of C-terminal lysines for maximal complement activation. MAbs; 2015. doi:10.1080/19420862.2015.1046665.
87. Liu H, Nowak C, Shao M, Ponniah G, Neill A. Impact of cell culture on recombinant monoclonal antibody product heterogeneity. Biotechnology progress. 2016;32(5):1103-12. doi:10.1002/btpr.2327.
88. Füssl F, Trappe A, Cook K, Scheffler K, Fitzgerald O, Bones J. Comprehensive characterisation of the heterogeneity of adalimumab via charge variant analysis hyphenated on-line to native high resolution Orbitrap mass spectrometry. MAbs; 2019. doi:10.1080/19420862.2018.1531664.
89. Sydow JF, Lipsmeier F, Larraillet V, Hilger M, Mautz B, Mølhøj M, Kuentzer J, Klostermann S, Schoch J, Voelger HR. Structure-based prediction of asparagine and aspartate degradation sites in antibody variable regions. PloS one. 2014;9(6):e100736. doi:10.1371/journal.pone.0100736.
90. Verma A, Ngundi MM, Burns DL. Mechanistic analysis of the effect of deamidation on the immunogenicity of anthrax protective antigen. Clinical and vaccine immunology: CVI. 2016;23(5):396. doi:10.1128/CVI.00701-15.
91. Mo J, Yan Q, So CK, Soden T, Lewis MJ, Hu P. Understanding the impact of methionine oxidation on the biological functions of IgG1 antibodies using hydrogen/deuterium exchange mass spectrometry. Analytical chemistry. 2016;88(19):9495-502. doi:10.1021/acs.analchem.6b01958.
92. Dashivets T, Stracke J, Dengl S, Knaupp A, Pollmann J, Buchner J, Schlothauer T. Oxidation in the complementarity-determining regions differentially influences the properties of therapeutic antibodies. MAbs; 2016. doi:10.1080/19420862.2016.1231277.
93. Liu-Shin LP-Y, Fung A, Malhotra A, Ratnaswamy G. Evidence of disulfide bond scrambling during production of an antibody-drug conjugate. MAbs; 2018. doi:10.1080/19420862.2018.1521128.
94. Moritz B, Stracke JO. Assessment of disulfide and hinge modifications in monoclonal antibodies. Electrophoresis. 2017;38(6):769-85. doi:10.1002/elps.201600425.
95. Wright A, Tao M, Kabat E, Morrison S. Antibody variable region glycosylation: position effects on antigen binding and carbohydrate structure. The EMBO journal. 1991;10(10):2717-23. doi:10.1002/j.1460-2075.1991.tb07819.x.
96. Leibiger H, WÜSTNER D, STIGLER R-D, MARX U. Variable domain-linked oligosaccharides of a human monoclonal IgG: structure and influence on antigen binding. Biochemical Journal. 1999;338(2):529-38. doi:10.1042/bj3380529.
97. van Bueren JJL, Rispens T, Verploegen S, van der Palen-Merkus T, Stapel S, Workman LJ, James H, van Berkel PH, van de Winkel JG, Platts-Mills TA. Anti-galactose-α-1, 3-galactose IgE from allergic patients does not bind α-galactosylated glycans on intact therapeutic antibody Fc domains. Nature biotechnology. 2011;29(7):574-6. doi:10.1038/nbt.1912.
98. Jefferis R. Posttranslational modifications and the immunogenicity of biotherapeutics. Journal of immunology research. 2016;2016. doi:10.1155/2016/5358272.
99. Muster W, Breidenbach A, Fischer H, Kirchner S, Müller L, Pähler A. Computational toxicology in drug development. Drug discovery today. 2008;13(7-8):303-10. doi:10.1016/j.drudis.2007.12.007.

100. Kuang Q, Wang M, Li R, Dong Y, Li Y, Li M. A systematic investigation of computation models for predicting Adverse Drug Reactions (ADRs). PloS one. 2014;9(9):e105889. doi:10.1371/journal.pone.0105889.
101. Bryson CJ, Jones TD, Baker MP. Prediction of immunogenicity of therapeutic proteins. BioDrugs. 2010;24(1):1-8. doi:10.2165/11318560-000000000-00000.
102. Baker M, Reynolds HM, Lumicisi B, Bryson CJ. Immunogenicity of protein therapeutics: The key causes, consequences and challenges. Self/nonself. 2010;1(4):314-22. doi:10.4161/self.1.4.13904.
103. Diao L, Meibohm B. Tools for predicting the PK/PD of therapeutic proteins. Expert opinion on drug metabolism & toxicology. 2015;11(7):1115-25. doi:10.1517/17425255.2015.1041917.
104. Wang J, Iyer S, Fielder PJ, Davis JD, Deng R. Projecting human pharmacokinetics of monoclonal antibodies from nonclinical data: comparative evaluation of prediction approaches in early drug development. Biopharmaceutics & drug disposition. 2016;37(2):51-65. doi:10.1002/bdd.1952.
105. Grinshpun B, Thorsteinson N, Pereira JN, Rippmann F, Nannemann D, Sood VD, Fomekong Nanfack Y. Identifying biophysical assays and in silico properties that enrich for slow clearance in clinical-stage therapeutic antibodies. Mabs; 2021. doi:10.1080/19420862.2021.1932230.
106. Desai DV, Kulkarni-Kale U. T-cell epitope prediction methods: an overview. Immunoinformatics. 2014:333-64. doi:10.1007/978-1-4939-1115-8_19.
107. Peters B, Nielsen M, Sette A. T cell epitope predictions. Annual review of immunology. 2020;38:123-45. doi:10.1146/annurev-immunol-082119-124838.
108. Reche PA, Glutting J-P, Zhang H, Reinherz EL. Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. Immunogenetics. 2004;56(6):405-19. doi:10.1007/s00251-004-0709-7.
109. Singh H, Raghava G. ProPred: prediction of HLA-DR binding sites. Bioinformatics. 2001;17(12):1236-7. doi:10.1093/bioinformatics/17.12.1236.
110. Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, Braxenthaler M, Gallazzi F, Protti MP, Sinigaglia F. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nature biotechnology. 1999;17(6):555-61. doi:10.1038/9858.
111. Nielsen M, Lundegaard C, Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC bioinformatics. 2007;8(1):1-12. doi:10.1186/1471-2105-8-238.
112. Wang P, Sidney J, Dow C, Mothé B, Sette A, Peters B. A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. PLoS Comput Biol. 2008;4(4):e1000048. doi:10.1371/journal.pcbi.1000048.
113. Zhang GL, Khan AM, Srinivasan KN, August JT, Brusic V. MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. Nucleic acids research. 2005;33(suppl_2):W172-W9. doi:10.1093/nar/gki452.
114. Andreatta M, Trolle T, Yan Z, Greenbaum JA, Peters B, Nielsen M. An automated benchmarking platform for MHC class II binding prediction methods. Bioinformatics. 2018;34(9):1522-8. doi:10.1093/bioinformatics/btx820.
115. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic acids research. 2020;48(W1):W449-W54. doi:10.1093/nar/gkaa379.
116. Nielsen M, Lundegaard C, Blicher T, Peters B, Sette A, Justesen S, Buus S, Lund O. Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. PLoS Comput Biol. 2008;4(7):e1000107. doi:10.1371/journal.pcbi.1000107.

117. Jacob L, Vert J-P. Efficient peptide–MHC-I binding prediction for alleles with few known binders. Bioinformatics. 2008;24(3):358-66. doi:10.1093/bioinformatics/btm611.
118. Zhang H, Lundegaard C, Nielsen M. Pan-specific MHC class I predictors: a benchmark of HLA class I pan-specific prediction methods. Bioinformatics. 2009;25(1):83-9. doi:10.1093/bioinformatics/btn579.
119. Yachnin BJ, Mulligan VK, Khare SD, Bailey-Kellogg C. MHCEpitopeEnergy, a Flexible Rosetta-Based Biotherapeutic Deimmunization Platform. Journal of Chemical Information and Modeling. 2021;61(5):2368-82. doi:10.1021/acs.jcim.1c00056.
120. Peng H-P, Lee KH, Jian J-W, Yang A-S. Origins of specificity and affinity in antibody–protein interactions. Proceedings of the National Academy of Sciences. 2014;111(26):E2656-E65. doi:10.1073/pnas.1401131111.
121. Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. Proceedings of the National Academy of Sciences. 1981;78(6):3824-8. doi:10.1073/pnas.78.6.3824.
122. Welling GW, Weijer WJ, van der Zee R, Welling-Wester S. Prediction of sequential antigenic regions in proteins. FEBS letters. 1985;188(2):215-8. doi:10.1016/0014-5793(85)80374-4.
123. Van Regenmortel M, De Marcillac GD. An assessment of prediction methods for locating continuous epitopes in proteins. Immunology letters. 1988;17(2):95-107. doi:10.1016/0165-2478(88)90076-4.
124. Pellequer J, Westhof E. PREDITOP: a program for antigenicity prediction. Journal of molecular graphics. 1993;11(3):204-10. doi:10.1016/0263-7855(93)80074-2.
125. Saha S, Raghava GPS. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins: Structure, Function, and Bioinformatics. 2006;65(1):40-8. doi:10.1002/prot.21078.
126. Gao J, Faraggi E, Zhou Y, Ruan J, Kurgan L. BEST: improved prediction of B-cell epitopes from antigen sequences. PloS one. 2012;7(6):e40104. doi:10.1371/journal.pone.0040104.
127. Singh H, Ansari HR, Raghava GP. Improved method for linear B-cell epitope prediction using antigen's primary sequence. PloS one. 2013;8(5):e62216. doi:10.1371/journal.pone.0062216.
128. Yao B, Zhang L, Liang S, Zhang C. SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity. PloS one. 2012;7(9):e45152. doi:10.1371/journal.pone.0045152.
129. Kulkarni-Kale U, Bhosle S, Kolaskar AS. CEP: a conformational epitope prediction server. Nucleic acids research. 2005;33(suppl_2):W168-W71. doi:10.1093/nar/gki460.
130. Zinsli LV, Stierlin N, Loessner MJ, Schmelcher M. Deimmunization of protein therapeutics–Recent advances in experimental and computational epitope prediction and deletion. Computational and Structural Biotechnology Journal. 2020. doi:10.1016/j.csbj.2020.12.024.
131. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced drug delivery reviews. 1997;23(1-3):3-25. doi:10.1016/S0169-409X(96)00423-1.
132. Lecerf M, Kanyavuz A, Lacroix-Desmazes S, Dimitrov JD. Sequence features of variable region determining physicochemical properties and polyreactivity of therapeutic antibodies. Molecular immunology. 2019;112:338-46. doi:10.1016/j.molimm.2019.06.012.
133. Hebditch M, Carballo-Amador MA, Charonis S, Curtis R, Warwicker J. Protein–Sol: a web tool for predicting protein solubility from sequence. Bioinformatics. 2017;33(19):3098-100. doi:10.1093/bioinformatics/btx345.

134. Hou Q, Kwasigroch JM, Rooman M, Pucci F. SOLart: a structure-based method to predict protein solubility and aggregation. Bioinformatics. 2020;36(5):1445-52. doi:10.1093/bioinformatics/btz773.
135. Sormanni P, Aprile FA, Vendruscolo M. The CamSol method of rational design of protein mutants with enhanced solubility. Journal of molecular biology. 2015;427(2):478-90. doi:10.1016/j.jmb.2014.09.026.
136. Sormanni P, Amery L, Ekizoglou S, Vendruscolo M, Popovic B. Rapid and accurate in silico solubility screening of a monoclonal antibody library. Scientific reports. 2017;7(1):1-9. doi:10.1038/s41598-017-07800-w.
137. Shan L, Mody N, Sormani P, Rosenthal KL, Damschroder MM, Esfandiary R. Developability assessment of engineered monoclonal antibody variants with a complex self-association behavior using complementary analytical and in silico tools. Molecular pharmaceutics. 2018;15(12):5697-710. doi:10.1021/acs.molpharmaceut.8b00867.
138. Wolf Pérez A-M, Sormanni P, Andersen JS, Sakhnini LI, Rodriguez-Leon I, Bjelke JR, Gajhede AJ, De Maria L, Otzen DE, Vendruscolo M. In vitro and in silico assessment of the developability of a designed monoclonal antibody library. MAbs; 2019. doi:10.1080/19420862.2018.1556082.
139. Kingsbury JS, Saini A, Auclair SM, Fu L, Lantz MM, Halloran KT, Calero-Rubio C, Schwenger W, Airiau CY, Zhang J. A single molecular descriptor to predict solution behavior of therapeutic antibodies. Science advances. 2020;6(32):eabb0372. doi:10.1126/sciadv.abb0372.
140. Hanke AT, Klijn ME, Verhaert PD, van der Wielen LA, Ottens M, Eppink MH, van de Sandt EJ. Prediction of protein retention times in hydrophobic interaction chromatography by robust statistical characterization of their atomic-level surface properties. Biotechnology progress. 2016;32(2):372-81. doi:10.1002/btpr.2219.
141. Goyon A, D'Atri V, Colas O, Fekete S, Beck A, Guillarme D. Characterization of 30 therapeutic antibodies and related products by size exclusion chromatography: Feasibility assessment for future mass spectrometry hyphenation. Journal of Chromatography B. 2017;1065:35-43. doi:10.1016/j.jchromb.2017.09.027.
142. Hebditch M, Roche A, Curtis RA, Warwicker J. Models for antibody behavior in hydrophobic interaction chromatography and in self-association. Journal of pharmaceutical sciences. 2019;108(4):1434-41. doi:10.1016/j.xphs.2018.11.035.
143. Sundberg EJ, Urrutia M, Braden BC, Isern J, Tsuchiya D, Fields BA, Malchiodi EL, Tormo J, Schwarz FP, Mariuzza RA. Estimation of the hydrophobic effect in an antigen− antibody protein− protein interface. Biochemistry. 2000;39(50):15375-87. doi:10.1021/bi000704l.
144. Almagro JC, Fransson J. Humanization of antibodies. Front Biosci. 2008;13(1):1619-33. doi:10.2741/2786.
145. Roguska MA, Pedersen JT, Keddy CA, Henry AH, Searle SJ, Lambert JM, Goldmacher VS, Blättler W, Rees AR, Guild BC. Humanization of murine monoclonal antibodies through variable domain resurfacing. Proceedings of the National Academy of Sciences. 1994;91(3):969-73. doi:10.1073/pnas.91.3.969.
146. Bartelds GM, Krieckaert CL, Nurmohamed MT, van Schouwenburg PA, Lems WF, Twisk JW, Dijkmans BA, Aarden L, Wolbink GJ. Development of antidrug antibodies against adalimumab and association with disease activity and treatment failure during long-term follow-up. Jama. 2011;305(14):1460-8. doi:10.1001/jama.2011.406.
147. Sekiguchi N, Kubo C, Takahashi A, Muraoka K, Takeiri A, Ito S, Yano M, Mimoto F, Maeda A, Iwayanagi Y. MHC-associated peptide proteomics enabling highly sensitive detection of immunogenic sequences for the development of therapeutic antibodies with low immunogenicity. MAbs; 2018. doi:10.1080/19420862.2018.1518888.

148. Karle AC. Applying MAPPs assays to assess drug immunogenicity. Frontiers in immunology. 2020;11:698. doi:10.3389/fimmu.2020.00698.
149. Abhinandan K, Martin AC. Analyzing the “degree of humanness” of antibody sequences. Journal of molecular biology. 2007;369(3):852-62. doi:10.1016/j.jmb.2007.02.100.
150. Gao SH, Huang K, Tu H, Adler AS. Monoclonal antibody humanness score and its applications. BMC biotechnology. 2013;13(1):1-12. doi:10.1186/1472-6750-13-55.
151. Thullier P, Huish O, Pelat T, Martin AC. The humanness of macaque antibody sequences. Journal of molecular biology. 2010;396(5):1439-50. doi:10.1016/j.jmb.2009.12.041.
152. Schmitz S, Soto C, Crowe Jr JE, Meiler J. Human-likeness of antibody biologics determined by back-translation and comparison with large antibody variable gene repertoires. Mabs; 2020. doi:10.1080/19420862.2020.1758291.
153. Chin M, Marks C, Deane CM. Humanization of antibodies using a machine learning approach on large-scale repertoire data. bioRxiv. 2021. doi:10.1093/bioinformatics/btab434.
154. Chailyan A, Tramontano A, Marcatili P. A database of immunoglobulins with integrated tools: DIGIT. Nucleic acids research. 2012;40(D1):D1230-D4. doi:10.1093/nar/gkr806.
155. Christley S, Scarborough W, Salinas E, Rounds WH, Toby IT, Fonner JM, Levin MK, Kim M, Mock SA, Jordan C. VDJServer: a cloud-based analysis portal and data commons for immune repertoire sequences and rearrangements. Frontiers in immunology. 2018;9:976. doi:10.3389/fimmu.2018.00976.
156. Corrie BD, Marthandan N, Zimonja B, Jaglale J, Zhou Y, Barr E, Knoetze N, Breden FM, Christley S, Scott JK. iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. Immunological reviews. 2018;284(1):24-41. doi:10.1111/imr.12666.
157. DeWitt WS, Lindau P, Snyder TM, Sherwood AM, Vignali M, Carlson CS, Greenberg PD, Duerkopp N, Emerson RO, Robins HS. A public database of memory and naive B-cell receptor sequences. PloS one. 2016;11(8):e0160853. doi:10.1371/journal.pone.0160853.
158. Rosenfeld AM, Meng W, Luning Prak ET, Hershberg U. Immunedb, a novel tool for the analysis, storage, and dissemination of immune repertoire sequencing data. Frontiers in immunology. 2018;9:2107. doi:10.3389/fimmu.2018.02107.
159. Zhang W, Wang L, Liu K, Wei X, Yang K, Du W, Wang S, Guo N, Ma C, Luo L. PIRD: Pan immune repertoire database. Bioinformatics. 2020;36(3):897-903. doi:10.1093/bioinformatics/btz614.
160. Marks C, Deane CM. How repertoire data are changing antibody science. Journal of Biological Chemistry. 2020;295(29):9823-37. doi:10.1074/jbc.REV120.010181.
161. Galson JD, Clutterbuck EA, Trück J, Ramasamy MN, Münz M, Fowler A, Cerundolo V, Pollard AJ, Lunter G, Kelly DF. BCR repertoire sequencing: different patterns of B-cell activation after two Meningococcal vaccines. Immunology and cell biology. 2015;93(10):885-95. doi:10.1038/icb.2015.57.
162. Miho E, Yermanos A, Weber CR, Berger CT, Reddy ST, Greiff V. Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires. Frontiers in immunology. 2018;9:224. doi:10.3389/fimmu.2018.00224.
163. Briney B, Inderbitzin A, Joyce C, Burton DR. Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature. 2019;566(7744):393-7. doi:10.1038/s41586-019-0879-y.
164. Slabodkin A, Chernigovskaya M, Mikocziova I, Akbar R, Scheffer L, Pavlović M, Bashour H, Snapkov I, Mehta BB, Weber CR. Individualized VDJ recombination predisposes the available Ig sequence space. bioRxiv. 2021. doi:10.1101/2021.04.19.440409.

165. Bashford-Rogers RJ, Smith KG, Thomas DC. Antibody repertoire analysis in polygenic autoimmune diseases. Immunology. 2018;155(1):3-17. doi:10.1111/imm.12927.
166. Liu J, Yang X, Lu X, Zhang L, Luo W, Cheng Y, Zhang L, Yang Y, Dai N, Xu Y. Impact of T-cell receptor and B-cell receptor repertoire on the recurrence of early stage lung adenocarcinoma. Experimental Cell Research. 2020;394(2):112134. doi:10.1016/j.yexcr.2020.112134.
167. Galson JD, Pollard AJ, Trück J, Kelly DF. Studying the antibody repertoire after vaccination: practical applications. Trends in immunology. 2014;35(7):319-31. doi:10.1016/j.it.2014.04.005.
168. Galson JD, Schaetzle S, Bashford-Rogers RJ, Raybould MI, Kovaltsuk A, Kilpatrick GJ, Minter R, Finch DK, Dias J, James LK. Deep sequencing of B cell receptor repertoires from COVID-19 patients reveals strong convergent immune signatures. Frontiers in immunology. 2020;11:3283. doi:10.3389/fimmu.2020.605170.
169. Richardson E, Galson JD, Kellam P, Kelly DF, Smith SE, Palser A, Watson S, Deane CM. A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-Pertussis toxoid antibodies. Mabs; 2021. doi:10.1080/19420862.2020.1869406.
170. Wong WK, Robinson SA, Bujotzek A, Georges G, Lewis AP, Shi J, Snowden J, Taddese B, Deane CM. Ab-Ligity: Identifying sequence-dissimilar antibodies that bind to the same epitope. mAbs; 2021. doi:10.1080/19420862.2021.1873478.
171. Schritt D, Li S, Rozewicki J, Katoh K, Yamashita K, Volkmuth W, Cavet G, Standley DM. Repertoire Builder: high-throughput structural modeling of B and T cell receptors. Molecular Systems Design & Engineering. 2019;4(4):761-8. doi:10.1039/C9ME00020H.
172. DeKosky BJ, Lungu OI, Park D, Johnson EL, Charab W, Chrysostomou C, Kuroda D, Ellington AD, Ippolito GC, Gray JJ. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. Proceedings of the National Academy of Sciences. 2016;113(19):E2636-E45. doi:10.1073/pnas.1525510113.
173. Raybould MI, Marks C, Kovaltsuk A, Lewis AP, Shi J, Deane C. Evidence of antibody repertoire functional convergence through public baseline and shared response structures. BioRxiv. 2020. doi:10.1101/2020.03.17.993444.
174. Robinson SA, Raybould MI, Marks C, Schneider C, Wong WK, Deane CM. Epitope profiling of coronavirus-binding antibodies using computational structural modelling. bioRxiv. 2021. doi:10.1101/2021.04.12.439478.
175. Kovaltsuk A, Raybould MI, Wong WK, Marks C, Kelm S, Snowden J, Trück J, Deane CM. Structural diversity of B-cell receptor repertoires along the B-cell differentiation axis in humans and mice. PLoS computational biology. 2020;16(2):e1007636. doi:10.1371/journal.pcbi.1007636.
176. Setliff I, Shiakolas AR, Pilewski KA, Murji AA, Mapengo RE, Janowska K, Richardson S, Oosthuysen C, Raju N, Ronsard L. High-throughput mapping of B cell receptor sequences to antigen specificity. Cell. 2019;179(7):1636-46. e15. doi:10.1016/j.cell.2019.11.003.
177. Carter PJ, Lazar GA. Next generation antibody drugs: pursuit of the'high-hanging fruit'. Nature Reviews Drug Discovery. 2018;17(3):197. doi:10.1038/nrd.2017.227.
178. Beck A, D'atri V, Ehkirch A, Fekete S, Hernandez-Alba O, Gahoual R, Leize-Wagner E, François Y, Guillarme D, Cianférani S. Cutting-edge multi-level analytical and structural characterization of antibody-drug conjugates: present and future. Expert review of proteomics. 2019;16(4):337-62. doi:10.1080/14789450.2019.1578215.
179. Khongorzul P, Ling CJ, Khan FU, Ihsan AU, Zhang J. Antibody–drug conjugates: a comprehensive review. Molecular Cancer Research. 2020;18(1):3-19. doi:10.1158/1541-7786.MCR-19-0582.

180. Leung D, Wurst JM, Liu T, Martinez RM, Datta-Mannan A, Feng Y. Antibody conjugates-recent advances and future innovations. Antibodies. 2020;9(1):2. doi:10.3390/antib9010002.
181. Wilton EE, Opyr MP, Kailasam S, Kothe RF, Wieden H-J. sdAb-DB: the single domain antibody database. 2018. doi:10.1021/acssynbio.8b00407.
182. Zavrtanik U, Hadži S. A non-redundant data set of nanobody-antigen crystal structures. Data in brief. 2019;24:103754. doi:10.1016/j.dib.2019.103754.
183. Zuo J, Li J, Zhang R, Xu L, Chen H, Jia X, Su Z, Zhao L, Huang X, Xie W. Institute collection and analysis of Nanobodies (iCAN): a comprehensive database and analysis platform for nanobodies. BMC genomics. 2017;18(1):1-5. doi:10.1186/s12864-017-4204-6.
184. Gentiluomo L, Svilenov HL, Augustijn D, El Bialy I, Greco ML, Kulakova A, Indrakumar S, Mahapatra S, Morales MM, Pohl C. Advancing therapeutic protein discovery and development through comprehensive computational and biophysical characterization. Molecular pharmaceutics. 2019;17(2):426-40. doi:10.1021/acs.molpharmaceut.9b00852.
185. Pollastri G, Przybylski D, Rost B, Baldi P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins: Structure, Function, and Bioinformatics. 2002;47(2):228-35. doi:10.1002/prot.10082.
186. Magnan CN, Baldi P. SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. Bioinformatics. 2014;30(18):2592-7. doi:10.1093/bioinformatics/btu352.
187. Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks–based regression. Proteins: Structure, Function, and Bioinformatics. 2004;56(4):753-67. doi:10.1002/prot.20176.
188. Yuan Z, Burrage K, Mattick JS. Prediction of protein solvent accessibility using support vector machines. Proteins: Structure, Function, and Bioinformatics. 2002;48(3):566-70. doi:10.1002/prot.10176.
189. Nepal R, Spencer J, Bhogal G, Nedunuri A, Poelman T, Kamath T, Chung E, Kantardjieff K, Gottlieb A, Lustig B. Logistic regression models to predict solvent accessible residues using sequence-and homology-based qualitative and quantitative descriptors applied to a domain-complete X-ray structure learning set. Journal of applied crystallography. 2015;48(6):1976-84. doi:10.1107/S1600576715018531.
190. Joo K, Lee SJ, Lee J. Sann: solvent accessibility prediction of proteins by nearest neighbor method. Proteins: Structure, Function, and Bioinformatics. 2012;80(7):1791-7. doi:10.1002/prot.24074.
191. Noé F, De Fabritiis G, Clementi C. Machine learning for protein folding and dynamics. Current opinion in structural biology. 2020;60:77-84. doi:10.1016/j.sbi.2019.12.005.
192. Raimondi D, Orlando G, Pancsa R, Khan T, Vranken WF. Exploring the sequence-based prediction of folding initiation sites in proteins. Scientific reports. 2017;7(1):1-11. doi:10.1038/s41598-017-08366-3.
193. Tan AC, Gilbert D, Deville Y. Multi-class protein fold classification using a new ensemble machine learning approach. Genome Informatics. 2003;14:206-17. doi:10.11234/gi1990.14.206.
194. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AW, Bridgland A. Improved protein structure prediction using potentials from deep learning. Nature. 2020;577(7792):706-10. doi:10.1038/s41586-019-1923-7.
195. Liu S, Liu C, Deng L. Machine learning approaches for protein–protein interaction hot spot prediction: Progress and comparative assessment. Molecules. 2018;23(10):2535. doi:10.3390/molecules23102535.

196. Zhang M, Su Q, Lu Y, Zhao M, Niu B. Application of machine learning approaches for protein-protein interactions prediction. Medicinal Chemistry. 2017;13(6):506-14. doi:10.2174/1573406413666170522150940.
197. Sarkar D, Saha S. Machine-learning techniques for the prediction of protein–protein interactions. Journal of biosciences. 2019;44(4):1-12. doi:10.1007/s12038-019-9909-z.
198. Melo R, Fieldhouse R, Melo A, Correia JD, Cordeiro MND, Gümüş ZH, Costa J, Bonvin AM, Moreira IS. A machine learning approach for hot-spot detection at protein-protein interfaces. International journal of molecular sciences. 2016;17(8):1215. doi:10.3390/ijms17081215.
199. Wang W, Yang Y, Yin J, Gong X. Different protein-protein interface patterns predicted by different machine learning methods. Scientific reports. 2017;7(1):1-13. doi:10.1038/s41598-017-16397-z.
200. Wang D, Liu D, Yuchi J, He F, Jiang Y, Cai S, Li J, Xu D. MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. Nucleic acids research. 2020;48(W1):W140-W6. doi:10.1093/nar/gkaa275.
201. Xu Y, Chou K-C. Recent progress in predicting posttranslational modification sites in proteins. Current topics in medicinal chemistry. 2016;16(6):591-603. doi:10.2174/1568026615666150819110421.
202. Bao W, Yuan C-A, Zhang Y, Han K, Nandi AK, Honig B, Huang D-S. Mutli-features prediction of protein translational modification sites. IEEE/ACM transactions on computational biology and bioinformatics. 2017;15(5):1453-60. doi:10.1109/TCBB.2017.2752703.
203. Sankar K, Hoi KH, Yin Y, Ramachandran P, Andersen N, Hilderbrand A, McDonald P, Spiess C, Zhang Q. Prediction of methionine oxidation risk in monoclonal antibodies using a machine learning method. MAbs; 2018. doi:10.1080/19420862.2018.1518887.
204. Lai P-K, Fernando A, Cloutier TK, Kingsbury JS, Gokarn Y, Halloran KT, Calero-Rubio C, Trout BL. Machine Learning Feature Selection for Predicting High Concentration Therapeutic Antibody Aggregation. Journal of Pharmaceutical Sciences. 2021;110(4):1583-91. doi:10.1016/j.xphs.2020.12.014.
205. Long X, Jeliazkov JR, Gray JJ. Non-H3 CDR template selection in antibody modeling through machine learning. PeerJ. 2019;7:e6179. doi:10.7717/peerj.6179.
206. Wong WK, Georges G, Ros F, Kelm S, Lewis AP, Taddese B, Leem J, Deane CM. SCALOP: sequence-based antibody canonical loop structure annotation. Bioinformatics. 2019;35(10):1774-6. doi:10.1093/bioinformatics/bty877.
207. Raimondi D, Orlando G, Vranken WF, Moreau Y. Exploring the limitations of biophysical propensity scales coupled with machine learning for protein sequence analysis. Scientific reports. 2019;9(1):1-11. doi:10.1038/s41598-019-53324-w.
208. Narayanan H, Dingfelder F, Butté A, Lorenzen N, Sokolov M, Arosio P. Machine learning for biologics: opportunities for protein engineering, developability, and formulation. Trends in pharmacological sciences. 2021. doi:10.1016/j.tips.2020.12.004.
209. Graves J, Byerly J, Priego E, Makkapati N, Parish SV, Medellin B, Berrondo M. A review of deep learning methods for antibodies. Antibodies. 2020;9(2):12. doi:10.3390/antib9020012.
210. Liberis E, Veličković P, Sormanni P, Vendruscolo M, Liò P. Parapred: antibody paratope prediction using convolutional and recurrent neural networks. Bioinformatics. 2018;34(17):2944-50. doi:10.1093/bioinformatics/bty305.
211. Deac A, VeliČković P, Sormanni P. Attentive cross-modal paratope prediction. Journal of Computational Biology. 2019;26(6):536-45. doi:10.1089/cmb.2018.0175.
212. Akbar R, Robert PA, Pavlović M, Jeliazkov JR, Snapkov I, Slabodkin A, Weber CR, Scheffer L, Miho E, Haff IH. A compact vocabulary of paratope-epitope interactions enables

predictability of antibody-antigen binding. Cell Reports. 2021;34(11):108856. doi:10.1016/j.celrep.2021.108856.
213. Liu G, Zeng H, Mueller J, Carter B, Wang Z, Schilz J, Horny G, Birnbaum ME, Ewert S, Gifford DK. Antibody complementarity determining region design using high-capacity machine learning. Bioinformatics. 2020;36(7):2126-33. doi:10.1093/bioinformatics/btz895.
214. Mason DM, Friedensohn S, Weber CR, Jordi C, Wagner B, Meng S, Gainza P, Correia BE, Reddy ST. Deep learning enables therapeutic antibody optimization in mammalian cells by deciphering high-dimensional protein sequence space. BioRxiv. 2019:617860. doi:10.1101/617860.
215. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, Shi J, Deane CM. SAbDab: the structural antibody database. Nucleic acids research. 2014;42(D1):D1140-D6. doi:10.1093/nar/gkt1043.
216. Allcorn LC, Martin AC. SACS—self-maintaining database of antibody crystal structure information. Bioinformatics. 2002;18(1):175-81. doi:10.1093/bioinformatics/18.1.175.
217. Ferdous S, Martin AC. AbDb: antibody structure database—a database of PDB-derived antibody structures. Database. 2018;2018. doi:10.1093/database/bay040.
218. Khetan R. Biopharma licensing and M&A trends in the 21st-century landscape. Journal of Commercial Biotechnology. 2020;25(3). doi:10.5912/jcb943.
219. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A. Highly accurate protein structure prediction with AlphaFold. Nature. 2021:1-11. doi:10.1038/s41586-021-03819-2.
220. Hansson T, Oostenbrink C, van Gunsteren W. Molecular dynamics simulations. Current opinion in structural biology. 2002;12(2):190-6.
221. Hospital A, Goñi JR, Orozco M, Gelpí JL. Molecular dynamics simulations: advances and applications. Advances and applications in bioinformatics and chemistry. 2015:37-47.
222. Halaby D, Poupon A, Mornon J-P. The immunoglobulin fold family: sequence analysis and 3D structure comparisons. Protein engineering. 1999;12(7):563-71.
223. Perrakis A, Sixma TK. AI revolutions in biology: The joys and perils of AlphaFold. EMBO reports. 2021;22(11):e54046.
224. Varadi M, Velankar S. The impact of AlphaFold Protein Structure Database on the fields of life sciences. Proteomics. 2023;23(17):2200128.
225. Bekker GJ, Ma B, Kamiya N. Thermal stability of single-domain antibodies estimated by molecular dynamics simulations. Protein Science. 2019;28(2):429-38. doi:10.1002/pro.3546.
226. Chakroun N, Hilton D, Ahmad SS, Platt GW, Dalby PA. Mapping the aggregation kinetics of a therapeutic antibody fragment. Molecular pharmaceutics. 2016;13(2):307-19.
227. Selbig J, Argos P. Relationships between protein sequence and structure patterns based on residue contacts. Proteins: Structure, Function, and Bioinformatics. 1998;31(2):172-85.
228. Capriotti E, Fariselli P, Calabrese R, Casadio R. Predicting protein stability changes from sequences using support vector machines. Bioinformatics. 2005;21(suppl_2):ii54-ii8.
229. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences. 2021;118(15):e2016239118.
230. Wouters OJ, McKee M, Luyten J. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. Jama. 2020;323(9):844-53.
231. Jacobs SA, Wu S-J, Feng Y, Bethea D, O'Neil KT. Cross-interaction chromatography: a rapid method to identify highly soluble monoclonal antibody candidates. Pharmaceutical research. 2010;27(1):65-71.

232. Haverick M, Mengisen S, Shameem M, Ambrogelly A. Separation of mAbs molecular variants by analytical hydrophobic interaction chromatography HPLC: overview and applications. MAbs; 2014.
233. Tessier PM, Sandler SI, Lenhoff AM. Direct measurement of protein osmotic second virial cross coefficients by cross-interaction chromatography. Protein science. 2004;13(5):1379-90.
234. Liu Y, Caffry I, Wu J, Geng SB, Jain T, Sun T, Reid F, Cao Y, Estep P, Yu Y. High-throughput screening for developability during early-stage antibody discovery using self-interaction nanoparticle spectroscopy. MAbs; 2014.
235. Sule SV, Sukumar M, Weiss WF, Marcelino-Cruz AM, Sample T, Tessier PM. High-throughput analysis of concentration-dependent antibody self-association. Biophysical journal. 2011;101(7):1749-57.
236. Patro SY, Przybycien TM. Self-interaction chromatography: a tool for the study of protein–protein interactions in bioprocessing environments. Biotechnology and bioengineering. 1996;52(2):193-203.
237. Kandari D, Bhatnagar R. Antibody engineering and its therapeutic applications. International Reviews of Immunology. 2023;42(2):156-83.
238. Jain T, Boland T, Vásquez M. Identifying developability risks for clinical progression of antibodies using high-throughput in vitro and in silico approaches. Mabs; 2023.
239. Fernández-Quintero ML, Ljungars A, Waibl F, Greiff V, Andersen JT, Gjølberg TT, Jenkins TP, Voldborg BG, Grav LM, Kumar S. Assessing developability early in the discovery process for novel biologics. MAbs; 2023.
240. Martin KP, Grimaldi C, Grempler R, Hansel S, Kumar S. Trends in industrialization of biotherapeutics: a survey of product characteristics of 89 antibody-based biotherapeutics. Mabs; 2023.
241. Ahmed L, Gupta P, Martin KP, Scheer JM, Nixon AE, Kumar S. Intrinsic physicochemical profile of marketed antibody-based biotherapeutics. Proceedings of the National Academy of Sciences. 2021;118(37):e2020577118.

CHAPTER 2

2 Methodology

The methodology of this thesis work can be summarized as using the clinical-stage antibodies as a reference for successful developability and obtaining developability assay cutoffs from their biophysical performance distributions. The next step has been to apply and demonstrate the utility and accuracy of these developability criteria on multiple antibody datasets such as the human immune repertoire dataset, phage display antibodies dataset, transgenic mice antibodies dataset, failed antibodies dataset and other case studies. This methodology is formulated by reviewing literature and other recent works that support the concept of using clinical or approved mAbs as reference sets and flag those candidates that lie at the extreme tail regions of these distributions.

2.1 Clinical-stage antibodies datasets:

We have used two datasets representing approved or clinical-stage antibodies – Jain Dataset[1] and TheraSabDab dataset[2]. These two comprehensive datasets were selected as they are widely recognized and annotated datasets for clinical-stage mAbs.

2.1.1 Jain Dataset

Supplementary information from the Jain publication[1] was extracted to obtain the sequence information and assay values on 12 biophysical assays for 137 antibodies. (Available at https://www.pnas.org/doi/abs/10.1073/pnas.1616408114) We used the Supplementary information Table S2: Sequence information for the 137 antibodies and Supplementary information Table S3: Results of 12 assays for 137 antibodies. We imported and created a merged Jain dataset file with the clinical-stage information, V_H and V_L sequence information and the measured experimental values for 12 assays.

Methodology of 12 biophysical assays selected and done by the Jain publication:[1]

1. HIC - The HIC assay was done with 5 μg IgG samples (1 mg/mL) that were spiked in with a mobile phase A solution (1.8 M ammonium sulfate and 0.1 M sodium phosphate at pH 6.5) to achieve a final ammonium sulfate concentration of about 1 M before analysis. A Sepax Proteomix HIC butyl-NP5 column was used with a liner gradient of mobile phase A and mobile phase B solution (0.1 M sodium phosphate, pH 6.5, flow rate -1 mL/min) over 20 min with UV absorbance monitoring at 280 nm.

2. SMAC - The SMAC assay was performed with 2 μg of samples that were injected into a Zenix SEC-300 column (213300-4630; Sepax Technologies). A flow rate of 0.35 mL/min with the running buffer containing 150 mM sodium phosphate at pH 7.0 was used. Retention time for each sample was assigned based on the major peak.

3. SGAC - The SGAC-SINS assay was performed with gold nanoparticles (15705; Ted Pella Inc.) coated with 80% capturing anti-human goat IgG Fc (109-005-098; Jackson ImmunoResearch) and 20% with polyclonal goat nonspecific antibody (005-000-003; Jackson ImmunoResearch). The antibodies of interest were then incubated with the particles for 30 min. The samples were then diluted with varying ranges of ammonium sulfate (300–1,000 mM in 100-mM steps) and incubated for an additional 1.5 h. The wavelength shift was measured using Molecular Devices SpectraMax M2 with SoftMax Pro6 software. The SGAC100 was obtained by graphing the wavelength shifts of a sample against the ammonium sulfate concentration and extrapolating the concentration at which the shift was 560 nm. For antibodies where the shift was below 560 nm at the the highest salt concentration a value of 1,000 mM was assigned.

4. CIC - The CIC column was prepared by coupling ∼30 mg of human serum polyclonal antibodies (I4506; Sigma) to a 1-mL HiTrap column (17-0716-01; GE Healthcare), followed by quenching with ethanolamine. Approximately 5 μg of each antibody was tested at a flow rate of 0.1 mL/min using PBS as a mobile phase on an Agilent 1100 series HPLC system.

5. CSI-BLI - The CSI-BLI assay was carried out as previous described (7). Briefly, human IgG was loaded to an AHQ biosensor (ForteBio) to ∼1 nm, followed by sensor blocking with human IgG1 Fc. The self-association was performed at 1 μM solution concentration of IgG for 300s on an Octet HTX system (ForteBio). The binding response from association was subtracted from that of a reference IgG (adalimumab).

6. AC-SINS - The AC-SINS assay was performed with gold nanoparticles (15705; Ted Pella Inc.) that were coated with 80% capturing anti-human goat IgG Fc (109-005-098; Jackson ImmunoResearch) and 20% with polyclonal goat nonspecific antibody (005-000-003; Jackson ImmunoResearch). The antibodies of interest were then incubated with the particles for 2 h and the wavelength shift was measured using Molecular Devices SpectraMax M2 with SoftMax Pro6 software. The self-interacting clones show a higher wavelength shift away from the PBS sample.

7. HEK titer expression - The antibodies were expressed in HEK293 cells. The V_H and V_L encoding gene fragments (Integrated DNATechnologies) were subcloned into the heavy- and light-chain pcDNA 3.4+vectors (ThermoFisher). All mAbs were expressed as IgG1 isotype. The corresponding vectors were cotransfected into HEK293 suspension cells. After 6 d of growth, the cell culture supernatant was harvested by centrifugation and passed over Protein A agarose (MabSelect SuRe; GE Healthcare).

8. PSR - The PSR assay was done with soluble membrane proteins prepared from CHO cells. The enriched membrane fraction was biotinylated using NHS-LCBiotin (Pierce, 21336; Thermo Fisher). This polyspecificity reagent was incubated with IgG-presenting yeast, followed by washing. Then secondary labeling mix (Extravidin-R-PE, antihuman LC-FITC, and propidium iodide) was added to the mixture. Samples were analyzed on FACSCanto (BD Biosciences) using an HTS sample injector. Flow cytometry data were analyzed for median fluorescence intensity (MFI) in the R-PE channel to assess nonspecific binding. MFI values were normalized from 0 to 1 based on three reference antibodies exhibiting low, medium, and high PSR MFI values.

9. ELISA - The ELISA protocol was done with six different antigens, cardiolipin (50 μg/mL, C0563; Sigma), KLH (5 μg/mL, H8283; Sigma), LPS (10 μg/mL, tlrl-eblps; InvivoGen), ssDNA (1 μg/mL, D8899; Sigma), dsDNA (1 μg/mL, D4522; Sigma), and insulin (5 μg/mL, I9278; Sigma), were coated onto ELISA plates (3369; Corning) individually at 50 μL per well overnight at 4 °C. Plates were blocked with PBS with 0.5% BSA at room temperature (RT) for 1 h, followed by three washes with PBST (PBS plus 0.1% Tween 20). Fifty microliters of 100 nM testing antibody solution was added to each well and incubated at RT for 1 h. The absorbance was read at 450 nm and score determined by normalizing absorbance with no test antibody control wells.

10. BVP - The BVP assay used 50 μL baculovirus particles (BlueSky Biotech) stock that was diluted with equal volume of 50 mM sodium carbonate (pH 9.6) per well and incubated on ELISA plates (3369; Corning) at 4 °C overnight. The next day, unbound BVPs were aspirated from the wells. All remaining steps were performed at room temperature. One hundred microliters of blocking buffer (PBS with 0.5% BSA) was added and let incubate for 1 h before three washes with 100 μL of PBS. Next, 50 μL of 1 μM testing antibodies in blocking buffer was added to the wells and incubated for 1 h followed by six washes with 100 μL of PBS. Fifty microliters of diluted anti-

human IgGHRP conjugate (81-7120; ZyMax) was added to the wells and incubated for 1 h followed by six washes as before. Finally, 50 μL of TMB substrate (34021; Fisher Scientific) was added to each well and incubated for 10–15 min. The reactions were stopped by adding 50 μL of 2 M sulfuric acid to each well. The absorbance was read at 450 nm and BVP score determined by normalizing absorbance by control wells with no test antibody.

11. DSF - The Tm with DSF was determined using a CFX96 Real-Time System from BioRad, based on the protocol described earlier (32). Briefly, 20 μL of 1 mg/mL sample was mixed with 10 μL of 20× SYPRO orange. The plate was scanned from 40 °C to 95 °C at a rate of 0.5 °C/2 min. The Fab Tm was assigned using the first derivative of the raw data from the BioRad analysis software.

12. ACC STAB - Samples were kept 1 mg/mL at 40 °C for 30 d in HBS (25 mM Hepes and 150 mM sodium chloride, pH 7.3). Time points were taken at day 0, 5, 20, and 30, and the samples were then analyzed by SEC (0022855; Tosoh Bioscience). For SEC analysis, the running buffer composition was 200 mM sodium phosphate and 250 mM sodium chloride, pH 7.0. A long-term stability slope was calculated from the percent aggregated, measured on the SEC that was used as the final ACC STAB score.

2.1.2 TheraSabDab Dataset

The Therapeutic Structural Antibody Database (TheraSabDab) tracks all antibody and nanobody-related therapeutics recognized by the World Health Organisation (WHO) with accompanying metadata. The TheraSabDab is available as a free online web server at https://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/therasabdab/search/. Raw TheraSabDab dataset was extracted from the online resource on 6 January 2022.

The downloaded TheraSabDab dataset had variable domain sequence information with accompanying metadata for 658 clinical-stage antibody-based biotherapeutics as of January 2022. TheraSabDab dataset contains the following columns - Name, Format, Clinical trial status, Target, Year, and sequence information for all antibodies. No data was excluded here to capture the full landscape of antibody therapeutics. The dataset visualization for Figure 10 and Figure 11 in Chapter 3 were performed using Power BI 2.118. 286.0. Here, advanced plotting was performed using the Power stacked bar graph feature with Clinical trial status in legend and format as the x-axis.

2.2 Human immune repertoire dataset:

The Observed Antibody Space (OAS) database contains annotated immune repertoires that cover one billion sequences from over 80 different studies. OAS is accessible via a web-based server at https://opig.stats.ox.ac.uk/webapps/oas/. The OAS search was customized for healthy non-vaccinated human immune repertoires. The search terms were human (Species); undefined (Age); all (BSource); all (BType); None (Vaccine); None (Disease); undefined (Subject). So, immune repertoires of recently vaccinated or individuals that were diseased at the time of sequencing were excluded. The OAS search returned 350,980 filtered paired sequences from two studies – Eccles and Jaffe. The raw variable domain sequence files were downloaded in 14 batch files in the csv format and then merged together manually into a single input file in a fasta format.

2.3 Multispecific format antibodies dataset:

We used a comprehensive manual search procedure using several online resources to create a new multispecific formats antibody dataset. Firstly, we used the International ImMunoGeneTics Information System (IMGT) database.[3] The IMGT/mAb-DB query search terms were modified for Development technology by selecting BiTE®, DuoBody®, Dual-Affinity Re-targeting (DART®), CrossMAb technology, Triomab, Pentambody™, and DVD-Ig™. The search yielded 132 antibody fragments across all the development technologies. There were 41 antibodies here with missing sequence information. Next, we checked for these missing mAb sequence information by online searches at AdisInsight (Link 1); Google Patents (Link 2); Antibody Resource (Link 3) for each antibody name. The company websites were also used where applicable to insert the sequence information. Finally, we excluded 29 antibody results with missing sequence information and obtained the final dataset of 103 engineered scFv fragments. The multispecific format antibodies dataset results are shown in Table 10, Chapter 4.

2.4 Phage display antibodies dataset:

We used a comprehensive manual search procedure using several online resources to create a new phage display antibodies dataset. Firstly, we used the International ImMunoGeneTics Information System (IMGT) database.[3] The IMGT/mAb-DB query search terms were modified for Development technology with prefix Antibody phage display by selecting the CAT phage library, MorphoSys's HuCAL®, Dyax library, and Artificial Human library. The search yielded 62 antibody fragments across all phage

display development technologies. Here, there were 27 antibody results with missing sequence information. Next, we checked for these missing mAb sequence information by online searches at the resources - AdisInsight (Link 1); Google Patents (Link 2); Antibody Resource (Link 3) for each antibody name. The company websites were also used where applicable to insert the sequence information. Finally, we excluded 22 antibody results with missing sequence information and obtained the final dataset of 40 phage display scFv fragments. The phage display antibodies dataset results with information on name, clinical status, and company are shown in Table 11, Chapter 4.

2.5 Transgenic mice antibodies dataset:

We used a comprehensive manual search procedure using several online resources to create a new transgenic mice antibodies dataset. Firstly, we used the International ImMunoGeneTics Information System (IMGT) database.[3] The IMGT/mAb-DB query search terms were modified for Development technology with prefix Transgenic mice by selecting Abgenix's XenoMouse®, Medarex's UltiMAb®, Medarex's HuMAb-Mouse®, and VelocImmune®. The search yielded 66 antibody fragments across all the development technologies. There were 25 antibody results here with missing sequence information. Next, we checked for these missing mAb sequence information by online searches at AdisInsight (Link 1); Google Patents (Link 2); Antibody Resource (Link 3) for each antibody name. The company websites were also used where applicable to insert the sequence information. Finally, we excluded 20 antibody results with missing sequence information and obtained the final dataset of 46 transgenic mice fragments. The transgenic mice antibodies dataset results are shown in the Table 12, Chapter 4.

2.6 AbPred Calculations and Application on datasets:

The AbPred tool calculates the predicted performance on 12 biophysical platforms, using machine learning algorithms trained on experimental data from the Jain dataset.[4] We have used the variable domain sequence information for separate datasets as inputs to the AbPred tool. The input fasta files were run on the Dockerhub platform available at https://hub.docker.com/r/maxhebditch/abpred that generates 12 output assay scores.

The web application for the sequence-based algorithms is also available online at the protein-sol webserver, at https://protein-sol.manchester.ac.uk/abpred, with models and virtualisation software available at https://protein-sol.manchester.ac.uk/software. The software is shell/perl based and should be simple to run on any Unix-like system.

Overall, the AbPred models have been developed from sequences of the (heavy and light chain) variable domains, using 35 sequence features namely the 20 amino acid compositions and 15 other sequence-derived features that represent physicochemical properties, with the machine-learning models being trained on the variation in those properties between the CDRs of the 137 mAbs from the Jain dataset.

Procedure to generate AbPred calculations for a dataset: (Docker Link)

Step 1 – Name the input fasta file as abpred.fasta in current directory.
Step 2 – Open Dockerhub. Import code with docker pull maxhebditch/abpred:latest.
Step 3 – Run docker code using run --rm -v $(pwd)/:/abpred/host maxhebditch/abpred.
Step 4 – Wait for output. Predictions from the machine learning models will be in a directory called abpred_outputs in your current directory.

AbPred Flow Diagram for Computational Developability Assessment:

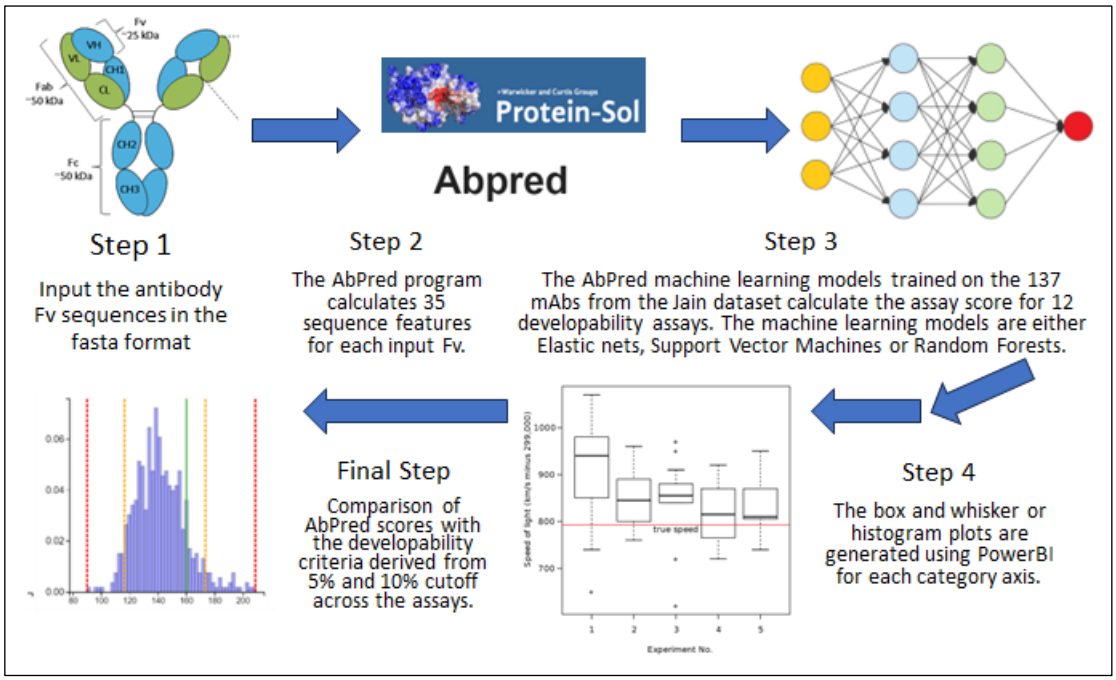


Figure 7: Application of AbPred for Computational Developability Assessment of antibodies. The AbPred machine-learning models trained on Jain dataset generate the 12 assay scores based on 35 sequence feature calculations of the input Fv sequences. These assay scores are finally compared with the clinical-stage computational developability assessment guidelines.

AbPred models have used 35 sequence features to understand the variance in 12 different biophysical characterisation assays often used in developability assessments. The 35 features are composed of the standard 20 amino acid propensities, followed by 7 amino acid composite scores (KmR = K-R, DmE = D-E, KpR = K+R, DpE = D+E, PmN = K+R-D-E, PpN = K+R+D+E, aro = F+W+Y) and further 8 sequence features,

fld = folding propensity[5], dis = disorder propensity[6], bet = beta strand propensities[7], mem = Kyte-Doolittle hydropathy[8], sequence entropy, pI, and absolute charge that are reflecting different biophysical calculations across the input variable sequence.

These 35 sequence features are used as input to the machine learning algorithms which are either the elastic net algorithm[9], or other non-linear algorithms like support vector machines[10] (SVM) and random forest algorithm[11]. These algorithms were selected as they had the lowest mean average error to the Jain experimental dataset. A summary of the algorithm performance used by AbPred for each assay is provided in Table 3. Here, the R^2 and p-value correspond to the overall fit of the machine learning models.

AbPred Assay	Algorithm	R^2	p-value
Hydrophobic Interaction Chromatography (HIC)	Elastic net	0.391	2.33E–17
Standup Monolayer Absorption Chromatography (SMAC)	Elastic net	0.353	7.33E–15
Cross-Interaction Chromatography (CIC)	SVM	0.306	4.46E–17
Affinity-Capture Self-Interaction Nanoparticle Spectroscopy (AC-SINS)	Elastic net	0.268	6.46E–14
Enzyme-Linked Immunosorbent Assay (ELISA)	Random forest	0.383	4.95E–77
Baculovirus Particle (BVP) assay	Random forest	0.355	6.85E–68
Salt-Gradient Affinity-Capture Spectroscopy (SGAC)	SVM	0.215	2.30E–39
Poly-Specificity Reagent (PSR) assay	SVM	0.316	2.39E–10
Expression Titer in HEK cells (HEK)	SVM	0.1121	1.87E–09
Differential Scanning Fluorimetry (DSF)	SVM	0.13	4.49E–08
Clone Self-Interaction by Biolayer Interferometry (CSI BLI)	SVM	0.169	1.24E–05
Accelerated Stability (ACC STAB) assay	SVM	0.086	2.82E–01

Table 3: AbPred machine learning algorithm summary for each of the 12 biophysical assays.

The feature selection stage of the machine learning methods gives an indication of the sequence-based features that are correlated with a particular developability assay. The heat map visualization of the Pearson correlation coefficient between the Fv sequence

composition scores (35 sequence features) used in the AbPred algorithms and the score on each of the 12 biophysical assays for the mAb137 Jain dataset are shown in Figure 8. Here, the dark red values indicate a stronger positive correlation, while the dark blue values indicate a stronger negative correlation. For instance, the HIC assay has a strong positive correlation to the sequence length and the aromatic amino acid content (aro). Other assays such as ELISA and PSR have a strong positive correlation to the charge-based features like lysine and arginine content (KpR) and the overall charge (PmN).

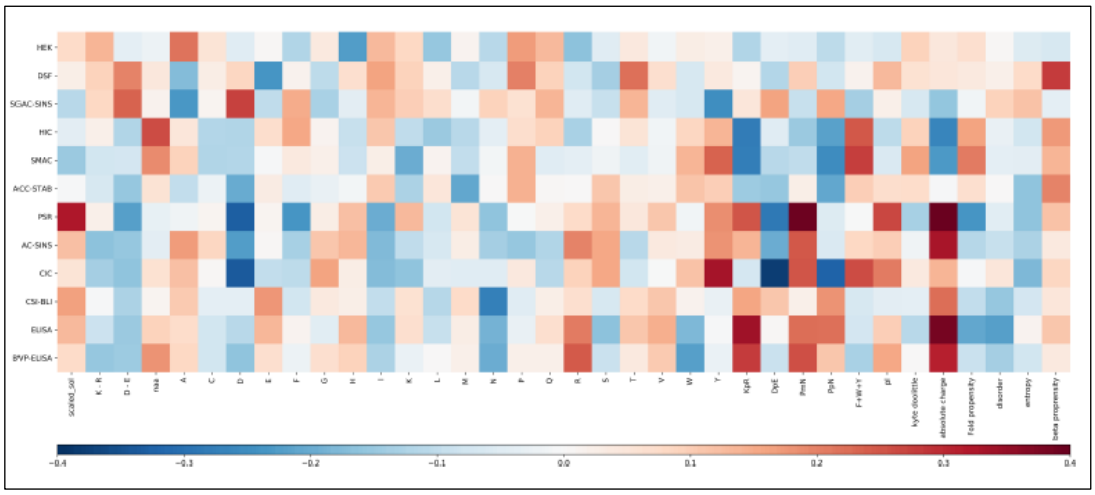


Figure 8: Heat map visual of the Pearson correlation coefficient between the Fv sequence composition scores (35 sequence features) used in the Abpred algorithms and the score on each of the 12 biophysical assays for the mAb137 Jain dataset. Dark red values indicate a stronger positive correlation, and dark blue values indicate a stronger negative correlation.

So, the general equation of an AbPred assay score is $\boldsymbol{Assay\ Score = \sum \beta_k(Property_k)}$ where k ranges from 1 to 35 to capture all sequence properties and β_k represents the coefficient assigned by the machine learning regression algorithm to each sequence property from the Jain experimental dataset fit. The full details of the machine learning equations and the optimization/minimization source code are available at Github .

A general machine learning regression algorithm estimates the mapping function *(f)* of output variables *(Y)* given input variables *(X)* that can be summarized by the general equation $Y = f(X)$. Here, we provide an example of the neural networks that are at the core of several machine learning algorithms such as random forests and artificial neural networks. The machine learning algorithms are usually composed of layers of nodes that attempt to simulate the decision process in the networks of neurons of the biological central nervous system. The first layer of a neural net is called the input layer, followed by hidden layers, then finally the output layer. Each node receives one

or more input signals. These input signals can come from either the raw data set or from neurons positioned at a previous layer of the neural net. The AbPred algorithms use the 35 sequence features as input signals. Each node then calculates an output signal based on either a linear combination of the input factors shown as $\sum x_n * w_n$ or through an activation function which limits the output amplitude depending on the machine learning algorithm subtype choice. Here, x_n represent the inputs to the model, while w_n represent the factor weights which multiply the corresponding input signal that are assigned by the model. The output signal is then either captured as the final output score as in the case of AbPred or sent to nodes deeper in the neural net in case of more complex models such as the artificial neural networks. So, each node in the AbPred algorithms processes an arriving signal by multiplying it to the factor weights (w_n) that may have a positive or negative multiplier value depending on the influence of each sequence feature towards the biophysical assay. A linear combiner then sums up the 35 input signals, weighted by the respective factor weights to generate the final output AbPred assay score. In case of AbPred tool, the factor weights were calculated from a previous model training step on the Jain experimental dataset[1] which decided the model architecture and the factor weight values for each assay. Figure 9 below has shown a general overview about the data processing in machine learning algorithms.

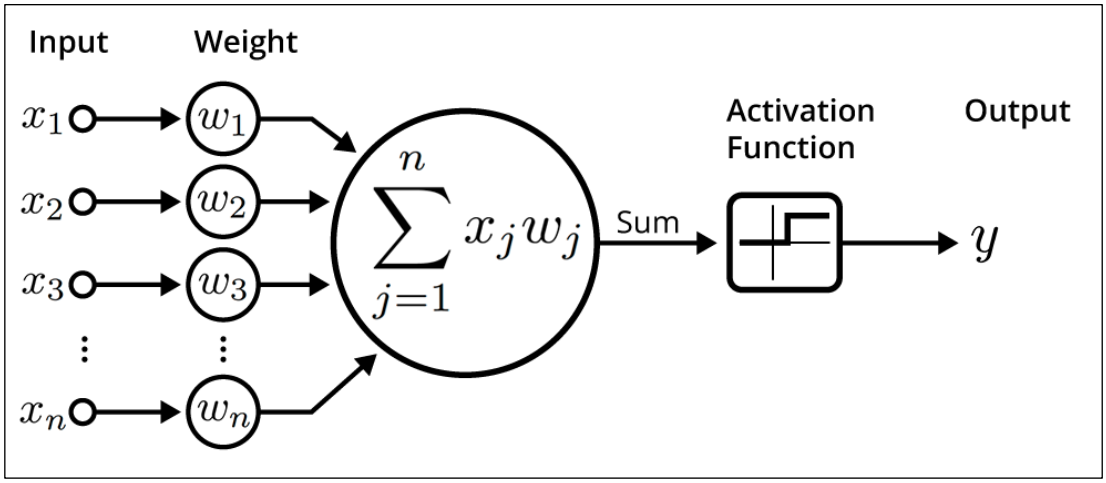


Figure 9: Overview of a general machine learning algorithm. x_n represent the inputs to the model, w_n represent the factor weights which multiply the corresponding input signal that are assigned by the model. Finally, the output signal (y) is either summed by a linear combination of the input factors $\sum x_n w_n$ or further connected to an activation function which limits the output amplitude depending on the algorithm choice. Image from www.freecodecamp.org.

Elastic net regression is a penalized linear regression model that is a statistical hybrid method which combines two regularized linear regression techniques - lasso and ridge, to deal with the multicollinearity issues that arise between predictor variables.[12] This

algorithm adds a penalty term to the standard least-squares objective function. Elastic net algorithm can perform robust feature selection by shrinking the coefficients of irrelevant variables to zero that results in a model with fewer variables, which is easier to interpret and less prone to overfitting. The elastic net regression is used for the HIC SMAC and AC-SINS assays in the AbPred tool. Next, the Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression.[13] The objective of the support vector machine algorithms is to minimize the margin that is the distance between the support vectors and hyperplane that distinctly classifies the data points in an N-dimensional space (N here represents the number of features). Since the SVM algorithms are very effective in high dimensional spaces, most of the AbPred assays namely – CIC, SGAC, PSR, HEK, DSF, CSI-BLI and ACC-STAB use the SVM algorithm. Finally, the random forest algorithm uses the decision tree framework to create multiple randomly drawn decision trees from the data for regression.[14] ELISA and BVP assays use random forests regression in AbPred.

2.7 TAP Calculations and Application on datasets:

The Therapeutic Antibody Profiler[15] (TAP) calculates five developability metrics for an input heavy and light chain sequence. These five metrics are namely - Total CDR Length; Patches of Surface Hydrophobicity (PSH) metric; Patches of Positive Charge (PPC) metric; Patches of Negative Charge (PNC) metric and the Structural Fv Charge Symmetry Parameter (SFvCSP). The heavy and light chain sequences were used as input to the TAP tool available online at https://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/sabpred/tap. TAP outputs a detailed profile of an antibody with a typical runtime of less than 30s with five metric scores, ABodyBuilder structural model and an interactive visual representation of the hydrophobic/charge patches on the antibody model. TAP uses ABodyBuilder to generate a model structure of the input antibody.

ABodyBuilder is a homology modelling program for antibody Fv modelling. It is a deep learning-based CDR loop structure prediction tool with the model trained on the position of the backbone atoms for all six CDR loops plus two anchor residues at either end. The TAP tool calculates the five scores across the CDR vicinity which comprises every surface-exposed IMGT-defined CDR and anchor residue, and all other surface-exposed residues with a heavy atom within a 4-A° radius. To calculate the five metric scores, the following charges were assigned by sequence: aspartic acid (−1); glutamic acid (−1); lysine (+1); arginine (+1); and histidine (+0.1). Salt-bridge residues were

assigned a charge of 0 and Tyrosine hydroxyl deprotonation was not considered. The PPC and PNC metrics are based on calculating the absolute value of the charge assigned to residue R represented as $|Q(R)|$. The general equation for both PPC and PNC score can be represented as $\sum|Q(R_1)|\ |Q(R_2)| \div r^2_{12}$ where R_1 and R_2 are any two surface-exposed residues with a closest heavy-atom distance, r_{12}, <7.5 A°. The PSH metric is calculated similarly with the normalized hydrophobicity score for residue R in scheme S represented as $H(R,S)$ replacing the $|Q(R)|$. Finally, the SFvCSP values were calculated as $\sum_{RH}|Q(R_H)|\ *\sum_{RL}|Q(R_L)|$ where R_H and R_L are surface-exposed V_H and V_L residues, respectively. Here, the residues defined as 'surface-exposed' have a >7.5% relative exposure across side-chain atoms, compared with the open-chain form alanine-R-alanine. The full details and model information are available at the GitHub repository available at https://github.com/orgs/oxpig/repositories.

2.8 **T20 Humanness Score Calculation:**

The T20 score analyzer is a tool that calculates the humanness of monoclonal antibody variable region sequences. The T20 score is scaled from 0 to 100, where a higher score is a more human-like antibody. In general, full-length sequences that score above 80 are considered human-like, while framework-only sequences that score above 85 are considered human-like. The online tool available at https://sam.curiaglobal.com/t20/ was used to calculate the T20 humanness score.

To calculate the T20 scores, an input variable region protein sequence is first assigned the Kabat numbering and CDR residues are identified. The full-length sequence or the framework only sequence with CDR residues removed is compared to every sequence in the respective antibody database using the blastp protein BLAST algorithm[16]. The sequence identity match between each pairwise comparison is isolated, and after every sequence in the database has been analyzed, the sequences are sorted from high to low based on the sequence identity to the input sequence. The percent identity of the Top 20 matched sequences is averaged to obtain the final T20 score. The output generated by the T20 tool include – the T20 score representing the humanness score on a 0-100 scale; BLAST results; and the FASTA sequences representing the FASTA formatted sequences of the top 20 antibodies that were used to calculate the T20 score.

2.9 **Master training dataset for machine learning classification:**

A combined excel sheet that had the calculated scores from AbPred and TAP were created to establish the training dataset for machine-learning algorithms. We evaluated a total of 52 properties that included Abpred assay scores (12), Protein-Sol sequence features (35), and finally the TAP scores (5). All these scores were evaluated for the TheraSabDab dataset which contains information on 658 antibodies in clinical trials or approved stages. This created 'Master training dataset' was a representation of the computational developability assessment results for the clinical-stage antibodies. This training dataset was imported in MATLAB and 5-fold cross-validation was performed to prepare the dataset for machine-learning training step. Here, the data is partitioned into 5 randomly chosen subsets (or folds) of roughly equal size. One subset is used to validate the machine learning model that is trained using the remaining subsets. This process is repeated 5 times so that we ensure that each subset is used exactly once for validation. MATLAB R2021a was used throughout this work for machine learning.

2.10 Application of machine learning classification algorithms for estimating the clinical trial progression of antibody therapeutics:

We have used the Machine Learning Toolbox™ and the new Classification Learner App in MATLAB R2021a (https://uk.mathworks.com/help/stats/classificationlearner-app.html) to train machine learning models of all the major classifiers: decision trees, discriminant analysis, support vector machines, logistic regression, nearest neighbors, naive Bayes, ensemble, and neural networks. The Classification Learner App performs supervised machine learning by supplying a known set of input data (observations or examples) and known responses to the data (labels or classes) to train a model that generates predictions for the response to new data. In our work, the known 'Clinical Trial Status' of antibodies from the TheraSabDab annotation were used as the labels while the master training dataset of 52 antibody variable region properties were used as the input data observations. The app also displays the results of the validated model. Diagnostic measures such as model accuracy, scatter plot or the confusion matrix chart reflect the validated model results.

Procedure to run machine-learning classification algorithms:

Step 1 – MATLAB Workspace >> Apps tab >> Machine Learning and Deep Learning group >> Click Classification Learner tab to open the Classification Learner app.

Step 2 – Choose a classifier. Learn tab >> Models section >> Click a classifier type. Additional information for Manual Classifier Training - To see all available classifier options, click the arrow on the far right of the Models section to expand the list of classifiers. The nonoptimizable model options in the Models gallery are preset starting points with different settings, suitable for a range of different classification problems. Step 3 – After selecting a classifier, we can train the model. Train section >> Click Train All >> Select Train Selected. Repeat the process on different classifiers. Finally, select one of the All options in the Models gallery to try all nonoptimizable models of the same or different types. To automatically tune hyperparameters of a specific model type, we select the corresponding Optimizable model and perform hyperparameter optimization. After training step, we compare and improve the Classification Models. Step 4 – Examine the Accuracy (Validation) score reported in the Models pane for each model. Click models in the Models pane and open the corresponding plots to explore the results. Compare model performance by inspecting the results in the plots. Compare the models by using Sort by >> Models pane. We select the best model in the Models pane and then try including and excluding different features in the model. Step 5 – Learn tab >> Options section >> Feature Selection. Use the available feature ranking algorithms to select features. We improve the model by removing the features with low predictive power and compare results among the models in the Models pane. Step 6 – We save the final machine-learning algorithms and compile the results.

We generated and tested over 38 machine learning classification models to predict clinical trial progression which were tested across TheraSabDab clinical-stage dataset. Overall, the models with best percentage accuracy (validation) score were optimized further using the manual classifier training procedures mentioned above. Diagnostic measures, such as model accuracy, and plots, such as a scatter plot or the confusion matrix chart, reflect the validated model results. Full detailed information is available at https://uk.mathworks.com/help/stats/classificationlearner-app.html.

We utilized a variety of classifier types across the 38 machine learning models. Firstly, we used the decision tree algorithms which are non-parametric supervised learning algorithms based on hierarchical tree structures that consists of a root node, branches, internal nodes and leaf nodes.[17] Here, three decision subtypes were explored - Coarse Tree, Medium Tree and Fine Tree. Next, we used Linear Discriminant and Quadratic Discriminant classifier types which assume that different classes generate data based

on different Gaussian distributions.[18] To train a classifier, the fitting function estimates the parameters of a Gaussian distribution for each class. Next, we used the Logistic Regression Classifiers that model the output class probabilities as a function of the linear combination of predictors. Next, Gaussian Naive Bayes and Kernel Naive Bayes classifiers were explored. The naive Bayes algorithms leverage the Bayes theorem and makes the assumption that predictors are conditionally independent, given the class.[19]

We then explored six subtypes of Support Vector Machines (SVMs) in Classification Learner: Linear SVM; Quadratic SVM; Cubic SVM; Fine Gaussian SVM; Medium Gaussian SVM; and Coarse Gaussian SVM. SVM classifies data by finding the best hyperplane that separates data points of one class from those of the other class.[20] The best hyperplane for an SVM means the one with the largest margin between the two classes. Margin means the maximal width of the slab parallel to the hyperplane that has no interior data points. Next, Nearest neighbor classifiers were used that categorize query points based on their distance to points or neighbors in a training dataset with use of various metrics to determine the distance.[21] Finally, Neural Network Classifiers were used for classification.[22] The first fully connected layer of the neural network has a connection from the network input (predictor data), and each subsequent layer has a connection from the previous layer. Each fully connected layer multiplies the input by a weight matrix and then adds a bias vector. An activation function follows each fully connected layer. The final fully connected layer and the subsequent softmax activation function produce the output, namely the classification scores and the predicted labels.

2.11 **Developability criteria assessment by Failed antibody dataset:**

We used a comprehensive manual search procedure using several online resources to create a new failed antibodies dataset. The failed antibodies dataset was focused on antibodies that were withdrawn or discontinued due to safety, low efficacy, or other reasons. Firstly, we used the International ImMunoGeneTics Information System (IMGT) database.[3] The IMGT/mAb-DB query search terms were modified for Development status by selecting Discontinued and Withdrawn. The search yielded 47 antibody fragments across both development status. However, there were 35 antibody results here with missing sequence information. The missing sequence information was a major problem encountered in building the failed antibodies dataset. So, next we checked for these missing mAb sequence information by manual online searches at AdisInsight (Link 1); Google Patents (Link 2); Antibody Resource (Link 3) for each

antibody name. We also added 18 new failed antibodies with sequence information by searches on the Clinical trial database (Link4) and Fierce Biotech (Link5). Finally, we excluded 27 antibody results with missing sequence information and obtained the final dataset of 38 failed antibody fragments. The failed antibodies dataset results are shown in the Table 16, Chapter 5. We then obtained AbPred scores for all these 38 antibodies.

The computational developability criteria performance was evaluated by calculating the 5% and 10% threshold cutoff values of clinical-stage antibodies and flagging the number of features failed by each failed antibody as shown in Figure 46. The threshold values for each assay are presented in Table 6. We next created in MATLAB a custom binary classification algorithm based on these cutoff values in Table 6. The AbPred scores for the failed antibodies dataset was then used as input to our custom binary classification algorithm. Finally, the confusion matrix plot was generated using the command confusionmat(g1,g2) in MATLAB. The final plot is shown in Figure 51.

2.12 **Kolmogorov-Smirnov test (K-S test) Statistics:**

The Kolmogorov-Smirnov test is a two-sample test to compare cumulative distribution functions. We have used K-S test to determine and validate if two samples appear to follow the same distributions. We used AbPred raw data input to the R-module code obtained from https://www.wessa.net/rwasp_Reddy-Moores%20K-S%20Test.wasp/. For shorter datasets like transgenic mice and phage display library dataset the online tool was used with each AbPred score data series entered as column delimited by a space or Tab. Chart options were selected for Width:600 and Height: 400. The tool returned two outputs - the K-S Test Statistic value and the P-value that were noted.

2.13 **Supplementary Information – MATLAB Codes**

Biopharmaceutical Informatics MATLAB Livescript

Computational Developability Assessment Codes
Part 1: Import the Data

This section imports the data using a function generated by the Import Tool.

Set up the Import Options and import the AbPred data

```
%PhD work Matlab Codes
%Created by Rahul Khetan.
%PhD Student at The University of Manchester
%Contact - rahul.khetan@manchester.ac.uk
```

```
 opts = spreadsheetImportOptions("NumVariables", 15);

 % Specify sheet and range
 opts.Sheet = "AbPred";
 opts.DataRange = "A2:O659";

 % Specify column names and types
 opts.VariableNames = ["SNo", "Therapeutic", "HIC", "SMAC", "SGAC",
"CIC", "CSIBLI", "ACSINS", "HEK", "PSR", "ELISA", "BVP", "DSF",
"ACCSTAB", "ClinicalTrial"];
 opts.VariableTypes = ["double", "string", "double", "double",
"double", "double", "double", "double", "double", "double", "double",
"double", "double", "double", "categorical"];

 % Specify variable properties
 opts = setvaropts(opts, "Therapeutic", "WhitespaceRule", "preserve");
 opts = setvaropts(opts, ["Therapeutic", "ClinicalTrial"],
"EmptyFieldRule", "auto");

 % Import the data
 TheraSabDabML1 = readtable("D:\Rahul
Khetan\Desktop\ACADEMICS\clubs\Curriculum vitae\Manchester
PhD\Computational Developability assessment\Year 3 work\MATLAB Data
Science and ML\TheraSabDab ML.xlsx", opts, "UseExcel", false)

 %Clear temporary variables
 clear opts
```

Set up the Import Options and import the TAP data

```
 opts = spreadsheetImportOptions("NumVariables", 6);

 % Specify sheet and range
 opts.Sheet = "Sheet1";
 opts.DataRange = "A2:F528";

 % Specify column names and types
 opts.VariableNames = ["CST", "TotalCDRLength", "PSH", "PPC", "PNC",
"SFvCSP"];
 opts.VariableTypes = ["string", "double", "double", "double",
"double", "double"];

 % Specify variable properties
 opts = setvaropts(opts, "CST", "WhitespaceRule", "preserve");
 opts = setvaropts(opts, "CST", "EmptyFieldRule", "auto");

 % Import the data
 TheraSabDabTAP = readtable("D:\Rahul
Khetan\Desktop\ACADEMICS\clubs\Curriculum vitae\Manchester
```

```
PhD\Computational Developability assessment\Year 3 work\MATLAB Data
Science and ML\TheraSabDab TAP.xlsx", opts, "UseExcel", false)

 %Clear temporary variables
 clear opts

 opts = spreadsheetImportOptions("NumVariables", 7);

 % Specify sheet and range
 opts.Sheet = "Sheet1";
 opts.DataRange = "A2:G423";

 % Specify column names and types
 opts.VariableNames = ["CST", "TotalCDRLength", "PSH", "PPC", "PNC",
"SFvCSP", "Status"];
 opts.VariableTypes = ["string", "double", "double", "double",
"double", "double", "categorical"];

 % Specify variable properties
 opts = setvaropts(opts, "CST", "WhitespaceRule", "preserve");
 opts = setvaropts(opts, ["CST", "Status"], "EmptyFieldRule", "auto");

 % Import the data
 TheraSabDabTAP2 = readtable("D:\Rahul
Khetan\Desktop\ACADEMICS\clubs\Curriculum vitae\Manchester
PhD\Computational Developability assessment\Year 3 work\MATLAB Data
Science and ML\TheraSabDab TAP2.xlsx", opts, "UseExcel", false)

 clear opts
```

Part 2: Visualizing Multidimensional Data

This section is used to visualize the distributions and relationships of biophysical properties and developability assays.

```
 % 12 biophysical assays + 35 ProteinSol properties
 % 5 Therapeutic Antibody Profiler (TAP)
 figure
 scatterhistogram(TheraSabDabML1,'HIC','SMAC')
 figure
 scatterhistogram(TheraSabDabML1,'ACSINS','SMAC')
 %12*12 = 144 such plots! Relationship between all assays

 %TheraSabDabML1.ClinicalTrial = categorical(ClinicalTrial,'Phase-
I','Phase-II','Phase-III','Approved','Preregistration','Phase-
I/II','Phase-II/III','Unknown');
 scatterhistogram(TheraSabDabML1,'HIC','SMAC',...
                     'GroupVariable','ClinicalTrial',...
                     'LegendVisible','on',...
                     'ScatterPlotProportion',0.6,...
```

```
                    'HistogramDisplayStyle',"bar");

 scatterhistogram(TheraSabDabML1,'ACSINS','SMAC',...
                    'GroupVariable','ClinicalTrial',...
                    'LegendVisible','on',...
                    'ScatterPlotProportion',0.6,...
                    'HistogramDisplayStyle',"bar");
 %12*12 = 144 such plots! Clinical Trial status as Legend.

 %Therapeutical Antibody profiler (TAP) plots
 %Creating Scatterplot matrix
 contvars = ["TotalCDRLength", "PSH", "PPC", "PNC", "SFvCSP"];
 X = TheraSabDabTAP:,contvars;
 gplotmatrix(X,[],[],[],("*"),[],[],'hist',contvars)

 %Scatterplot matrix with Clinical Trial legend
 contvars = ["TotalCDRLength", "PSH", "PPC", "PNC", "SFvCSP"];
 X = TheraSabDabTAP2(:,contvars);
 gplotmatrix(X,[],TheraSabDabTAP2.Status,[],("*"),[],[],'hist',contvar
s)

 %Input assay pairs of your choice to assess other plots
```

Part 3: Feature Engineering

To obtain features that are better predictors of developability than original AbPred/TAP variables alone. We apply the following three feature generation approaches: transforming variables, discretization, and summarizing groups.

```
 % Visualizing each assay in a separate plot
   %stackedplot(TheraSabDabML1);
   StackedTheraSabDab = TheraSabDabML1(:,3:14);
   figure();
   stackedplot(StackedTheraSabDab)
   xlabel("TheraSabDab antibody number")
   %ylabel("Developability Assay")

 % k-means clustering algorithm
   %idxC5 =
kmeans(TheraSabDabML1,5,"Distance","sqeuclidean","Replicates",15);
   %[silh5,h] = silhouette(X,idxC5,"sqEuclidean");
   %idxCL5 =
kmeans(TheraSabDabTAB,5,"Distance","sqeuclidean","Replicates",15);
   %[silh5,h] = silhouette(X,idxCL5,"sqEuclidean");
   %clustevQ = evalclusters(X,"kmeans","silhouette","KList",2:6)
   %kmeanbest = clustev.OptimalK

 % Variance Thresholding for Continuous Features
   %figure;
   %bar(V);
```

```
   %hold on; box on; grid on;
   %plot(V,"r.-","LineWidth",2,"MarkerSize",20);
   %hold off;
   %set(gca,"Xtick",1:length(cvars),"XTickLabel",cvars(idx));
   %title("Variance proportion by feature")

 % Principal Component Analysis
  X = StackedTheraSabDab;
  mu = mean(X); % Row vector of column (feature) means
  r = range(X); % Row vector of column ranges
  X = (X-mu)./r; % Scaled feature matrix
  [P,S,V] = pca(X);
  Vnorm = V/sum(V); % Normalize the variance so the total is 1
  figure; hold on;
  bar(Vnorm);
  %stairs(0.5:59.5,cumsum(Vnorm));
  grid on; box on; hold off;
  set(gca,"XTick",0:5:60)
  %legend(Component Variance,"FontSize",12)
  %xlabel(Component)

  VOrigNorm = var(X);
  [VOrigNorm,idx] = sort(VOrigNorm,"descend");
  VOrigNorm = VOrigNorm/sum(VOrigNorm); % Normalize the variance so
the total is 1
  figure; hold on;
  bar(VOrigNorm);
  %stairs(0.5:59.5,cumsum(VOrigNorm));
  grid on; box on; hold off;

%set(gca,"XTick",0.5:59.5,"XTickLabel",tbl.Properties.VariableNames(id
x),"XTickLabelRotation",45)
  set(gcf,"units","normalized","OuterPosition",[0 0 1 1])
  %legend(Cumulative Variance,"FontSize",12,"Location","east")
  %xlabel(Feature)
  %ylim([0,1])

  %figure;

%heatmap(P,"XDisplayLabels","P"+(1:60),"YDisplayLabels",tbl.Properties
.VariableNames,"Title","Component Weights","Colormap",hot(50));
  %set(gcf,"units","normalized","outerposition",[0 0 1 1])

  figure;
  scatter3(S(:,1),S(:,2),S(:,3))
  xlabel("$P^1$","FontSize",16,"Interpreter","latex")
  ylabel("$P^2$","FontSize",16,"Interpreter","latex")
  zlabel("$P^3$","FontSize",16,"Interpreter","latex")
  axis equal;
```

Part 4: Machine Learning

We apply several supervised machine learning approaches for classification of: 1. Clinical Trial Status from Developability Assays/Features and 2. Development Technology from Biophysical Data.

```
 %%Classification 1
 %%Clinical Trial Status
 % Set up the Import Options and import the data
 opts = spreadsheetImportOptions("NumVariables", 20);
 opts.Sheet = "Master Dataset";
 opts.DataRange = "A2:T452";
 opts.VariableNames = ["SNo", "Therapeutic", "HIC", "SMAC", "SGAC",
"CIC", "CSIBLI", "ACSINS", "HEK", "PSR", "ELISA", "BVP", "DSF",
"ACCSTAB", "TotalCDRLength", "PSH", "PPC", "PNC", "SFvCSP",
"ClinicalTrial"];
 opts.VariableTypes = ["double", "string", "double", "double",
"double", "double", "double", "double", "double", "double", "double",
"double", "double", "double", "double", "double", "double", "double",
"double", "categorical"];
 opts = setvaropts(opts, "Therapeutic", "WhitespaceRule", "preserve");
 opts = setvaropts(opts, ["Therapeutic", "ClinicalTrial"],
"EmptyFieldRule", "auto");
 ML1MasterSheet = readtable("D:\Rahul
Khetan\Desktop\ACADEMICS\clubs\Curriculum vitae\Manchester
PhD\Computational Developability assessment\Year 3 work\MATLAB Data
Science and ML\ML1 Master Sheet.xlsx", opts, "UseExcel", false);
 clear opts

 %Classification Learner app
 [trainedClassifier, validationAccuracy] =
trainClassifier(trainingData)
 % Extract predictors and response
 inputTable = trainingData;
 predictorNames = ['HIC', 'SMAC', 'SGAC', 'CIC', 'CSIBLI', 'ACSINS',
'HEK', 'PSR', 'ELISA', 'BVP', 'DSF', 'ACCSTAB', 'TotalCDRLength',
'PSH', 'PPC', 'PNC', 'SFvCSP'];
 predictors = inputTable(:, predictorNames);
 response = inputTable.ClinicalTrial;
 isCategoricalPredictor = [false, false, false, false, false, false,
false, false, false, false, false, false, false, false, false, false,
false];

 % Train a classifier
 % This code specifies all the classifier options and trains the
classifier.
 classificationTree = fitctree(...
     predictors, ...
     response, ...
     'SplitCriterion', 'gdi', ...
```

```
    'MaxNumSplits', 4, ...
    'Surrogate', 'off', ...
    'ClassNames', categorical({'Approved'; 'Phase-I'; 'Phase-I/II';
'Phase-II'; 'Phase-II/III'; 'Phase-III'; 'Preregistration';
'Unknown'}x));

% Create the result struct with predict function
predictorExtractionFcn = @(t) t(:, predictorNames);
treePredictFcn = @(x) predict(classificationTree, x);
trainedClassifier.predictFcn = @(x)
treePredictFcn(predictorExtractionFcn(x));

% Add additional fields to the result struct
trainedClassifier.RequiredVariables = {'ACCSTAB', 'ACSINS', 'BVP',
'CIC', 'CSIBLI', 'DSF', 'ELISA', 'HEK', 'HIC', 'PNC', 'PPC', 'PSH',
'PSR', 'SFvCSP', 'SGAC', 'SMAC', 'TotalCDRLength'};
trainedClassifier.ClassificationTree = classificationTree;
trainedClassifier.About = 'This struct is a trained model exported
from Classification Learner R2021a.';
trainedClassifier.HowToPredict = sprintf('To make predictions on a
new table, T, use: \n  yfit = c.predictFcn(T) \nreplacing ''c'' with
the name of the variable that is this struct, e.g. ''trainedModel''.
\n \nThe table, T, must contain the variables returned by: \n
c.RequiredVariables \nVariable formats (e.g. matrix/vector, datatype)
must match the original training data. \nAdditional variables are
ignored. \n \nFor more information, see <a
href="matlab:helpview(fullfile(docroot, ''stats'', ''stats.map''),
''appclassification_exportmodeltoworkspace'')">How to predict using an
exported model</a>.');

% Extract predictors and response
% This code processes the data into the right shape for training the
% model.
inputTable = trainingData;
predictorNames = ['HIC', 'SMAC', 'SGAC', 'CIC', 'CSIBLI', 'ACSINS',
'HEK', 'PSR', 'ELISA', 'BVP', 'DSF', 'ACCSTAB', 'TotalCDRLength',
'PSH', 'PPC', 'PNC', 'SFvCSP'];
predictors = inputTable(:, predictorNames);
response = inputTable.ClinicalTrial;
isCategoricalPredictor = [false, false, false, false, false, false,
false, false, false, false, false, false, false, false, false, false,
false];

% Perform cross-validation
partitionedModel = crossval(trainedClassifier.ClassificationTree,
'KFold', 3);

% Compute validation predictions
```

```
 [validationPredictions, validationScores] =
kfoldPredict(partitionedModel);

 % Compute validation accuracy
 validationAccuracy = 1 - kfoldLoss(partitionedModel, 'LossFun',
'ClassifError');
```

Machine Learning Classification Algorithms

```
 %%K-Nearest Neighbor(KNN) Algorithm
 classificationKNN = fitcknn(...
     predictors, ...
     response, ...
     'Distance', 'Euclidean', ...
     'Exponent', [], ...
     'NumNeighbors', 100, ...
     'DistanceWeight', 'Equal', ...
     'Standardize', true, ...
     'ClassNames', categorical({'Approved'; 'Phase-I'; 'Phase-I/II';
'Phase-II'; 'Phase-II/III'; 'Phase-III'; 'Preregistration';
'Unknown'}));

 %%Support Vector Machine (SVM) Algorithm
 template = templateSVM(...
     'KernelFunction', 'linear', ...
     'PolynomialOrder', [], ...
     'KernelScale', 'auto', ...
     'BoxConstraint', 1, ...
     'Standardize', true);
 classificationSVM = fitcecoc(...
     predictors, ...
     response, ...
     'Learners', template, ...
     'Coding', 'onevsone', ...
     'ClassNames', categorical({'Approved'; 'Phase-I'; 'Phase-I/II';
'Phase-II'; 'Phase-II/III'; 'Phase-III'; 'Preregistration';
'Unknown'}));

 %%Neural Network Algoritm
 classificationNeuralNetwork = fitcnet(...
     predictors, ...
     response, ...
     'LayerSizes', 25, ...
     'Activations', 'relu', ...
     'Lambda', 0, ...
     'IterationLimit', 1000, ...
     'Standardize', true, ...
     'ClassNames', categorical(['Approved'; 'Phase-I'; 'Phase-I/II';
'Phase-II'; 'Phase-II/III'; 'Phase-III'; 'Preregistration';
'Unknown']));
```

```
 predictorExtractionFcn = @(t) t(:, predictorNames);
 neuralNetworkPredictFcn = @(x) predict(classificationNeuralNetwork,
x);
 trainedClassifier.predictFcn = @(x)
neuralNetworkPredictFcn(predictorExtractionFcn(x));

 %%Naive Bayes Classification
 if any(strcmp(distributionNames,'Kernel'))
     classificationNaiveBayes = fitcnb(...
         predictors, ...
         response, ...
         'Kernel', 'Normal', ...
         'Support', 'Unbounded', ...
         'DistributionNames', distributionNames, ...
         'ClassNames', categorical({'Approved'; 'Phase-I'; 'Phase-
I/II'; 'Phase-II'; 'Phase-II/III'; 'Phase-III'; 'Preregistration';
'Unknown'}));
 else
     classificationNaiveBayes = fitcnb(...
         predictors, ...
         response, ...
         'DistributionNames', distributionNames, ...
         'ClassNames', categorical({'Approved'; 'Phase-I'; 'Phase-
I/II'; 'Phase-II'; 'Phase-II/III'; 'Phase-III'; 'Preregistration';
'Unknown'}));
 end

 predictorExtractionFcn = @(t) t(:, predictorNames);
 naiveBayesPredictFcn = @(x) predict(classificationNaiveBayes, x);
 trainedClassifier.predictFcn = @(x)
naiveBayesPredictFcn(predictorExtractionFcn(x));

 %Contact Rahul Khetan for full MATLAB codes of all ML algorithms
 %Total 38 machine learning models trained on ML1MasterSheet.xlsx
```

2.14 **References:**

1. Jain T, Sun T, Durand S, Hall A, Houston NR, Nett JH, Sharkey B, Bobrowicz B, Caffry I, Yu Y. Biophysical properties of the clinical-stage antibody landscape. Proceedings of the National Academy of Sciences. 2017;114(5):944-9. doi:10.1073/pnas.1616408114.
2. Raybould MI, Marks C, Lewis AP, Shi J, Bujotzek A, Taddese B, Deane CM. Thera-SAbDab: the therapeutic structural antibody database. Nucleic acids research. 2020;48(D1):D383-D8. doi:10.1093/nar/gkz827.
3. Poiron C, Wu Y, Ginestoux C, Ehrenmann F, Duroux P, Lefranc M. IMGT/mAb-DB: the IMGT® database for therapeutic monoclonal antibodies. Poster no101. 2010;11:382.
4. Hebditch M, Warwicker J. Charge and hydrophobicity are key features in sequence-trained machine learning models for predicting the biophysical properties of clinical-stage antibodies. PeerJ. 2019;7:e8199. doi:10.7717/peerj.8199.

5. Uversky VN, Gillespie JR, Fink AL. Why are “natively unfolded” proteins unstructured under physiologic conditions? Proteins: structure, function, and bioinformatics. 2000;41(3):415-27.
6. Linding R, Russell RB, Neduva V, Gibson TJ. GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic acids research. 2003;31(13):3701-8.
7. Costantini S, Colonna G, Facchiano AM. Amino acid propensities for secondary structures are influenced by the protein structural class. Biochemical and biophysical research communications. 2006;342(2):441-51.
8. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. Journal of molecular biology. 1982;157(1):105-32.
9. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2005;67(2):301-20.
10. Drucker H, Wu D, Vapnik VN. Support vector machines for spam categorization. IEEE Transactions on Neural networks. 1999;10(5):1048-54.
11. Ho TK. Random decision forests. Proceedings of 3rd international conference on document analysis and recognition; 1995.
12. Zhang Z, Lai Z, Xu Y, Shao L, Wu J, Xie G-S. Discriminative elastic-net regularized linear regression. IEEE Transactions on Image Processing. 2017;26(3):1466-81.
13. Noble WS. What is a support vector machine? Nature biotechnology. 2006;24(12):1565-7.
14. Segal MR. Machine learning benchmarks and random forest regression. 2004.
15. Raybould MI, Marks C, Krawczyk K, Taddese B, Nowak J, Lewis AP, Bujotzek A, Shi J, Deane CM. Five computational developability guidelines for therapeutic antibody profiling. Proceedings of the National Academy of Sciences. 2019;116(10):4025-30. doi:10.1073/pnas.1810576116.
16. Tatusova TA, Madden TL. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS microbiology letters. 1999;174(2):247-50.
17. Kwok SW, Carter C. Multiple decision trees. Machine intelligence and pattern recognition. 9: Elsevier; 1990. p. 327-35.
18. Randles RH, Broffitt JD, Ramberg JS, Hogg RV. Generalized linear and quadratic discriminant functions using robust estimates. Journal of the American Statistical Association. 1978;73(363):564-8.
19. Patil AS, Pawar B. Automated classification of web sites using Naive Bayesian algorithm. Proceedings of the international multiconference of engineers and computer scientists; 2012.
20. Tang Y, Zhang Y-Q, Chawla NV, Krasser S. SVMs modeling for highly imbalanced classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2008;39(1):281-8.
21. Chang C-L. Finding prototypes for nearest neighbor classifiers. IEEE Transactions on Computers. 1974;100(11):1179-84.
22. Richard MD, Lippmann RP. Neural network classifiers estimate Bayesian a posteriori probabilities. Neural computation. 1991;3(4):461-83.

CHAPTER 3

3 Computational developability assessment framework and guidelines based on clinical-stage antibody therapeutics

3.1 Introduction

Clinical-stage antibodies serve as benchmarks for acceptable biophysical properties desired in antibody therapeutics and represent the practical guidelines for antibody drug candidates in preclinical development. Approved and clinical-stage antibodies, therefore, serve as a reference for successful developability, providing the desired thresholds for biophysical properties. In this chapter, we have leveraged two datasets of clinical-stage antibodies namely – the Jain dataset (137 antibodies) and the TheraSAbDab dataset (658 antibodies) for computational developability assessment.

The main objective of this chapter is to establish benchmark thresholds for biophysical assay performance based on the approved and clinical-stage mAbs which serve as a reference for successful developability. These biophysical assay benchmarks would serve as the computational developability assessment criteria for antibodies. This chapter also aims to evaluate the developability of natural human antibodies and validate the developability potential of human immune repertoire using a case study.

This chapter starts with a general overview of how clinical-stage antibodies can be employed as benchmarks for antibody informatics studies. The two main datasets representative of approved and clinical-stage antibodies used in this study are outlined in section 3.3. Next, two main informatics tools for computational developability assessment benchmarking are discussed in section 3.4. We then provide computational developability assessment results and benchmark thresholds in section 3.5. Next, the developability criteria based on clinical-stage antibodies are compared to the human immune repertoire dataset in section 3.6. A case study on True Human™ antibody therapeutics is presented to demonstrate the unique developability considerations for natural human antibodies. Finally, the developability criteria are summarized and a computational developability assessment framework is proposed in the remainder of the report outlined in section 3.7. Overall, the developability benchmarks established in this chapter will inform future developability predictions of preclinical antibodies and other engineered antibody formats in early R&D and lead optimization stages.

3.2 Methods

Preparation and curation of Clinical-stage datasets: Raw TheraSabDab data was extracted from https://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/therasabdab/ on 6 February 2022. Jain dataset sequence information and 12 assays experimental datasets were extracted from the supplementary information in the Jain publication. Dataset visualization in Figure 10 and Figure 11 were performed using Power BI 2.118. 286.0.

AbPred measurements on clinical-stage datasets: VHVL sequence information for separate clinical-stage datasets were saved as input fasta files. The Abpred predictions were generated from the dockerhub source code available at docker pull maxhebditch/abpred using run command docker run --rm -v $(pwd)/:/abpred/host maxhebditch/abpred. More details at https://hub.docker.com/r/maxhebditch/abpred.

TAP measurements on clinical-stage datasets: We used web sequence submission form and the GitHub repositories at https://github.com/orgs/oxpig/repositories to get the five metric values for 658 input sequences from the TAP tool available at https://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/sabpred. Also, the homology Fv models generated by ABodyBuilder2 were downloaded for future structural analysis.

Kolmogorov-Smirnov test (K-S test) Statistics: The Kolmogorov-Smirnov Test R module from http://www.wessa.net/rwasp_Reddy-Moores%20K-S%20Test.wasp was used to generate the K-S test Statistic and P-value from histogram raw data input.

Creation of Human Immune Repertoire Dataset: The Observed Antibody Space (OAS) database contains annotated immune repertoires that cover one billion sequences from over 80 different studies. OAS is accessible via a web server at https://opig.stats.ox.ac.uk/webapps/oas/. The OAS search was customized for healthy non-vaccinated human immune repertoires. The OAS search returned 350,980 filtered paired sequences from two studies – Eccles and Jaffe. The same procedure of AbPred and TAP predictions was carried out on this human repertoire dataset for comparison.

Principal Component Analysis (PCA) of biophysical assay features: The PCA analysis was carried out in MATLAB using the command coeff = pca(X) that returns the principal component coefficients, also known as loadings, for the n-by-p data matrix X. Each column of coeff contains coefficients for one principal component, and the columns are in descending order of component variance. By default, pca centres the data and uses the singular value decomposition (SVD) algorithm.

3.3 **Datasets representing clinical-stage antibody therapeutics**

3.3.1 Jain Dataset: Biophysical performance of clinical-stage antibodies

The Jain dataset[1] contains experimental assay measurements across 12 different biophysical characterization platforms for 137 late-stage clinical therapeutics. Jain Dataset comprises of 137 clinical-stage antibodies including 48 approved antibodies. Jain *et al.* have provided experimental biophysical measurements on a dozen assays commonly used to evaluate the 'developability' of antibodies.[1] These observed metrics for clinical-stage antibodies serve as benchmarks for acceptable biophysical properties desired in antibody therapeutics and represent the practical guidelines for antibody drug candidates in preclinical development.

The Jain dataset has 48 approved antibodies, 46 phase II stage antibodies, and the remaining 43 antibodies out of the total 137 antibodies are in phase III clinical trials which provides a balanced representation of approved and advanced clinical stage antibodies. The most common mAb isotype is the "IgG1" format with a total of 87 antibodies out of 137 total (63.5%). Other mAb formats captured in Jain dataset are "IgG2" (21), "IgG4" (23), and six other formats such as "IgM", "Fab" and "scFv-Fc". Therefore, novel synthetic formats are underrepresented in this dataset. The dataset has 124 antibodies of kappa (κ) isotype, while the remaining 13 antibodies have lambda (λ) light chains. The Jain clinical-stage antibodies originate from diverse sources – 58 (42%) are classified as "fully human" (with -UMAB suffix), 67 (49%) are "humanized" (with –ZUMAB suffix), and 12 (9%) have "chimeric" origin with at least one fully non-human variable region (with –XIMAB or –XIZUMAB suffix).

The twelve biophysical characterization assays evaluated in the Jain dataset are well-established for therapeutic antibody characterization. It is likely that these or related assays are generally used in the early stages of an antibody discovery cascade. An overview and the biophysical significance of each assay have been provided below:

Hydrophobic Interaction Chromatography (HIC): The hydrophobic interaction chromatography (HIC) is a powerful technique used for the analytical characterization of monoclonal antibodies in a panel of developability assessments and other liquid chromatographic applications. HIC is a good technique for separating the different populations of antibody molecules while maintaining biological activity due to the use of conditions and matrices that operate under less denaturing conditions. So, the main

advantage of HIC compared to other known chromatography procedures is that it is non-denaturating, so the native forms of the proteins are expected to be maintained.

HIC separates molecules based on differences in their surface hydrophobicity.[2-4] The HIC media are composed of alkyl or aryl ligands coupled to an inert, porous matrix which are then packed into a chromatography column in a packed bed arrangement. Samples are loaded on the column using a mobile phase with a high salt concentration and then salt concentration is gradually lowered to start eluting proteins. HIC utilizes a reversible interaction between the proteins and the hydrophobic ligand of a HIC resin. The most hydrophobic protein is bound to a hydrophobic ligand on the HIC resin, so the protein with the lowest degree of hydrophobicity is eluted first. Therefore, the manipulation of salt gradients allows differential elution of the proteins where the antibodies are separated in the order of increasing surface hydrophobicity.

The Jain experimental values for HIC are measurements of the HIC retention time in the order of minutes. Therefore, a lower retention time corresponding to a low degree of hydrophobicity is desired for an antibody drug candidate. A high hydrophobic interaction chromatography assay value is unfavourable and represents potential concerns regarding the high hydrophobicity that may lead to formation of aggregates and other consequences of sticky hydrophobic interactions.

Standup Monolayer Absorption Chromatography (SMAC): Standup monolayer absorption chromatography (SMAC) is an alternative high-performance liquid chromatography (HPLC) based screening method to assess non-specific interactions and other correlated developability factors.[5] The particular resin in a SMAC assay is a hydrophobic standup monolayer with terminal hydrophilic groups that cause delayed retention of ‘sticky’ antibodies. Any non-specific interactions of injected antibodies with the column matrix cause high retention time in standup monolayer adsorption chromatography (SMAC). So, the antibodies prone to precipitation or aggregation are retained longer on the column with broader peaks. The retention times in SMAC assay for antibodies are inversely related to their colloidal stability. A previous study has shown that the CamSol scores show a strong correlation with the SMAC measurements.[6] Therefore, SMAC is an indirect method for predicting protein solubility and generic aggregation propensity for an antibody drug candidate.

Salt-Gradient Affinity-Capture Spectroscopy (SGAC): The Salt-gradient affinity capture self-interaction nanoparticle spectroscopy (SGAC-SINS) is an assay to quantify the interaction between antibodies bound to the surface of gold nanoparticles by measuring the wavelength shift with varying ranges of ammonium sulfate (300 – 1,000 mM in 100-mM steps). The SGAC100 is obtained by graphing the wavelength shifts of a sample against the ammonium sulfate concentration and extrapolating the concentration at which the shift was 560 nm. For antibodies where the shifts are below 560 nm at the highest salt concentration a value of 1,000 mM is assigned. SGAC assay measures the salt concentration (mM) as an indirect estimate of the wavelength shift. The mAbs with low self-association have lower wavelength shifts and high salt concentration values. Therefore, high salt concentration (mM) measurements reflect optimal biophysical profiles for an antibody candidate while low SGAC assay values correspond to unfavourable developability properties. This trend is opposite to that observed in other assays where low assay values implied better biophysical properties.

Cross-Interaction Chromatography (CIC): The cross-interaction chromatography (CIC) assay is designed to measure the weak cross-interactions of a mAb with polyclonal human serum antibodies that are bound to the stationary phase of a chromatography resin.[7] Antibodies with high cross-interaction propensity due to exposed interaction-prone surfaces are eluted later. Such antibodies with high retention time are likely to interact with several different *in vivo* targets and represent antibodies with a low degree of specificity.

Low retention times are desired for antibody drug candidates in a CIC assay. So, a high CIC assay value demonstrates an unfavourable developability profile. A previous study has established strong correlations between cross-interaction assay retention time measurements and clearance rates for human IgG1 antibodies.[8]

Clone Self-Interaction by Biolayer Interferometry (CSI BLI): A reliable assay for real-time observation of self-association and dissociation of antibodies has been clone self-interaction-biolayer interferometry (CSI BLI). It is a high-throughput method that uses a label-free technology to measure self-interaction for mAbs.[9] Control antibodies are loaded onto a biosensor tip followed by directional capture of antibody Fc region. Next, the binding response of the mAb is captured by an internal reflection interference pattern. The binding response from the association step is subtracted from

that of reference clinical-stage IgG - adalimumab. Finally, the interference pattern shifts by an amount proportional to the change in thickness of the biological layer when self-interaction occurs. Therefore, CSI BLI allows for direct monitoring of antibody self-binding rather than relying on the cumulative effects of interactions on retention time in the chromatographic methods.

CSI BLI assay is measured in terms of BLI response units. Antibodies with low degree of self-association generate a low self-binding response in this assay and generally have high solubility. On the other hand, antibodies with strong self-association behaviour have high BLI response unit values and exhibit poor solubility. Therefore, low CSI BLI assay values are desired for antibody drug candidates in preclinical stages as low BLI assay values correspond to optimal developability profiles.

Affinity-Capture Self-Interaction Nanoparticle Spectroscopy (AC-SINS): Another powerful technique targeting self-interaction measurements for selection of mAbs with excellent biophysical properties during early antibody discovery is the affinity-capture self-interaction nanoparticle spectroscopy (AC-SINS). In this assay, gold nanoparticles coated with anti-Fc or polyclonal antibodies specific for human mAbs are added to dilute antibody solutions. Next, the wavelength of maximum absorbance for antibody–gold conjugates is measured that shifts as the distance between particles is reduced due to attractive mAb self-interactions. The mAbs prone to self-association cause clustering of nanoparticles, which can be monitored by plasmon wavelength shift.[10, 11] Antibodies with low degree of self-association have lower plasmon wavelength shift in AC-SINS assay, therefore, lower AC-SINS assay values are desired of therapeutic antibodies while very high AC-SINS assay values indicate potential self-interaction liabilities for antibody drug candidates.

Expression Titer in HEK cells (HEK): The degree of expression of an antibody in human embryonic kidney cells is representative of the ease of scalability for high-throughput *in vitro* manufacture of the antibody. A high expression level corresponding to a high concentration of the harvested supernatant is desired for any antibody drug candidate. Therefore, a low HEK assay value indicates an antibody with poor expression levels with unfavourable developability.

Poly-Specificity Reagent (PSR): The poly-specificity reagent (PSR) binding assay evaluates the antigen binding properties and non-specific interactions for an

antibody.[12] Overall, this method leverages the high-throughput capacity of flow cytometry to profile antigen and non-specific binding in parallel. This approach employs a polyspecificity reagent (PSR) that is generated via biotinylation of soluble membrane proteins from eukaryotic cell lines. This polyspecificity reagent is incubated with IgG-presenting yeast, followed by washing. Next, fluorescence-activated cell sorting (FACS) is used to determine the median fluorescence intensity (MFI). Finally, the MFI values were normalized from 0 to 1 based on three reference antibodies exhibiting low, medium, and high PSR MFI values. Polyspecificity is a highly undesirable property that has been linked to poor antibody pharmacokinetics. A high MFI value in a PSR assay corresponds to polyspecificity and non-specific binding. So, a low PSR assay value is desired for an antibody drug candidate while high PSR values are flags for non-specific antigen binding.

Enzyme-Linked Immunosorbent Assay (ELISA): Enzyme-linked immunosorbent assay (ELISA) is a very popular assay that is widely used to identify antibody-antigen binding. ELISA is a labelled immunoassay to analyze antibody-antigen interactions.[13] In an ELISA protocol, a set of antigens are fixed on the wells of the plate. Next, antibody candidate coupled with enzyme is added into the plate wells and incubated for about an hour before washing step. Finally, the fluorescence intensity of the plate is detected by an absorbance reading at 450 nm after an appropriate substrate is added to the sample. A high ELISA assay value represents multiantigen nonspecificity issues. Therefore, a low ELISA assay value is desired for a therapeutic antibody drug candidate while high ELISA values are indicative of polyspecificity liabilities.

Baculovirus Particle (BVP) assay: The baculovirus particle enzyme-linked immunosorbent assay is a similar ELISA-based approach that uses the membrane proteins presented on the surface of a baculovirus particle (BVP) as a reagent to capture mAbs with cross-interaction propensity.[14] A low BVP assay value is desired in this polyspecificity screening assay for antibody drug candidates. Kelly *et al.* have demonstrated strong correlations of non-specificity ELISA binding score using baculovirus particles (BVP) with PSR assay and antibody clearance rates in humans.[8] Therefore, BVP ELISA is closely linked to other assays measuring cross-reactivity and polyspecificity of antibodies.

Differential Scanning Fluorimetry (DSF): Differential scanning fluorimetry (DSF) is a powerful high-throughput assay widely used to evaluate the thermal stability of an antibody. DSF measures protein unfolding by monitoring changes in fluorescence as a function of temperature. Firstly, the target antibody is mixed with a fluorescent dye specific for hydrophobic regions such as SPYRO orange. Next, the temperature is increased and as the protein begins to unfold, buried hydrophobic residues become exposed and the level of fluorescence is measured to get the melting temperature (Tm). An antibody with low thermal stability will likely spontaneously unfold and become immunogenic. Therefore, a high Tm is desired for any antibody drug candidate while low DSF assay values represent low thermodynamic stability.

Accelerated Stability (ACC STAB): The percentage of monomeric species assessed by size-exclusion chromatography (SEC) in the context of an accelerated stability study (ACC STAB) is a reliable technique to measure the size, aggregation propensity, and long-term stability of antibodies. In ACC STAB, samples are stored for 30 days and analyzed by SEC at separate intervals to detect the fraction of monomeric protein. Finally, the long-term stability slope is calculated from the percent aggregated measured on the SEC. Size exclusion chromatography separates antibody molecules based on their size. The monomeric antibodies are trapped in the stationary phase pore system while the aggregated antibodies will flow through the column more rapidly.

ACC STAB assay measurement for the Jain dataset is the long-term stability slope. So, a low ACC STAB assay value is desired for antibody drug candidates as a low slope value represents minimal aggregation and long-term stability.

3.3.2 TheraSabDab: A database of clinical-stage antibody therapeutics

The Therapeutic Structural Antibody Database (TheraSabDab) tracks all antibody and nanobody-related therapeutics recognized by the World Health Organisation (WHO) with accompanying metadata. The TheraSabDab is available as a free online web server at https://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/therasabdab/search/.

TheraSabDab Dataset contains the names, sequences, molecular formats, and clinical trial status information for 658 clinical-stage antibody-based biotherapeutics as of 2023. Of them, 96 (14.5%) have been “approved” for human use, while it has 123 (18.7%) antibodies in “phase-I” trials, 29 (4.4%) antibodies in “phase-I/ II” trials, 230

(34.9%) antibodies in “phase-II” trials, 12 (1.8%) antibodies in “phase-II/ III” trials and 84 (12.8%) antibodies in phase-III clinical trials. The remaining antibodies are either being investigated in uncategorized clinical trials or have unknown clinical trial status due to simultaneous trials for multiple indications. Figure 10 summarizes the distribution of antibodies as per their clinical trial status in TheraSabDab.

In the TheraSabDab dataset, 584 (89%) are full-length monoclonal antibodies (Whole mAb) and 74 (11%) are antibody fragments in several molecular formats such as Fab Fusion antibodies, bispecific scFvs and nanobodies. The clinical trial status for different antibody formats is shown in Figure 11. Bispecific antibody formats such as bispecific mAbs and bispecific scFvs are mostly in Phase-I or Phase-II early-stage clinical trials due to the growing popularity of bispecific antibody formats in recent years. Engineered scFvs have the highest approval rate (25%) among all antibody formats which demonstrates the higher clinical efficiency achieved for these formats due to customized optimization of developability liabilities.

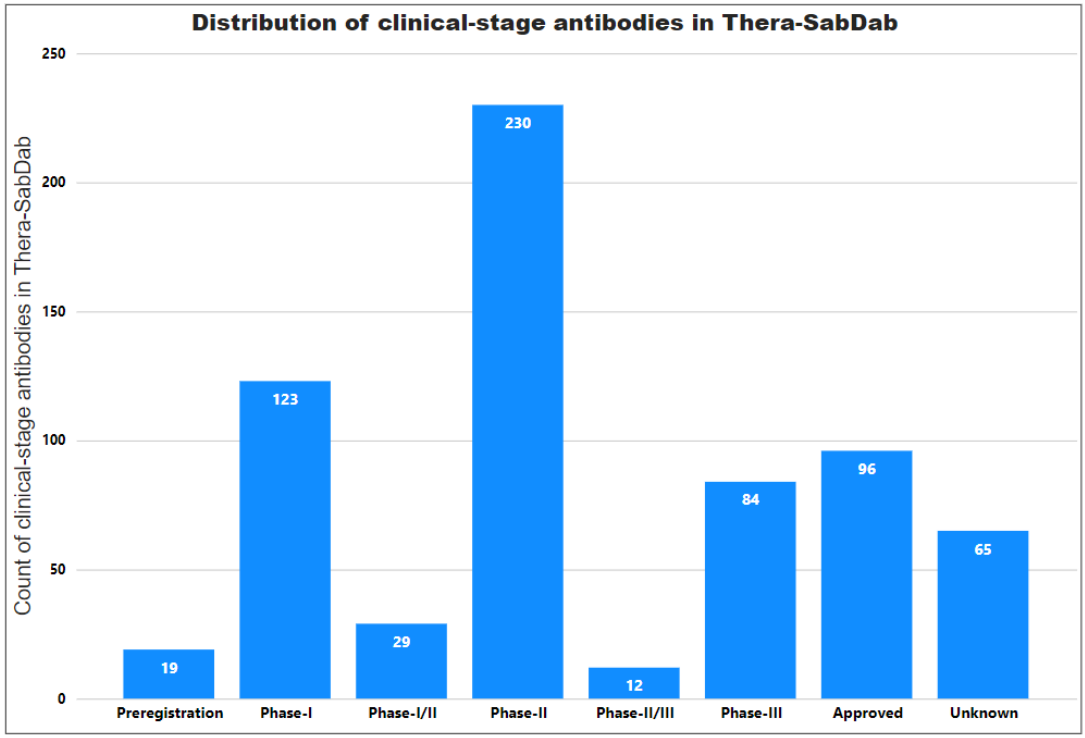


Figure 10: Clinical Trial Status of the TheraSabDab dataset of 658 clinical-stage antibodies and antibody fragments. (TheraSabDab dataset as of February 2023).

The TheraSabDab biotherapeutics in advanced clinical stages are being investigated for treatment in several therapeutic areas including oncology, autoimmune disorders, infectious diseases, and chronic diseases, among others. They have diverse molecular

targets which provide a comprehensive representation of all the possible applications of antibody therapeutics.

The TheraSAbDab is updated whenever a new WHO International Non-proprietary Name (INN) list is released, adding all therapeutics with an accompanying variable domain sequence. Also, the clinical trial status is updated for all actively-developed therapeutics according to the latest updates on AdisInsight. The up-to-date lists of therapeutic sequences with metadata are available online at the TheraSAbDab search page. It also provides additional functionality by allowing users to search for sequence identity to other known therapeutics in the Structural Antibody Database (SAbDab).

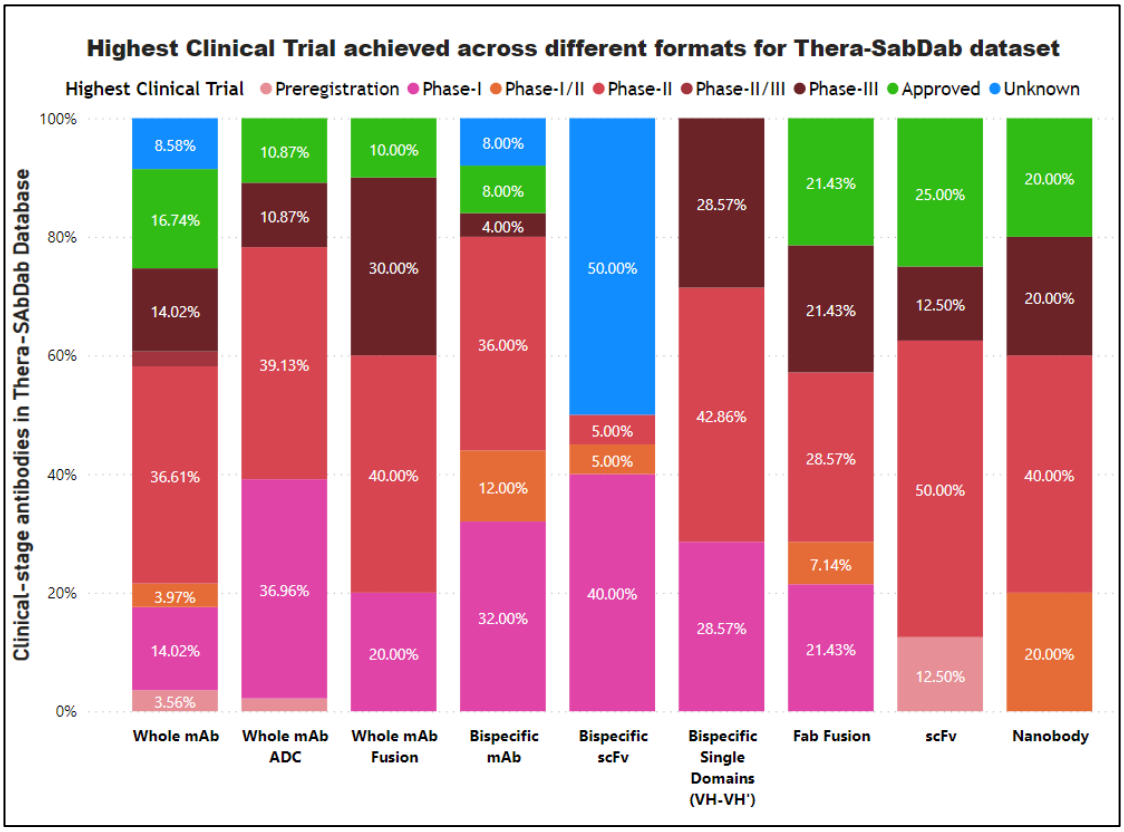


Figure 11: Distribution of clinical-stage antibodies in TheraSabDab database according to their clinical trial status for different antibody formats shown along the x-axis. (Feb 2023).

3.4 **Antibody informatics tools for evaluating clinical-stage mAbs**

3.4.1 AbPred – Machine learning algorithms on the Jain dataset

AbPred[15] utilizes machine learning algorithms trained on the Jain dataset. AbPred tool is based on sequence trained models for all 12 biophysical measurements which make it a reliable antibody informatics tool for predicting the performance on biophysical

characterization platforms. Finally, the tool also provides a meta score that provides an average rank by combining scaled rankings from multiple biophysical platforms.

Hebditch *et al.* have previously described the 35 sequence features for predicting protein solubility from sequence in Protein–Sol tool.[16] These 35 features are composed of 20 amino acid compositions; 7 composite scores of amino acid combinations (KmR = K-R, DmE = D-E, KpR = K+R, DpE = D+E, PmN = K+R-D-E, PpN = K+R+D+E, aro = F+W+Y); and 8 other sequence features (length, pI, kyte-doolittle hydropathy, absolute charge at pH 7, fold propensity, disorder propensity, sequence entropy, and β-strand propensity. So, in addition to charge-based features like KmR, DmE, DpE, PmN, and PpN the model also contains non-polar features such as aromatic (F+W+Y) composition and eight known sequence property features. The Abpred tool is trained on 12 biophysical characterization assays for 137 clinical-stage antibodies in the Jain dataset based on these same sequence features from the Protein-Sol tool.

The algorithms with the lowest mean average error were chosen for each assay based on training on sequence composition scores of 137 Jain dataset variable region sequences in CDRs and transformed experimental measurements provided by the Jain dataset. Overall, the HIC, SMAC, and AC-SINS have been trained on Elastic net algorithms, ELISA and BVP have been captured using Random Forest algorithms while the remaining assays have been trained using Support Vector Machine (SVM) algorithms. A summary of chosen machine-learning algorithm for each assay has been provided in Table 3. The AbPred predictions of 12 biophysical assays for Jain dataset antibodies are reliable for all assays except DSF and ACC STAB. The AbPred predictions for enzyme-linked immunosorbent assay (ELISA) and baculovirus particle (BVP) assay have the highest R^2 values of 0.9215 and 0.8963 respectively.

The AbPred tool to predict these assay values from variable chain sequence is now openly available as a web application at https://protein-sol.manchester.ac.uk/abpred.

3.4.2 TAP – Five developability properties based on TheraSabDab

The Therapeutic Antibody Profiler[17] (TAP) is a computational tool for comparing selected antibody variable domain structural properties with post phase-I clinical-stage antibodies. TAP tool utilizes five developability guidelines based on variable region

properties derived from clinical-stage therapeutics in the TheraSabDab database. The tool is available at https://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/sabpred/tap.

It uses antibody variable domain sequence (both heavy and light chain sequence) input to generate a structural variable fragment (Fv) model using structural modelling algorithms based on deep-learning known as ABodyBuilder.[18] Next, the TAP tool calculates five structural measures linked to poor developability namely - Total CDR Length (L), Patches of Surface Hydrophobicity Metric (PSH), Patches of Positive Charge Metric (PPC), Patches of Negative Charge Metric (PNC) and Structural Fv Charge Symmetry Parameter (SFvCSP). Finally, it compares these scores against guideline thresholds of 658 clinical-stage therapeutic Fv domains. It also reports potential sequence liabilities that relate to post-translational modifications, non-CDRH3 loop canonical forms, and 3D structural visualization of the variable regions.

The CDR length feature links to developability insights as it can capture the binding-site shape and CDRH3 loop features. The Patches of Surface Hydrophobicity Metric (PSH) indicates the level of clustering of hydrophobic residues with high scores being representative of large hydrophobic patches in the CDR Vicinity. The CDR Vicinity encompasses all surface-exposed IMGT CDR and anchor residues, as well as other surface exposed residues with at least one heavy atom within a radius of 4Å. The PSH metric thus provides a measure of hydrophobicity in the CDR regions which can be used to estimate aggregation propensity of antibodies. The surface charge metrics used here are patches of positive charge (PPC) and patches of negative charge (PNC) which represent the normalized sum of charged residues for a target protein sequence. Both these charge-based metrics are linked to several developability related measures such as colloidal stability, polyspecificity, and self-association. Finally, the Structural Fv charge symmetry parameter (SFvCSP) is the product of surface-exposed net charges of V_H and V_L chains. A previous study by Sharma *et al.* has revealed that oppositely charged V_H and V_L chains are linked to high viscosity and poor biophysical performance.[19] Therefore, highly negative SFvCSP scores indicate that the target antibody may exhibit charge asymmetry and poor developability.

The correlation between both the antibody informatics tools are captured in Table 4 which are calculated from the TheraSabDab dataset. Overall, there are no major direct correlations that suggest that these two tools are independent and measure different

features. However, as expected for biophysical property measurements we observed overlap among certain features. Firstly, the CDR length and PSH metric from the TAP tool are fairly correlated to the HIC and SMAC assays from the AbPred tool. This is expected as all these features are known to be dependent on hydrophobicity and other biophysical behaviour related to stickiness of an antibody. Next, the SFvCSP feature from TAP tool is somewhat predictive of self-association and cross-reactivity assays namely CIC, CSI BLI, AC SINS, and PSR assays from the AbPred tool. A possible explanation is that the surface-exposed net charges of the variable regions in SFvCSP calculation are decided by the charged amino acid composite scores which have a high weightage in AbPred model for most of self-association and cross-reactivity assays.

	Total CDR Length	*PSH*	*PPC*	*PNC*	*SFvCSP*
HIC	0.345598	0.254939	-0.17925	-0.10234	-0.27599
SMAC	0.407831	0.318361	-0.18971	-0.04017	-0.38347
SGAC	-0.09477	-0.1126	-0.10293	0.150299	-0.26735
CIC	0.035274	0.12275	0.116725	-0.42885	0.418167
CSI BLI	-0.16406	-0.05696	0.12726	-0.06966	0.383577
AC SINS	0.027678	0.136738	0.159366	-0.22365	0.391554
HEK	-0.0309	0.021705	-0.00182	0.009956	-0.00791
PSR	-0.12559	-0.02575	0.202852	-0.1974	0.568225
ELISA	0.148152	0.059025	0.138027	0.041761	0.255604
BVP	0.282374	0.174563	0.113287	0.057952	0.190414
DSF	0.087722	0.004956	0.03165	0.052377	0.102555
ACC STAB	0.09367	0.084402	-0.11669	0.0011	-0.00151

Table 4: Correlation between AbPred assay scores and TAP five metric scores for mAbs. The 12 biophysical assays are presented as rows and 5 TAP metrics are presented as columns. The five TAP metrics used here are Total CDR Length, Patches of Surface Hydrophobicity Metric (PSH), Patches of Positive Charge Metric (PPC), Patches of Negative Charge Metric (PNC) and Structural Fv Charge Symmetry Parameter (SFvCSP).

Finally, the total CDR length feature from TAP tool is fairly correlated to most liquid chromatography assays in AbPred. This high influence of antibody sequence length derived features on the chromatography assays can be because larger molecules have more binding regions and are retained longer in the chromatography column. Overall,

we conclude that a combination of sequence and structural features influence the target monoclonal antibody biophysical performance on all major developability assays.

3.5 **Developability criteria based on clinical-stage antibodies**

In our effort to establish suitable developability criteria for clinical-stage antibodies, we tested multiple biopharmaceutical informatics tools listed in Table 2. The search for right antibody informatics tools for benchmarking clinical mAbs was challenging. AbPred and TAP emerged as reliable tools in our analysis as they are unambiguous, appropriate, complete, and reflective of biophysical performance.[15, 17] So, firstly, we used AbPred tool predictions on the larger TheraSabDab dataset of 658 clinical-stage antibody therapeutics to establish the thresholds of biophysical assay performance.

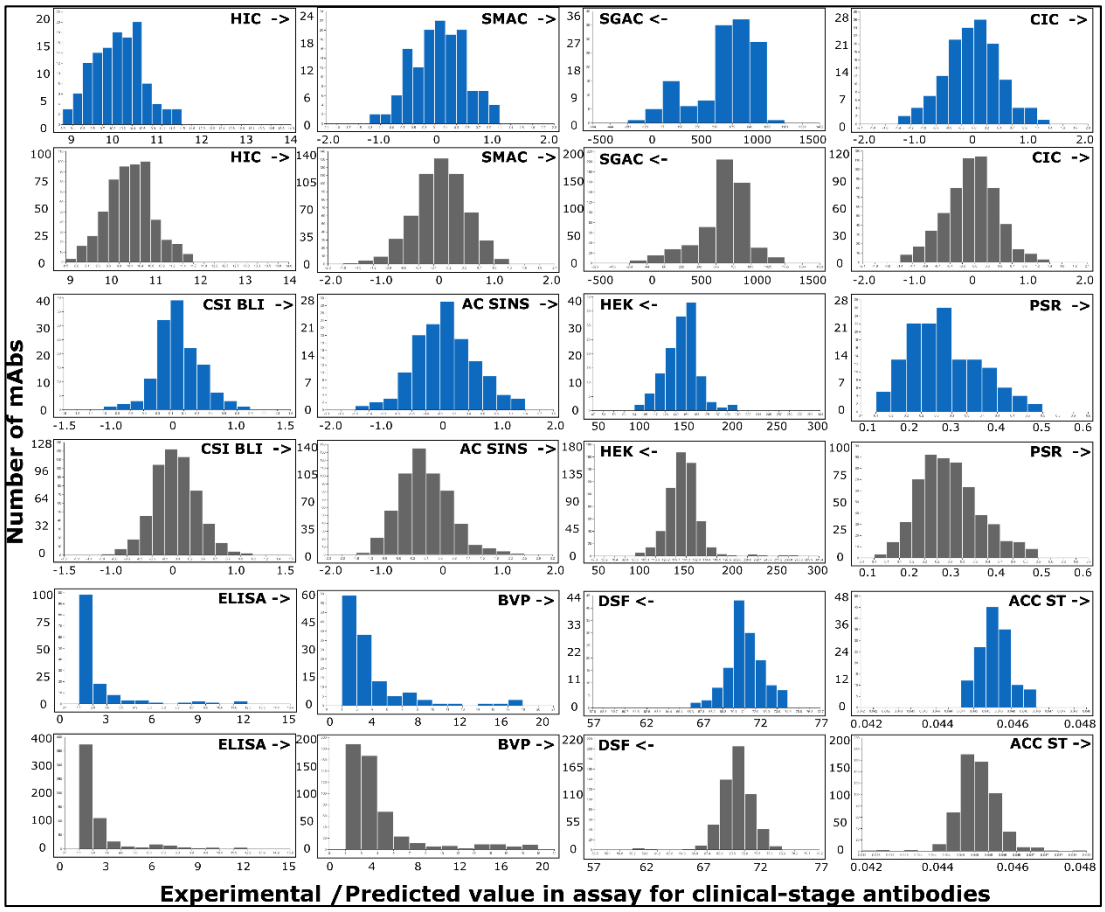


Figure 12: Histograms of 12 biophysical assay values for clinical-stage antibodies. The blue histograms represent the experimental values (Abpred transformed) of 137 Jain antibodies, while the grey histograms represent the assay values of 658 TheraSabDab antibodies. The arrows next to assay names indicate the direction of unfavourable values for each assay.

We observe that the histogram distributions of all 12 biophysical characterization assays are asymmetrically long-tailed for the clinical-stage antibodies dataset. Most of these distributions are asymmetrically long-tailed in the unfavourable direction as

shown in the Figure 12. Here, an arrow at the top corner of each chart represents the direction of unfavourable values. A high assay value is desired for three assays namely - SGAC, HEK, and DSF indicated by a left arrow while a low assay value is favourable for the remaining nine assays shown by a right-facing arrow in Figure 12.

Most importantly, these distributions are consistent between the Jain dataset of 137 antibodies (shown in blue) and the TheraSabDab dataset of 658 antibodies (shown in grey). Previous studies have demonstrated the accuracy and precision of the AbPred tool with the Jain dataset experimental measurements which confirm the consistency of AbPred assay predictions with the observed experimental reality.[15] Our approach in this study has been to use these reliable assay predictions on a larger dataset of TheraSabDab clinical-stage antibodies which is much more comprehensive compared to the Jain dataset. This approach provides a strong rationale and evidence to reliably benchmark clinical-stage therapeutic antibodies for establishing a new computational developability assessment focusing on antibody therapeutics.

We used the Kolmogorov-Smirnov test (K-S test) for statistical validation of our result that the TheraSabDab distributions are consistent with the Jain experimental results.[20]

Assay	K-S Test Statistic	P-value	Mean Jain dataset	Mean TheraSabDab	Difference
HIC	0.094891	0.56814	10.116	10.047	0.068838
SMAC	0.065693	0.92899	0.01022	-0.038099	0.048319
SGAC	0.168274	0.05872	610.46	611.11	-0.64415
CIC	0.062485	0.91572	-0.06228	-0.01872	-0.043554
CSI – BLI	0.11679	0.30756	-0.00061	0.02032	-0.020932
AC SINS	0.13942	0.11846	0.023368	-0.038516	0.061884
HEK	0.10219	0.47177	146.84	146.43	0.41275
PSR	0.11679	0.30756	0.27655	0.28993	-0.013382
ELISA	0.092483	0.54925	1.9909	2.0959	-0.10497
BVP	0.12879	0.22365	3.5762	3.6255	-0.049276
DSF	0.13139	0.18775	71.109	70.743	0.3664
ACC STAB	0.12574	0.04522	0.045229	0.045156	7.2886e-05

Table 5: Kolmogorov-Smirnov test statistical analysis result for comparison of histograms. A high P-value is desired which proves the consistency between the two histogram distributions. All assays except ACC STAB have a K-S test p-value > 0.05 which prove consistency between the Jain dataset (actual experimental reality) and the TheraSabDab dataset calculations.

The two-sample Kolmogorov-Smirnov test is a well-recognized nonparametric test that compares the cumulative distributions of two data sets. The null hypothesis in the K-S test is that both input data groups were sampled from populations with identical distributions. It tests for any violation of that null hypothesis in terms of differences in medians, different variances, or different distributions. So, a high P-value is desired while a small P-value means that the two input data groups represent populations with different distributions which may differ in median, variability, or the shape of the distribution. The used R module based statistical tool[21] provided the following results shown in Table 5. Most of the assays have a high P-value as desired which proves the consistency between the two histogram distributions. Only the SGAC and ACC STAB assays have small P-values which indicate substantial differences between shape and spread for Jain and TheraSabDab dataset assay output histograms in these two assays. These may be due to lower AbPred prediction accuracy for these assays. However, the P-values for these two assays are not too far from the significance threshold of 0.05.

We used 10% and 5% as the two guideline cutoffs for each assay. This was to capture the extremes of each histogram distribution. Here, a 10% cutoff is equivalent of a soft threshold while a 5% cutoff is hard threshold to flag mAbs with developability issues. These cutoffs have been provided in the Table 6. A high difference between the 5% cutoff and 10% cutoff indicates an assay distribution with high kurtosis: a measure of tailedness or the likelihood of extreme outcomes. This high kurtosis is observed for SGAC, ELISA, and BVP distributions which concludes that these assays have wide tail distributions and are the most sensitive to selection of the % criteria for the cutoffs.

Biophysical Assay	**10% Cutoff**	**5% Cutoff**
Hydrophobic Interaction Chromatography (HIC)	HIC > 10.8482	HIC > 11.1490
Standup Monolayer Absorption Chromatography (SMAC)	SMAC > 0.6047	SMAC > 0.7943
Salt-Gradient Affinity-Capture Spectroscopy (SGAC)	SGAC < 234.1128	SGAC < 78.0115
Cross-Interaction Chromatography (CIC)	CIC > 0.6425	CIC > 1.0554
Clone Self-Interaction by Biolayer Interferometry (CSI BLI)	CSI BLI > 0.4564	CSI BLI > 0.5635
Affinity-Capture Self-Interaction Nanoparticle Spectroscopy (AC-SINS)	AC-SINS > 0.9858	AC-SINS > 1.5725
Expression Titer in HEK cells (HEK)	HEK < 128.2454	HEK < 121.6744
Poly-Specificity Reagent (PSR) assay	PSR > 0.4204	PSR > 0.4396
Enzyme-Linked Immunosorbent Assay (ELISA)	ELISA > 3.8382	ELISA > 6.8259
Baculovirus Particle (BVP) assay	BVP > 7.4012	BVP > 14.1383
Differential Scanning Fluorimetry (DSF)	DSF < 68.7440	DSF < 68.0450
Accelerated Stability (ACC STAB)	A STAB > 0.0458	A STAB > 0.0461

Table 6: Computational Developability assessment criteria for clinical-stage antibodies based on Abpred biophysical assay thresholds. Worst 10% cutoff and 5% cutoff values are provided.

We tested our hypothesis that the therapeutic antibodies that fall within the 10% cutoff (and 5% cutoff) values exhibit optimal developability on trastuzumab (Herceptin). We have used trastuzumab as a reference because it is well-recognized among industry professionals to possess good biophysical properties and optimal developability. The assay values for trastuzumab are respectively 9.9091 (HIC); -0.0191 (SMAC); 763.3616 (SGAC); -0.1244 (CIC); 0.0953 (CSI BLI); 0.0267 (AC SINS); 157.3354 (HEK); 0.2681 (PSR); 1.3187 (ELISA); 1.7005 (BVP); 72.8277 (DSF) and 0.0452 (ACC STAB). All of these assay values are well within the threshold limits provided in Table 6. We also have further validated these threshold limits with other datasets of engineered antibodies, human antibodies, and failed antibodies in later chapters.

The approved and clinical-stage antibodies capture a broad spectrum of biophysical properties. Envafolimab; a single domain antibody against programmed death ligand 1 (PD-L1) fused with human Fc, has the lowest hydrophobicity prediction among our results with the lowest HIC score of 2.91. Also, Envafolimab has the lowest SMAC score of – 5.94 which indicates lowest aggregation propensity and a very high colloidal stability. These assay scores explain how Envafolimab became the first and only globally approved subcutaneously injectable PD-L1 antibody therapeutic as optimal hydrophobicity and colloidal stability features are supportive of the subcutaneous administration. While, on the contrary, Lesofavumab; a monoclonal antibody for the treatment of Influenza B infection has the highest HIC score of 11.68.

Similarly, a detailed analysis of other assays revealed important insights and safety information. For instance, teclistamab (Tecvayli), a therapy for relapsed or refractory multiple myeloma had the maximum CIC score of 1.40 which obviously lies above the 5% cutoff in our benchmark (CIC > 1.0554). In a previous phase 1/2 study, teclistamab demonstrated frequent grade 3 and above common adverse events such as cytokine release syndrome (72.1% patients); neutropenia (70.9% patients); anemia (52.1% patients); thrombocytopenia (40.0% patients) and infections (76.4% patients) with five deaths overall related to teclistamab.[22] The observed cytokine release syndrome and immune effector cell-associated neurotoxicity are most likely due to high cross-reactivity and off-target binding of teclistamab to immune cells. Thus, our developability assessment could have flagged such therapeutics due to breach of 5% threshold in CIC assay and saved phase 1-2 trial cost and more importantly lives of those five patients in the study. Teclistamab was also flagged by other related assays

like CSI BLI and AC SINS where the biophysical assay scores were among the worst 5% scores among all clinical-stage therapeutic antibodies.

Next, we have calculated 10% cutoff and 5% cutoff values for the five developability metrics proposed by Therapeutic Antibody Profiler (TAP).[17] These threshold cutoffs for CDR length, PSH, PPC, PNC, and SFvCSP metrics are presented in Table 7. Our benchmarking is different from the amber flag region and red flag region cutoffs proposed in previous work by Raybould *et al*. The red flag cutoff represents the maximum or minimum value observed while the amber flag region represents a region of extreme outcomes equivalent to our 5% cutoff. We have changed the benchmarking framework as their red flag benchmarking is too strict, where the use of maximum or minimum values is very sensitive to the clinical stage mAb outliers and may also lead to many false negatives. The biophysical assay measures for new mAb candidates in advance stages are often overlapping with the observed ranges for clinical-stage mAbs as the drug discovery and optimization process is derived from clinical-stage mAb backbones with the same underlying main building blocks. So, the use of 5% and 10% cutoffs instead of amber and red flags would ensure that we don't miss out on many promising candidates in our computational developability screening process for mAbs.

TAP Property	10% Cutoff	5% Cutoff
Total CDR Length (L)	53 ≤ L or L ≤ 44	54 ≤ L or L ≤ 42
Patches of Surface Hydrophobicity (PSH)	144.63 ≤ PSH	156.20 ≤ PSH
Patches of Positive Charge (PPC)	1.14 ≤ PPC	1.25 ≤ PPC
Patches of Negative Charge (PNC)	1.30 ≤ PNC	1.84 ≤ PNC
Structural Fv Charge Symmetry Parameter (SFvCSP)	SFvCSP ≤ - 4.00	SFvCSP ≤ - 6.30

Table 7: Computational Developability assessment criteria for clinical-stage antibodies based on TAP scores for five poor developability metrics. Worst 10% cutoff and 5% cutoff values are provided. The CDR length metric has both an upper threshold and lower threshold value.

Patches of positive charge (PPC) represents regions or areas on the surface of a biomolecule that have a net positive charge which are characterized by an excess of positively charged amino acid residues or groups, such as lysine (Lys) or arginine (Arg) in the protein structure. Highly positively charged patches cause aggregation. Otilimab, a monoclonal antibody therapy for rheumatoid arthritis had the highest score in PPC Metric with a value of 3.58 which is much higher than our threshold cutoffs of 1.14 (10% cutoff) and 1.25 (5% cutoff) respectively. GSK recently terminated the

Otilimab program and declined the regulatory submissions as it failed to reach its primary endpoint in the Phase III ContRAst-3 study.[23] Therefore, a very high PPC score not consistent with other clinical-stage mAbs could have been indicative to GSK team of a possible attrition in clinical trials during the developability assessment steps.

Another example is brolucizumab, a proposed treatment for neovascular age-related macular degeneration which had the second highest PNC score with a value of 3.72. Several serious adverse events were observed in clinical trials for brolucizumab such as intraocular inflammation and retinal vasculitis.[24] Novartis later halted three phase 3 studies of brolucizumab in patients with retinal diseases due to these safety concerns. The presence of several patches of high negative charge indicated by high PNC score could have triggered these inflammatory responses by the immune system in patients. A high PNC score can trigger inflammatory cells such as blood cells that may adhere to the negative surface of the antibody which will facilitate the recruitment of immune cells to the site of cell adhesion. A high PNC score may also activate the complement system or activate cytokine storm by interaction with the positively charged cells. Therefore, these computational developability assessment tools can serve as valuable resource to mitigate clinical trial failures and flag associated developability concerns.

3.6 **Human immune repertoire dataset**

Human antibody repertoires have emerged as a reliable and comprehensive resource to probe the diversity of human antibodies. The entire set of antibodies produced in an individual is called their antibody repertoire, also referred to as Immunoglobulin (Ig) repertoire or the B-cell receptor (BCR) repertoire.[25, 26]

Billions of unique antibodies are secreted every day by different white blood cells in the human body. This comprehensive molecular diversity of circulating antibodies is fundamental to our immune system to ward off against infectious diseases and toxic threats. Such antibody diversity is possible by V(D)J recombination – a unique genetic mechanism that enables mammalian cells to generate an almost unlimited number of different light and heavy chains in a remarkably economical way. An unlimited array of antibody repertoire is formed from B lymphocytes by rearranging, recombining, and mutating the genetic code. This genetic mechanism is a hallmark of the immune physiology of most vertebrates to achieve multiple antigen-binding sites.

An elegant feature of vertebrate immune physiology is the selection of antibody proteins that are specific, safe, and tolerated by the body. Studies have shown that the B lymphocytes which produce "good" and acceptable antibodies are stimulated to proliferate while B lymphocytes that shuffled gene sequences to produce "bad", intolerable or autoreactive (targeting the body's own tissue) antibodies were subjected to programmed cell death.[27, 28] This selection is a fundamental step in human antibody engineering to enable the body to produce an enormous antibody library without creating antibody molecules that cause any harm or trigger immunogenic reaction.

A previous work had compared the post phase-1 therapeutics to Vander Heiden's human immunoglobulin gene sequencing (Ig-seq) models.[17] Overall, Raybould *et al.* concluded that a subset of natural human antibodies were unsuitable for therapeutic use. They observed shorter mean CDRH3 loop length for therapeutics than human-expressed antibodies with an increase in the CDR lengths with more humanization. Also, the therapeutic antibodies had lower mean hydrophobicity than natural human antibodies which was the main differentiation between the two datasets. However, the charge-based structural features like PPC, PNC, and SFvCSP were similar between therapeutic antibodies and natural human antibody repertoire.

3.6.1 OAS: Observed Antibody Space database

The Observed Antibody Space (OAS) database contains annotated immune repertoires that cover one billion sequences from over 80 different studies.[26] OAS is accessible via a web server at https://opig.stats.ox.ac.uk/webapps/oas/. It contains both unpaired and paired (V_H/V_L) sequence data with filters on sequence attributes such as chain type, species, and disease state. Thus, OAS repertoires cover diverse immune states, organisms (primarily human and mouse), and individuals. It serves as an excellent tool for the data mining of immune repertoires for an improved understanding of immune response and the development of better biotherapeutics.

We have selected a subset representing the natural human immune system from the available millions of sequences in BCR-seq datasets. The chosen filters were selective for healthy non-vaccinated human immune repertoires in the entire OAS dataset. The search yielded 350,980 filtered paired sequences from two previous BCR-seq studies namely Eccles_2020[29] and Jaffe_2022[30]. The search results are presented in Table 8.

Details	DS Name	#Unique Sequences	Organism	Isotype	Chain	Disease	Vaccine	Individual
Details	Eccles_2020	100	human	All	Paired	None	None	Healthy-1
Details	Eccles_2020	47	human	All	Paired	None	None	Healthy-1
Details	Eccles_2020	624	human	All	Paired	None	None	Healthy-1
Details	Jaffe_2022	889	human	All	Paired	None	None	Donor-3
Details	Jaffe_2022	18273	human	All	Paired	None	None	Donor-3
Details	Jaffe_2022	8945	human	All	Paired	None	None	Donor-3
Details	Jaffe_2022	12381	human	All	Paired	None	None	Donor-3
Details	Jaffe_2022	19179	human	All	Paired	None	None	Donor-3
Details	Jaffe_2022	9757	human	All	Paired	None	None	Donor-3
Details	Jaffe_2022	18920	human	All	Paired	None	None	Donor-3

Table 8: OAS search results for healthy non-vaccinated human immune repertoires. The OAS search returned 350,980 filtered paired sequences from two studies – Eccles and Jaffe.

It is important to note the huge size of this human immune repertoire dataset with over 350,000 sequences which is more than 2500 times the Jain dataset and more than 500 times the TheraSabDab dataset. It makes the human immune repertoire dataset very computationally expensive and unique in size and storage requirements.

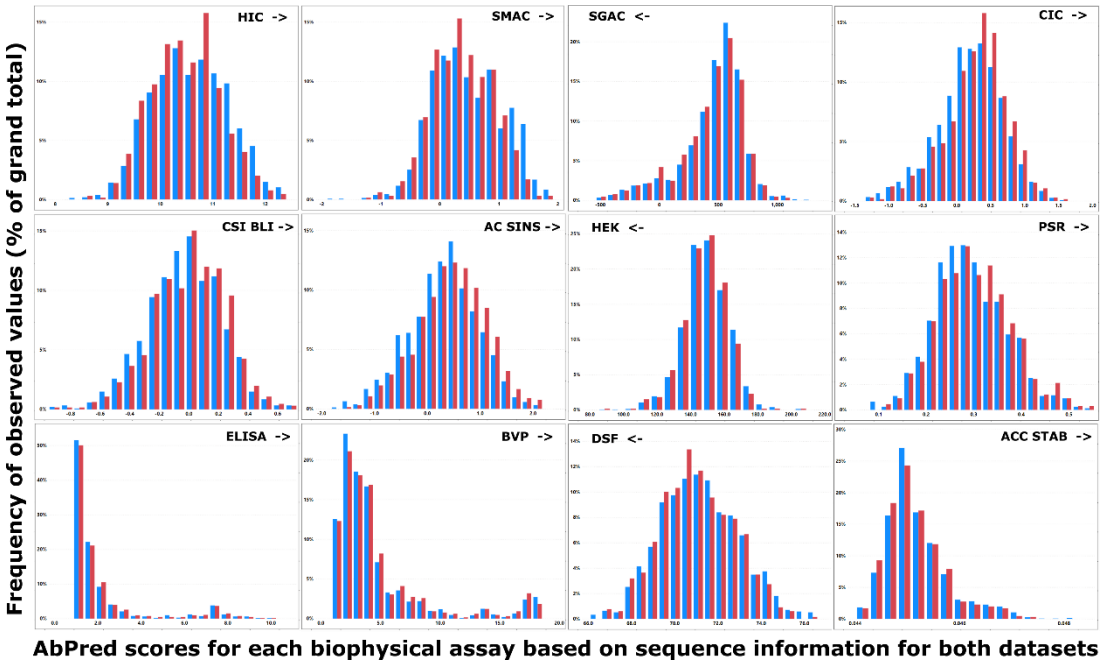


Figure 13: Histograms of 12 biophysical assay values for human immune repertoire (blue) and clinical-stage antibodies (red). The arrows next to assay names in top corner indicate the direction of unfavourable values. The frequency on y-axis is expressed as % of total count.

In this work, we have employed AbPred predictions on this compiled human immune repertoire dataset and compared the assay scores with the TheraSabDab clinical-stage

antibodies. This provided us insights into the difference in the biophysical properties between natural human antibodies and commercial antibody therapeutics in advanced clinical stages. The final results are shown in Figure 13. The histogram distributions show the human immune repertoire dataset in blue and clinical-stage antibodies in red. We observe that the natural human antibodies distribution spanned the range covered by TheraSabDab antibodies which confirm that the variation and diversity of human antibodies is quite vast and not restrictive to a particular biophysical range. Overall, the natural human antibodies were also asymmetrically long-tailed but exhibited different skewness and peak location in the histogram distributions.

We used again the Kolmogorov-Smirnov test (K-S test) for statistical comparison of the two distributions. Interestingly, for most of the assays, we observed a high K-S test statistic and a low P-value which rejects the standard null hypothesis that both are similar distributions. The high difference in medians for both distributions for most assays except one further bolster this trend. From this, we conclude that human antibodies differ from clinical-stage antibody therapeutics in most biophysical assays.

Assay	K-S Test Statistic	P-value	Median Human OAS	Median TheraSabDab	Difference
HIC	0.28607	1.0325e-14	10.5375	10.1444	0.3931
SMAC	0.34498	0	0.4572	0.0222	0.4350
SGAC	0.23404	4.4409e-16	554.8119	645.0108	-90.1989
CIC	0.35714	0	0.4070	0.0127	0.3943
CSI – BLI	0.12614	5.6773e-05	0.0585	-0.0120	0.0705
AC SINS	0.30547	0	0.5677	0.0493	0.5185
HEK	0.18693	2.068e-10	152.7021	147.8285	4.8737
PSR	0.19301	4.5241e-11	0.3077	0.2734	0.0343
ELISA	0.050152	0.37952	1.6454	1.6855	-0.0402
BVP	0.1383	6.8452e-06	3.9264	3.3635	0.5630
DSF	0.089666	0.010081	71.2242	70.9381	0.2861
ACC STAB	0.093607	0.043077	0.0452	0.0452	0

Table 9: Kolmogorov-Smirnov test results for comparison of human and clinical-stage mAbs. A low P-value is observed for all assays except ELISA which proves that human antibodies differ from clinical-stage antibody therapeutics histograms in most biophysical assays.

However, ELISA was an outlier with this trend with a high and statistically significant P-value of 0.379 which indicates strong similarity in ELISA assay scores for both

distributions. ELISA, which is a measure of the antigen specificity for an antibody has an overlap between natural human antibodies and clinical-stage antibodies. It is a very interesting result which proves that the natural human immune repertoire antibodies perform at par with clinical-stage antibodies in the specificity assays. This is indeed expected as the natural human antibodies are designed by our immune system to be very specific for a particular antigen target through V(D)J recombination.

A detailed analysis of each assay distribution reveals the proportion of human immune repertoire above the proposed developability cutoffs. 18.06% of human antibodies are above the 5% threshold cutoff (11.149) for the HIC assay. This subset of natural human antibodies has a very high hydrophobicity making some of them unsuitable for commercial use. 23.16% of human antibodies are above the 5% threshold cutoff (0.7943) for the SMAC assay; 10.38% are below the 5% threshold cutoff (78.011) for the SGAC assay; 3.93% are above the 5% threshold cutoff (1.0554) for the CIC assay; 1.22% are above the 5% threshold cutoff (0.5635) for the CSI BLI assay; 3.09% are above the 5% threshold cutoff (1.5725) for the AC SINS assay; 1.48% are below the 5% threshold cutoff (121.674) for the HEK assay; 4.13% are above the 5% threshold cutoff (0.4396) for the PSR assay; 7.09% are above the 5% threshold cutoff (6.8259) for the ELISA assay; 7.54% are above the 5% threshold cutoff (14.1383) for the BVP assay; 1.74% are below the 5% threshold cutoff (68.045) for the DSF assay and 10.77% are above the 5% threshold cutoff (0.0461) for the ACC STAB assay.

A lower percentage than the 5% cutoff means that the natural immune repertoire antibodies have better developability for that assay as less proportion of antibodies breach the threshold criteria. While a higher percentage than the 5% cutoff means that the human antibodies assay value distribution is skewed in the unfavourable direction and represents poor developability properties compared to clinical-stage therapeutic mAbs. Therefore, the above analysis and comparison with threshold cutoffs for the clinical-stage therapeutics suggest that the natural human antibodies have better biophysical properties in assays like CIC, CSI BLI, AC SINS, and PSR that measure cross-interaction, self-association, and other binding properties. Human antibodies are also predicted to display better expression levels in HEK titer with only 1.48% subset of human antibodies below the clinical-stage benchmark. However, we observed poor developability for human antibodies in terms of assays measuring hydrophobicity and stability such as HIC, SMAC, SGAC, and ACC STAB. Especially, with over 18% of

human antibodies falling above the 5% threshold in HIC and SMAC, hydrophobicity looks to be the key feature to target in antibody engineering for biotherapeutics.

3.6.2 Case Study: True Human™ antibody therapeutics

True Human™ antibodies reproduce the extraordinary genetic diversity among the human antibody genes by utilizing the incredibly diverse repertoire of unique B lymphocytes created by the fundamental biology of natural antibody production. Bermekimab (commonly known as MABp1 or Xilonix™) is a first-in-class True Human™ monoclonal antibody targeting anti-interleukin-1-alpha (IL-1α) that is being evaluated in phase III trials as of 2023 for late-stage colorectal cancer by XBiotech.[31]

MABp1 is a natural antibody derived from an affinity-matured *in vivo* human immune response without any sequence modifications. This human monoclonal autoantibody is generated by Epstein-Barr virus-immortalization and CD40-activation of B cells from an individual with circulating anti-IL-1α that acts as a high-affinity IL-1α specific inhibitor.[32] True Human™ antibodies, therefore, harness the naturally occurring immunity sustained by billions of different antibodies that circulate through our blood. XBiotech has employed high stringency antibody mining technologies to identify a single clinically relevant antibody from billions of antibody molecules present in a blood donor sample.

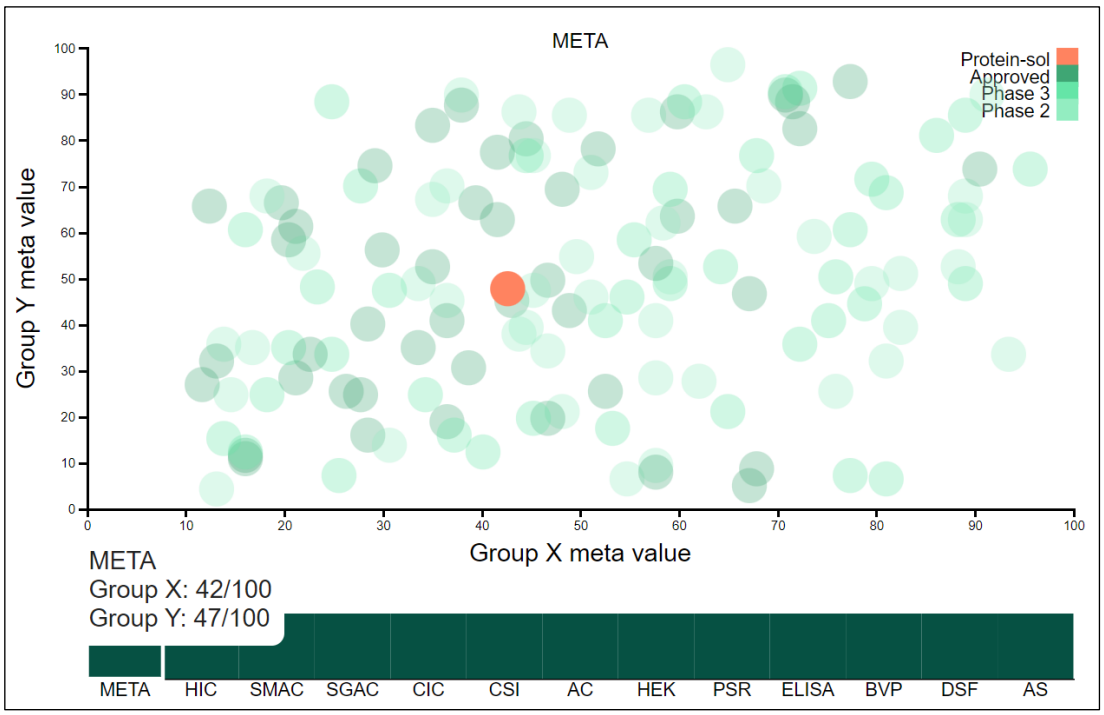


Figure 14: AbPred prediction results for Bermekimab. The provided Meta score combines and averages multiple biophysical platforms. Overall heatmap (green) shown in the bottom.

We checked the developability of this lead True Human™ candidate on AbPred. The results have been provided in Figure 14. Overall, the True Human™ antibody had an excellent developability performance in all assays. This is captured in the Meta score which combines and averages multiple biophysical platforms. It had a good rank of 42/100 in the Group X assays (charge-based assays) – ELISA, BVP ELISA, PSR, CSI, ACC STAB, CIC, and a rank of 47/100 in Group Y assays (two hydrophobicity-based assays described in AbPred tool) – HIC and SMAC. It also had all the assay scores within the 10% developability thresholds confirming no biophysical liabilities.

The safety, tolerability, and pharmacokinetic profile of True Human™ antibody therapeutics have been characterized in several previous clinical studies.[33, 34] The first study showed that True Human™ antibodies are well tolerated in metastatic cancer patients with no dose-limiting toxicities or immunogenicity.[33] The pharmacokinetic data for MABp1 were consistent at all dose levels and showed no evidence of accumulation or increased clearance at increasing doses. Also, there were no serious treatment-related adverse events for all 42 patients in the study. The next study assessed the treatment of *Staphylococcus aureus* Bacteremia infection with 514G3 – a True Human™ antibody targeting Staphylococcus Protein A which was isolated from a healthy human donor.[34] 514G3 was safe and well tolerated at all dose levels tested in this phase I, double blind, multicenter, randomized, placebo controlled, dose escalation study in patients hospitalized with *S. aureus* infections. Finally, other True Human™ antibodies by XBiotech are also being investigated in advanced clinical trials for multiple clinical indications such as Atopic Dermatitis, IL-1α-related inflammations, Hidradenitis Suppurativa and Type 2 Diabetes. Therefore, the optimal developability profile is clinically validated for True Human™ antibodies.

A similar optimal developability profile was observed in the TAP results. The overall results are shown in Figure 15. The TAP scores for the five metrics were respectively: 49 for Total CDR length; 160.0458 for PSH; 0.2813 for PPC score; 0.0 for PNC score and 6.71 for SFvCSP score. All these values lie within the amber and red flag threshold regions as per the five computational developability guidelines. Figure 15 shows the histogram distributions for each metric. So, we conclude that the lead True Human™

antibody candidate displays the right CDR length, optimal charge, and hydrophobicity in the CDR vicinity, and charge symmetry between the heavy and light chain arms.

Figure 15: Therapeutic Antibody Profiler (TAP) results for Bermekimab. Part A shows the plots for five structural metrics and the TAP score (green line) along with the amber and red flags. Part B shows the TAP score values for each assay and the corresponding flag colour.

The computational developability assessment results and clinical data for True Human™ antibody therapeutics suggest a unique safety profile for biotherapeutics cloned from the natural human immune response. Such antibodies extracted from human immune repertoire have the potential to emerge as the best tolerated therapies in oncology and beyond, making them ideally suited for treating patients with reduced tolerance to other immunogenically sensitive therapies in current practice.

True Human™ antibody therapeutics created by engineering natural human antibodies have high developability. So, this case study demonstrates that naturally occurring human immune repertoire antibodies can be successfully engineered as commercial therapeutics that display optimal biophysical properties. We expect more such novel approaches like True Human™ technology to emerge as we gain deeper understanding of natural human immunity and advance new antibody mining technologies focused on immune repertoires to identify more clinically relevant antibodies.

3.7 **Conclusion**

We have analyzed several biophysical features linked to developability across known clinical-stage therapeutics and human immune repertoire datasets. The main outcome

was to establish benchmark thresholds for biophysical assay performance based on approved and clinical-stage mAbs which are assumed to have good developability characteristics for therapeutic development.

This analysis concluded that the biophysical property distributions are asymmetrically long-tailed in the unfavourable direction for clinical-stage mAbs. We then proposed new developability criteria derived from the worst 10% and 5% cut-off values for twelve biophysical assays and five variable region properties for TheraSabDab clinical-stage antibodies. Finally, we used these threshold criteria to compare natural human antibodies to commercial therapeutics and validated the developability potential of human immune repertoire using a case study.

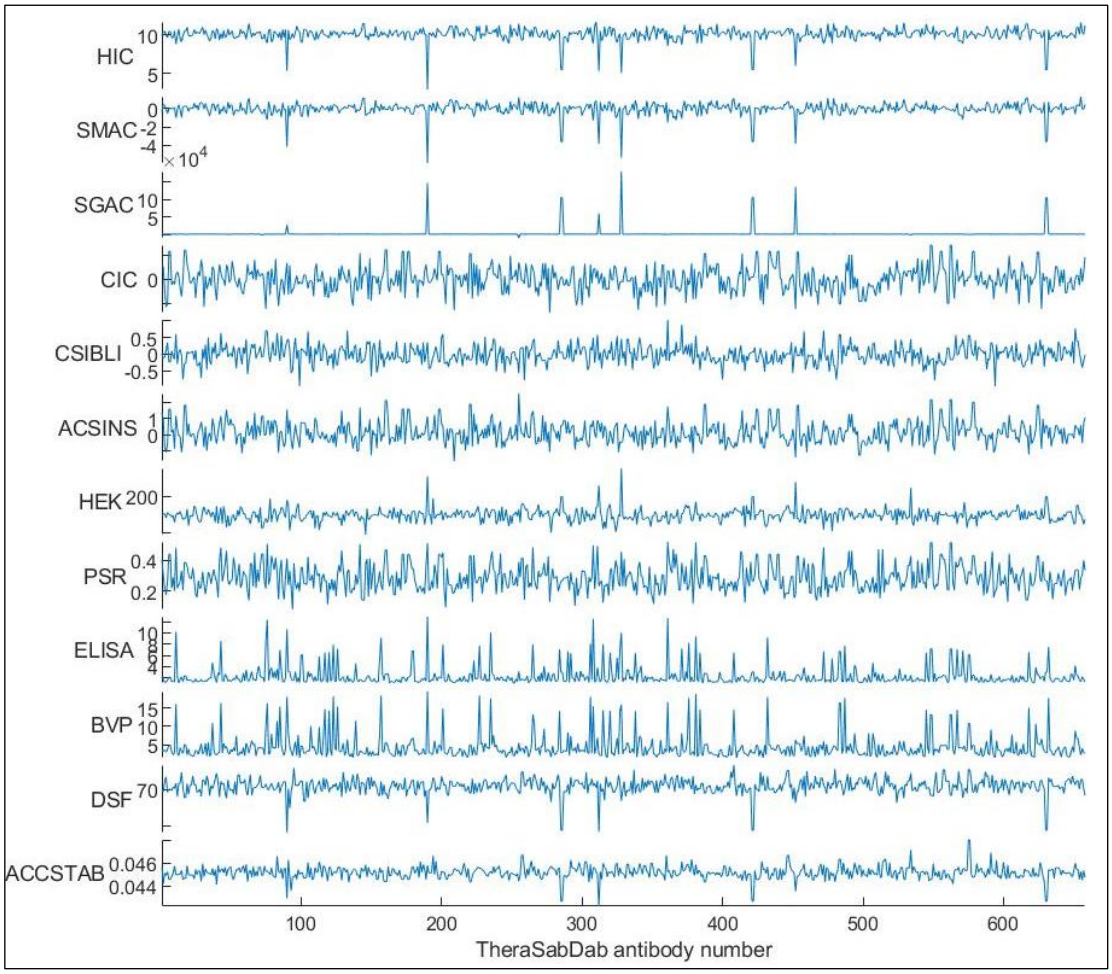


Figure 16: Stacked plot of biophysical assay values for TheraSabDab clinical stage antibodies

Figure 16 provides a summary of the biophysical assay performance for TheraSabDab clinical-stage antibodies. Overall, the stacked plot and Pearson correlation coefficients (R-squared > 0.7 for each pair) show that the assay groupings are Group A: HIC, SMAC and SGAC; Group B: CIC, CSI BLI and AC SINS; Group C: ELISA and BVP-

ELISA; and Group D: DSF and ACC STAB for the TheraSabDab dataset. However, HEK and PSR are independent assays that have unique trends and don't belong to any of the above groupings. This clustering into the above groups is possibly explained by the measurement of similar underlying biophysical properties for each group. For instance, the Group A assays - HIC, SMAC and SGAC are all known to measure the "stickiness" of the antibodies either directly or indirectly. The Group B assays - CIC, CSI BLI and AC SINS are all well-established assays to measure the interaction and association properties of the antibodies. So, the underlying self-association or binding measurements overlap between these assays. Next, Group C assays - ELISA and BVP-ELISA are both labelled immunoassays with the same setup with the only difference in using baculovirus particle (BVP) as the reagent in BVP-ELISA assay. So, the key underlying biophysical measurement is polyspecificity for both the assays. Finally, ACC and DSF both measure unfolding stability and are grouped together in Group D.

This is also confirmed by the results of Principal Component Analysis (PCA). PCA is a dimensionality reduction method that extracts features that successively maximize variance. We have used the PCA function in MATLAB which uses the singular value decomposition (SVD) algorithm to rank the feature columns in the descending order of component variance. The PCA results for the entire matrix of 12 biophysical assay values for TheraSabDab antibodies are shown in Figure 17. We conclude that the four assay groups and two distinct assays make a total of six assay groups that explain over 75% data variance. The MATLAB codes for PCA are documented in Chapter 2.

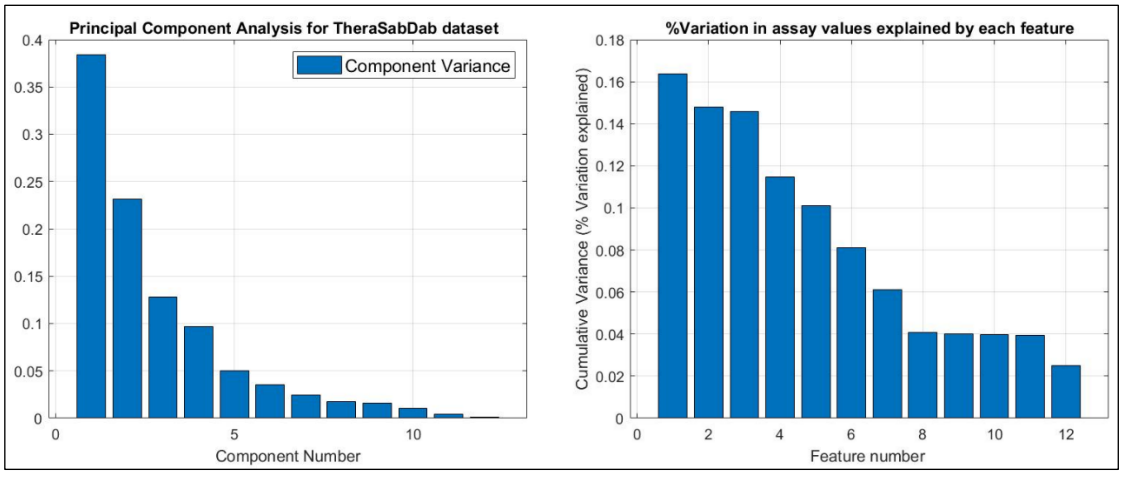


Figure 17: Principal Component Analysis (PCA) results for 12 biophysical assay features

The comparison of natural human antibodies with clinical-stage therapeutics suggests that human immune repertoire derived antibodies have exceptional specificity, binding

and interaction properties but a subset may have poor hydrophobicity and long-term stability. So, human-expressed antibodies are highly vulnerable to protein unfolding and aggregation under a shear stress environment and high-concentration storage conditions often encountered during manufacturing and shipping of therapeutic mAbs. Evidence from our case study on True Human™ antibody therapeutic further supports successful engineering of the human immune repertoire antibodies towards clinically relevant commercial therapeutics that display optimal biophysical properties.

Finally, this work has demonstrated several real-world examples where the proposed computational developability criteria were able to predict clinical trial attritions, flag developability concerns, or even predict the likelihood of success towards approval. This new developability criteria established in Chapter 3 has been further validated in Chapter 5 to evaluate a failed antibodies dataset and detailed statistical metrics on the model performance of a binary classification algorithm have been provided in Chapter 5. Our detailed insights into TheraSabDab, Jain assays, and new OAS subset can guide experimental design or future computational developability assessment frameworks.

3.8 References

1. Jain T, Sun T, Durand S, Hall A, Houston NR, Nett JH, Sharkey B, Bobrowicz B, Caffry I, Yu Y. Biophysical properties of the clinical-stage antibody landscape. Proceedings of the National Academy of Sciences. 2017;114(5):944-9. doi:10.1073/pnas.1616408114.
2. Lienqueo ME, Mahn A, Salgado JC, Asenjo JA. Current insights on protein behaviour in hydrophobic interaction chromatography. Journal of Chromatography B. 2007;849(1-2):53-68.
3. Queiroz J, Tomaz C, Cabral J. Hydrophobic interaction chromatography of proteins. Journal of biotechnology. 2001;87(2):143-59.
4. McCue JT. Theory and use of hydrophobic interaction chromatography in protein purification applications. Methods in enzymology. 2009;463:405-14.
5. Kohli N, Jain N, Geddie ML, Razlog M, Xu L, Lugovskoy AA. A novel screening method to assess developability of antibody-like molecules. MAbs; 2015. doi:10.1080/19420862.2015.1048410.
6. Wolf Pérez A-M, Sormanni P, Andersen JS, Sakhnini LI, Rodriguez-Leon I, Bjelke JR, Gajhede AJ, De Maria L, Otzen DE, Vendruscolo M. In vitro and in silico assessment of the developability of a designed monoclonal antibody library. MAbs; 2019. doi:10.1080/19420862.2018.1556082.
7. Jacobs SA, Wu S-J, Feng Y, Bethea D, O'Neil KT. Cross-interaction chromatography: a rapid method to identify highly soluble monoclonal antibody candidates. Pharmaceutical research. 2010;27(1):65-71.
8. Kelly RL, Sun T, Jain T, Caffry I, Yu Y, Cao Y, Lynaugh H, Brown M, Vásquez M, Wittrup KD. High throughput cross-interaction measures for human IgG1 antibodies correlate with clearance rates in mice. MAbs; 2015.
9. Sun T, Reid F, Liu Y, Cao Y, Estep P, Nauman C, Xu Y. High throughput detection of antibody self-interaction by bio-layer interferometry. MAbs; 2013.

10. Liu Y, Caffry I, Wu J, Geng SB, Jain T, Sun T, Reid F, Cao Y, Estep P, Yu Y. High-throughput screening for developability during early-stage antibody discovery using self-interaction nanoparticle spectroscopy. MAbs; 2014.
11. Wu J, Schultz JS, Weldon CL, Sule SV, Chai Q, Geng SB, Dickinson CD, Tessier PM. Discovery of highly soluble antibodies prior to purification using affinity-capture self-interaction nanoparticle spectroscopy. Protein Engineering, Design and Selection. 2015;28(10):403-14.
12. Xu Y, Roach W, Sun T, Jain T, Prinz B, Yu T-Y, Torrey J, Thomas J, Bobrowicz P, Vásquez M. Addressing polyspecificity of antibodies selected from an in vitro yeast presentation system: a FACS-based, high-throughput selection and analytical tool. Protein Engineering, Design & Selection. 2013;26(10):663-70.
13. Lequin RM. Enzyme immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA). Clinical chemistry. 2005;51(12):2415-8.
14. Hötzel I, Theil F-P, Bernstein LJ, Prabhu S, Deng R, Quintana L, Lutman J, Sibia R, Chan P, Bumbaca D. A strategy for risk mitigation of antibodies with fast clearance. MAbs; 2012.
15. Hebditch M, Warwicker J. Charge and hydrophobicity are key features in sequence-trained machine learning models for predicting the biophysical properties of clinical-stage antibodies. PeerJ. 2019;7:e8199. doi:10.7717/peerj.8199.
16. Hebditch M, Carballo-Amador MA, Charonis S, Curtis R, Warwicker J. Protein–Sol: a web tool for predicting protein solubility from sequence. Bioinformatics. 2017;33(19):3098-100. doi:10.1093/bioinformatics/btx345.
17. Raybould MI, Marks C, Krawczyk K, Taddese B, Nowak J, Lewis AP, Bujotzek A, Shi J, Deane CM. Five computational developability guidelines for therapeutic antibody profiling. Proceedings of the National Academy of Sciences. 2019;116(10):4025-30. doi:10.1073/pnas.1810576116.
18. Leem J, Dunbar J, Georges G, Shi J, Deane CM. ABodyBuilder: Automated antibody structure prediction with data–driven accuracy estimation. MAbs; 2016. doi:10.1080/19420862.2016.1205773.
19. Sharma VK, Patapoff TW, Kabakoff B, Pai S, Hilario E, Zhang B, Li C, Borisov O, Kelley RF, Chorny I. In silico selection of therapeutic antibodies for development: viscosity, clearance, and chemical stability. Proceedings of the National Academy of Sciences. 2014;111(52):18601-6.
20. Massey Jr FJ. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American statistical Association. 1951;46(253):68-78.
21. Holliday I. Kolmogorov-Smirnov test (v1. 0.4) in free statistics software (v1. 2.1). Office for Research Development and Education, https: wessa net/rwasp_Reddy-Moores% 20K-S% 20Test wasp. 2017.
22. Moreau P, Garfall AL, van de Donk NW, Nahi H, San-Miguel JF, Oriol A, Nooka AK, Martin T, Rosinol L, Chari A. Teclistamab in relapsed or refractory multiple myeloma. New England Journal of Medicine. 2022;387(6):495-505.
23. Release GP. GSK provides update on ContRAst phase III programme for otilimab in the treatment of moderate to severe rheumatoid arthritis: GSK Media; 2022 [
24. Khanani AM, Zarbin MA, Barakat MR, Albini TA, Kaiser PK, Guruprasad B, Agashivala N, Justin SY, Wykoff CC, MacCumber MW. Safety outcomes of brolucizumab in neovascular age-related macular degeneration: results from the IRIS registry and komodo healthcare map. JAMA ophthalmology. 2022;140(1):20-8.
25. Chaudhary N, Wesemann DR. Analyzing immunoglobulin repertoires. Frontiers in immunology. 2018;9:462.

26. Olsen TH, Boyles F, Deane CM. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. Protein Science. 2022;31(1):141-6.
27. Wang F, Ekiert DC, Ahmad I, Yu W, Zhang Y, Bazirgan O, Torkamani A, Raudsepp T, Mwangi W, Criscitiello MF. Reshaping antibody diversity. Cell. 2013;153(6):1379-93.
28. Eyerman MC, Zhang X, Wysocki LJ. T cell recognition and tolerance of antibody diversity. The Journal of Immunology. 1996;157(3):1037-46.
29. Eccles JD, Turner RB, Kirk NA, Muehling LM, Borish L, Steinke JW, Payne SC, Wright PW, Thacker D, Lahtinen SJ. T-bet+ memory B cells link to local cross-reactive IgG upon human rhinovirus infection. Cell reports. 2020;30(2):351-66. e7.
30. Jaffe DB, Shahi P, Adams BA, Chrisman AM, Finnegan PM, Raman N, Royall AE, Tsai F, Vollbrecht T, Reyes DS. Functional antibodies exhibit light chain coherence. Nature. 2022:1-6.
31. Hong DS, Janku F, Naing A, Falchook GS, Piha-Paul S, Wheler JJ, Fu S, Tsimberidou AM, Stecher M, Mohanty P. Xilonix, a novel true human antibody targeting the inflammatory cytokine interleukin-1 alpha, in non-small cell lung cancer. Investigational new drugs. 2015;33:621-31.
32. Garrone P, Djossou O, Fossiez F, Reyes J, Ait-Yahia S, Maat C, Ho S, Hauser T, Dayer J-M, Greffe J. Generation and characterization of a human monoclonal autoantibody that acts as a high affinity interleukin-1α specific inhibitor. Molecular immunology. 1996;33(7-8):649-58.
33. Hong DS, Hui D, Bruera E, Janku F, Naing A, Falchook GS, Piha-Paul S, Wheler JJ, Fu S, Tsimberidou AM. MABp1, a first-in-class true human antibody targeting interleukin-1α in refractory cancers: an open-label, phase 1 dose-escalation and expansion study. The Lancet Oncology. 2014;15(6):656-66.
34. Huynh T, Stecher M, Mckinnon J, Jung N, Rupp ME. Safety and tolerability of 514G3, a true human anti-protein a monoclonal antibody for the treatment of S. aureus bacteremia. Open Forum Infectious Diseases; 2016.

CHAPTER 4

4 Computational developability assessment of engineered antibodies and next-generation biotherapeutics

4.1 Abstract

The evolution of antibody engineering to alter antibody size, shape, and valency has yielded a large diversity of new antibody formats. These new improvements have greatly increased the chance to engineer antibodies with desired biophysical properties while achieving optimal drug-like properties and high-potency. This chapter aims to characterize and compare the developability and biophysical behaviour of multiple categories of new engineered antibody formats. We have used the computational developability assessment criteria established in the previous chapter to assess multispecific formats such as DuoBody®; BiTEs; CrossMabs; DART® that represent different structures of antibody therapeutics, and then two technologies namely – (1) Phage display formats such as CAT phage library; MorphoSys HuCAL®; Dyax library, and next (2) Transgenic mice formats such as XenoMouse®; UltiMAb®; VelocImmune® and HuMAb-Mouse® using the available antibody informatics tools. Our results highlight the best platform for the desired application in each case.

4.2 Introduction

Novel antibody discovery and optimization technologies have emerged in the past two decades.[1] This development of next generation of antibody-based biologics is driven by the desired enhancement in functional affinity and modality of current therapeutics. These next-generation antibody therapeutics span multiple new formats such as IgG, Fc fusion proteins, scFv, recombinant antibody fragments, antibody-drug conjugates, immune cells expressing CAR antibodies, immunocytokines, radioimmunoglobulins, and other new engineered variants. These formats are referred to as 'engineered' antibodies throughout this chapter to differentiate these new alternative formats.

The first wave of antibody engineering approaches were focused on the manipulation of the variable regions for humanization and affinity-maturation.[2-4] The next wave of approaches were to generate different types of antibody fragments such as scFvs[5], single domain antibodies[6], diabodies[7], TandAbs[8], and PEGylated Fabs[9]. An important engineering approach to improve the functional affinity of antibodies was to introduce multivalency in antibody formats by combining multiple binding domains using techniques such as domain swapping, antibody domain fusion, or self-assembly of

heavy and light chains into new formats.[10-12] These new formats also provide novel mechanisms for disease intervention by their inherent ability to bind multiple antigens. For example, novel bispecific antibodies enable the use of complex mechanisms such as immune cell redirected tumour killing[13], receptor cross-linking[14], and enhanced cell specificity[15] in cancer and autoimmunity. Over the years, several technologies have been implemented in order to generate new alternate engineered therapeutics such as scFv-based formats, hetero-dimerization domains, quadroma technology, human phage display technology, and transgenic mice technology.

The desirable attributes in an engineered antibody include high solubility, low aggregation, high thermal stability, low viscosity, and high chemical stability. A major rate-limiting step in the adoption of these new engineered therapeutic formats is the validation of the developability potential for these new antibody formats. Therefore, the developability characteristics and validation studies of novel engineered antibodies and antibody fragments are receiving increased attention.

This chapter starts with a general overview of the engineered antibodies and the special developability considerations required for these new and emerging special therapeutic modalities. Next, computational developability assessment results are presented for a compiled dataset of engineered antibodies created from the IMGT® database - the International ImMunoGeneTics Information System® database in section 4.4. Here, we have compared and discussed the developability trends among different categories of engineered antibodies to capture their differentiation in key biophysical features. A special case study on all available bispecific antibodies is also presented to explore the biophysical performance of bispecific formats in section 4.4.2. We use this framework to comment on the developability and clinical success of a new engineered antibody drug discovery platform - Azymetric™ technology in our case study in section 4.4.3.

A similar strategy is employed with a human antibody phage display library dataset in section 4.5 and transgenic mice antibodies dataset in section 4.6 to gain understanding of the unique developability profile of various engineered antibody platforms. Finally, we provide insights and conclusions from these biopharmaceutical informatics results about the challenges of using engineered antibody formats as therapeutic candidates.

4.3 **Methods**

Creation of antibody datasets: We have used International ImMunoGeneTics Information System (IMGT) database and manual online resources to create three datasets of antibody therapeutics – (Part 1) multispecific formats such as DuoBody®; BiTEs; CrossMabs; DART®, and (Part 2.1) phage display formats such as CAT phage library; MorphoSys HuCAL®; Dyax library, and (Part 2.2) transgenic mice formats such as XenoMouse®; UltiMAb®; VelocImmune® and HuMAb-Mouse® using the relevant keyword search filters. IMGT is available at https://www.imgt.org/mAb-DB/.

AbPred measurements on engineered mAb datasets: VHVL sequence information for separate engineered datasets were saved as input fasta files. The Abpred predictions were generated from the dockerhub source code available at docker pull maxhebditch/abpred using run command docker run --rm -v $(pwd)/:/abpred/host maxhebditch/abpred. More details at https://hub.docker.com/r/maxhebditch/abpred.

TAP measurements on engineered datasets: We used web sequence submission form and the GitHub repositories at https://github.com/orgs/oxpig/repositories to get the five metric values for the input sequences from the TAP tool available at https://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/sabpred. Also, the homology Fv models generated by ABodyBuilder2 were downloaded for future structural analysis.

Kolmogorov-Smirnov test (K-S test) Statistics: The Kolmogorov-Smirnov Test R module from http://www.wessa.net/rwasp_Reddy-Moores%20K-S%20Test.wasp was used to generate the K-S test Statistic and P-value from histogram raw data input.

Azymetric™ platform technology case study: The antibody sequence information was extracted from Patent number: 11306156 (Modified antigen binding polypeptide constructs and uses thereof). https://patents.google.com/patent/WO2015181805A1.

Protein-Sol measurements on bispecific case study dataset: Full antibody sequence information was saved as input fasta files. These fasta files were used on the web-based Protein-Sol tool available at https://protein-sol.manchester.ac.uk/. The Protein-Sol sequence algorithm calculates 35 sequence features and provides an output table in excel. A docker container was also used to create a local instance of the protein-sol solubility algorithm. The patches tool was used to visualize the surface patches of potential and hydrophobicity for PDB structure of bispecific antibodies in Figure 23.

Part 1 – Evaluating different antibody structural formats:

4.4 Engineered antibody fragments dataset

A dataset of over a hundred engineered antibodies for multiple clinical indications and originating from diverse development technologies was extracted from the IMGT® database. We then compiled the sequence and 3D structure information for these antibodies from several online resources. The engineered antibodies with missing sequence information were discontinued for further analysis while the scFv fragment from each engineered antibody was used as a separate query in computational developability assessments. Finally, we had an annotated dataset of 103 engineered scFv fragments for analysis using antibody informatics tools such as ProteinSol, AbPred, and TAP. These were applied sequentially to the entire antibody dataset.

Engineered antibodies dataset (Table 10) collates publicly available information on various engineered antibody therapeutics that were available in the market or were being investigated in human clinical trials as of 2023. Figure 18 shows the clinical trial status for different categories of engineered antibodies. Among these 103 antibodies, majority of the antibodies (70%) are in early-stage clinical trials with 58 antibodies in phase - I clinical trials and 14 in phase - II clinical trials. It is expected since engineered antibodies are recent innovations departing from the conventional mAbs that have been launched in clinical development only in recent years. Therefore, only six of these engineered antibodies in the dataset have been approved and have received the formal marketing authorizations namely – blinatumomab (BLINCYTO®), faricimab (VABYSMO™), catumaxomab (REMOVAB®), amivantamab (RYBREVANT™), emicizumab (HEMLIBRA®) and tebentafusp (KIMMTRAK®).

INN (International Non-proprietary Name)	Common name / Proprietary name	Company	Clinical indication	Highest Clinical Trial
Development Technology : **BiTE® (Bispecific T cell Engager) technology** **Antibody Format:** (scFv - heavy - kappa) - (scFv - heavy - lambda) – scFc				
acapatamab	AMG-160	Amgen (Thousand Oaks CA USA)	Cancers, prostate, metastatic	Phase I
blinatumomab	BLINCYTO®, AMG103, BITE MT-103, bscCD19xCD3 MEDI-538, MT103	MedImmune (Gaithersburg MD USA) (US) / Amgen (Thousand Oaks CA USA) (US) / AstraZeneca (London UK)	Lymphoblastic leukemia (B cell ALL)	Approved
eluvixtamab	AMG-330, MT-114	Amgen (Thousand Oaks CA USA)	Acute myeloid leukemia (AML)	Phase I
emerfetamab	AMG-673	Amgen (Thousand Oaks CA USA)	Acute myeloid leukemia (AML)	Phase I

emirodatamab	AMG 427	Amgen (Thousand Oaks CA USA)	Acute myeloid leukemia (AML)	Phase I
etevritamab	AMG 596	Amgen (Thousand Oaks CA USA)	Glioblastoma	Phase I
gresonitamab	AMG 910	Amgen (Thousand Oaks CA USA)	GE Junction Cancer	Phase I
pacanalotamab	AMG-420, BI-836909	Amgen (Thousand Oaks CA USA)	Multiple myeloma (MM)	Phase I
pasotuxizumab	AMG 212, BAY 2010112	Amgen (Thousand Oaks CA USA) / Bayer HealthCare Pharmaceuticals (Leverkusen Germany) / Micromet Inc. (Munich Germany)	Cancers, prostate	Phase I
pavurutamab	AMG-701, BCMA HLE-BiTE	Amgen (Thousand Oaks CA USA)	Multiple myeloma (MM)	Phase I
solitomab	AMG 110	Micromet AG (Munich Germany)/ Amgen (Thousand Oaks CA USA)	Solid tumours	Phase I
tarlatamab	AMG-757	Amgen (Thousand Oaks CA USA)	Cancers, small cell lung (SCLC)	Phase I
vepsitamab	AMG 199	Amgen (Thousand Oaks CA USA)	GE Junction Cancer	Phase I
N/A	AMG 211	Amgen (Thousand Oaks CA USA)	Adeno carcinoma	Phase I
Development Technology : **CrossMAb technology** **Antibody Format:** IgG1 - kappa - lambda with half-IG VL-CH1/VH-CK crossover				
faricimab	VABYSMO™	Roche, F. Hoffmann-La Roche Ltd. (Basel Switzerland)	Age-related macular degeneration (AMD)	Approved
vanucizumab	RG-7221	Roche, F. Hoffmann-La Roche Ltd. (Basel Switzerland)	Cancers, colorectal (CRC)	Phase II
Development Technology : **Dual Variable Domain immunoglobulin (DVD-Ig™)** **Antibody Format:** [VH - VH' - H-Gamma1_VL - VL' - C-kappa] - dimer				
lutikizumab	ABT-981	AbbVie Inc. (North Chicago IL USA)	Osteoarthritis (OA)	Phase II
remtolumab	ABT-122, A-1230717	AbbVie Inc. (North Chicago IL USA)	Rheumatoid arthritis (RA), Psoriatic arthritis (PSA)	Phase II
romilkimab	SAR156597	Sanofi (Paris France)	Idiopathic pulmonary fibrosis	Phase II
Development Technology : **Dual-Affinity Re-targeting (DART®)** **Antibody Format:** V-Lambda - VH _ V-Kappa - VH' or [V-kappa'-VH-h-CH2-CH3_V-kappa-VH']2				
duvortuxizumab	JNJ-64052781, MGD011, RES192M1.2, hBU12(2.4)-hXR32-MP3 M1.2	MacroGenics Inc. (Rockville MD USA) / Janssen Pharmaceuticals, inc. (Titusville NJ USA)	lymphoblastic leukemia (B cell ALL)	Phase I
flotetuzumab	MGD-006, RES234	MacroGenics Inc. (Rockville MD USA)	Acute myeloid leukemia (AML)	Phase I
lorigerlimab	MGD 019	MacroGenics Inc. (Rockville MD USA)	Solid tumours	Phase I
obrindatamab	MGD009, RES281 M1.1	MacroGenics Inc. (Rockville MD USA)	Solid tumours, Advanced Cancers	Phase I
Development Technology : **DuoBody®** **Antibody Format:** IgG1 - lambda – kappa or IgG1 – kappa – kappa				
acasunlimab	BNT-311, DuoBody-PD-L1x4-1BB, GEN-1046	Genmab A/S (Copenhagen Denmark) / BioNTech SE (Mainz Germany)	Solid tumours	Phase I/II
tecaginlimab	GEN-1042, BNT-312	Genmab A/S (Copenhagen Denmark)	Solid tumours	Phase I/II
Development Technology : **Nanobody®** **Antibody Format:** VH - VH' and other formats.				

gefurulimab	ALXN-1720, CON-9978	Alexion Pharmaceuticals Inc. (New Haven CT USA)	Complement component deficiency	Phase I
isecarosmab	M-6495	Merck KgaA (Darmstadt Germany)	Solid tumours	Phase I
ozoralizumab	ATN-103	Ablynx (Ghent/Zwijnaarde Belgium)	Rheumatoid arthritis (RA)	Phase III
sonelokimab	M-1095, MSB-0010841	Merck & Co., Inc. (Whitehouse Station NJ USA)	Psoriasis	Phase I
vobarilizumab	ALX-0061, 20A11-9mer-ALB11	Ablynx (Ghent/Zwijnaarde Belgium) / AbbVie Inc. (North Chicago IL USA)	Rheumatoid arthritis (RA), Inflammatory conditions	Phase II
Development Technology : **Other formats such as Triomab®, Pentambody™, ADAPTIR™ Bispecific etc.** **Antibody Format:** [scFv]2 - Fc - [scFv]2, IgG2a - kappa / G2b – lambda and other multispecific formats.				
zenocutuzumab	MCLA-128, PB4188, R040517	Merus NV (Utrecht Netherlands)	Solid tumours	Phase I/II
N/A	ES414	Emergent Biosolutions (Rockville MD USA)	Cancers, prostate	Phase I
fidasimtamab	BH-2950, IBI-315	Hanmi Pharmaceutical (Seoul Korea) / Innovent Biologics (Suzhou China)	Solid tumours	Phase I
vixtimotamab	AMV-564, TandAb T564	Amphivena Therapeutics (South San Francisco CA USA)	Acute myeloid leukemia (AML)	Phase I
catumaxomab	REMOVAB®, TPBs01, Musmus G2a kappa	Fresenius Biotech GmbH (Bad Homburg Germany) / TRION Pharma (Munich Germany)	Cancers, ovarian, Gastric Cancers	Approved
ertumaxomab	REXOMUN®	Fresenius Biotech GmbH (Bad Homburg Germany) / TRION Pharma (Munich Germany)	Cancers, breast	Approved
alnuctamab	CC-93269, EM-901	Celgene corporation (Summit USA)	Multiple myeloma (MM)	Phase I
amivantamab	RYBREVANT™, JNJ-61186372, JNJ-6372, amivantamab-vmjw	Janssen Research & Development, LLC (Raritan NJ USA)	Cancers, non-small cell lung (NSCLC)	Approved
bafisontamab	EMB-01, FIT-013a	EpimAb Biotherapeutics (Shangai China)	Cancers, non-small cell lung (NSCLC)	Phase I/II
bavunalimab	XmAb-22841	Xencor Inc. (Monrovia CA USA)	Solid tumours	Phase I
cadonilimab	AK-104	Akeso Biopharma, Inc. (Fremont CA USA)	Cancers, non-small cell lung (NSCLC)	Phase II
cevostamab	BFCR-4350-A, BFCR-4350A, RG-6160, RO-7187797	Genentech Inc. (San Francisco CA USA)	Multiple myeloma (MM)	Phase I
cibisatamab	RO-6958688, CEA-TCB, RG-7802	Roche Ltd. (Basel Switzerland) / Genentech Inc. (San Francisco CA USA)	Solid tumours Cancers	Phase I
cinrebafusp alfa	PRS-343	Pieris Pharmaceuticals, Inc. (Boston MA USA)	Solid tumours	Phase I
dilpacimab	ABT-165, PR-1283233	AbbVie Inc. (North Chicago IL USA)	Solid tumours	Phase I
efdamrofusp alfa	ACVP-1, IBI 302	Innovent Biologics (Suzhou China)	Age-related degeneration (AMD)	Phase I
elranatamab	PF-06863135, PF-3135	Pfizer (New York NY USA)	Multiple myeloma (MM)	Phase I
emicizumab	HEMLIBRA®, ACE-910, RG6013,	Chugai Pharmaceutical Co., Ltd. (Tokyo Japan) / Roche, F. Hoffmann-La Roche Ltd. (Basel Switzerland)	Coagulation factor VIII deficiency (Hemophilia A)	Approved
ensomafusp alfa	CD19-4-1BBL, RG 6076, RO-7227166	Roche, F. Hoffmann-La Roche Ltd. (Basel Switzerland)	Lymphoma, B cell	Phase I

epcoritamab	GEN3013	Genmab A/S (Copenhagen Denmark)	B-cell malignancies	Phase II
erfonrilimab	KN-046	Alphamab Co.,Ltd (Suzhou China)	Cancers, non-small cell lung (NSCLC)	Phase II
glofitamab	CD20-TCB (2:1), RG-6026	Roche, F. Hoffmann-La Roche Ltd. (Basel Switzerland)	Lymphoma, B cell	Phase I
gremubamab	MEDI3902	MedImmune (Gaithersburg MD USA)	Nosocomial pneumonia	Phase I
istiratumab	MM-005, MM-141	Merrimack Pharmaceuticals (Cambridge MA USA)	Hepatocellular carcinoma (HCC)	Phase II
ivonescimab	AK 112	Akeso Biopharma, Inc. (Fremont CA USA)	Cancers, non-small cell lung (NSCLC)	Phase II
izuralimab	XmAb-23104	Xencor Inc. (Monrovia CA USA)	Solid tumours	Phase I
mosunetuzumab	BTCT4465A, RG-7828, RO7030816	Roche Ltd. (Basel Switzerland) / Genentech Inc. (San Francisco CA USA)	Cancers, blood Lymphoma, follicular (FL)	Phase I
navicixizumab	OMP-305B83	OncoMed Pharmaceuticals (Redwood City CA USA)	Solid tumours	Phase I
nivatrotamab	Hu3F8-BsAb	Memorial Sloan-Kettering Cancer Center (NY USA)	Neuroblastoma	Phase I/II
odronextamab	REGN-1979	Regeneron Pharmaceuticals Inc. (Tarrytown NY USA)	Chronic lymphocytic leukemia (CLL)	Phase I
petosemtamab	MCLA-158	Merus NV (Utrecht Netherlands)	Cancers, colorectal (CRC)	Phase I
plamotamab	XmAb-13676	Xencor Inc. (Monrovia CA USA)	Hematologic-blood cancer	Phase I
rovelizumab	LEUKARREST™, Hu23F2G	Eli Lilly (Indianapolis IN USA)	Multiple sclerosis (MS), Ischemic stroke	Phase III
runimotamab	BTRC-4017A, RG-6194	Genentech Inc. (S. San Francisco CA USA)	Solid tumours	Phase I
simridarlimab	IBI 322	Innovent Biologics (Suzhou China)	Solid tumours	Phase I
sonelokimab	M-1095, MSB-0010841	Merck & Co., Inc. (Whitehouse Station NJ USA)	Psoriasis	Phase I
talquetamab	JNJ-64407564	Janssen Research & Development, LLC (Raritan NJ USA)	Relapsed multiple myeloma	Phase I
tebentafusp	KIMMTRAK®	Immunocore Ltd (Abingdon UK)	Melanoma, malignant	Approved
tebotelimab	MGD-013	MacroGenics Inc. (Rockville MD USA)	Solid tumours, Hematologic-blood cancer	Phase I
teclistamab	JNJ-64007957	Janssen Research & Development, LLC (Raritan NJ USA)	Multiple myeloma (MM)	Phase I
tepoditamab	MCLA-117, PB9122	Merus NV (Utrecht Netherlands)	Acute myeloid leukemia (AML)	Phase I/II
tibulizumab	LY 3090106	Eli Lilly (Indianapolis IN USA)	Sjögren's syndrome (SjS)	Phase I
tidutamab	XmAb-18087	Xencor Inc. (Monrovia CA USA)	Cancers, gastrointestinal	Phase I
ubamatamab	REGN4018	Regeneron Pharmaceuticals Inc. (Tarrytown NY USA)	Cancers, ovarian	Phase II
vibecotamab	XmAb14045	Xencor Inc. (Monrovia CA USA)	Acute lymphocytic leukemia (ALL)	Phase I
voxalatamab	JNJ-63898081, JNJ-8081	Janssen Research & Development, LLC (Raritan NJ USA)	Solid tumours	Phase I

vudalimab	XmAb-20717	Xencor Inc. (Monrovia CA USA)	Solid tumours	Phase I
N/A	4F2	MRC Technology (London UK) / Merck KgaA (Darmstadt Germany)	Cancers	Phase I
N/A	4D5-8, UCHT1, BsF(ab')2 v1, F(ab')2 4D5-8/UCHT1-v1	Charing Cross Sunley Research Centre (UK) / Genentech Inc. (S. San Francisco CA USA)	Cancers (overexpressing ERBB2)	Phase II
N/A	MDX-H210, 520C9XH22	Medarex (Princeton NJ USA)	Cancers, prostate	Phase II
N/A	MDX-447, F(ab')2 H425 (anti-EGFR)/H22 (anti-FcγRI)	Genmab A/S (Copenhagen Denmark) / Medarex (Princeton NJ USA)	Cancers	Phase II
N/A	MDX-220, F(ab')2 HCC49 (anti-TAG-72)/H22 (anti-FcγRI)	Genmab A/S (Copenhagen Denmark) / Medarex (Princeton NJ USA)	Acute myeloid leukemia (AML)	Phase I
N/A	MM-111	Merrimack Pharmaceuticals (Cambridge MA USA)	Cancers, gastric	Phase II
N/A	REGN1979	Regeneron Pharmaceuticals Inc. (Tarrytown NY USA)	Lymphoma, diffuse large B cell (DLBCL)	Phase I

Table 10: Engineered antibodies dataset. The information is extracted from publicly available online resources such as AdisInsight, IMGT® database, and ClinicalTrials.gov database.

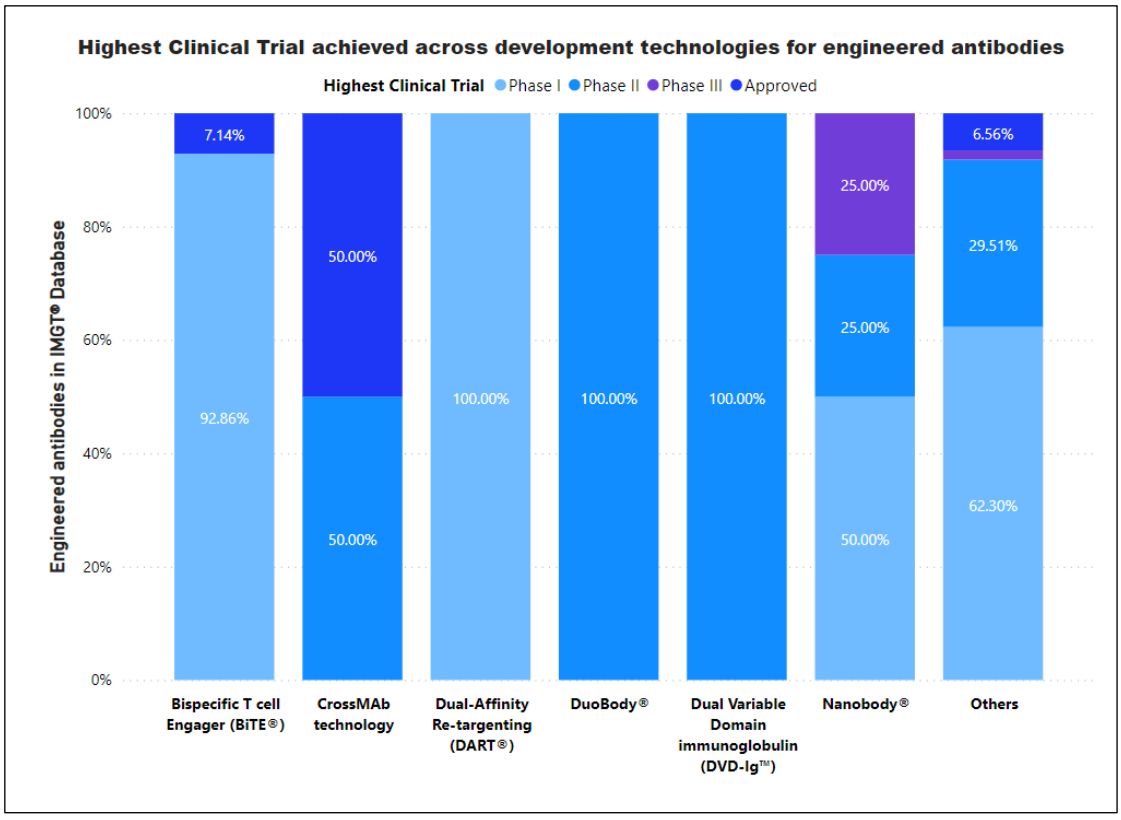


Figure 18: Distribution of engineered antibodies according to their clinical trial status for different categories of novel development technologies extracted from the IMGT® database.

Among the full-length mAbs, 27% are IgG1s, 8% are IgG2s, and 7% are IgG4s. An explanation of this trend can be the differences in the molecular attributes and effector

functions between these IgG subclasses. IgG1 has the highest FcγR-binding affinity, followed by IgG3, IgG2, and IgG4.[16] So, these IgG subclasses differ in triggering FcγR-expressing cells, which result in phagocytosis or antibody-dependent cell-mediated cytotoxicity and activating complement. These four subclasses - IgG1, IgG2, IgG3, and IgG4, are highly conserved, but differ in their constant region, particularly in their hinges and upper CH2 domains. A previous study has demonstrated that IgG1 mAbs are more prone to hinge region fragmentation compared to IgG2 and IgG4 mAbs under heat, and pH-related stress conditions.[17] So, IgG2 and IgG4 formats would be preferred under high thermal stress and long-term cold storage conditions. The remaining engineered biotherapeutics are various engineered Fabs and engineered scFv formats unique to each platform technology. The light chains for 78% of the 103 engineered antibody therapeutics are of kappa (κ) isotype, and the remaining 22% are lambda (λ) isotype. The engineered biotherapeutics serve several disease areas and clinical indications including cancers, solid tumours, multiple myeloma, acute myeloid leukaemia, and rheumatoid arthritis. A summary of all the major engineered antibody platform technologies has been provided below.

BiTE® (Bispecific T cell Engager): BiTE (bispecific T-cell engager) technology is a targeted immuno-oncotherapy platform that recruits patients' own cytotoxic T cells to tumour cells.[18] BiTEs consist of one arm with designed specificity towards antigen on tumour cells and other arm engineered to bind with a surface molecule on T cells linked by a peptide linker. BiTEs direct T cells' cytotoxic activity to eliminate tumour cells when both targets are engaged by their respective scFv arms. Blinatumomab was the first canonical BiTE molecule that targets CD19 surface antigens on B cells approved for the treatment of acute lymphocytic leukaemia (ALL) in December 2014.[19] The tumour-binding arm can be engineered to target different types of cancer. A considerable number of related bispecific T cell-recruiting antibodies which are potentially effective in tumour immunotherapy have been derived from BiTEs.[20]

CrossMAb technology: CrossMab technology uses crossover of different domains within the Fab-fragment within one arm of a bispecific IgG antibody to enable correct chain association while employing knob-into-hole technology to achieve correct heterodimerization of the heavy chains.[21, 22] CrossMabs have no chemical linkers or connectors. Overall, the CrossMab technology results in the correct assembly of the desired bispecific antibody especially due to use of domain crossover in the Fab region

to enforce correct light chain pairing. CrossMab technology has evolved in the past decade to be one of the most versatile and broadly applied technologies with nearly 20 bispecific antibodies developed by Roche and others have advanced to clinical trials.[23] Notably, the Ang-2/VEGF bispecific antibody Faricimab marketed as Vabysmo™ has been approved in 2022 and CD20/CD3 T cell bispecific antibody Glofitamab is currently in pivotal Phase 3 trials.

Dual Variable Domain immunoglobulin (DVD-Ig™): The DVD-Ig is an IgG-like molecule designed by inserting two variable domains in tandem through a short peptide linkage in the heavy and light chains.[24, 25] Overall, the target-binding variable domains of two mAbs are combined to create a dual-targeting tetravalent molecule. The fusion orientation of the two variable domains and the choice of linker sequence are critical to the functional activity and efficient expression of the DVD-Ig antibody. The domain flexibility in the DVD-Igs allows antigen binding with minimal steric hindrance. A DVD-Ig has many desirable properties of a normal IgG-like mAb such as high *in vivo* stability, excellent physicochemical and pharmacokinetic properties, good expression in mammalian cells, and amenability to large-scale manufacturing.[26]

Dual-Affinity Re-targeting (DART®): DART is an alternative bispecific antibody platform each Fv is formed by the association of a VL partner on one chain with a VH partner on the second chain in a VL_A-VH_B + VL_B-VH_A configuration.[27] Therefore, the heterobispecific DART structure consists of two covalently linked chains each with a unique binding site. MacroGenics' internal pipeline has over 100 DART molecules in preclinical and early clinical trials for multiple clinical indications such as cancer, autoimmune disorders, and infectious diseases. Previous studies have demonstrated an increased level of potency, higher magnitude of T cell activation, favourable safety, and pharmacokinetic properties for DART antibodies.[28] However, most of the DART molecules have been discontinued in Phase 2 and Phase 3 trials by the company.

DuoBody®: The Genmab DuoBody® platform is a versatile technology for creating bispecific antibodies by employing controlled fab arm exchange to combine two distinct binding specificities within the same molecule.[29] This controlled fab arm exchange is achieved by single matched mutation in the CH3 region of each parental IgGs that ultimately results in the heterodimerization of the heavy chain-light chain pairs. DuoBody antibodies, therefore, retain native IgG structure and are compatible

with additional Fc engineering strategies to form high purity bispecifics capable of large-scale manufacturing. In 2021, RYBREVANT™ received the U.S. FDA approval which is the first among the therapies created using DuoBody technology platform.[30]

Nanobody®: Nanobodies or V_{HH} fragments are the recombinant variable domains of heavy-chain-only antibodies with many unique properties such as small size, superior stability and solubility, minimal cross-reactivity, and deep tissue penetration.[31, 32] Nanobody molecules are derived from llamas, alpacas, and other species that have only "heavy-chain" peptides. Such heavy chain fragments when connected like beads on a string, exhibit multivalent binding to many different targets at once.

Several other mature commercial technology platforms are available in the market for generation of multispecific antibodies.[33] Some of the key other formats from over 30 formats are ADAPTIR, ART-Ig, BEAT, DAF, DutaFab, Hetero-Ig, IgG-scFv, Multiclonics, Pentambody, Tandab, Triomab, XmAb, VELOCI-Bi, and WuxiBODY.

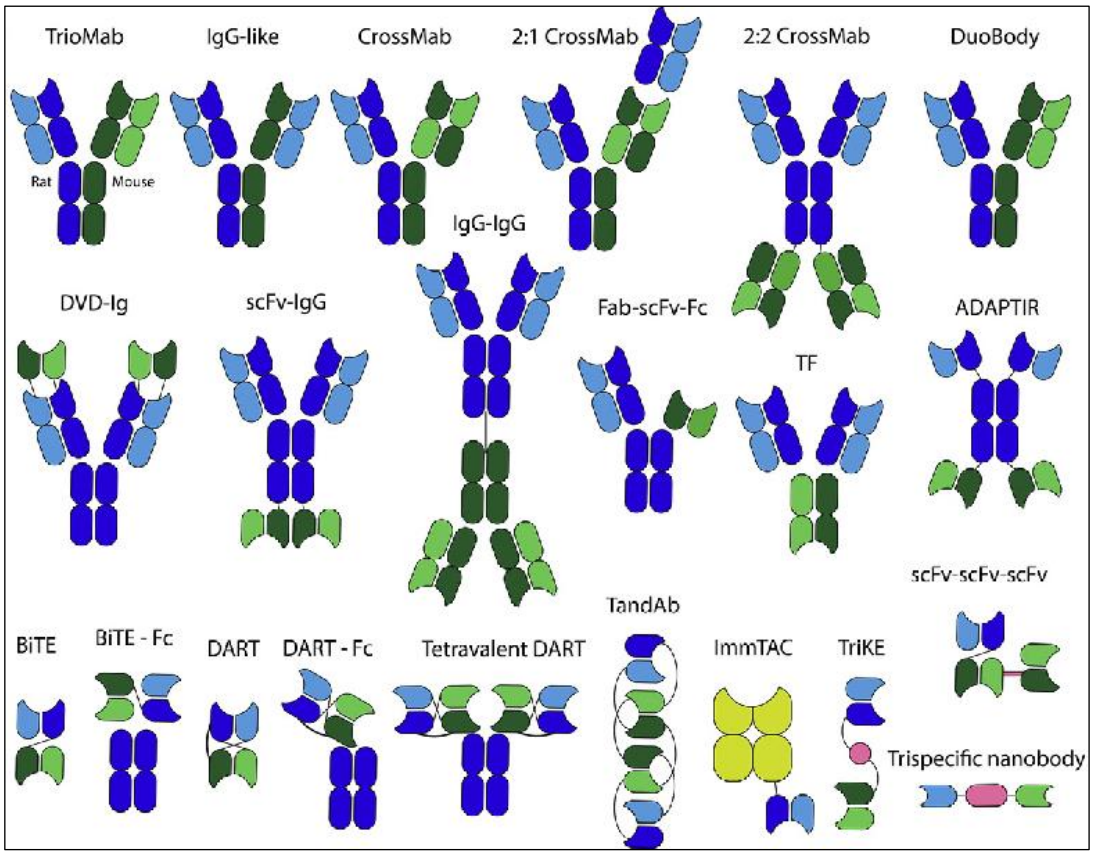


Figure 19: Common engineered bispecific and multispecific antibody formats. The dark blue and dark green represent the heavy chains, while the lights blue and green colour chains represent the light chains. Image adapted from Suurs, Frans V., et al review.[34]

The first subset of engineered antibody formats is the engineered antibody fragments which use Fabs, Fc region, and scFvs as the building block. These engineered antibody

fragments are linked by short peptide sequences which can be made into bivalent, trivalent, or tetravalent formats. Antibody discovery technologies such as dual affinity re-targeting (DART), bispecific T-cell engager (BiTE), tandem antibodies (TandAbs), and nanobody platforms fall in this category.

The next engineered antibodies subset contains multivalent IgG-like antibodies with heterodimeric heavy chains. The knobs-into-holes (kih), Duobody, Triomab, and other similar technologies created to promote correct heavy and light chain pairing fall in this category. Next category is IgG fusion antibodies and chemically coupled antibody fragments. These are comprised of IgGs with other binding domains fused to either the N or C terminus of either the heavy or light chains. So, these may take form of IgG-scFv; IgG-dAb; MAbtyrin, and DVD-Ig antibodies. Finally, fusion proteins and related formats such as Fc-fusions, kih-Fc fusions, and bispecific-Fc fusions are very popular as they provide the ability to combine additional receptor binding sites within the final antibody structure. These engineered formats are shown in Figure 19.

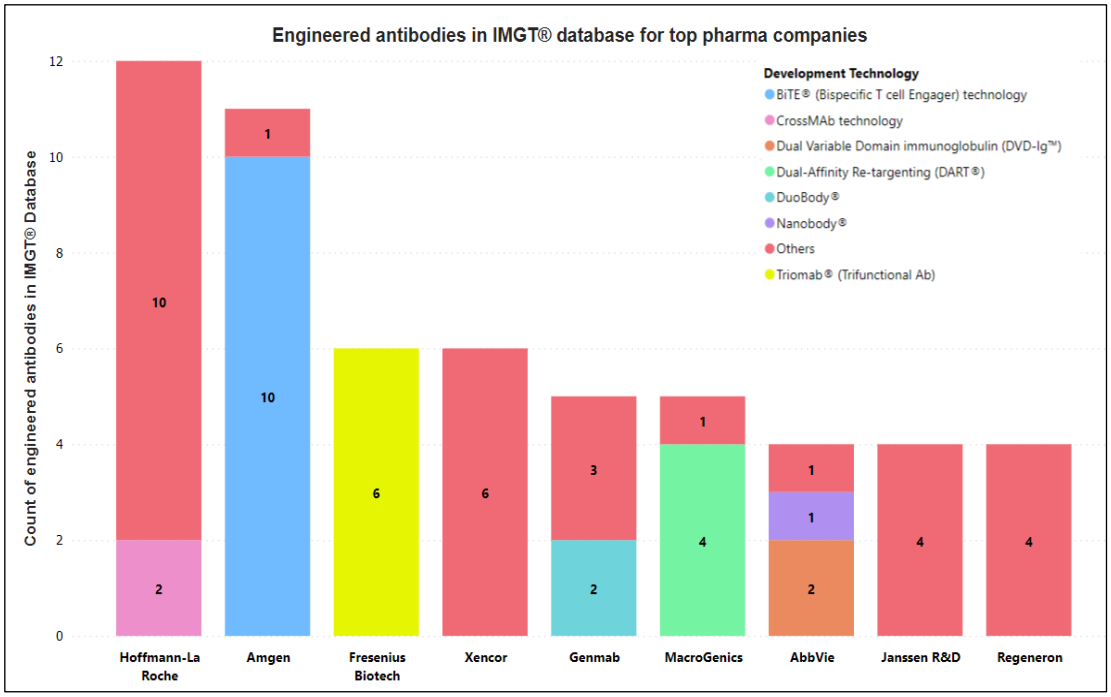


Figure 20: Engineered antibodies platforms in IMGT® database for top pharma companies.

Our created engineered antibodies dataset has been sourced from various development technologies that are proprietary platform technologies in antibody discovery for top pharma and biotechnology companies. Figure 20 above shows the allotment of various antibody platform technologies between the pharma and biotech companies. Roche

and Amgen are the key players in the development of engineered antibodies. Amgen is pioneering BiTE® technology to advance the immuno-oncology field and bring new therapeutic approaches to patients. BiTE® (Bispecific T-cell Engagers) are the most abundant engineered antibodies in clinical trials representing 14% of antibodies in this dataset. Currently, Amgen is designing BiTE® molecules with additional features to extend the serum half-life by a few hours.

Roche has new CrossMAb antibodies under development and has been also advancing other formats such as DutaMab™ technology acquired from Dutalys in 2014. Genmab has also developed a robust pipeline of products using DuoBody®, HexaBody®, and DuoHexaBody® platforms. MacroGenics is also advancing multi-specific platforms such as DART® and TRIDENT®. Several other novel formats have been included as well in our dataset under others category. So, overall, the dataset is well balanced with sufficient representation from all major platform technologies in commercial use.

4.4.1 Computational developability analysis of biophysical performance

We started the computational developability assessment of the engineered antibodies dataset with the AbPred tool. Figure 21 shows the AbPred scores for some biophysical assays across different engineered antibody formats. It has different categories on the x-axis in the order starting from left namely - BiTE®, CrossMAb, DART®, DuoBody®, DVD-Ig™, Nanobody®, and others. The y-axis shows assay scores with the arrow on y-axis indicating the direction of unfavourable assay values. Also, the 5% and 10% cutoff values from our previous developability assessment benchmarks are shown in yellow and red colour respectively for each biophysical assay.

We observe interesting developability insights for each technology from below results. CrossMAb and Nanobody® antibodies displayed very high hydrophobicity in the HIC assay. All the CrossMab format antibodies breached the 10% cutoff (in yellow) from clinical-stage therapeutics with 75% of the CrossMab antibodies above even the 5% cutoff (in red). All other platform technologies except nanobodies had excellent HIC performance with most antibodies within the 10% cutoff threshold value (10.848) – BiTE (100%); DART (100%); Duobody (100%) and DVD-Ig (100%). A similar trend was observed for the SMAC assay with CrossMab and Nanobody® above 10% cutoff.

However, the SGAC assay values were comparable and optimal within the 10% cutoff for all platforms. So, the SGAC assay scores indicate that all engineered antibody platforms have very low self-association and thus we expect good colloidal stability for these engineered antibodies. We hypothesize that this optimal self-association property may be explained by the inherent engineering of binding domains to not self-interact often through a constrained pairing of heavy and light chains. So, there are fewer unpaired binding sites and 'charge hotspots' in the engineered antibodies that result in minimal self-association.

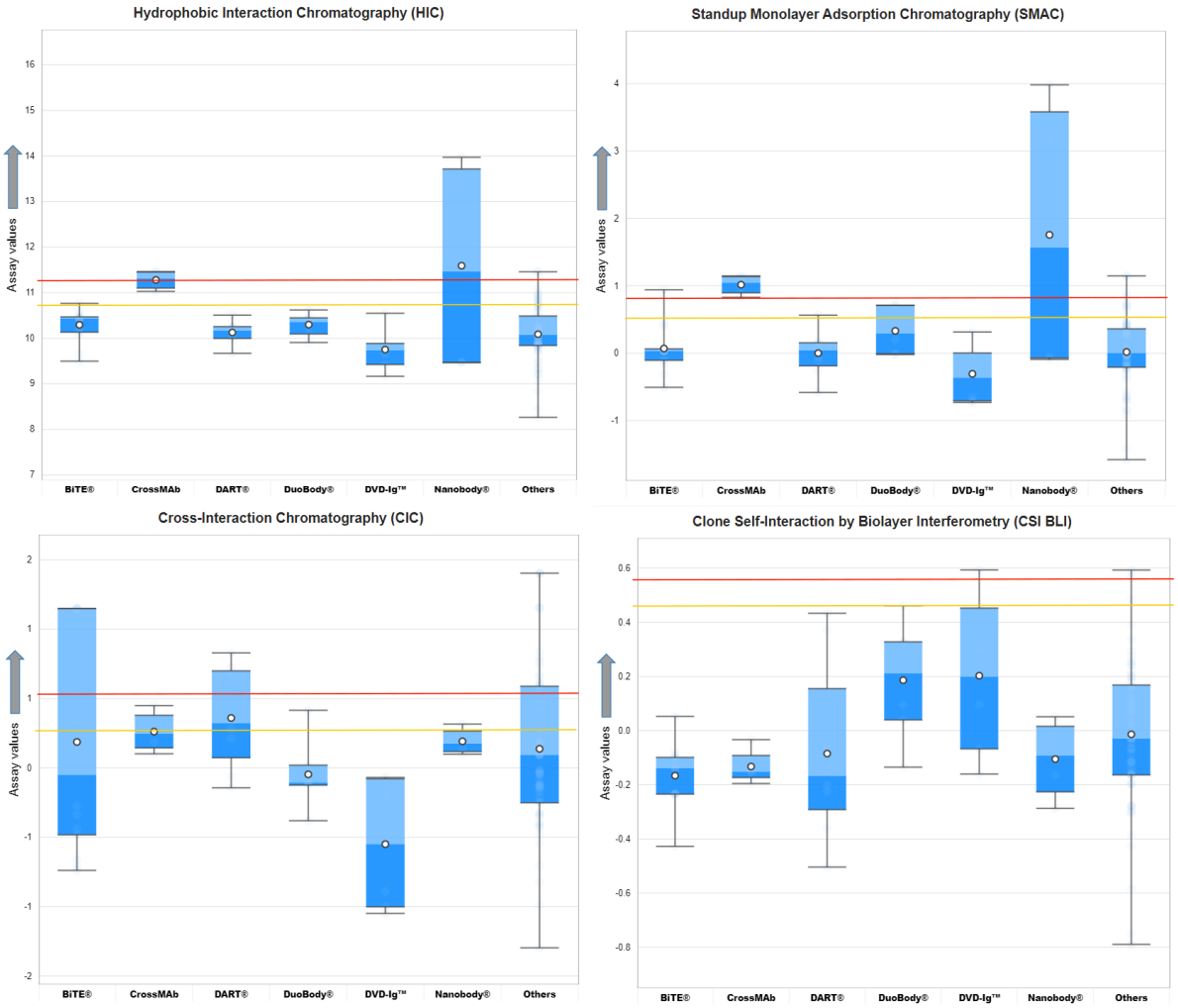


Figure 21: AbPred scores for four developability assays for different categories of engineered antibodies shown on x-axis. The arrow on y-axis indicates the direction of unfavorable values.

Further evidence of low self-association in engineered antibodies is also seen in CSI BLI assay results where all platforms have values under the 10% cutoff representing comparable self-association to successful clinical-stage antibodies. For the PSR assay measuring the polyspecificity, we observe that a fraction (21%) of the BiTE antibodies were above the 10% cutoff but not above the 5% cutoff which indicates a concern for

off-target effects in clinical trials for BiTEs. Interestingly, the antibody platform with the most optimal PSR assay performance was CrossMAb with possibly the lowest off-target binding. CrossMab antibodies also had the best performance in the ELISA assay, which is another measure of multiantigen nonspecificity. So, this implies that the CrossMAb engineering has designed the antibody molecules for lower off-target binding but at a corresponding risk of high hydrophobicity, established earlier.

The DVD-Ig platform had the best overall developability profile among all categories because it was mostly within the 10% threshold cutoffs for all assays while the other formats breached the 10% cutoffs in more than one assay. It had the lowest values in the CIC assay representing the lowest cross-reactivity among all formats. DVD-Ig antibodies had the lowest mean (-0.55); quartile 1 (-1.00); median (-0.55); quartile 3 (-0.08) and upper whisker (-0.07) values for the CIC assay as shown in Figure 22.

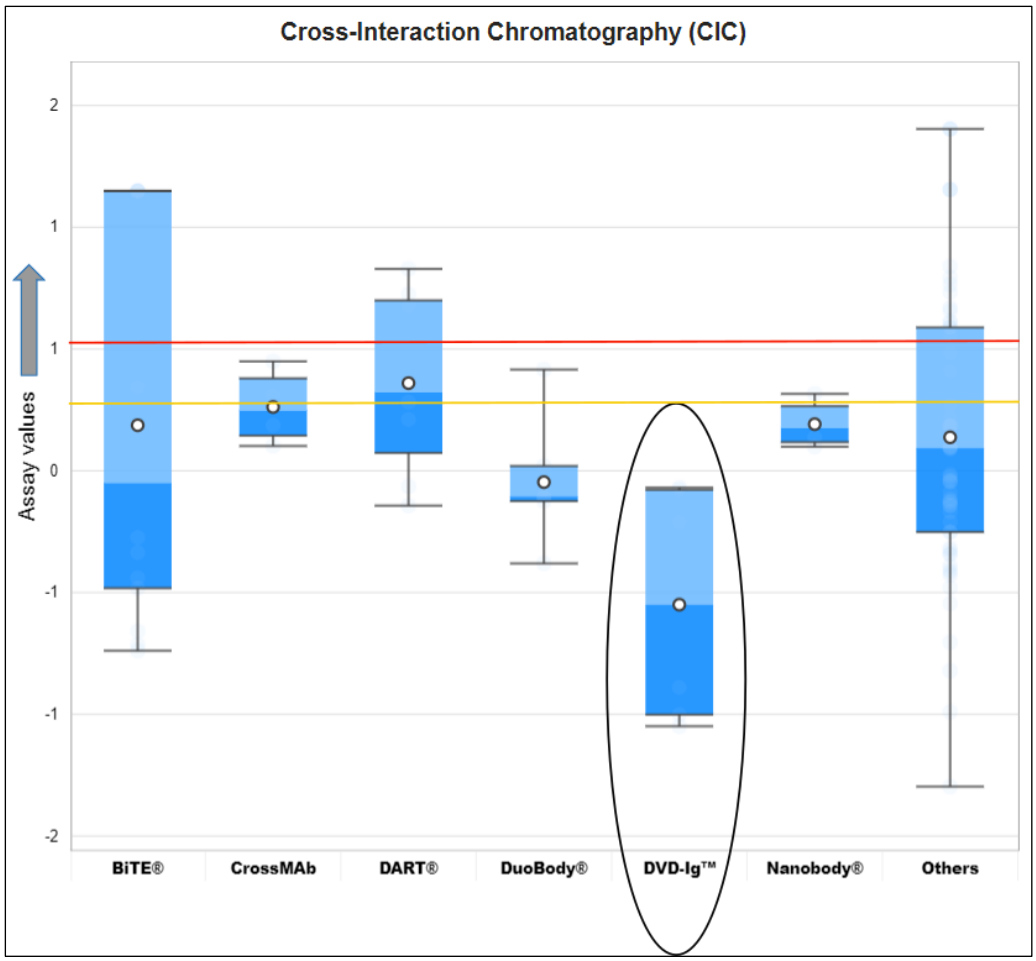


Figure 22: Cross-Interaction Chromatography (CIC) assay values for different categories of engineered antibodies. The arrow on y-axis indicates the direction of unfavorable values.

A similar trend is observed for the AC-SINS assay with DVD-Ig platform showing the lowest self-association behaviour. This validates an excellent interaction and binding profile for the DVD-Ig antibodies. Overall, there were no developability issues identified for DVD-Ig antibodies from our developability assessment.

We therefore conclude that DVD-Ig™ platform has lowest non-specific interactions and cross-reactivity reflecting best overall developability profile among all engineered antibody platforms. A possible explanation is that the scFv-based constructs may have constraints imparted by the linker sequences and a tendency to form aggregates due to domain exchange of the variable regions with partners from other molecules leading to a poor developability profile. While DVD-Ig antibodies have a similar configuration to a conventional IgG with IgG-like features and exceptional domain flexibility.[35]

An extensive literature review revealed a possible mechanistic insight into the superior developability profile for the DVD-Ig™ platform. A previous study has suggested the role of the additional variable domain region via linkers that imposes limits on target size and location, and limits conformational changes (stabilization) upon target binding.[24] Therefore, by design since DVD-Ig antibodies have an addition of the first variable domain to the second variable domain via a flexible linker sequence, they achieve lower non-specific interactions with overall excellent developability.

4.4.2 Case Study 1: Bispecific antibody formats

This case study aims to characterize the biophysical behaviour of bispecific antibody formats specifically such as dual-scFv, BiTEs, CrossMab Fabs, and DART molecules using available developability assessment tools. In this case study, we evaluated the structural features such as surface patches, charge, and hydrophobicity for a custom bispecific antibody dataset. We have then used antibody informatics tools such as ProteinSol and AbPred to identify the key biophysical features that decide bispecific antibody developability. The results were used to interpret the sequence and structural liabilities which are then used to guide future antibody engineering approaches.

An extensive search was performed on PDB for engineered antibody fragments and 10 engineered antibody fragments which have bispecific functionality were shortlisted for analysis. We characterized the charge and hydrophobicity of all available 10 bispecific antibody fragments.

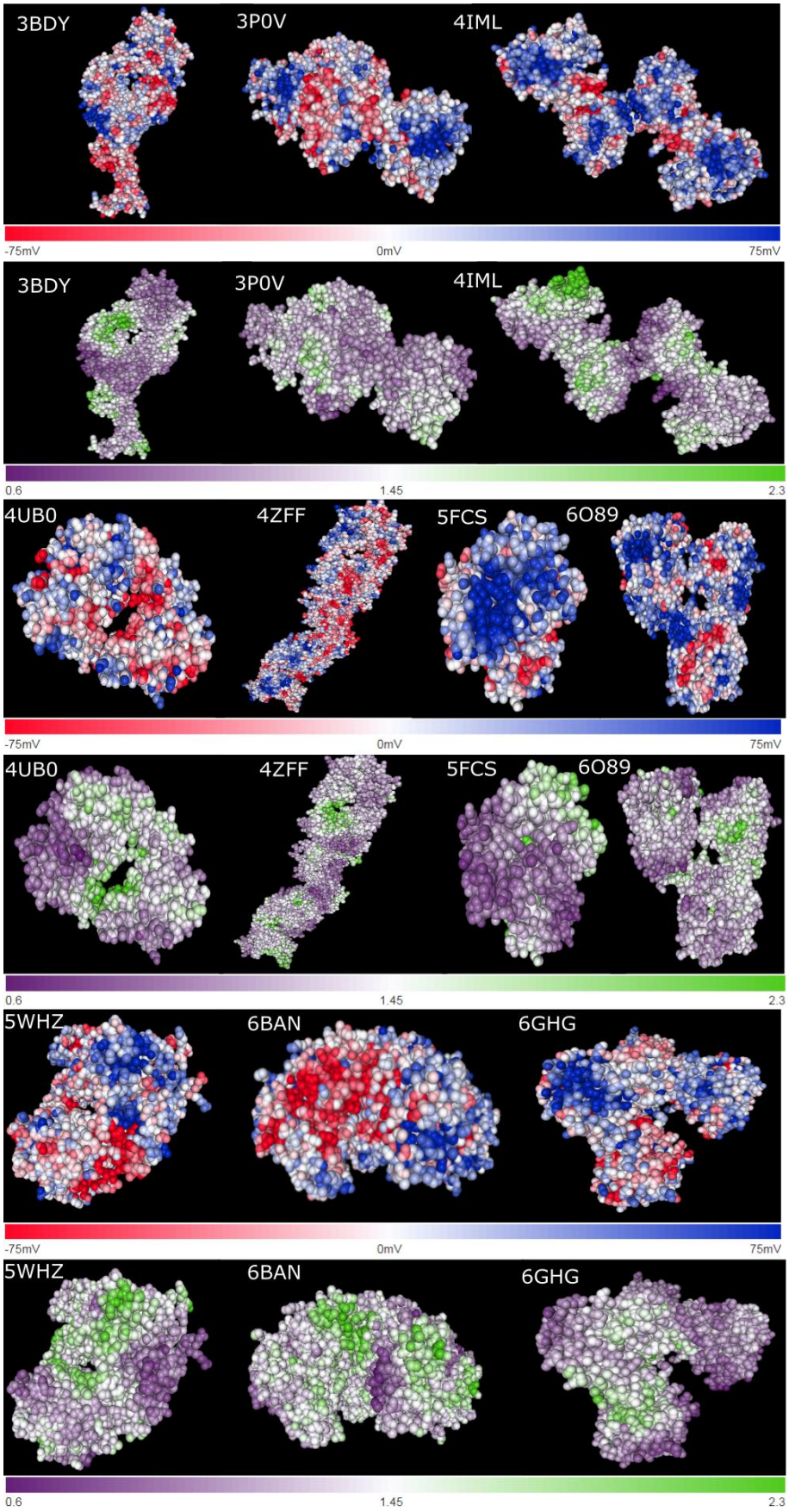


Figure 23: Protein-Sol visualization of charged and hydrophobic surface patches on each bispecific antibody. (A) The Fab is colour-coded from negatively charged (red) to positively

charged (blue). (B) The Fab is colour-coded from polar (purple) to non-polar (green) where the scale value represents the patch NPP ratio.

These are represented by the following PDB IDs:

1. 3BDY - Dual specific bH1 Fab (HER2 X VEGF)
2. 3P0V - Anti-EGFR/HER3 Fab DL11
3. 4IML - Anti-Ang2 CrossFab
4. 4UB0 - Monovalent bispecific IgG 'DuetMab'
5. 4ZFF - Dual-acting Fab 5A12 (Two-in-One VEGF/angiopoietin 2)
6. 5FCS - PF-06671008, Anti-P-cadherin/Anti-CD3 Bispecific DART Molecule
7. 5WHZ - PGDM1400-10E8v4 CODV Bispecific Fab
8. 6BAN – ROR1 BiTE, T-Cell Engaging Bispecific Antibody
9. 6GHG – Roche CrossMab
10. 6O89 - Anti-CD28xCD3 CODV Fab

Figure 23 presents the visualization of charged and hydrophobic surface patches for each bispecific antibody. These figures are visualized using the embedded NGL viewer on the ProteinSol patches software available online at Protein-Sol tool.[36]

We have used the ProteinSol patches software to quickly identify hotspots of relative hydrophobicity (higher NPP ratio). The 10 antibody fragments have been colour-coded red for negatively charged to blue for positively charged surface patches. Also, for the hydrophobicity visuals the antibody fragments are colour-coded purple for polar to green for hydrophobic patches where the scale represents the polarity ratio.

Next, we used AbPred tool to predict biophysical properties for 12 known biophysical platforms. Figure 24 shows the final heatmap image for all 10 bispecific antibodies. The heatmap is colour coded for each scFv dependent on a threshold value calculated by taking the worst 10% cutoff for the predicted Jain dataset values. If the predicted value is above the threshold value for the experiment, the corresponding square is coloured red, otherwise coloured green. So, here green squares represent good developability while red squares indicate poor developability profile.

The protein-sol sequence algorithm calculates 35 sequence features:- Composition of the standard 20 amino acids, sequence length (len), Lysine minus Arginine (KmR), Aspartic Acid minus Glutamic Acid (DmE), Lysine plus Arginine (KpR), Aspartic

Acid plus Glutamic Acid (DpE), K+R-D-E (PmN), K+R+D+E (PpN), Phenylalanine + Tryptophan + Tyrosine (aro), folding propensity (fld), Disorder propensity (dis), Beta strand propensities (bet), Kyte-Doolittle hydropathy (mem), pI, Sequence entropy (ent), Absolute charge at pH 7 (abs).[37] So, we have characterized the sequence feature scores of our selected 10 bispecific fragments based on these 35 sequence features calculated by the ProteinSol algorithm. A previous work by Hebditch *et al.* has correlated the 35 sequence features to the 12 Abpred assays which will be used to interpret the underlying sequence features that are the reason for poor developability.[38]

Figure 24: Predicted performance on 12 Jain biophysical platforms for bispecific antibody fragments. The 10 bispecific formats under study are represented by their 4-digit PDB codes.

We conclude that three bispecific antibodies namely 3P0V (Anti-EGFR/HER3 Fab DL11), 4ZFF (Dual-acting Fab 5A12), and 5FCS (Anti-P-cadherin/Anti-CD3 DART molecule) have developability liabilities. Another bispecific antibody in our dataset, 5WHZ (PGDM1400-10E8v4 CODV Bispecific Fab) is also predicted to have poor HEK titer expression and thermal stability but optimal other biophysical features. The anti-EGFR/HER3 Fab (3P0V) is predicted to have an unfavourable performance on SGAC-SINS, CSI-BLI, AC-SINS, ELISA, BVB, and DSF. So, using the Pearson

correlation coefficient matrix[38] by Hebditch *et al* we attribute the poor developability to high net absolute charge (abs), low content of aspartic acid (D), and high arginine-lysine content (KpR). This is further validated by a high value of KpR (1.312) and a high value of absolute charge (+0.148) from the Protein-Sol score. We also find these positive charge hotspots for 3P0V on the patches visualization. So, future protein-engineering efforts for this anti-EGFR/HER3 Fab must be directed towards increasing the aspartic acid content, lowering arginine-lysine residues, and thus lowering charge.

The dual-acting Fab 5A12 (4ZFF) is predicted to have an unfavourable performance on HIC, SMAC, and SGAC assays. So, using the Pearson correlation coefficient matrix we attribute the poor developability to a high aromatic amino acid (aro=F+W+Y) content with a score of 2.547 and an excess of Tyrosine (Y). So, future protein-engineering efforts for 4ZFF must be directed towards lowering the aromatic amino acid content to achieve optimal biophysical properties.

The anti-P-cadherin/anti-CD3 DART bispecific antibody (5FCS) is predicted to have a poor performance on SMAC, SGAC, CIC, AC-SINS, and PSR. So, using the Pearson correlation coefficient matrix we attribute the poor developability to net absolute charge (abs), the content of tyrosine (Y) and aromatic amino acids (aro), the content of aspartic acid (D) and PmN (K+R-D-E). Low Protein-Sol scores in aromatic amino acid content (0.713) and aspartic acid (-1.75) for this DART antibody were the key sequence liabilities. So, proposed further optimization in DART format pipeline is to add aromatic amino acids and aspartic acid residues to the Fab or linker region.

Finally, the 10E8v4 CODV Bispecific Fab (5WHZ) is predicted to have a poor performance on HEK and DSF assays. So, using the Pearson correlation coefficient matrix we attribute the problem to beta-strand propensity (bet) and low alanine (A), proline (P), and glutamine (Q). Low Protein-Sol scores of alanine (-2.132), glutamine (-0.66), and beta-strand propensity (-0.462) for 5WHZ bispecific confirm the sequence liabilities. So, future protein-engineering efforts must be directed towards increasing beta-strand propensity along with alanine and glutamine content.

4.4.3 Case Study 2: Azymetric™ antibody therapeutics

Zymeworks' Azymetric™ drug discovery platform enables the transformation of monospecific antibodies into bispecific and multispecific antibodies using proprietary

amino acid modifications.[39] The Azymetric™ technology is based on a spontaneous assembly of two different Fab domains consisting of unique heavy-chain and light-chain pairings into a single molecule based on a library of proprietary amino acid substitutions. Such novel bispecific antibodies designed using Azymetric™ platform resemble conventional monospecific antibodies in structural features while having a bispecific functionality. Azymetric™ bispecific technology enables the development of biotherapeutics with unique mechanisms of action not accessible through typical monospecific antibodies. The engineered bispecific dual-targeting properties are used for synergistically blocking multiple signalling pathways, increasing tumour-specific targeting, recruiting immune cells to tumours, and enhancing receptor clustering.

Bispecific antibodies sourced from the first-generation engineered mAb platforms have a significantly divergent structure from usual monoclonal antibodies and often require complex manufacturing processes. In contrast, Azymetric™ bispecifics retain the desirable developability properties and qualities such as high stability, long serum half-life, and low immunogenicity risk which are similar to their monospecific counterparts. Azymetric™ bispecifics in the preclinical development have excellent manufacturability as they are compatible with standard manufacturing processes with high yields and purity, which accelerates manufacturing timelines and reduces costs.

In this case study, we have performed a computational developability assessment on Zanidatamab – a HER2 x HER2 Azymetric™ bispecific lead that is currently being evaluated in Phase 1 / Phase 2 clinical trials as a best-in-class treatment for patients with human epidermal growth factor receptor 2 (HER2) expressing cancers, including biliary tract, gastroesophageal adenocarcinomas, breast, and other tumour types.[40]

The Abpred scores and ranking relative to Jain clinical-stage mAbs for Zanidatamab are shown in Figure 25. We observe an excellent rank of this Azymetric™ bispecific antibody in all 12 assays – HIC (51/100); SMAC (48/100); SGAC (50/100); CIC (38/100); CSI BLI (43/100); AC-SINS (51/100); HEK (43/100); PSR (79/100); ELISA (67/100); BVP (34/100); DSF (37/100) and ACC STAB (51/100). These are normalized ranks scored out of 100 relative to Jain dataset clinical-stage antibodies. We, therefore, conclude that the Azymetric™ bispecific antibody performs at par with the clinical-stage antibodies in all biophysical assays. So, the transformation from a monospecific to a bispecific format did not hamper the biophysical performance.

Zanidatamab Azymetric™ antibody also has excellent META scores of 38/100 for the Group X assays and 36/100 for the Group Y assays as shown in Figure 26. Here, the META score combines and averages multiple biophysical platforms. The meta score is calculated by ranking the Jain dataset results in order from best to worst rank, and then calculating where the input candidate sequence falls within that ranking for each biophysical platform. Group X consists of charge-based assays namely ELISA, BVP ELISA, PSR, CSI, ACC STAB, and CIC. Group Y consists of hydrophobicity-based assays SMAC and HIC. The green heatmap also provides the confirmation that all Azymetric™ antibody scores are within the worst 10% cutoff for Jain dataset values.

Next, we evaluated Zanidatamab using the Therapeutic Antibody Profiler (TAP). The TAP results are shown in Figure 27. Zanidatamab had an excellent score in all five structural TAP metrics – Total CDR length (47); PSH (121.464); PPC (0.3201); PNC (0.1336) and SFvCSP (8.0). These scores are also within the amber flag and red flag cutoffs shown in Figure 27 as colour-coded vertical lines on the x-axis.

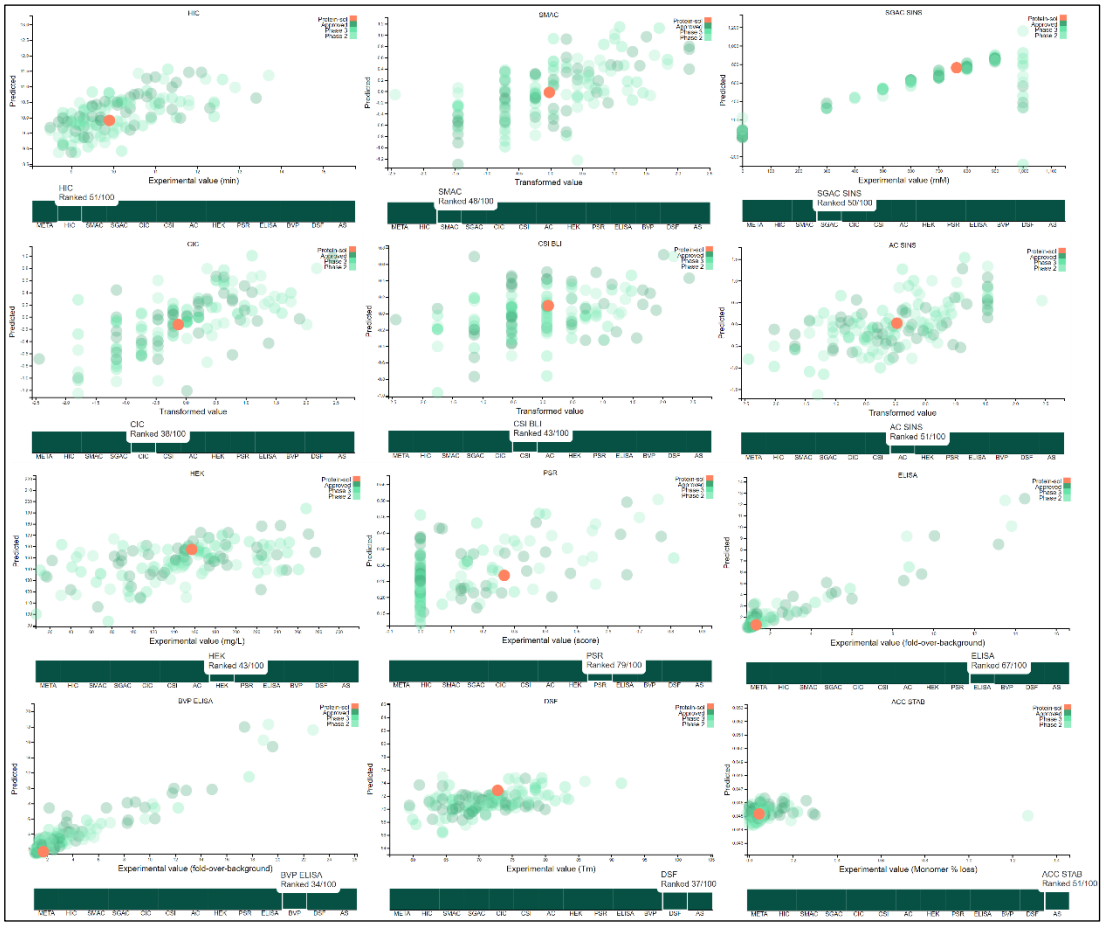


Figure 25: AbPred scores for Azymetric™ antibody on 12 developability assays. Heatmap rank provided at the bottom for each assay. Zanidatamab – the lead Azymetric™ antibody is shown in red while Jain clinical stage antibodies are shown in green in each scatter plot.

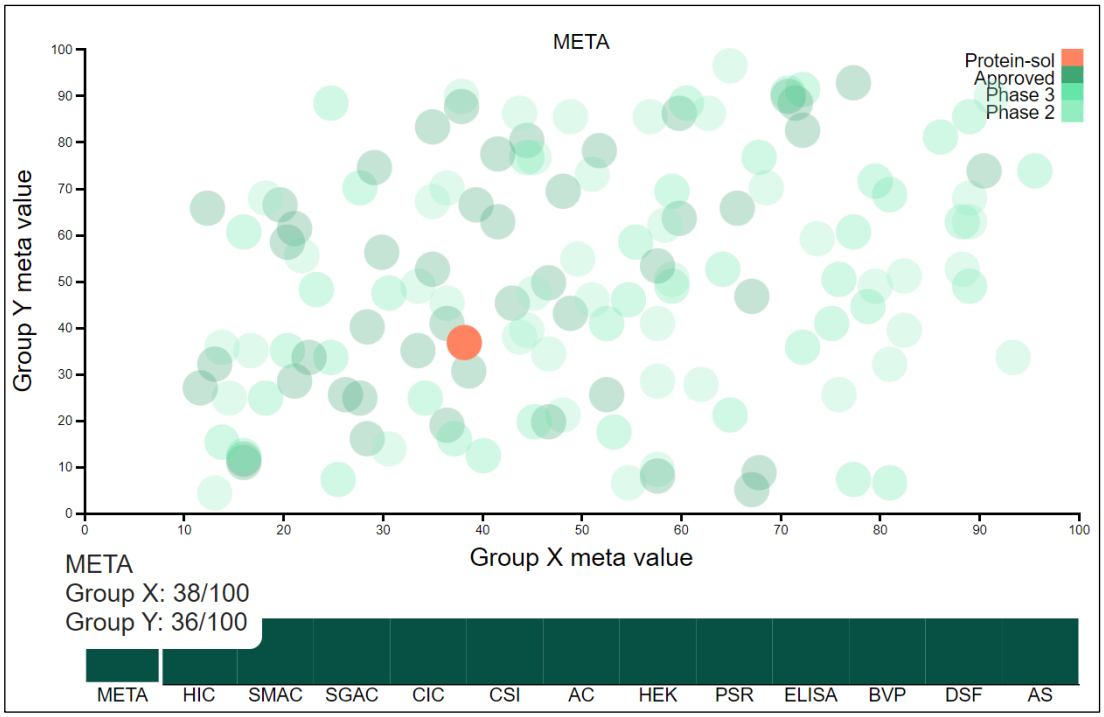


Figure 26: Meta score which combines and averages multiple biophysical platforms for the Azymetric™ antibody. Group Y: HIC and SMAC. Group X: Other Charge based assays.

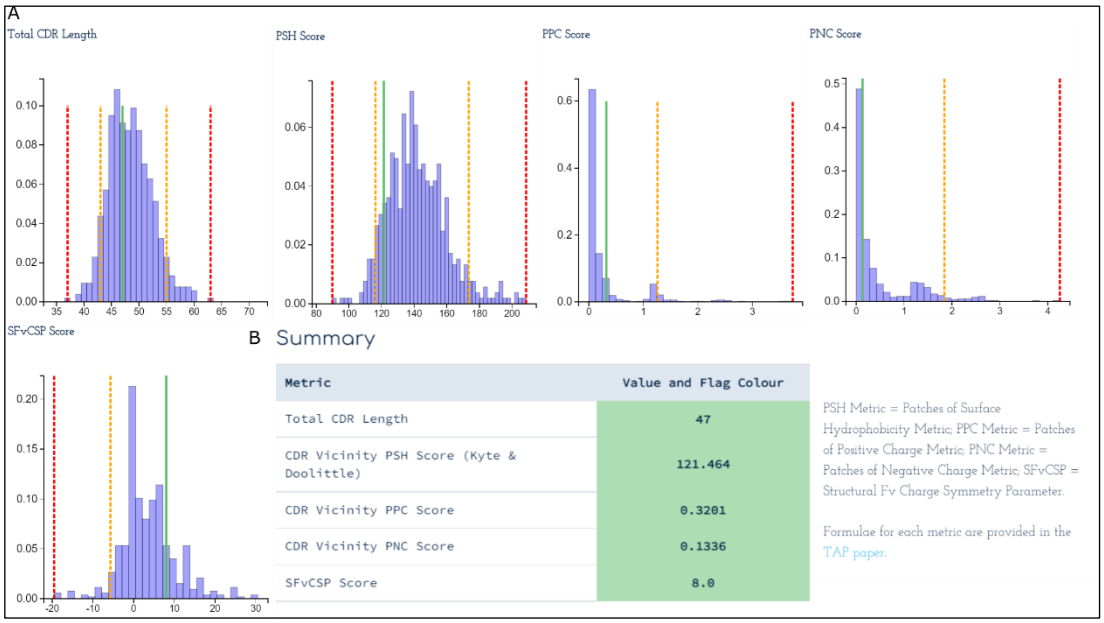


Figure 27: TAP results for the Azymetric™ antibody. Part A – Score histograms for all five developability metrics. Part B – Summary of TAP scores and flag colour for Zanidatamab.

Therefore, we conclude from our computational developability assessment results that the Azymetric™ antibodies have an excellent developability profile. No sequence and structural liabilities were observed across any tool. So, we predict favourable clinical trial success towards approval for Zanidatamab in current phase II clinical trials. Such traction is evident in Zymeworks' licensing deals with key pharma/biotech partners.

Part 2 – Evaluating different antibody discovery technologies:

4.5 Antibody phage display library dataset

Phage display has emerged as a powerful platform for the discovery of therapeutic antibodies since antibody phage display was first developed by George P. Smith and Sir. Gregory Winter in the beginning of the 1990s, who were eventually awarded the Nobel Prize in Chemistry in 2018 for their work on development of phage display of peptide and antibodies. Antibody phage display libraries involve the isolation of fully human-derived mAbs from large Ig gene repertoires displayed on the surface of bacteriophages. Therefore, antibody phage display technology creates combinatorial antibody libraries on filamentous phages for generating desired antigen specific mAbs.

The phage display technology is based on the fundamental biology of filamentous phages which are able to display a peptide of interest on their surfaces after inserting a foreign DNA fragment into the filamentous phage coat protein gene.[41] Phage libraries generated from human rearranged V-gene repertoires are constructed from mRNA or RNA extracted from B cells of immunized or native donors. Construction of immunized or native libraries involves using reverse transcription polymerase chain reaction (RT-PCR) to prepare the cDNA template. This is followed by the amplification of the repertoire of VL and VH genes by PCR, before cloning into the phagemid.[42] Antibody phage display is a versatile, *in vitro* selection technology that can be utilized to discover high-affinity antibodies specific to a wide variety of antigens. A previous review has detailed all the approved mAbs derived from phage display technology.[43] However, it is important to note that the success of isolating useful antibodies in phage display is highly dependent on the quality and the nature of the targeted antigen used in biopanning and the size and quality of the library.

A dataset of over fifty antibodies derived from phage display libraries for multiple clinical indications and originating from diverse development technologies was extracted from the IMGT® database. We then compiled the V_H-V_L sequence and 3D structure information for these antibodies from several online resources. The phage display antibodies with missing sequence information were removed for further analysis while scFv fragment from each phage display antibody was used as a separate query in the computational developability assessments. Finally, we had a dataset of 40 phage display scFv fragments shown in Table 11 for evaluating developability.

Our antibody phage display library dataset provides publicly available information on various phage display antibody therapeutics that were available in the market or were being investigated in human clinical trials as of 2023. In this dataset, there are a total of 40 antibodies with 11 (27.5%) in phase-I trials, 17 (42.5%) in phase–II trials, and 7 (17.5%) in phase-III clinical trials. Five phage display antibodies have been approved and have received the regulatory and marketing authorizations for various clinical indications – adalimumab (HUMIRA®), romiplostim (NPLATE®), tralokinumab (ADBRY™), atezolizumab (TECENTRIQ®), and tafasitamab (MONJUVI®).

INN (International Nonproprietary Name)	Common name / Proprietary name	Company	Clinical indication	Highest Clinical Trial
Artificial human phage display library				
lesabelimab	LDP	Dragon Sail Pharmaceutical Co. Ltd. (Shanghai China)	Cancers	Phase II
reozalimab	IBI 318	Innovent Biologics (Suzhou China)	Cancers, NSCLC	Phase I
atezolizumab	TECENTRIQ®, MPDL3280A	Roche Ltd. (Basel Switzerland) / Genentech Inc. (San Francisco CA USA)	Cancers, breast Solid tumours	Approved
fazpilodemab	BFKB8488A	Roche Ltd. (Basel Switzerland)	Diabetes mellitus (DM)	Phase I
trontinemab	RG-6102, RO7126209	Roche Ltd. (Basel Switzerland)	Dementia, Alzheimers	Phase I/II
Cambridge Antibody Technology (CAT) human antibody phage display library				
adalimumab	HUMIRA® (EU/US)	Cambridge Antibody Technology (Cambridge UK)	Crohn's disease, Rheumatoid arthritis (RA)	Approved
belimumab	BENLYSTA®, LymphoStat-B	GlaxoSmithKline (Brentford UK) / Human Genome Sciences Inc. (HGSI) (Rockville MD USA)	Systemic lupus erythematosus (SLE), Vaculitis	Approved
bertilimumab	CAT-213, iCo-008	iCo Therapeutics Inc. (Vancouver BC Canada)	Crohn's disease, Colitis, ulcerative (UC)	Phase II
fresolimumab	GC-1008	Genzyme Corp. (Cambridge MA USA) / Cambridge Antibody Technology (Cambridge UK)	Idiopathic pulmonary fibrosis	Phase I
mapatumumab	HGS-ETR1, TRM-1	Cambridge Antibody Technology (Cambridge UK) / Human Genome Sciences Inc. (HGSI) (Rockville MD USA)	Cancers, colorectal (CRC) Non-Hodgkin's lymphoma (NHL)	Phase II
mavrilimumab	CAM-3001	MedImmune (Gaithersburg MD USA) / AstraZeneca (London UK)	Rheumatoid arthritis (RA)	Phase II
tralokinumab	ADBRY™, CAT-354,	MedImmune (Gaithersburg MD USA) / AstraZeneca (London UK) / LEO pharma (France Versailles)	Atopic dermatitis (AD), Idiopathic pulmonary fibrosis	Approved
Dyax human antibody phage display library				
cixutumumab	IMC-A12, LY3012217	Eli Lilly (Indianapolis IN USA) / ImClone Systems Inc. (Somerville NJ USA)	Cancers, non-small cell lung (NSCLC)	Phase II
necitumumab	PORTRAZZA™, IMC-11F8, LY3012211	Eli Lilly (Indianapolis IN USA) / ImClone Systems Inc. (Somerville NJ USA)	Cancers, colorectal (CRC)	Approved

opicinumab	BIIB-033	Biogen, Inc. (Cambridge MA USA)	Multiple sclerosis (MS)	Phase II
ramucirumab	CYRAMZA™, 1121B, IMC-1121B, LY3009806	Eli Lilly (Indianapolis IN USA) / ImClone Systems Inc. (Somerville NJ USA)	Cancers, bladder Solid tumours	Approved
seribantumab	MM-121, SAR256212	Sanofi (Paris France) Merrimack Pharmaceuticals (Cambridge MA USA)	Cancers, ovarian Cancers, breast	Phase II
Other human antibody phage display technologies				
adecatumumab	MT201	Micromet AG (Munich Germany)	Breast Cancer, Prostate Cancer	Phase II
avelumab	BAVENCIO®, MSB-0010718C, MSB0010682	Merck Serono International S.A. (Geneva Switzerland)	Cancers, non-small cell lung (NSCLC)	Phase III
camoteskimab	AEVI-007, AVTX-007, CERC-007, CERC007	Avalo Therapeutics (Wayne PA USA)	Adult-onset Still's disease (AOSD)	Phase I
carlumab	CNTO 888	Centocor Inc. (Horsham PA USA)	Cancers, Pulmonary Fibrosis	Phase II
namilumab	MT203	Micromet Inc. (Munich Germany) / Nycomed (Zurich Switzerland)	Rheumatoid arthritis (RA), Multiple sclerosis (MS)	Phase I
orticumab	BI-204, R-7418	BioInvent (Lund Sweden)	Atherosclerosis, Acute coronary syndrome (ACS)	Phase II
izalontamab	SI-1, SI-1X6.4, SI-B001	Systimmune Inc (Redmond WA USA)	Cancers, colorectal (CRC)	Phase II
MorphoSys's HuCAL® phage library technology				
anetumab ravtansine	BAY 94-9343	ImmunoGen Inc. (Cambridge MA USA)	Solid tumours, Mesothelioma	Phase I
bimagrumab	BYM338	Novartis Pharmaceuticals Corp. (East Hanover NJ USA) / MorphoSys (Martinsried/Pla Germany)	Type 2 diabetes (T2 DM), Musculoskeletal diseases	Phase III
elgemtumab	LJM716, NVS201010	Novartis Pharmaceuticals Corp. (East Hanover NJ USA)	Cancers, breast Cancers, gastric	Phase I
gantenerumab	R1450	Roche Ltd. (Basel Switzerland) / MorphoSys (Martinsried/Pla Germany)	Alzheimer's disease (AD)	Phase III
guselkumab	TREMFYA™, CNTO-1959	MorphoSys (Martinsried/Pla Germany)	Psoriatic arthritis (PSA)	Approved
otilimab	GSK3196165, MOR-04357, MOR103	GlaxoSmithKline (Brentford UK) / MorphoSys (Martinsried/Pla Germany)	Rheumatoid arthritis (RA), Osteoarthritis (OA)	Phase II
setrusumab	BPS-804, MOR-05813	Mereo Biopharma (London UK)	Osteogenesis imperfecta (OI)	Phase II
tarextumab	OMP-59R5	GlaxoSmithKline (Brentford UK) / OncoMed Pharmaceuticals (Redwood City CA USA)	Cancers, pancreatic Cancers, small cell lung (SCLC)	Phase I
tesidolumab	LFG 316	Novartis Pharmaceuticals Corp. (East Hanover NJ USA) / Alcon Laboratories, Inc. (Fort Worth TX USA)	Age-related macular degeneration (AMD)	Phase II
utomilumab	PF-05082566, PF-2566	Pfizer (New York NY USA)	Non-Hodgkin's lymphoma (NHL)	Phase I
vantictumab	OMP-18R5	OncoMed Pharmaceuticals (Redwood City CA USA)	Solid tumours	Phase I
xentuzumab	BI 836845	Boehringer Ingelheim Pharmaceuticals (Ridgefield CT USA)	Solid tumours	Phase I
Felzartamab	MOR202	MorphoSys (Martinsried/Pla Germany)	Multiple myeloma (MM)	Phase II

tafasitamab	MONJUVI®, MOR-00208, XENP-5574, XmAb®5574,	MorphoSys (Martinsried/Pla Germany) / Xencor Inc. (Monrovia CA USA)	Chronic lymphocytic leukemia (CLL)	Approved
Peptide phage display library technology				
romiplostim	NPLATE® , AMG 531	Amgen (Thousand Oaks CA USA)	Chronic ITP	Approved
trebananib	AMG 386, 2xCon4C	Amgen (Thousand Oaks CA USA)	Cancers, ovarian	Phase III

Table 11: Antibody phage display library dataset. The information is extracted from publicly available online resources such as AdisInsight, IMGT® database, and ClinicalTrials.gov.

The majority of the phage display antibodies are generated by three company-owned libraries - Cambridge Antibody Technology (CAT), Dyax, and MorphoSys's human combinatorial antibody libraries (HuCAL®). A summary of major human antibody phage display technologies is provided below.

Cambridge Antibody Technology (CAT) human antibody phage display library: Cambridge Antibody Technology (CAT) displays naïve scFv antibody domains on the surface of fd bacteriophage to generate combinatorial libraries. CAT involves creating a large library of phages, each of which displays a different human antibody fragment on its surface. In the CAT libraries, variable chain (V) gene repertoires are made *in vitro* by combining unrearranged V genes with D and J segments and cloning the final gene sequence in a bacteriophage.[44] The final result is a highly diverse collection of human antibody fragments that are expressed on the surface of phage. The library is entirely derived from human antibody sequences, which reduces the risk of immune reactions when used as therapeutic agents. Cambridge Antibody Technology (CAT) has the highest number of approved mAbs including several blockbuster drugs like HUMIRA®, BENLYSTA®, ADBRY™, Lumoxiti™, ABthrax®, and Gamifant®.

Dyax human antibody phage display library: Dyax's human antibody phage display libraries combine immunoglobulin gene fragments from human donors with strategically designed synthetic DNA to generate semi-synthetic Fabs. Dyax's state-of-the-art antibody phage display library contains over 10 billion unique clones that allow for rapid isolation of fully human target-specific antibodies. Dyax library has been used to identify high-affinity human antibodies that bind to numerous therapeutic targets. Dyax has made its technology widely available through nonexclusive licenses in three areas: therapeutic products; *in vitro* diagnostics; and research products to major firms like Genzyme, Merck, and Novagen.[45] Kalbitor®, Takhzyro®, Portrazza®, and Cyramza™ are some of the approved therapeutics derived from the Dyax library.

MorphoSys's HuCAL® phage display library technology: MorphoSys's Human Combinatorial Antibody Libraries (HuCAL) technology is an advanced version of phage display technology to generate HuCAL PLATINUM® library that contains approximately 45 billion different fully human antibodies. The unique feature about this library is that all CDRs here are diversified by trinucleotide-directed mutagenesis method (TRIM) that has yielded up to a 25-fold greater diversity.[46] Also, additional sequence optimization has been carried out in HuCAL library to enhance mammalian cell expression levels and eliminate the undesirable motifs that limit expression rates. MorphoSys has created a pipeline of more than 60 drug candidates. MorphoSys's HuCAL has the highest number of mAbs in clinical development among all categories with Tremfya™, Monjuvi®, and Ilumya® as the approved therapeutics.

Other artificial human phage display libraries: Earlier, the commercial use of phage display was restricted to only a few selected biopharmaceutical companies with rights to the phage display intellectual property. However, most of the key patents covering phage display technology have expired in the US and Europe, providing incentives to academic and biotech start-ups to design, construct and screen their own artificial human phage display libraries. Therefore, several other artificial human phage display libraries such as the SuperHuman™ library have been introduced that are based on a set of modular framework master genes with highly diversified CDRs to capture the structural immune repertoire.

Several other antibody drugs are developed by *in vitro* affinity maturation using phage display technology. In this approach, synthetic gene libraries are constructed using random mutagenesis or site-specific mutagenesis in the antigen-binding regions such as the V_H and V_L sequences. These libraries are then displayed on the phages, and finally antibodies with high affinity for the antigens are isolated by biopanning. We have included such company reported affinity maturation subsets in "others" category.

4.5.1 Computational developability analysis of biophysical performance

We started the computational developability assessment of the phage display library dataset with the AbPred tool. The AbPred scores were evaluated for all 12 biophysical assays across different antibody phage display platform technologies. We observed interesting insights for each phage display library from the below results. The HIC assay values were comparable and mostly within the 10% cutoff value (10.848) for all

antibody phage display platforms – CAT (100%); Dyax (80%); MorphoSys (85%) and Artificial (100%). A similar trend was observed for the SMAC assay with most phage display antibodies below the 10% cutoff value.

However, CAT phage display antibodies displayed highest scores in the SGAC assay which indicate low self-association and excellent colloidal stability for CAT phage display antibodies. The minimum SGAC score of 656.45 for CAT antibodies is well above the 10% cutoff value (< 234.11) which suggests minimal self-association. The relative comparison of SGAC assay values between antibody phage display platforms is shown in Figure 28. Here, the y-axis shows assay scores with the arrow on the y-axis indicating the direction of unfavourable assay values. Also, the 5% and 10% cutoff values from our previous developability assessment benchmarks are shown in yellow and red colour horizontal lines respectively for each biophysical assay.

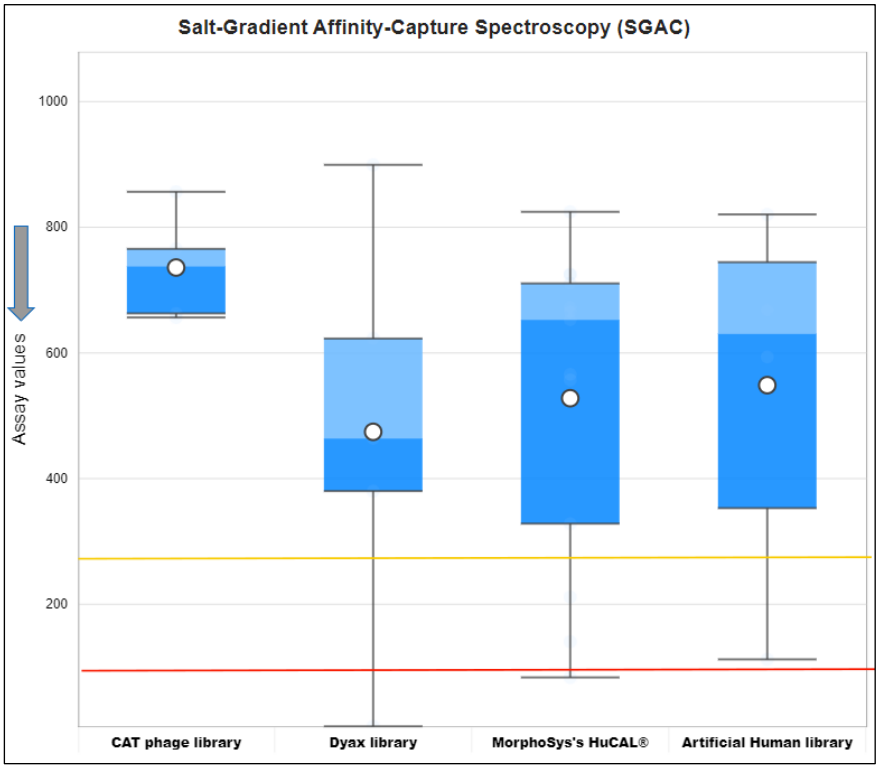


Figure 28: Salt-Gradient Affinity-Capture Spectroscopy (SGAC) values for different antibody phage display platforms. The arrow on y-axis indicates the direction of unfavorable values.

A similar trend was observed for the Cross-Interaction Chromatography (CIC) assay. All CAT antibodies had negative scores (normalized AbPred score) in the CIC assay which represent quick elution time in the chromatography column and therefore minimal cross-reactivity. The relative CIC assay values between all antibody phage display platforms are shown in Figure 29.

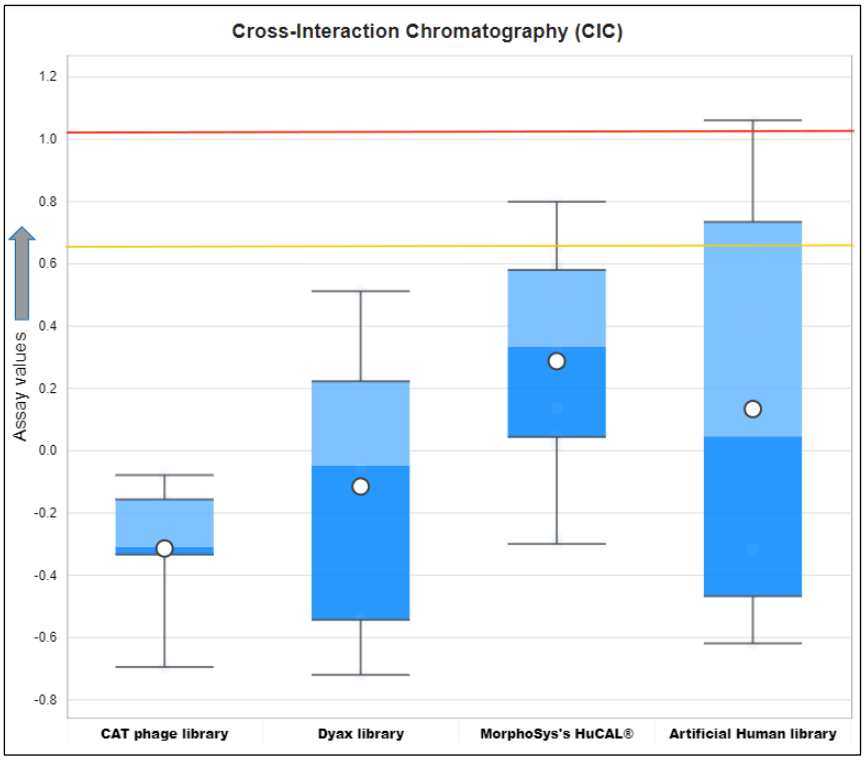


Figure 29: Cross-Interaction Chromatography (CIC) assay values for different antibody phage display platforms. The arrow on y-axis indicates the direction of unfavorable values.

Further evidence of the lowest self-association in CAT phage display antibodies is seen in AC-SINS results. The relative AC-SINS results are captured in Figure 30.

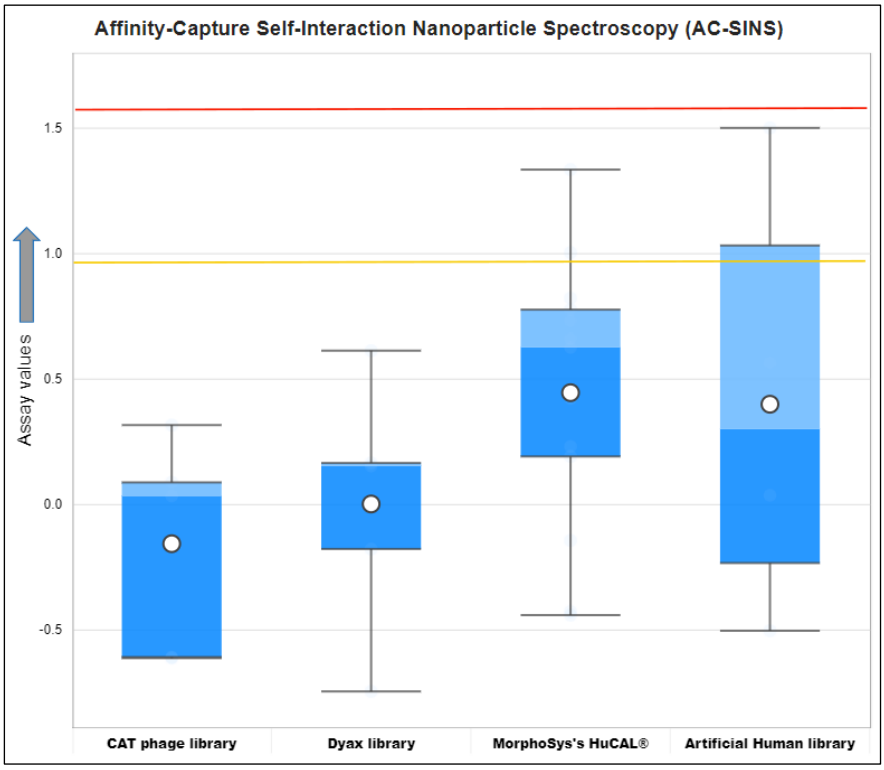


Figure 30: Affinity-Capture Self-Interaction Nanoparticle Spectroscopy (AC-SINS) assay values for different categories of antibody phage display platforms. The arrow on y-axis indicates the direction of unfavorable values.

However, CAT antibodies had the worst HEK assay performance which suggests a low degree of expression in human embryonic kidney cells. So, the CAT phage display antibodies may relatively have high-throughput manufacturability issues that make CAT technology a commercially less viable option among phage display technologies.

Overall, all CAT antibodies are above the 10% cutoff value (< 128.25) which indicates that this is an acceptable titer value in comparison to overall clinical-stage antibodies. In general, this high degree of expression for phage display library antibodies may be attributed to the natural functional expression of bacteriophage genes for coat proteins and surface markers which result in high HEK titer values.

CAT phage display antibodies also had a good performance in the PSR assay measuring polyspecificity with all antibodies within 10% cutoff values. Interestingly, the phage display platform with the most optimal PSR assay performance was the Dyax library. Dyax library also showed slightly better performance than CAT library in ELISA and BVP assays. It implies Dyax library antibodies by design have the lowest off-target binding and lowest multiantigen nonspecificity.

The CAT phage display platform had the best overall developability profile among all categories. A clear relative advantage is observed for three assays - SGAC, CIC, and AC-SINS. However, other phage display technologies probably have a better HEK expression profile and overall commercial manufacturability. Therefore, future CAT technology phage display optimization and engineering approaches should be directed towards increasing the degree of expression to improve ease of manufacturability. There were no major developability issues identified for CAT antibodies from our AbPred assessment as even the HEK liability was insignificant in the overall scenario.

4.5.2 TAP: Therapeutic Antibody Profiler results

We next calculated the Therapeutic Antibody Profiler (TAP) scores for each phage display platform. The full results for TAP scores are shown in Figure 31. All phage display technologies have similar scores in the CDR length, Patches of Surface Hydrophobicity Metric (PSH), and Patches of Positive Charge (PPC) metric. Each

phage display platform antibodies are within the 10% cutoff value (53 ≤ L or L ≤ 44) for CDR length - CAT (100%); Dyax (79%); MorphoSys (77%) and Artificial (100%).

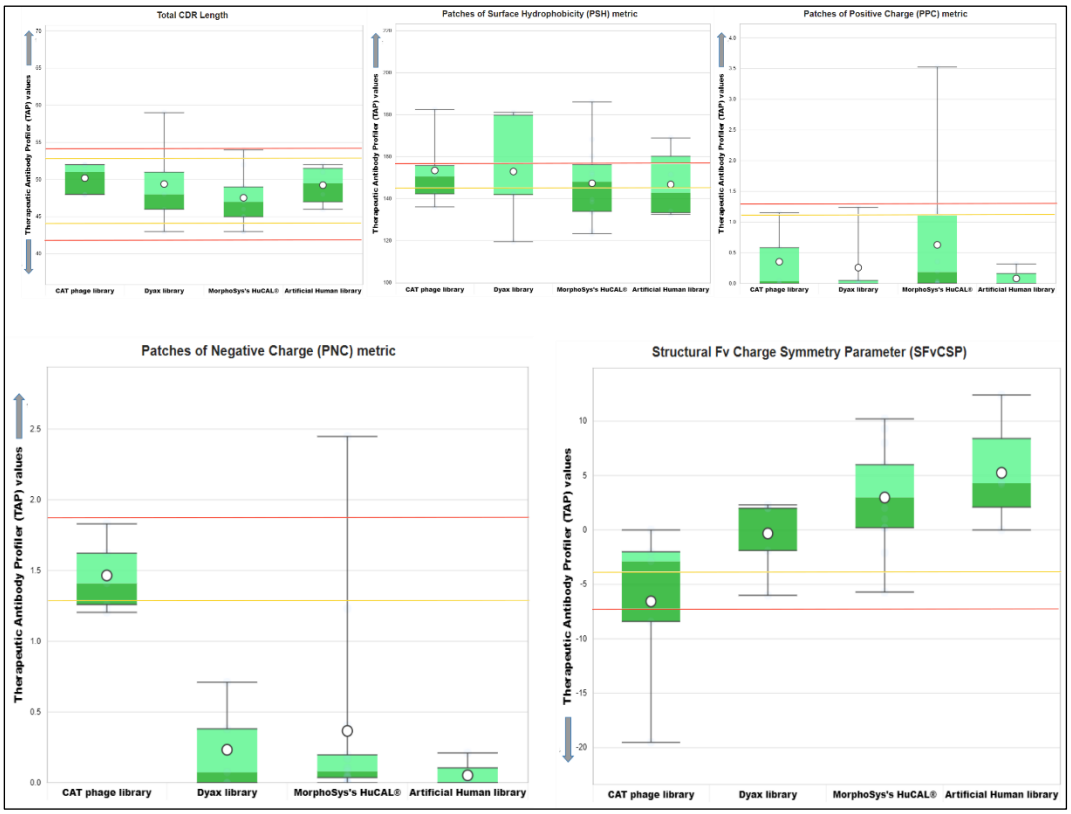


Figure 31: TAP scores for five structural metrics related to developability for different categories of phage display antibodies shown on x-axis. TAP scores are shown on the y-axis.

However, phage display antibodies breached the 10% cutoff value from clinical-stage therapeutics for PSH (144.63 ≤ PSH). We observe that 60% of CAT; 80% of Dyax; 54% of MorphoSys and 50% of Artificial antibodies are above this 10% cutoff. Also, many phage display antibodies even breached the 5% cutoff value from clinical-stage therapeutics (156.20 ≤ PSH). We observe that 20% of CAT; 40% of Dyax; 31% of MorphoSys and 25% of Artificial antibodies are above this 5% cutoff. It indicates that the phage display antibodies in general have high patches of surface hydrophobicity. But this result is in contrast to our equivalent Abpred predictions for hydrophobicity in HIC and SMAC assays. Earlier, we observed optimal and similar hydrophobicity profiles for all phage display categories.

TAP results also diverge in the Patches of Negative Charge (PNC) metric. Here, CAT antibodies display the worst performance with over 60% of the dataset above the 10% threshold (1.30 ≤ PNC), while mAbs from the other phage display technologies are mostly within this 10% threshold. The PNC metric scores for different phage display

antibodies are shown in Figure 32. As per these TAP results, CAT antibodies have many patches of negative charge that should result in lower colloidal stability and poor biophysical performance. But the AbPred predictions earlier were quite opposite to this with no developability issues identified for CAT antibodies. This inconsistency is also seen for the Structural Fv Charge Symmetry Parameter (SFvCSP) where the CAT antibodies again display the worst performance with over 40% of the dataset below the 10% cutoff (SFvCSP ≤ - 4.00). The SFvCSP scores for different phage display platforms are shown in Figure 33.

A possible explanation for this inconsistency is a potential homology modelling error in ABodyBuilder. Since the TAP tool uses ABodyBuilder - a deep-learning based structural modelling algorithm to generate a structural homology model for the input antibody variable domain sequence, we believe that this algorithm is not accurate for phage display antibody datasets. It can be because of the absence of phage display libraries in the training subset for this deep-learning algorithm. Therefore, TAP is not fit for use on antibody phage display libraries, a concept which has emerged recently.

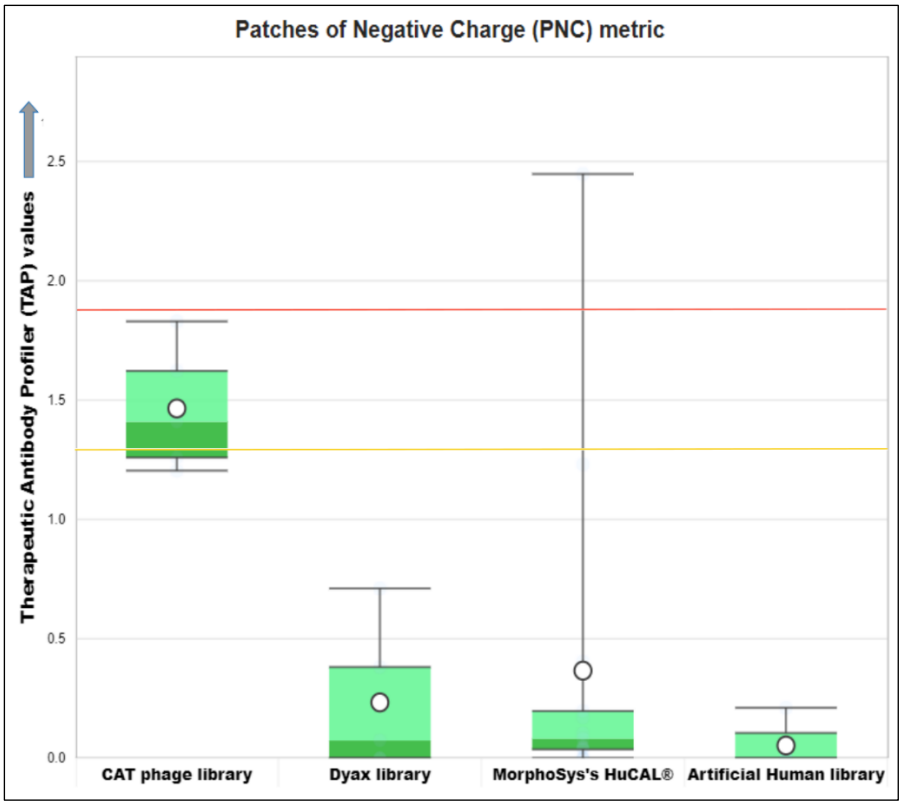


Figure 32: Patches of Negative Charge (PNC) metric values for different categories of phage display antibodies. PNC is calculated across the CDR vicinity.

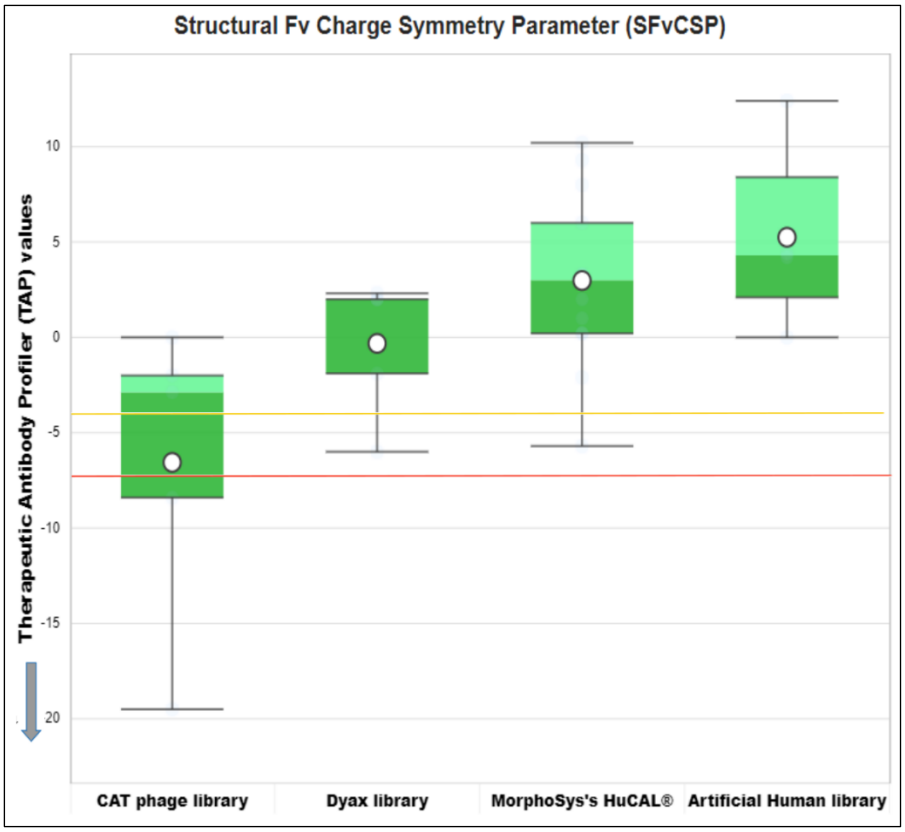


Figure 33: Charge Symmetry (SFvCSP) values for different phage display antibodies.

We conclude from the above computational developability assessment results that the Cambridge Antibody Technology (CAT) phage display antibodies have the best overall developability profile among all phage display categories. They demonstrate lowest self-associations and cross-reactivity with optimal biophysical performance. However, the only tradeoff with CAT antibodies is a relatively low expression titer in HEK assay compared to other phage display technologies but acceptable and fair enough in comparison to clinical-stage benchmarks. However, Therapeutic Antibody Profiler has contrasting results where CAT antibodies are predicted to have patches of negative charge and charge asymmetry in the heavy and light chain that usually leads to high viscosity and colloidal instability. We believe this discrepancy arises due to low homology modelling accuracy for phage display antibodies in TAP.

We observed that the Dyax library antibodies had the best performance in assays measuring polyspecificity – PSR, ELISA, and BVP. A detailed literature search revealed that Dyax libraries combine immunoglobulin sequences from human donors with strategically designed synthetic DNA.[47] This provides Dyax antibodies by design with synthetic targeting in key antigen contact sites in the heavy-chain complementary determining regions CDR1 and CDR2. So, we expect higher efficacy and lower

adverse events being reported for Dyax library antibodies in advanced clinical trials. For instance, we expect clinical-trial success towards approval for Opicinumab - a Dyax antibody in phase II studies that is being investigated as a treatment to promote myelin repair in Multiple Sclerosis. Our results show that it had the lowest PSR assay score (0.1953); lowest ELISA score (1.1558) and the lowest BVP score (1.4846) among the entire phage display antibody dataset.

MorphoSys's HuCAL library antibodies demonstrate highest expression level in the HEK assay compared to all other platforms as evidenced by the highest Maximum and Quartile 3 values of 184.35 mg/L and 171.09 mg/L respectively. This result justifies the HuCAL library approach to remove the sequence liabilities that limit expression in the HuCAL GOLD® and HuCAL PLATINUM® libraries. Therefore, our AbPred predictions are consistent with the company claims of a more optimized phage library.

The differences can also be attributed to the inherent differences in the antibody structure formats for different phage display libraries. For instance, mAbs isolated from the CAT libraries belong to two IgG subclasses, IgG1 and IgG4, with the majority being IgG1-λ. On the other hand, mAbs from Dyax libraries belong to IgG1 and IgG2 with the majority being IgG1-κ. The selection of desired phage display platform technology for antibody discovery depends on the preferences in the required biophysical performance and the intended application of the selected antibody therapeutic. Antibody discovery scientists need to consider the trade-off between colloidal stability, polyspecificity, and expression levels in making the phage display platform selection.

4.6 **Transgenic mice antibodies dataset**

Murine-derived mAbs have potential problems such as limited therapeutic efficacy and immunogenicity in human trials. Previous studies have indicated that patients treated with murine-derived mAbs may develop a human antimouse antibody (HAMA) response, which accelerates mAb clearance and could result in undesirable allergic reactions upon repeated administration.[48] Also, mouse-derived mAbs can cause serious adverse events such as hypersensitivity. Therefore, antibody engineering techniques have been subsequently utilized to create chimeric or humanized antibodies by combining the murine CDRs or antigen-binding regions with human Fc constant

regions. Such antibodies with engineered human constant regions are known to maintain target specificity as well as reduce the HAMA response.[49-51]

Fully human antibodies are now generated using hybridoma technology in transgenic mice models whereby the mouse immunoglobulin (Ig) gene loci have been replaced with human loci within the transgenic mouse genome.[52-54] The transgenic mice technology harnesses the natural recombination and affinity maturation machinery to generate high-affinity functional human antibodies. A plethora of human antibodies from transgenic mice are now in preclinical and early clinical stages.[55] Currently, 74% of fully human mAb therapies approved by US Food and Drug Administration (FDA) are derived from transgenic animal platforms.[56] The real-world data accumulated from patients to date suggests that these antibodies are meeting the expectations for lack of immunogenicity and pharmacokinetic characteristics, firmly establishing transgenic mice drug discovery platforms within the pharmaceutical industry.

A dataset of over fifty antibodies derived from transgenic mice platforms for multiple clinical indications was extracted from the IMGT® database. We then compiled the V_H-V_L sequence and 3D structure information for these antibodies from several online resources. The transgenic mice antibodies with missing sequence information were removed for further analysis while the scFv fragment from each transgenic mice antibody was used as a separate query in computational developability assessments. Finally, we had a dataset of 46 transgenic mice scFv fragments shown in Table 12 for evaluating developability using available biopharmaceutical informatics tools.

Our transgenic mice antibodies dataset provides publicly available information on various transgenic mice antibody therapeutics that were available in the market or were being investigated in human clinical trials as of 2023. In this dataset, there are a total of 46 antibodies with 7 (15.2%) in phase-I trials, 9 (19.5%) in phase–II trials, and 9 (19.5%) in phase - III clinical trials. Since the transgenic mice technology has been a well-known method of antibody discovery and generation, there are 21 (45.6%) transgenic mice antibodies that have been approved for various clinical indications. These include several popular blockbuster drugs such as brodalumab (SILIQ™) for the treatment of psoriasis, erenumab (AIMOVIG™) for migraine, daratumumab (DARZALEX™) for multiple myeloma, and durvalumab (IMFINZI™) for cancers.

INN (International Nonproprietary Name)	Common name / Proprietary name	Company	Clinical indication	Highest Clinical Trial
Abgenix's XenoMouse® technology				
brodalumab	SILIQ™, AMG827	MedImmune (Gaithersburg MD USA) AstraZeneca (London UK)	Psoriasis	Approved
conatumumab	AMG 655, TRAIL-R2mAb	Amgen (Thousand Oaks CA USA)	Cancers, non-small cell lung (NSCLC)	Phase II
denosumab	XGEVA®, PROLIA®	Amgen (Thousand Oaks CA USA)	Osteoporosis, skeletal-related events	Approved
durvalumab	IMFINZI™, MEDI4736	MedImmune (Gaithersburg MD USA) / AstraZeneca (London UK)	Cancers, non-small cell lung (NSCLC)	Approved
erenumab	AIMOVIG™, AMG 334	Novartis Pharmaceuticals (NJ USA) / Amgen (Thousand Oaks CA USA)	Migraine	Approved
evolocumab	REPATHA™, AMG 145	Amgen (Thousand Oaks CA USA)	Hyperlipidemia, Stroke	Approved
fulranumab	4D4, AMG-403, JNJ-42160443	Johnson & Johnson (PA USA) / Amgen (Thousand Oaks CA USA)	Osteoarthritis (OA)	Phase III
glembatumumab vedotin	CDX-011, CR011-vcMMAE	Celldex Therapeutics, Inc. (Needham MA USA)	Cancers, breast	Phase III
lucatumumab	HCD122, CHIR-12	Novartis Pharmaceuticals Corp. (East Hanover NJ USA) / Xoma (Berkeley CA USA)	Chronic lymphocytic leukemia (CLL)	Phase I
panitumumab	VECTIBIX®, ABX-EGF	Amgen (Thousand Oaks CA USA)	Cancers, colorectal (CRC)	Approved
prezalumab	AMG-557	Amgen (Thousand Oaks CA USA) / AstraZeneca (London UK)	Systemic lupus erythematosus (SLE)	Phase I
rilotumumab	AMG102	Amgen (Thousand Oaks CA USA)	Cancers, gastric	Phase III
secukinumab	COSENTYX®	Novartis Pharmaceuticals Corp. (East Hanover NJ USA)	Psoriatic arthritis (PSA)	Approved
tremelimumab	CP-675	MedImmune (Gaithersburg MD USA) / AstraZeneca (London UK)	Melanoma	Phase III
vixarelimab	KPL-716	Kiniksa Pharmaceuticals Ltd. (Bermuda USA)	Prurigo	Phase II
Medarex's UltiMAb® technology				
anifrolumab	SAPHNELO™, MDX-1333	MedImmune (Gaithersburg MD USA) / AstraZeneca (London UK)	Systemic lupus erythematosus (SLE)	Approved
canakinumab	ILARIS®, ACZ885	Novartis Pharmaceuticals Corp. (East Hanover NJ USA)	Systemic juvenile idiopathic athritis (SJIA)	Approved
eldelumab	BMS-936557, MDX-1100	Bristol-Myers Squibb (Princeton NJ USA)	Rheumatoid arthritis (RA)	Phase II
golimumab	SIMPONI®, CNTO 148	Centocor Ortho Biotech Inc. (Horsham PA USA)	Rheumatoid arthritis (RA)	Approved
inclacumab	LC1004-002, RO4905417	Roche Ltd. (Basel Switzerland)	Myocardial infarction	Phase II
ipilimumab	YERVOY®, BMS-734016	Bristol-Myers Squibb (Princeton NJ USA)	Renal cell carcinoma (RCC)	Approved
iratumumab	MDX-060	Medarex (Princeton NJ USA)	Hodgkin's disease (HD)	Phase II
lirilumab	BMS-986015	Bristol-Myers Squibb (Princeton NJ USA)	Acute myeloid leukemia (AML)	Phase II
nivolumab	OPDIVO®, BMS-936558	Bristol-Myers Squibb (Princeton NJ USA)	Cancers	Approved

olaratumab	LARTRUVO™	Eli Lilly (Indianapolis IN USA) / ImClone Systems Inc. (Somerville NJ USA)	Soft-tissue sarcoma (STS)	Approved
teprotumumab	TEPEZZA®, RO4858696-000	Genmab A/S (Copenhagen Denmark) / Horizon Therapeutics (Dublin Ireland)	Thyroid eye disease (TED)	Approved
urelumab	BMS-663513	Bristol-Myers Squibb (Princeton NJ USA)	Tumors	Phase II
ustekinumab	STELARA®, CNTO 1275	Medarex (Princeton NJ USA) / Janssen Biotech, Inc (Horsham PA USA)	Crohn's disease	Approved
Medarex's HuMAb-Mouse® technology				
actoxumab	MBL-CDA1, CDA-1, MDX-066,	Merck & Co., Inc. (NJ USA)	Clostridium difficile diarrhea	Phase III
bezlotoxumab	ZINPLAVA™, CDB-1, MDX-1388, MK-6072	Merck & Co., Inc (NJ USA)	Clostridium difficile diarrhea	Approved
daratumumab	DARZALEX™, HuMax-CD38	Genmab A/S (Copenhagen Denmark) / Janssen Biotech, Inc (Horsham PA USA)	Multiple myeloma (MM)	Approved
ofatumumab	ARZERRA®, KESIMPTA®, HuMax-CD20®	Genmab A/S (Copenhagen Denmark) / GlaxoSmithKline (Brentford UK)	Chronic lymphocytic leukemia (CLL)	Approved
zalutumumab	HuMaX-EGFR™	Genmab A/S (Copenhagen Denmark)	Cancers, head, and neck	Phase III
zanolimumab	HuMax-CD4®	Genmab A/S (Copenhagen Denmark) / TenX BioPharma (Philadelphia USA)	Lymphoma, cutaneous T cell (CTCL)	Phase III
Other transgenic mice technologies				
inezetamab	AMG-994	Amgen (Thousand Oaks CA USA)	Solid tumors	Phase I
pelgifatamab	BAY-2315158	Bayer AG (Leverkusen Germany)	Cancers, prostate, metastatic	Phase I
VelocImmune® technology				
alirocumab	PRALUENT® REGN727, SAR-236553	Regeneron Pharmaceuticals Inc. (Tarrytown NY USA)	Atherosclerosis	Approved
dupilumab	DUPIXENT®, CD124, REGN668,	Regeneron Pharmaceuticals Inc. (Tarrytown NY USA)	Asthma, Atopic dermatitis (AD)	Approved
enoticumab	REGN-421, SAR153192	Regeneron Pharmaceuticals Inc. (Tarrytown NY USA)	Cancers	Phase I
fasinumab	REGN475, SAR164877	Regeneron Pharmaceuticals Inc. (Tarrytown NY USA)	Osteoarthritis (OA)	Phase III
intetumumab	CNTO 095, CNTO-95	Johnson & Johnson (Langhorne PA USA)	Solid tumors	Phase II
nesvacumab	REGN-910	Regeneron Pharmaceuticals Inc. (Tarrytown NY USA)	Solid tumors (Treatment)	Phase I
sarilumab	KEVZARA®	Regeneron Pharmaceuticals Inc. (Tarrytown NY USA) / Sanofi (Paris France)	Rheumatoid arthritis (RA)	Approved
suptavumab	REGN-2222, SAR-438584	Regeneron Pharmaceuticals Inc. (Tarrytown NY USA)	Respiratory Syncytial Virus (RSV) infection	Phase III
trevogrumab	REGN-1033	Regeneron Pharmaceuticals Inc. (Tarrytown NY USA) / Sanofi (Paris France)	Sarcopenia	Phase II

Table 12: Transgenic mice antibodies dataset. The information is extracted from publicly available online resources such as AdisInsight, IMGT® database, and ClinicalTrials.gov.

A major advantage of the transgenic mice approach is that natural diversification and selection can be exploited under the control of the animal immune system. The integrated human immunoglobulin (Ig) loci can undergo normal biological processes

of DNA rearrangement and hypermutation. So, antibodies derived from transgenic mice or other living animals have passed the nature-selection. A summary of major transgenic mice antibody platforms is provided below.

Abgenix's XenoMouse® technology: Abgenix's approach to generating fully human antibodies employs genetically engineered strains of mice in which endogenous mouse antibody gene expression is suppressed by removing the J-chain, effectively disabling the antibody-generating system of the host.[57] The XenoMouse® technology is based on the introduction of human germline loci by Yeast Artificial Chromosomes (YACs) into the mouse germline with inactivated mouse antibody machinery. Finally, these YAC transgenes are integrated into the mouse chromosome with superior genetic stability. XenoMouse® technology has evolved to be a highly reliable antibody discovery platform since the approval of first transgenic mice antibody - Panitumumab (Vectibix®) a fully human antibody directed against epidermal growth factor receptor (EGFR) for the treatment of advanced colorectal cancer.

Medarex's UltiMAb® technology: Medarex's new proprietary UltiMAb® technology employs engineered transgenic mice for producing the entire spectrum of human antibodies ranging from IgG1, IgG2, IgG3, IgG4, and IgA antibodies, all from a single fusion. In brief, the UltiMAb platform uses yeast artificial chromosome encompassing heavy-chain and light-chain transgenes to express the human IgGκ repertoire instead of the murine immunoglobulin repertoire.[58] The transgenic mice are then immunized with a recombinant human antigen (eg IL-1β) to generate hybridomas which are selected and cloned to purify the final antibody by affinity chromatography on a protein A column. This validated technology platform has produced over 8 approved compounds which are currently approved and marketed therapies like SIMPONI™, STELARA™, and ILARIS®. Medarex's UltiMab® platform has produced more approved drugs than any other human antibody platform in the industry.

Medarex's HuMAb-Mouse® technology: Medarex's HuMAb-Mouse® technology now owned by Bristol-Myers-Squibb has the proven ability to generate fully human antibodies with affinities in the picomolar range in response to immunization. In HuMAb-Mouse®, the mouse immunoglobulin genes are disrupted by homologous recombination and human heavy/light chain transgenes, including constant (C), variable (V), diversity (D), and joining (J) regions.[59] HuMAb-Mouse® has been a very

successful platform for antibody discovery with three superstar marketing approvals such as ZINPLAVA™, DARZALEX™, and ARZERRA®.

VelocImmune® technology: VelocImmune® a proprietary technology by Regeneron creates a multitude of optimized antibody drug candidates efficiently and directly from immunized mice. Unlike the other transgenic mic antibody platforms with inactivated antibody machinery, the VelocImmune mice have a completely normal immune system and are indistinguishable from wild-type mice in antigen response. In VelocImmune®, only the mouse variable region is replaced with human heavy and light chain counterparts. However, VelocImmune mice retain the mouse heavy chain constant regions, therefore, preserving the normal mouse B-cell signalling and maturation.[60] VelocImmune technology has been used to create multiple antibodies including Libtayo® (Cemiplimab), Praluent® (Alirocumab), and Kevzara® (Sarilumab) which are approved in multiple countries around the world.[55] Recently in 2022, FDA approved Dupixent® (Dupilumab) for the treatment of moderate-to-severe Atopic Dermatitis in children.

Other transgenic mice technologies: One issue that has hampered wide adoption of transgenic mice as platforms for antibody discovery is the commercial rights to the proprietary technology. However, other new *in vivo* mouse technologies such as AlivaMab® Mouse, MeMo® Transgenic Mouse, Kymouse™, and Crescendo Mouse have established themselves as successors in transgenic mice technologies.

4.6.1 Computational developability analysis of biophysical performance

We started the computational developability assessment of the transgenic mice dataset with the AbPred tool. AbPred scores were evaluated for all 12 biophysical assays across different transgenic mice platform technologies. We observe some interesting developability insights for each transgenic mice platform from the above results. The HIC assay values were comparable and mostly within the 10% cutoff value (10.848) for all transgenic mice platforms – XenoMouse® (87%); HuMAb-Mouse® (100%); UltiMAb® (77%); VelocImmune® (100%) and Others (100%). A similar trend was observed for the SMAC assay with most transgenic mAbs below the 10% cutoff value.

However, VelocImmune® transgenic mice antibodies displayed the highest scores in the SGAC assay which indicate low self-association and excellent colloidal stability

for VelocImmune® antibodies. The minimum SGAC score of 573.95 for VelocImmune® antibodies is well above the 10% cutoff value (< 234.11) which suggests that these have minimal self-association. The relative comparison of SGAC assay values between phage display platforms is shown in Figure 34. VelocImmune® antibodies have a mean SGAC assay score of 731.51 with an interquartile range (IQR) of 215.99 which are the best SGAC scores among all transgenic mice antibodies.

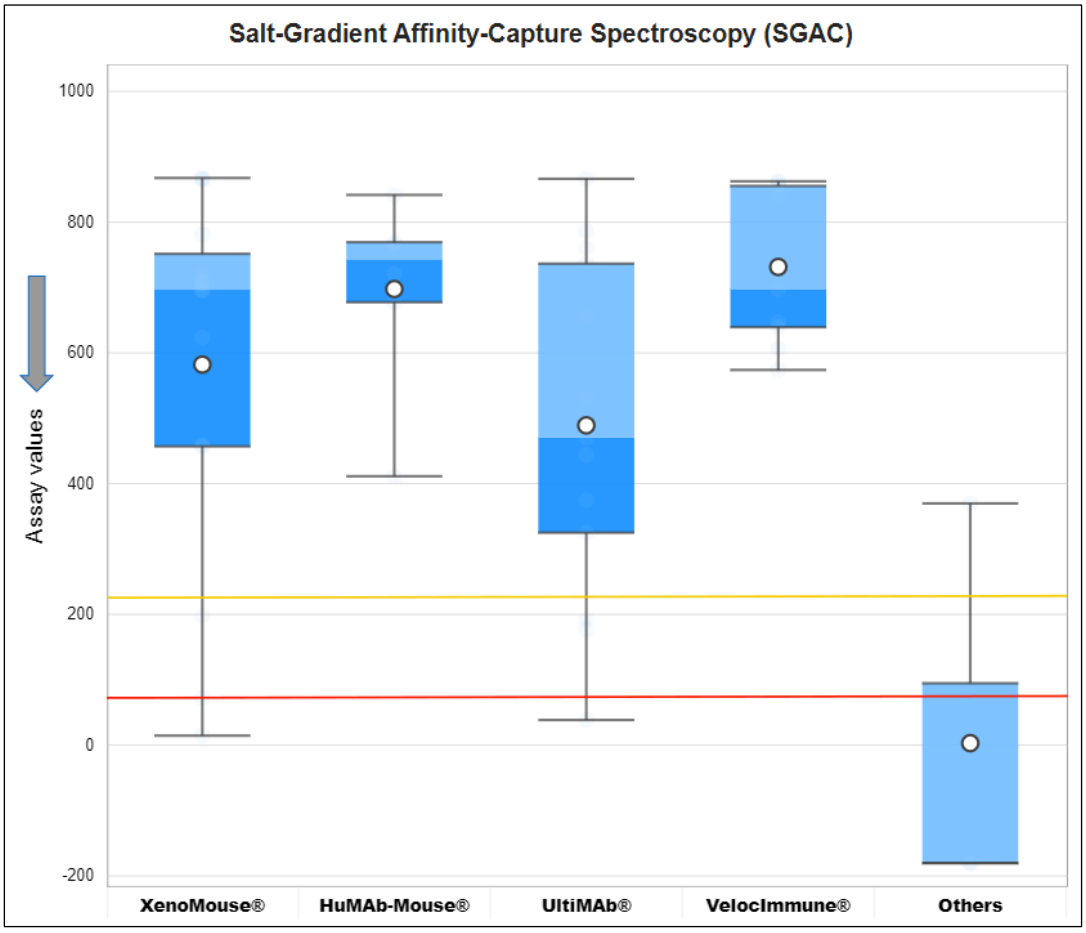


Figure 34: Salt-Gradient Affinity-Capture Spectroscopy (SGAC) assay values for transgenic mice platform categories. The arrow on y-axis indicates the direction of unfavorable values.

A similar trend was observed for the Cross-Interaction Chromatography (CIC) assay. Most VelocImmune® antibodies had negative scores (normalized AbPred score) in the CIC assay which represents quick elution time and therefore minimal cross-reactivity. The relative CIC assay values between transgenic mice antibody platforms are shown in Figure 35. It is important to also note that all other transgenic mice platforms mostly have values within the 10% cutoffs for CIC, CSI BLI, AC-SINS, and HEK assays that suggest comparable interaction and expression profiles to successful clinical-stage antibodies. VelocImmune® antibodies also had an excellent performance in the PSR

assay that measures polyspecificity. All transgenic mice antibodies were within the 10% cutoff value (PSR > 0.4204). The relative PSR assay values between transgenic mice antibody platforms are also shown in Figure 35.

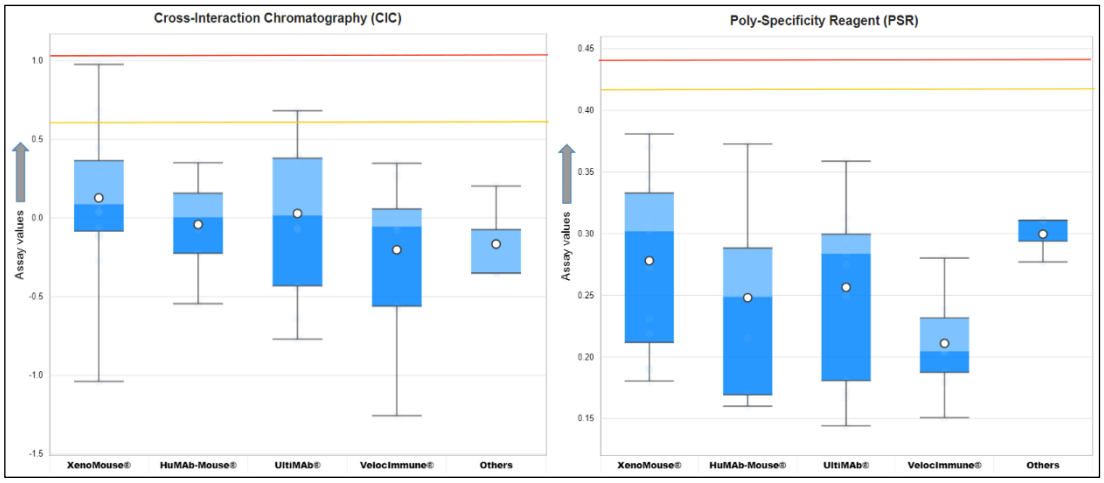


Figure 35: Cross-Interaction (CIC) and Poly-Specificity (PSR) assay values for different transgenic mice platforms. The arrow on y-axis indicates the direction of unfavorable values.

The VelocImmune® platform had the best overall developability profile among all the transgenic mice categories. A clear relative advantage is observed for three assays – PSR, SGAC, and CIC. A possible explanation is that in VelocImmune mice, only a small 6Mb variable portion of the mouse immunoglobulin (Ig) loci is humanized that retains mouse heavy chain constant regions.[61] Therefore, all humanized mouse lines preserve the normal mouse maturation and Fc-mediated effector functions. There were no developability issues identified for VelocImmune® antibodies from our assessment.

4.6.2 TAP: Therapeutic Antibody Profiler results

The TAP scores were comparable for all transgenic mice platforms across all five developability metrics namely total CDR length; patches of surface hydrophobicity (PSH); patches of positive charge (PPC); patches of negative charge (PNC) and structural Fv charge symmetry parameter (SFvCSP) as shown in Figure 36. Overall, the TAP results were inconclusive for any relative distinction between transgenic mice platforms. Also, all transgenic mice antibodies were within the 10% cutoff guidelines derived from clinical-stage antibodies - $53 \leq L$ or $L \leq 44$; $144.63 \leq PSH$; $1.14 \leq PPC$; $1.30 \leq PNC$ and $SFvCSP \leq -4.00$. This optimal developability profile for transgenic mice antibodies can be attributed to 'ELISA-based selection and purification' of the target antibody in the final step of the transgenic mice antibody discovery process.

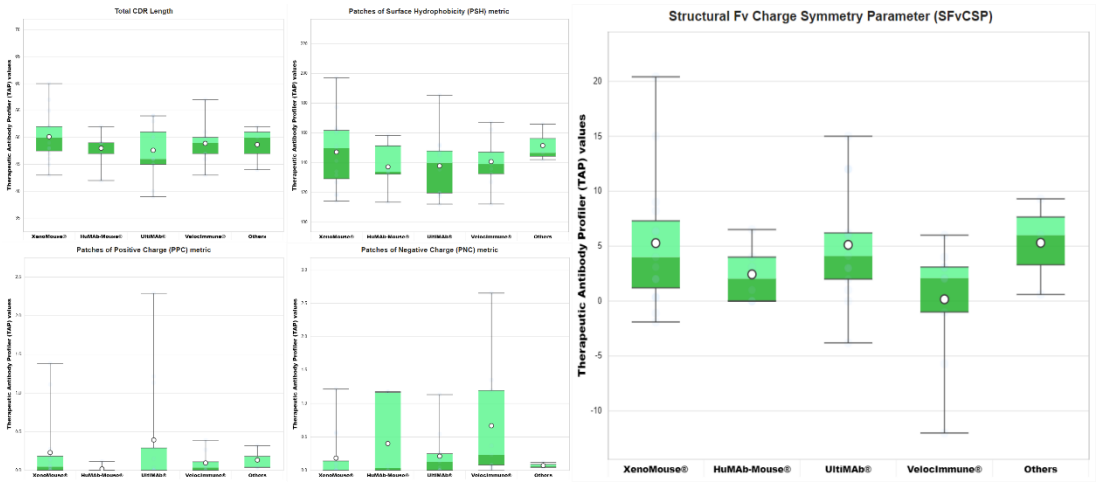


Figure 36: TAP scores for five structural metrics related to developability for different categories of transgenic mice antibodies shown on x-axis. TAP scores are shown on y-axis.

4.7 **Conclusion**

A major challenge in developing engineered antibody therapeutics is the selection of the ideal molecular format from many structurally diverse alternatives that can support a wide range of different biophysical and pharmacological properties. The selected format is best chosen to match the proposed mechanisms of action and the specific clinical application.[62] The final antibody discovery platform and format are chosen after detailed *in vitro* and *in vivo* functional characterizations. In this work, we have evaluated the developability of major antibody formats across three major antibody discovery technologies – multispecific engineering, phage display, and transgenic mice. Our analysis has compared the key developability considerations across various mAb formats which can accelerate the functional characterization and final selection.

The computational developability assessment results show that the DVD-Ig™ platform has the best overall developability profile among all multispecific engineered antibody platforms with the lowest non-specific interactions and cross-reactivity. It is because DVD-Ig™ antibodies have a similar configuration to conventional IgGs while the scFv-based constructs have linker sequence constraints and a tendency to form aggregates due to domain exchange of the variable regions. The phage display platform selection depends on the trade-off between colloidal stability, polyspecificity, and expression levels. The CAT antibodies are preferable for therapeutic applications requiring high colloidal stability, while Dyax antibodies are

preferred for low polyspecificity and MorphoSys's HuCAL® antibodies are suitable for obtaining high expression levels.

The VelocImmune® mouse technology antibodies have the best developability profile among the transgenic mice platforms owing to their lowest polyspecificity and non-specific interactions in PSR, CIC, and SGAC assays. This can be attributed to normal maturation and Fc-mediated effector functions retained in the VelocImmune® edits. Antibodies derived from phage display technology offer several advantages compared to other antibody discovery platforms such as the ability to generate conformation-specific antibodies, bypassing animal immunization, and most importantly the ability to isolate antibodies against toxic or non-immunogenic antigens. However, the vast majority of the approved therapeutic antibodies are derived from immunized mice technologies despite these several advantages of antibody phage display.

The post-translational modifications and tuning process imposed by the mammalian immune system creates antibodies with better biophysical attributes compared to phage display antibodies. For instance, *E. coli* derived phage display antibodies are not glycosylated which results in poor binding and pharmacokinetics when used as therapeutics in humans. We conclude that antibodies directly discovered by phage display exhibit significant developability risks compared to those derived from immunized mice. A previous investigation had found that phage display derived therapeutic antibodies have higher self-interaction and poly-reactivity due to the higher percentage of aliphatic residues in their CDRs compared to the non-phage derived antibodies.[63] Our results are consistent with these previous findings as the phage display library dataset antibodies have the assay scores skewed towards an unfavourable direction in AC-SINS and PSR assays respectively.

Naive phage antibody libraries generally contain antibodies in either the scFv or Fab format. The Fab antibody format is generally preferred over the scFv format. There is a stronger association between the two chains of a Fab fragment compared to the pairing of the variable regions of antibody heavy and light chains (VH and VL) in a scFv fragment essential for effective antigen binding. The spatial separation and orientation of the antigen-binding sites between different formats are key features that decide overall developability. Therefore, screening new formats for their PPC, PNC, and SCFvCSP scores from TAP tool can be important for developability assessment.

New emerging antibody modification technologies and novel discovery platforms are anticipated to significantly expand the repertoire of engineered antibody therapeutics against the vast range of diseases. There is tremendous potential for most engineered antibody formats as next-generation biopharmaceuticals with superior developability profiles or as robust diagnostic reagents in other applications such as biosensors.

4.8 References

1. Strohl WR. Current progress in innovative engineered antibodies. Protein & cell. 2018;9(1):86-120.
2. Queen C, Schneider WP, Selick HE, Payne PW, Landolfi NF, Duncan JF, Avdalovic NM, Levitt M, Junghans RP, Waldmann TA. A humanized antibody that binds to the interleukin 2 receptor. Proceedings of the National Academy of Sciences. 1989;86(24):10029-33.
3. Junghans R, Waldmann T, Landolfi N, Avdalovic N, Schneider W, Queen C. Anti-Tac-H, a humanized antibody to the interleukin 2 receptor with new features for immunotherapy in malignant and immune disorders. Cancer Research. 1990;50(5):1495-502.
4. Weiss U, Rajewsky K. The repertoire of somatic antibody mutants accumulating in the memory compartment after primary immunization is restricted through affinity maturation and mirrors that expressed in the secondary response. The Journal of experimental medicine. 1990;172(6):1681-9.
5. Bird RE, Hardman KD, Jacobson JW, Johnson S, Kaufman BM, Lee S-M, Lee T, Pope SH, Riordan GS, Whitlow M. Single-chain antigen-binding proteins. Science. 1988;242(4877):423-6.
6. Ward ES, Güssow D, Griffiths AD, Jones PT, Winter G. Binding activities of a repertoire of single immunoglobulin variable domains secreted from Escherichia coli. Nature. 1989;341(6242):544-6.
7. Holliger P, Prospero T, Winter G. " Diabodies": small bivalent and bispecific antibody fragments. Proceedings of the National Academy of Sciences. 1993;90(14):6444-8.
8. Kipriyanov SM, Moldenhauer G, Schuhmacher J, Cochlovius B, Von der Lieth C-W, Matys ER, Little M. Bispecific tandem diabody for tumor therapy with improved antigen binding and pharmacokinetics. Journal of molecular biology. 1999;293(1):41-56.
9. Choy E, Hazleman B, Smith M, Moss K, Lisi L, Scott D, Patel J, Sopwith M, Isenberg D. Efficacy of a novel PEGylated humanized anti-TNF fragment (CDP870) in patients with rheumatoid arthritis: a phase II double-blinded, randomized, dose-escalating trial. Rheumatology. 2002;41(10):1133-7.
10. Miller A, Carr S, Rabbitts T, Ali H. Multimeric antibodies with increased valency surpassing functional affinity and potency thresholds using novel formats. MAbs; 2020.
11. Wu X, Yuan R, Bacica M, Demarest SJ. Generation of orthogonal Fab-based trispecific antibody formats. Protein Engineering, Design and Selection. 2018;31(7-8):249-56.
12. Joosten V, Lokman C, van Den Hondel CA, Punt PJ. The production of antibody fragments and antibody fusion proteins by yeasts and filamentous fungi. Microbial Cell Factories. 2003;2:1-15.
13. Huehls AM, Coupet TA, Sentman CL. Bispecific T-cell engagers for cancer immunotherapy. Immunology and cell biology. 2015;93(3):290-6.
14. Weidle UH, Kontermann RE, Brinkmann U. Tumor-antigen–binding bispecific antibodies for cancer treatment. Seminars in oncology; 2014.

15. Mazor Y, Hansen A, Yang C, Chowdhury PS, Wang J, Stephens G, Wu H, Dall'Acqua WF. Insights into the molecular basis of a bispecific antibody's target selectivity. MAbs; 2015.
16. Vidarsson G, Dekkers G, Rispens T. IgG subclasses and allotypes: from structure to effector functions. Frontiers in immunology. 2014;5:520.
17. Cain P, Huang L, Tang Y, Anguiano V, Feng Y. Impact of IgG subclass on monoclonal antibody developability. Mabs; 2023.
18. Dao T, Pankov D, Scott A, Korontsvit T, Zakhaleva V, Xu Y, Xiang J, Yan S, de Morais Guerreiro MD, Veomett N. Therapeutic bispecific T-cell engager antibody targeting the intracellular oncoprotein WT1. Nature biotechnology. 2015;33(10):1079-86.
19. Goebeler M-E, Bargou R. Blinatumomab: a CD19/CD3 bispecific T cell engager (BiTE) with unique anti-tumor efficacy. Leukemia & lymphoma. 2016;57(5):1021-32.
20. Tian Z, Liu M, Zhang Y, Wang X. Bispecific T cell engagers: an emerging therapy for management of hematologic malignancies. Journal of Hematology & Oncology. 2021;14(1):1-18.
21. Klein C, Schaefer W, Regula JT. The use of CrossMAb technology for the generation of bi-and multispecific antibodies. MAbs; 2016.
22. Klein C, Schaefer W, Regula JT, Dumontet C, Brinkmann U, Bacac M, Umaña P. Engineering therapeutic bispecific antibodies using CrossMab technology. Methods. 2019;154:21-31.
23. Surowka M, Schaefer W, Klein C. Ten years in the making: application of CrossMab technology for the development of therapeutic bispecific antibodies and antibody fusion proteins. MAbs; 2021.
24. Jakob CG, Edalji R, Judge RA, DiGiammarino E, Li Y, Gu J, Ghayur T. Structure reveals function of the dual variable domain immunoglobulin (DVD-Ig™) molecule. MAbs; 2013.
25. Wu C, Ying H, Grinnell C, Bryant S, Miller R, Clabbers A, Bose S, McCarthy D, Zhu R-R, Santora L. Simultaneous targeting of multiple disease mediators by a dual-variable-domain immunoglobulin. Nature biotechnology. 2007;25(11):1290-7.
26. Gu J, Ghayur T. Generation of dual-variable-domain immunoglobulin molecules for dual-specific targeting. Methods in enzymology. 502: Elsevier; 2012. p. 25-41.
27. Moore PA, Zhang W, Rainey GJ, Burke S, Li H, Huang L, Gorlatov S, Veri MC, Aggarwal S, Yang Y. Application of dual affinity retargeting molecules to achieve optimal redirected T-cell killing of B-cell lymphoma. Blood, The Journal of the American Society of Hematology. 2011;117(17):4542-51.
28. LaMotte-Mohs R, Shah K, Smith D, Gorlatov S, Ciccarone V, Tamura J, Li H, Rillema J, Licea M. MGD013, a bispecific PD-1 x LAG-3 Dual-Affinity Re-Targeting (DART®) protein with T-cell immunomodulatory activity for cancer treatment. Cancer Res. 2016;76(Suppl 14):3217.
29. Pillarisetti K, Baldwin E, Babich A, Majewski N, Barone L, Li Y, Zhang X, Chin D, Luistro L, Mendonça M. Development of a new BCMAxCD3 Duobody® antibody for multiple myeloma. 2016.
30. Syed YY. Amivantamab: first approval. Drugs. 2021;81(11):1349-53.
31. Bao G, Tang M, Zhao J, Zhu X. Nanobody: a promising toolkit for molecular imaging and disease therapy. EJNMMI research. 2021;11(1):1-13.
32. De Meyer T, Muyldermans S, Depicker A. Nanobody-based products as research and diagnostic tools. Trends in biotechnology. 2014;32(5):263-70.
33. Ma J, Mo Y, Tang M, Shen J, Qi Y, Zhao W, Huang Y, Xu Y, Qian C. Bispecific antibodies: from research to clinical application. Frontiers in Immunology. 2021:1555.
34. Suurs FV, Lub-de Hooge MN, de Vries EG, de Groot DJA. A review of bispecific antibodies and antibody constructs in oncology and clinical challenges. Pharmacology & therapeutics. 2019;201:103-19.

35. DiGiammarino E, Ghayur T, Liu J. Design and generation of DVD-Ig™ molecules for dual-specific targeting. Therapeutic Proteins: Methods and Protocols. 2012:145-56.
36. Hebditch M, Warwicker J. Web-based display of protein surface and pH-dependent properties for assessing the developability of biotherapeutics. Scientific reports. 2019;9(1):1-9.
37. Niwa T, Ying B-W, Saito K, Jin W, Takada S, Ueda T, Taguchi H. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. Proceedings of the National Academy of Sciences. 2009;106(11):4201-6.
38. Hebditch M, Warwicker J. Charge and hydrophobicity are key features in sequence-trained machine learning models for predicting the biophysical properties of clinical-stage antibodies. PeerJ. 2019;7:e8199. doi:10.7717/peerj.8199.
39. Ng G, Weisser N, Wickman G, Spreter von Kreudenstein T. ZW38, a bispecific CD3 x CD19 azymetric antibody to deplete human leukemic B cells by the “controlled” activation of T cells. 2015.
40. Meric-Bernstam F, Hanna DL, El-Khoueiry AB, Kang Y-K, Oh D-Y, Chaves JM, Rha SY, Hamilton EP, Pant S, Javle MM. Zanidatamab (ZW25) in HER2-positive biliary tract cancers (BTCs): Results from a phase I study. 2021.
41. Kehoe JW, Kay BK. Filamentous phage display in the new millennium. Chemical reviews. 2005;105(11):4056-72.
42. Hoogenboom HR, de Bruıne AP, Hufton SE, Hoet RM, Arends J-W, Roovers RC. Antibody phage display technology and its applications. Immunotechnology. 1998;4(1):1-20.
43. Alfaleh MA, Alsaab HO, Mahmoud AB, Alkayyal AA, Jones ML, Mahler SM, Hashem AM. Phage display derived monoclonal antibodies: from bench to bedside. Frontiers in Immunology. 2020:1986.
44. McCafferty J, Griffiths AD, Winter G, Chiswell DJ. Phage antibodies: filamentous phage displaying antibody variable domains. nature. 1990;348(6301):552-4.
45. Glaser V. Conflicts brewing as phage display gets complex. Nature Biotechnology. 1997;15(6):506-.
46. Prassler J, Thiel S, Pracht C, Polzer A, Peters S, Bauer M, Nörenberg S, Stark Y, Kölln J, Popp A. HuCAL PLATINUM, a synthetic Fab library optimized for sequence diversity and superior performance in mammalian expression systems. Journal of molecular biology. 2011;413(1):261-78. doi:10.1016/j.jmb.2011.08.012.
47. Hoet RM, Cohen EH, Kent RB, Rookey K, Schoonbroodt S, Hogan S, Rem L, Frans N, Daukandt M, Pieters H. Generation of high-affinity human antibodies by combining donor-derived and synthetic complementarity-determining-region diversity. Nature biotechnology. 2005;23(3):344-8.
48. Legouffe E, Liautard J, Gaillard J, Rossi J, Wijdenes J, Bataille R, Klein B, Brochier J. Human anti-mouse antibody response to the injection of murine monoclonal antibodies against IL-6. Clinical & Experimental Immunology. 1994;98(2):323-9.
49. Morrison SL, Oi VT. Genetically engineered antibody molecules. Advances in immunology. 1989;44:65-92.
50. Studnicka GM, Soares S, Better M, Williams RE, Nadell R, Horwitz AH. Human-engineered monoclonal antibodies retain full specific binding activity by preserving non-CDR complementarity-modulating residues. Protein Engineering, Design and Selection. 1994;7(6):805-14.
51. Riechmann L, Clark M, Waldmann H, Winter G. Reshaping human antibodies for therapy. Nature. 1988;332(6162):323-7.

52. Lonberg N, Taylor LD, Harding FA, Trounstine M, Higgins KM, Schramm SR, Kuo C-C, Mashayekh R, Wymore K, McCabe JG. Antigen-specific human antibodies from mice comprising four distinct genetic modifications. Nature. 1994;368(6474):856-9.
53. Mendez MJ, Green LL, Corvalan JR, Jia X-C, Maynard-Currie CE, Yang X-d, Gallo ML, Louie DM, Lee DV, Erickson KL. Functional transplant of megabase human immunoglobulin loci recapitulates human antibody response in mice. Nature genetics. 1997;15(2):146-56.
54. Lonberg N. Human antibodies from transgenic animals. Nature biotechnology. 2005;23(9):1117-25.
55. Moran N. Mouse platforms jostle for slice of humanized antibody market. Nature Biotechnology. 2013;31(4):267-9.
56. Ma B, Osborn M. Transgenic animals for the generation of human antibodies. Introduction to Antibody Engineering: Springer; 2021. p. 97-127.
57. Jakobovits A, Amado RG, Yang X, Roskos L, Schwab G. From XenoMouse technology to panitumumab, the first fully human antibody product from transgenic mice. Nature biotechnology. 2007;25(10):1134-43.
58. Fishwild DM, O'Donnell SL, Bengoechea T, Hudson DV, Harding F, Bernhard SL, Jones D, Kay RM, Higgins KM, Schramm SR. High-avidity human IgGκ monoclonal antibodies from a novel strain of minilocus transgenic mice. Nature biotechnology. 1996;14(7):845-51.
59. Mukherjee J, Chios K, Fishwild D, Hudson D, O'Donnell S, Rich SM, Donohue-Rolfe A, Tzipori S. Production and characterization of protective human antibodies against Shiga toxin 1. Infection and immunity. 2002;70(10):5896-9.
60. Murphy AJ, Macdonald LE, Stevens S, Karow M, Dore AT, Pobursky K, Huang TT, Poueymirou WT, Esau L, Meola M. Mice with megabase humanization of their immunoglobulin genes generate antibodies as efficiently as normal mice. Proceedings of the National Academy of Sciences. 2014;111(14):5153-8.
61. Devoy A, Bunton-Stasyshyn RK, Tybulewicz VL, Smith AJ, Fisher EM. Genomically humanized mice: technologies and promises. Nature Reviews Genetics. 2012;13(1):14-20.
62. Spiess C, Zhai Q, Carter PJ. Alternative molecular formats and therapeutic applications for bispecific antibodies. Molecular immunology. 2015;67(2):95-106.
63. Kaleli NE, Karadag M, Kalyoncu S. Phage display derived therapeutic antibodies have enriched aliphatic content: insights for developability issues. Proteins: Structure, Function, and Bioinformatics. 2019;87(7):607-18.

CHAPTER 5

5 Machine learning approaches to estimate clinical trial success from computational developability assessments

5.1 Abstract

This chapter aims to use computational developability assessments to estimate the clinical trial progression of candidate antibodies. We have used machine learning and other data science algorithms to determine the biophysical features which determine the clinical trial progression of therapeutic mAbs. Firstly, we have tested the ability of the 50+ sequence or structural features employed in our computational developability assessment analysis to classify mAbs as per their clinical trial progression. We have then used feature engineering techniques and other data transformation approaches to optimize the biophysical assay features to predict the clinical trial progression. Next, we have explored other additional features and tools known in biopharmaceutical informatics to obtain the T20 score as a reliable estimate of clinical trial progression for mAbs. Finally, we have used a new dataset of failed antibodies to flag mAbs with a low clinical trial success rate based on our computational developability framework.

5.2 Introduction

A major focus of this chapter is to use biopharmaceutical informatics resources to predict the clinical trial success of therapeutic mAbs. Monoclonal antibodies undergo several stages of clinical trials before they are approved for use. These stages are designed to evaluate the safety, efficacy, and dosage regimens of the mAbs. The starting work to support clinical development and ensure patient safety of mAbs are the mandatory nonclinical safety studies.[1] These preclinical studies are strategically designed to anticipate the dose, concentration, schedule, route of administration and duration to be used in the clinical studies. A Pre-IND meeting with the Food and Drug Administration (FDA) or other regulatory agencies prior to the initiation of nonclinical safety studies is also common and highly recommended in drug development.

The data from laboratory research and animal studies are then submitted to regulatory authorities for approval to proceed to clinical trials. The Phase 1 trials involve a small number of healthy volunteers or patients and focus primarily on assessing the safety and tolerability of the mAb. These are often dose escalation studies with investigation of potential side effects on around 10 to 50 people. The primary objectives of Phase 1 trials are to determine the maximum tolerated dose (MTD), pharmacokinetics, and

pharmacodynamics of the mAb. Phase 1 trials also provide initial evidence of the target mAb's potential efficacy and help identify any side effects or adverse reactions.

The Phase 2 trials involve a larger group of patients who have the condition or disease targeted by the mAb. Phase 2 trials may also explore different patient populations, dosing regimens, and potential combinations with other treatments. These trials aim to further evaluate the safety and efficacy of the mAb, determine the optimal dosage, and identify potential adverse effects. Phase 2 trials are usually randomized, controlled studies evaluating the safety and efficacy of a drug for a particular condition and involve participants selected using narrow criteria, to allow close monitoring of a relatively homogenous patient population. Phase 2 is more focused on the therapeutic efficacy in a particular patient population to establish whether or not the drug may ultimately benefit patients and provides a basis for decision-making regarding the mAb's further development in Phase 3 clinical trials.

Phase 3 trials involve a larger population of patients and are designed to confirm the mAb's efficacy, further assess its safety profile, and evaluate its overall risk-benefit ratio. Phase 3 trials involve comparing the mAb to a placebo or an existing standard of care to generate statistically significant data on the mAb's therapeutic benefits, optimal dosage, and potential adverse events. Sometimes the Phase 3 trials involve thousands of people in many different hospitals and even different countries. The results from Phase 3 trials are crucial for regulatory submissions and determining the mAb's approval for commercial use in each jurisdiction.

The manufacturer can submit a new Biologics License Application (BLA) or a similar regulatory submission to the appropriate regulatory authorities such as U.S. Food and Drug Administration (FDA) or the European Medicines Agency (EMA). After review, if the regulatory agency determines that the benefits outweigh the risks, they grant approval for the mAb's marketing and use. Post-marketing surveillance may then be conducted to gather additional information about the approved mAb's long-term safety, effectiveness, and optimal use in larger patient populations. Full detailed information and clinical trials guidance documents for US submission are available at https://www.fda.gov/regulatory-information/search-fda-guidancedocuments/clinical-trials-guidance-documents and at https://euclinicaltrials.eu/guidance-and-q-as/ for new submissions in the European Union (EU) region.

5.3 **Methods**

Creation of master training dataset: We created a combined Excel sheet that had the calculated scores from AbPred, Protein-Sol, and TAP. We evaluated a total of 52 properties that included Abpred assay scores (12), Protein-Sol sequence features (35), and the TAP scores (5). All these scores were evaluated for the TheraSabDab dataset which consists of information on 658 antibodies in clinical trials or approved stages. This sheet was a representation of the computational developability assessment results for clinical-stage antibodies. This training dataset was imported in MATLAB and 5-fold cross-validation was performed on this training dataset. Here, the data is partitioned into 5 randomly chosen subsets (or folds) of roughly equal size. One subset is used to validate the machine learning model trained using the remaining subsets. This process is repeated 5 times so that each subset is used exactly once for validation.

Machine Learning Classification algorithms for clinical trial progression: We have used the Machine Learning Toolbox™ and the new Classification Learner App in MATLAB to train machine learning models of all major classifiers: decision trees, discriminant analysis, support vector machines, logistic regression, nearest neighbors, naive Bayes, ensembles, and neural networks. The code generated and tested over 38 machine learning models which were tested across the TheraSabDab clinical-stage dataset. The features used in the model are Abpred assay scores (12), Protein-Sol sequence features (35), and the TAP scores (5). The models with the best percentage accuracy (validation) score were optimized further using manual classifier training procedures. Diagnostic measures, such as model accuracy, and plots, such as a scatter plot or the confusion matrix chart, reflect the validated model results. Full detailed information is available at https://uk.mathworks.com/help/stats/train-classification-models-in-classification-learner-app.html. and https://uk.mathworks.com/help.html.

Performance assessment using Failed antibody dataset: Failed antibodies dataset was created manually by compiling information from PubMed, Clinical trial database (www.clinicaltrials.gov), patents, regulatory filings, company websites, media news, and the ImMunoGeneTics information system. Antibodies withdrawn or discontinued due to safety, low efficacy or other strategic reasons were included in the dataset. See Table 16. The computational developability criteria performance was evaluated by calculating the 5% and 10% threshold cutoff values of clinical-stage antibodies and flagging the number of features failed by each failed antibody shown in Figure 46.

Scatterplot matrix of clinical trial status: We used gplotmatrix(X,[],group) function in MATLAB to obtain a matrix of scatter plots for the AbPred, Protein-Sol, and TAP results respectively. Each off-diagonal plot in the resulting figures is a scatter plot of a column of X against another column of X. Also, each diagonal plot represented the histogram for each feature under consideration. We have also highlighted clinical trial progression categories in different colours in the legend in all scatterplot figures.

AbPred measurements on antibody datasets: VHVL sequence information for separate antibody datasets were saved as input fasta files. The Abpred predictions were generated from the dockerhub source code available at docker pull maxhebditch/abpred using run command docker run --rm -v $(pwd)/:/abpred/host maxhebditch/abpred. More details at https://hub.docker.com/r/maxhebditch/abpred.

TAP measurements on antibody datasets: We used web sequence submission form and the GitHub repositories at https://github.com/orgs/oxpig/repositories to get the five metric values for the input sequences from the TAP tool available at https://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/sabpred. Also, the homology Fv models generated by ABodyBuilder2 were downloaded for future structural analysis.

Protein-Sol measurements on antibody datasets: We extracted the full antibody sequence information from the TheraSabDab database which was saved as input fasta files. The docker container was then used to create a local instance of the protein-sol solubility algorithm. The fasta files having sequence information were used as input on the docker program for the Protein-Sol algorithm that calculates 35 sequence features and provides an output table in Excel. The web-based Protein-Sol tool is also available at https://protein-sol.manchester.ac.uk/ for additional information.

Biopharmaceutical Informatics tools Calculations: Sequence information or the corresponding structural homology models from ABodyBuilder2 (inbuilt in the TAP calculations) were used as inputs to respective informatics tools outlined in Table 15.

Humanness Score Calculations: The T20 score analyzer is a tool that calculates the humanness of monoclonal antibody variable region sequences. The T20 score is scaled from 0 to 100, where a higher score is a more human-like antibody. In general, full-length sequences that score above 80 are considered human-like, while framework-only sequences that score above 85 are considered human-like. We used the online web tool available at https://sam.curiaglobal.com/t20/ to calculate the T20 score.

5.4 **Developability assessments for estimating clinical trial success**

In our endeavour to use developability features for estimating the clinical trial success and progression, we started with the five Therapeutic Antibody Profiler (TAP) features namely the Total CDR Length (L), Patches of Surface Hydrophobicity Metric (PSH), Patches of Positive Charge Metric (PPC), Patches of Negative Charge Metric (PNC) and Structural Fv Charge Symmetry Parameter (SFvCSP).

Each of these features is likely related to an underlying biophysical and developability aspect. For instance, the CDR length feature links to developability insights as it can capture the binding-site shape and CDRH3 loop features. A possible hypothesis is that a shorter CDR length is representative of a concave shape while a longer CDR length tends to have a convex binding site during epitope interaction. Further experimental studies such as antibody-antigen binding imaging and mapping studies for multiple CDR length profiles can help unravel more insights into the role of CDR length on developability. To evaluate the role of these features, we have created, refined, and implemented various machine-learning classification algorithms in MATLAB for estimating the clinical trial progression of TheraSabDab antibodies. Firstly, we have used the scatterplot matrices tool in MATLAB to visualize this multidimensional data.

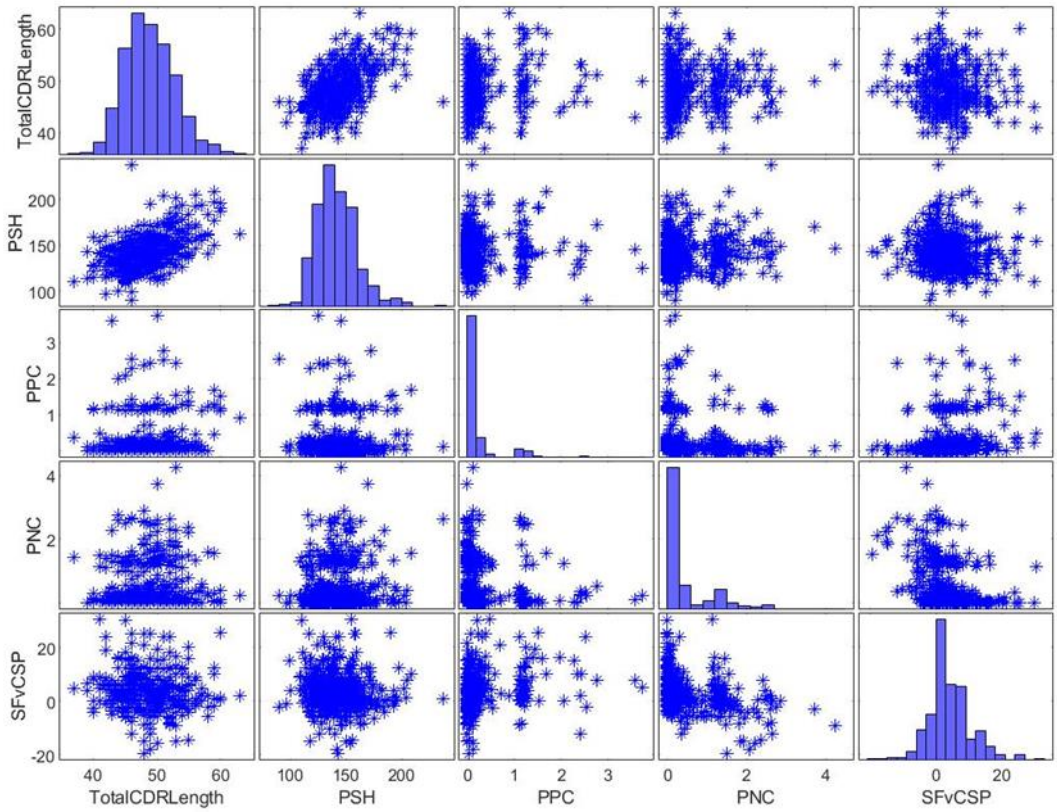


Figure 37: Scatterplot matrix of TAP metrics for TheraSabDab clinical-stage antibodies.

Figure 37 shows the scatterplot matrix of TAP metrics for TheraSabDab clinical-stage antibodies dataset. It depicts the relationship between the five Therapeutic Antibody Profiler (TAP) parameters for clinical-stage antibodies dataset. We observe a positive linear correlation between 'Total CDR Length' and 'PSH' features. It is likely because there is increase in the number or size of hydrophobic patches on the surface of the longer CDR as the length of the CDRs increase. A higher PSH score is representative of a higher aggregation propensity of antibodies. High PSH scores due to hydrophobic surface patches lead to unfavourable hydrophobic interactions with the antigen and formulation buffer that promote self-association, clustering, and aggregation of the antibody. The binding affinity, specificity and stability are negatively impacted by the presence of these hydrophobic surface patches which decrease overall stability profile.

Other feature pairs exhibited weak or no correlation with distributed scatter plots. We found an interesting observation that for many scatter plots the antibodies dataset TAP scores cluster into 2-3 distinct groups. This trend was highly evident in the plots for Patches of Positive Charge (PPC) and Patches of Negative Charge (PNC) scores.

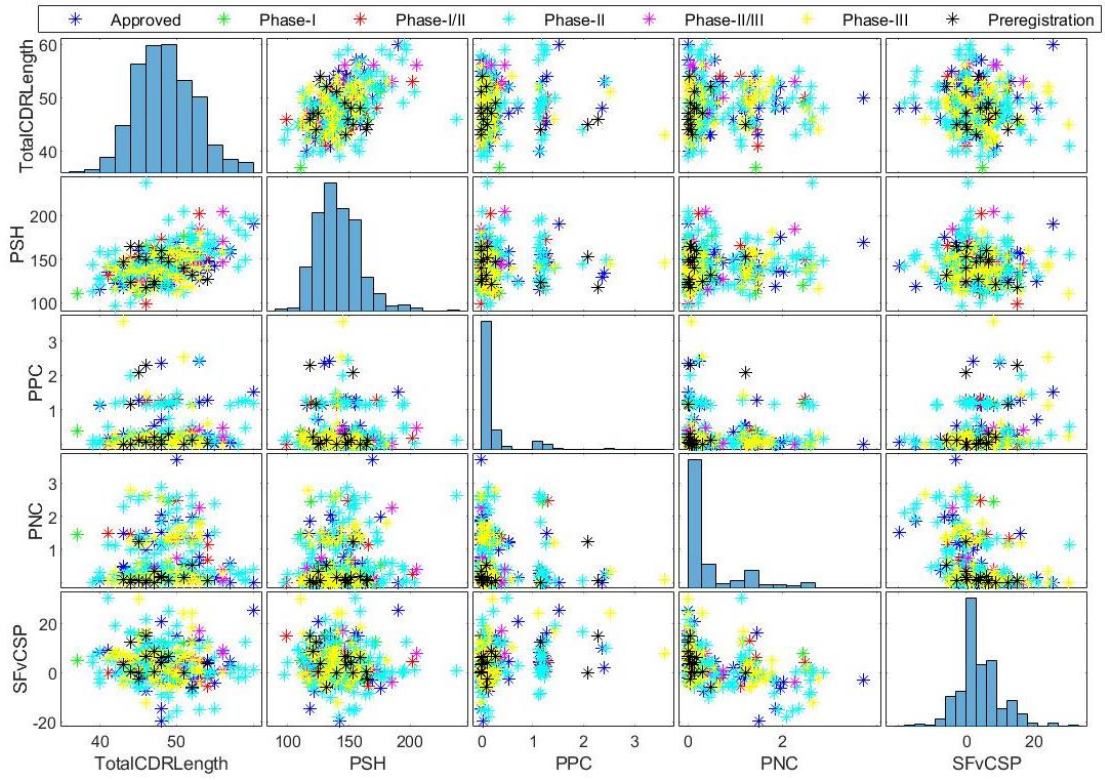


Figure 38: Scatterplot matrix of TAP features for TheraSabDab as per clinical trial status.

An initial hypothesis was that these clusters represent antibodies in different stages of clinical trials namely Phase 1, Phase 2, and Phase 3. Such a relationship if true would

have suggested the importance of positive or negative charge categories to classify clinical progression. Therefore, we annotated the clinical trial stage for each antibody in Figure 38. It shows the TAP scores along with the different stages of clinical trial progression in the legend namely – Approved, Phase 1, Phase 1/2, Phase 2, Phase 2/3, Phase 3, and others. Here we observe overlap among different categories of clinical trial stages for Therapeutic Antibody Profiler (TAP) scores. There are no visible distinct clusters or boundaries that separate TheraSabDab antibodies in different phases of clinical trials in Figure 38.

We also checked the overlap among the histogram distributions for different clinical trial stages with the Kolmogorov-Smirnov (K-S) test. The p-values from the K-S test for each of the five TAP property was less than 0.05 which suggests overlap in the score distributions. We, therefore, conclude that TAP scores cannot be solely used to predict the possible clinical trial stage outcome for a target antibody candidate. Next, we extended our analysis to the 12 AbPred developability assay features as well. We evaluated 12 biophysical assay feature scores namely HIC, SMAC, SGAC, CIC, CSI BLI, AC SINS, HEK, PSR, ELISA, BVP, DSF, and ACCSTAB from the AbPred tool.

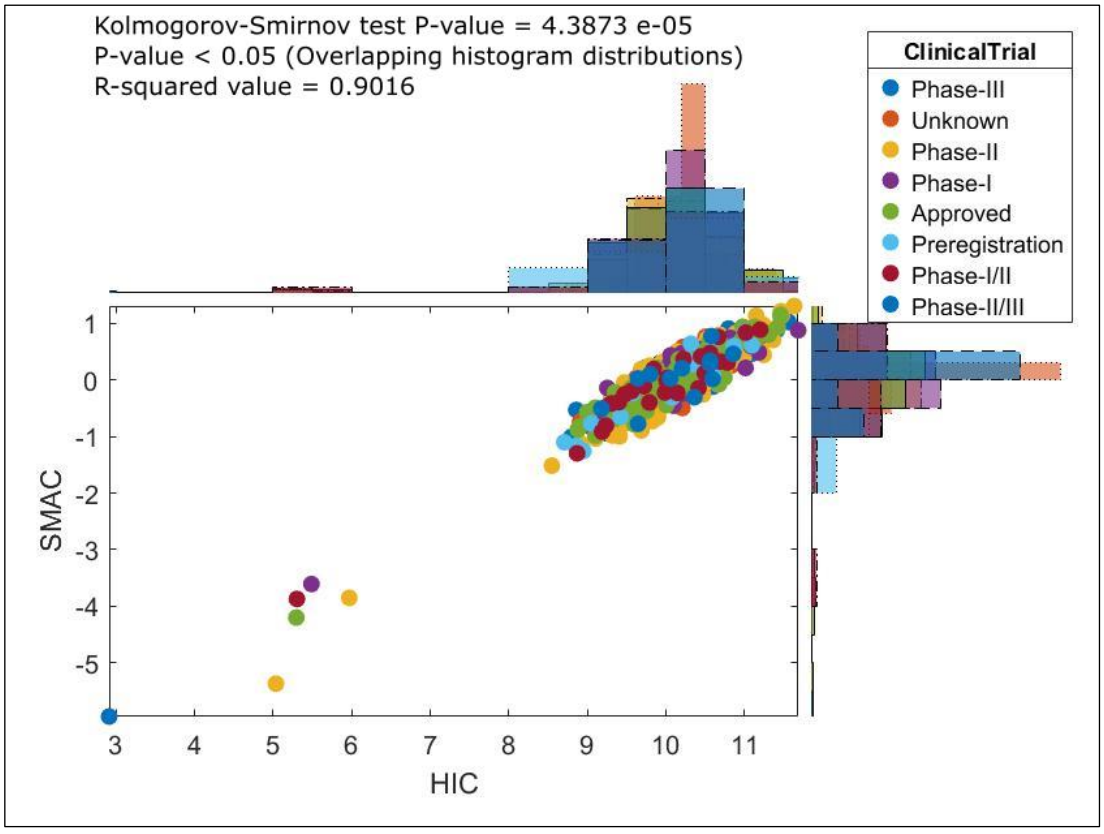


Figure 39: A sample scatter histogram for HIC and SMAC assay features for TheraSabDab.

The AbPred scatterplot matrices were created and analyzed for a total of 144 (12*12) pairs of biophysical assays. Here, we again had an overlap among the different stages of clinical trials with a low K-S test statistic p-value. So, we infer that none of these 12 AbPred assay features can solely estimate clinical trial progression. For example, Figure 39 shows a pair scatterplot histogram of the HIC and SMAC assay feature pair.

The HIC scores shown in Figure 39 are mostly between 9 to 11 while the SMAC scores are mostly between 1 to -1 for TheraSabDab clinical-stage antibodies. We can clearly see that histogram distributions for the approved (shown in green), phase 1 (shown in purple), phase 2 (shown in yellow), and phase 3 (shown in blue) are overlapping in these score ranges with a low K-S test statistic p-value of 4.3873 e-05 (p-value <0.05).

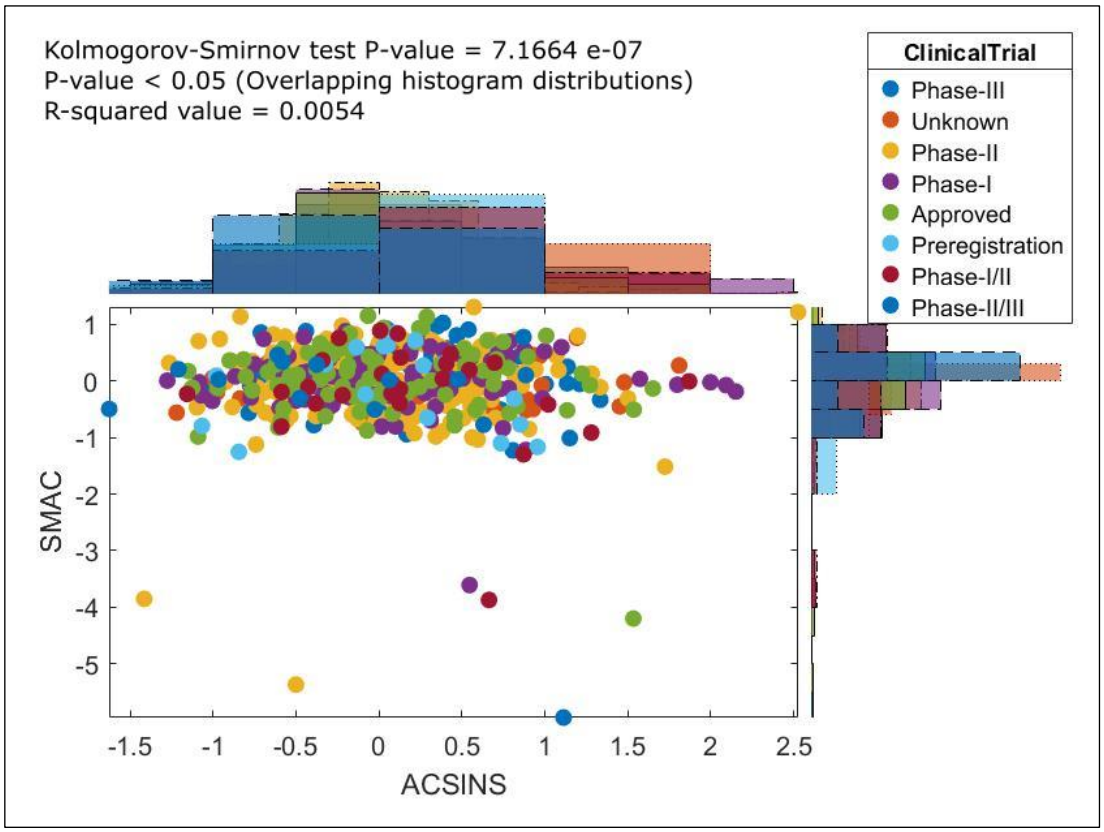


Figure 40: A sample scatter histogram for SMAC and AC-SINS features for TheraSabDab.

A similar trend of overlapping histograms was seen in all other 144 assay feature pairs. For example, Figure 40 shows overlapping distributions for SMAC and ACSINS assay pairs. These 12 assay features represent a range of biophysical properties ranging from hydrophobicity, aggregation propensity, solubility, colloidal stability, self-interaction, and long-term stability. However, none of these biophysical assay features can be used to predict the clinical trial stage of the target mAb candidate. A possible explanation

is that the assay biophysical features play a key role in preclinical safety and stability evaluation during the screening and optimization stages. However, once the mAb is ready for the clinical studies, it already is selected for the best manufacturability.

We also performed an analysis of 35 Protein-Sol sequence feature scores for all 658 TheraSabDab clinical-stage antibodies. These 35 Protein-Sol features are composed of 20 amino acid compositions; 7 composite scores of amino acid combinations (KmR = K-R, DmE = D-E, KpR = K+R, DpE = D+E, PmN = K+R-D-E, PpN = K+R+D+E, aro = F+W+Y); and 8 other sequence features (Length, pI, Kyte-Doolittle hydropathy, Absolute charge at pH 7, Fold propensity, Disorder propensity, Sequence entropy and β-strand propensity). Here we again observed an overlap among the sequence feature scores for mAbs in different clinical trial stages. So, we conclude that sequence-based features such as amino acid compositions and sequence-derived properties are not solely a predictor of clinical trial progression for therapeutic mAbs.

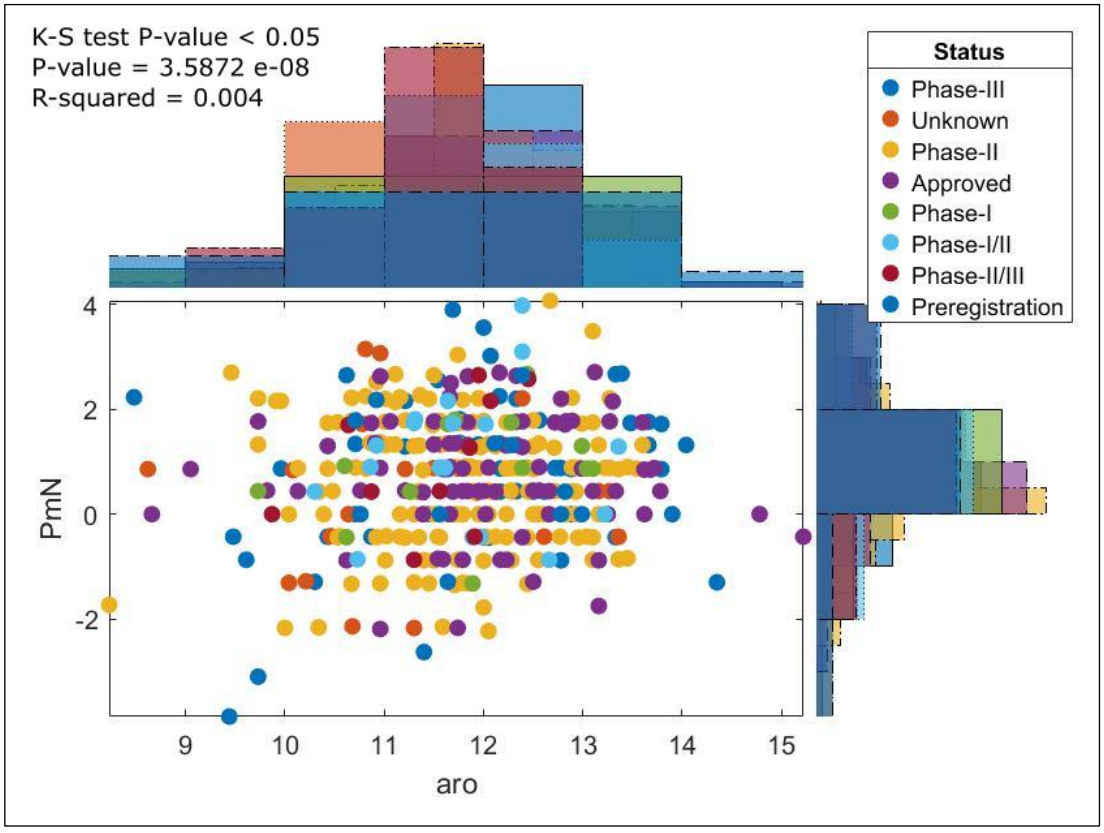


Figure 41: A sample scatter histogram for 'Charge' and 'aromatic content' for TheraSabDab.

We evaluated a total of 1225 sequence feature pairs (35*35) for all clinical-stage antibodies. A sample scatter histogram for charge (PmN) and aromatic amino acid content (aro) has been shown in Figure 41. Previous studies by our group have

suggested that charge and hydrophobicity based assays are key features for predicting the biophysical properties of clinical-stage antibodies.[2] We, therefore, checked in this study if we could predict the clinical trial progression stage based on the charge and hydrophobicity sequence features. However, we observe an overlap in the histogram distribution with a low K-S test p-value of 3.5872 e-08. This suggests that charge and hydrophobicity features calculated from sequence information are not sufficient to reliably estimate the clinical trial progression stage of monoclonal antibodies.

We have tabulated the K-S test p-value statistical measure for all features in Table 13. Overall, we conclude that all 50+ developability features used in our computational developability assessment workflow are unable to estimate clinical trial progression stage of therapeutic mAbs. It is therefore possible that many other factors apart from developability features and biophysical assay performance influence the likelihood of clinical trial success for any antibody therapeutic. We hypothesize from these results that it is not possible to predict a clinical trial stage based on just biophysical features.

Feature	P-value	Feature	P-value	Feature	P-value
CDR Length	2.5817 e -05	Cysteine (C)	5.3526 e -08	Tyrosine (Y)	5.7418 e -05
PSH	1.7389 e -06	Glycine (G)	2.9369 e -08	K minus R	1.2862 e -08
PPC	3.8632 e -02	Histidine (H)	3.6392 e -07	D minus E	1.3648 e -08
PNC	5.7916 e -02	Isoleucine (I)	7.1225 e -06	K plus R	3.8465 e -07
SFvCSP	6.7527 e -05	Leucine (L)	1.4263 e -08	D plus E	6.7253 e -08
HIC	4.3383 e -05	Methionine (M)	9.1412 e -08	PmN (Charge)	1.8229 e -08
SMAC	4.9097 e -07	Proline (P)	8.6045 e -09	PpN	7.8212 e -05
SGAC	7.2803 e -07	Serine (S)	2.2853 e -06	Aro (F + W + Y)	1.7643 e -08
CIC	4.3095 e -06	Threonine (T)	7.2752 e -07	folding (fld)	2.9332 e -08
CSI BLI	8.1164 e -06	Valine (V)	8.8264 e -08	disorder (dis)	5.8657 e -07
AC SINS	2.2567 e -07	Aspartate (D)	5.3318 e -06	beta strand (bet)	8.1334 e -08
HEK	3.8273 e -06	Glutamate (E)	3.7528 e -08	Length	4.1092 e -05
PSR	5.2681 e -06	Phenylalanine (F)	9.4193 e -08	pI	3.5818 e -06
ELISA	1.2903 e -07	Lysine (K)	4.8169 e -07	entropy (ent)	7.1295 e -08
BVP	4.3529 e -06	Asparagine (N)	8.1827 e -08	Charge ph7	6.4205 e -07
DSF	5.8267 e -05	Glutamine (Q)	7.2482 e -06	Kyte-Doolittle	1.7304 e -08
ACC STAB	6.7914 e -06	Arginine (R)	5.4292 e -08		
Alanine (A)	9.4822 e -08	Tryptophan (W)	3.4138 e -08		

Table 13: K-S test statistic p-value for used tool features. All have p-values < 0.05.(Overlap)

Indeed, some previous studies have demonstrated that clinical trial design, therapeutic area, location of trial, primary outcome measure and are some other key factors beyond

biophysical properties that influence the outcome of clinical trials.[3-5] Other parameters such as disease type, industry sponsor, biomarker, lead indication status, and time are also important in deciding the clinical trial outcome.[6] These can be subjective features that are unrelated to any developability measure and can be highly variable between different clinical trials. So, these other factors need to be accounted for to estimate the clinical trial progression for an antibody candidate. However, we know from industry application perspective that only developability-related features can be controlled in the antibody design and engineering process to optimize the clinical stage outcome.

5.5 Feature engineering and machine learning classification of the developability assay properties for clinical trial progression

Feature engineering is the process of selecting, manipulating, and transforming raw data and variables into features that can be used for creating a predictive model using machine learning. Since our 50+ sequence or structural features used in computational developability assessments were not able to predict clinical trial progression of mAbs, in this section we have used feature engineering techniques to create new variables that can potentially predict clinical trial stage outcomes of therapeutic mAbs. We used three major feature generation approaches: transforming variables, discretization, and summarizing groups. We utilized the feature selection functions available in Statistics and Machine Learning Toolbox™ in MATLAB and the Classification Learner App.

The feature selection algorithms search for a subset of predictors that optimally models measured responses, subject to constraints such as required or excluded features and the size of the subset. In Classification Learner, we evaluated the different features (or predictors) to include in the model while also testing multiple machine learning classification model types. The Classification Learner app has in-built algorithms to train machine learning models of all major classifiers: decision trees, discriminant analysis, support vector machines, logistic regression, nearest neighbors, naive Bayes, ensembles, and neural networks. Finally, the best percentage accuracy (validation) score was compared for different ML algorithms and the results were summarized.

The automated classifier trained a selection of different types of classification models on our full dataset for 52 features using 5-fold cross-validation. We created and tested over 38 machine-learning models. (See methods). These included basic linear models such as linear discriminant to complex models such as trilayered neural networks and

cubic KNN method. We observed that the maximum achieved accuracy was 43.7% for coarse KNN, linear discriminant, SVM, and ensemble methods. Neural network methods performed comparatively poorly with a maximum model accuracy achieved of 41.2%. Figure 42 presents the neural network model 3.1 results for our dataset with data categorized as per their clinical trial progression. We observe that most of the predictions are incorrect (shown in a cross) rather than antibodies classified correctly (shown in dots). Even with the manual fine-tuning, we were unable to find a reliable machine-learning classification algorithm that could predict clinical trial progression.

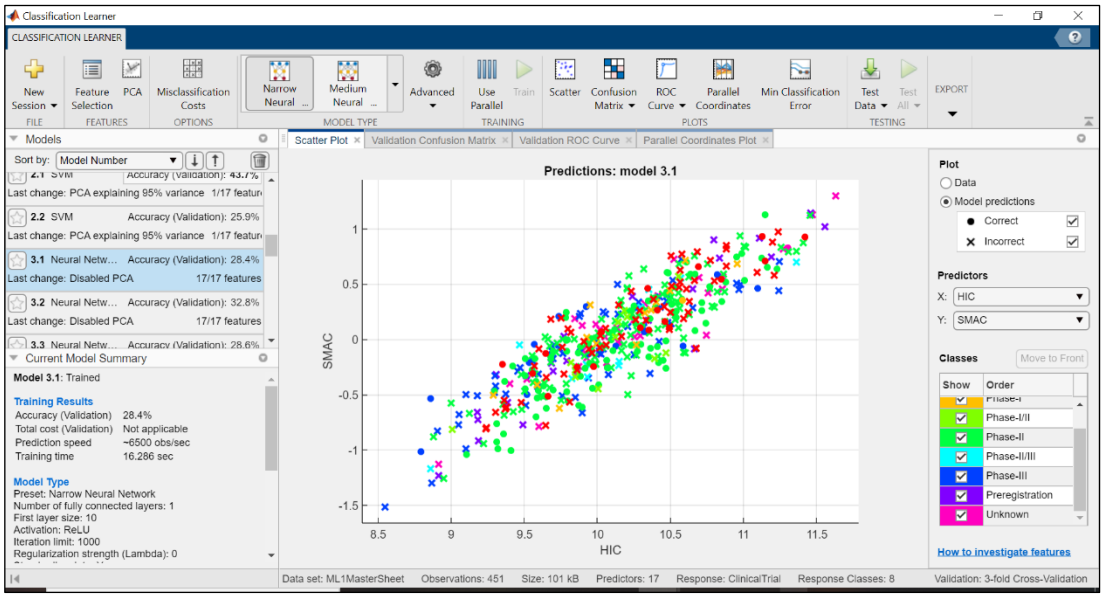


Figure 42: Classification Learner interface in MATLAB for creating and analyzing machine learning algorithms for the clinical-stage antibodies developability dataset. The HIC vs SMAC plot from the neural network model 3.1 is shown in the above figure that is classified according to different stages of clinical trials. Model accuracy and other information is shown on left.

The Receiver Operator Characteristic (ROC) curve is a well-established evaluation metric for classification problems. It is a probability curve that plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at the various threshold values. True Positive Rate (TPR) is also known as sensitivity or recall. TPR represents the proportion of actual positive instances correctly classified as positive by the model. It is calculated as TP / (TP + FN), where TP is the number of true positives and FN is the number of false negatives. The False Positive Rate (FPR) is the proportion of actual negative instances incorrectly classified as positive by the model. It is calculated as FP / (FP + TN), where FP is the number of false positives and TN is the number of true negatives. The Area under the ROC Curve (AUC) metric measures the entire two-dimensional area underneath the entire ROC curve. The AUC-ROC curve is created

by plotting the TPR on the y-axis against the FPR on the x-axis, with each point representing a different classification threshold. The curve ranges from (0,0) to (1,1), where (0,0) represents a perfect classifier and (1,1) represents a random classifier.

The area under the curve (AUC) is a metric used to summarize the performance of the model. A perfect classifier has an AUC of 1, while a completely random classifier has an AUC of 0.5. Generally, a higher AUC indicates that the model is more accurate at distinguishing between positive and negative instances. Figure 43 shows the output AUC-ROC curves for the best performing machine learning classification algorithms created from the Classification Learner app in MATLAB. We observe that the AUC values for the models are low and near 0.5. For instance, the Naïve Bayes Classifier has an AUC of 0.56 for clinical-stage dataset. It suggests that our classification models have no discrimination capacity to distinguish between different clinical trial stages.

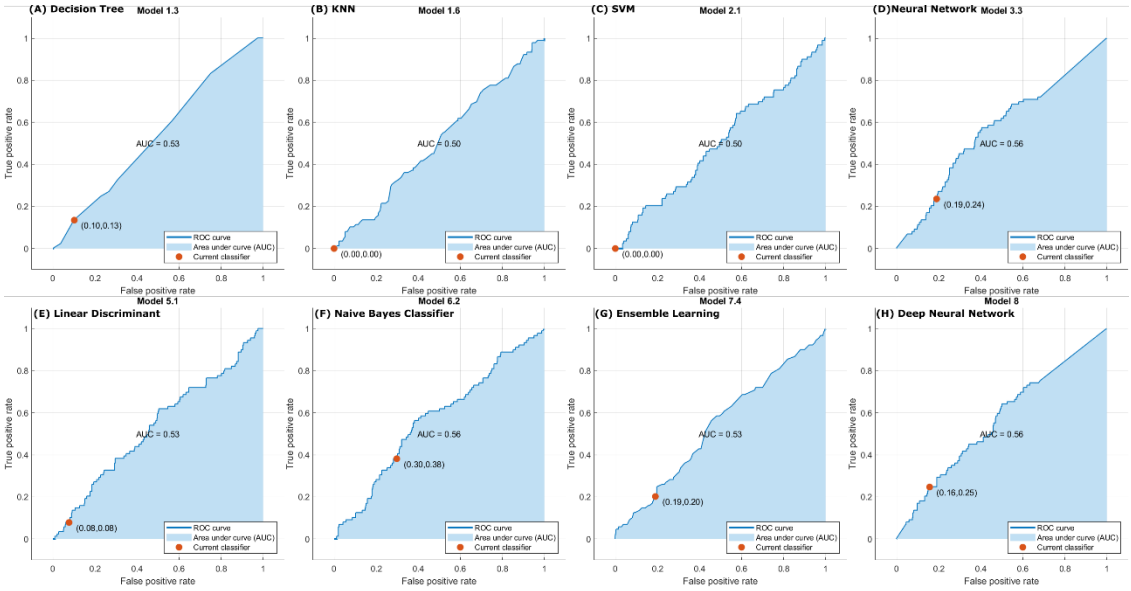


Figure 43: Receiver Operating Characteristic (ROC) curves and model performance for all machine learning algorithm types implemented using the Classification learner in MATLAB.

A confusion matrix is another performance evaluation tool commonly used in machine learning to assess the accuracy of a classification model. It is particularly useful when dealing with multi-class classification problems. Figure 44 shows the confusion matrix for the fine Gaussian SVM classifier for our computational developability assessment dataset on clinical-stage mAbs. The y-axis in the figure shows the actual 'true' labels while the x-axis shows the 'predicted' clinical-stage labels by the SVM classifier.

The confusion matrix is a table that summarizes the performance of a classification model by comparing the predicted class labels with the actual class labels of the test data. We observe high False Negative Rates (FNR) for all categories of clinical-stage antibodies. The approved antibodies had a false negative rate of 61.8% which implied that our algorithm would miss the true positives 61.8% of the time with being accurate in the prediction of true positives only 38.2% of the time which is very low. The Phase 1 labelled antibodies had FNR of 100% which means that our algorithms completely classified the actual Phase 1 antibodies wrongly into other clinical-stage categories. The Phase 2 labelled antibodies had a high FNR of 54.3% while the Phase 3 labelled antibodies also had a very high FNR of 85.5%. Therefore, we conclude our machine learning algorithms perform very poorly in classifying therapeutic mAbs as per their clinical trial progression status. This poor performance of our classifiers is very likely due to other features beyond developability that impact the clinical trial progression.

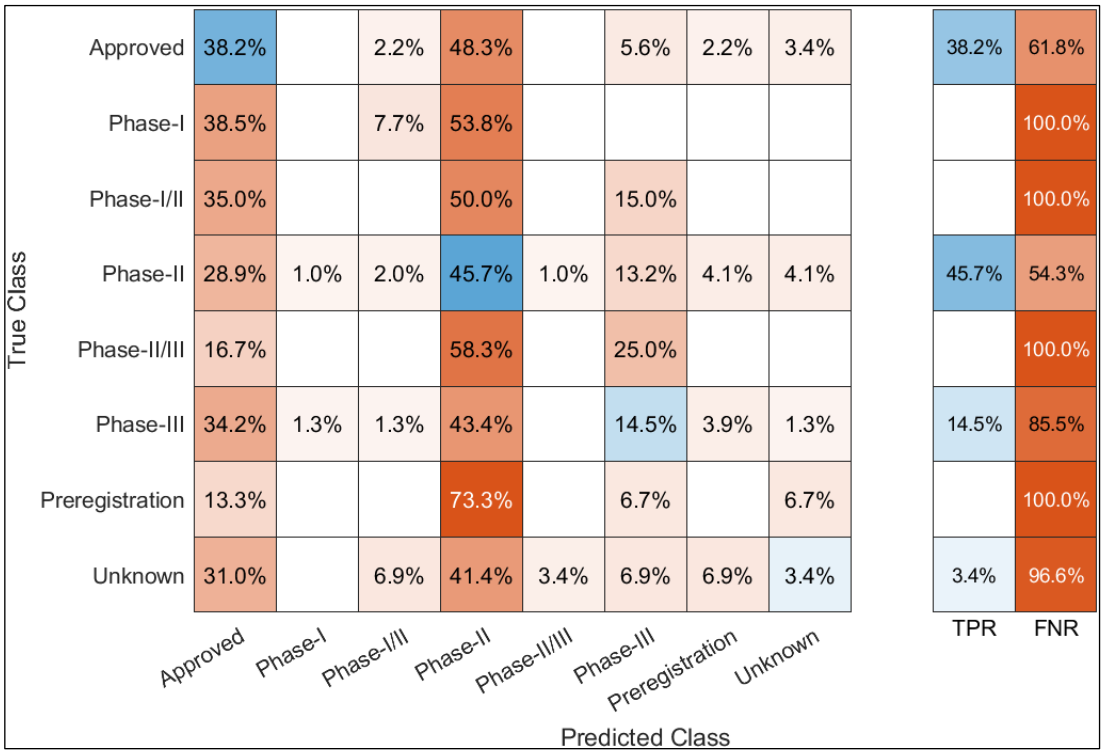


Figure 44: Confusion matrix of clinical stage outcomes for the fine Gaussian SVM algorithm.

Overall, the feature engineering and machine learning classification results suggest that the features derived from biophysical properties or developability assays are not reliable to classify therapeutic antibodies as per their clinical-trial progression stage. Previous studies have shown that several other factors such as clinical trial design, therapeutic area, location of trial, and primary outcome measure impact the clinical

trial progression of therapeutic antibodies.[3-5] Also other external elements associated with biology, mechanism of action, and risk vs benefit profile of a biological drug candidate for an indication can also significantly affect its clinical trial progression.

Therefore, we conclude that clinical trial design, location, disease area, drug biology, primary outcome measure, and others are some factors beyond biophysical properties that may influence the outcome of clinical trials. Only the developability scores and features are not solely predictive of clinical trial progression as verified by ROC curves and model performance for our machine learning classification algorithms. Any future machine learning algorithms should incorporate these additional factors to obtain a reliable classification of the clinical trial progression for therapeutic antibodies.

5.5.1 Evaluation of multiple biopharmaceutical informatics tools

We evaluated new features from other biopharmaceutical informatics tools to compare the predictive ability of these new features to estimate clinical trial progression since the previous tools were not successful in estimating clinical trial progression. These new tools explore a range of biophysical properties such as aggregation, solubility, post-translational modifications, and immunogenicity. Table 14 shows the new tools used to predict clinical trial progression of therapeutic antibodies.

Biopharmaceutical Informatics tools evaluated for estimating Clinical Trial progression			
➢ CamSol	➢ SAP	➢ AggScore	➢ SCWRL4.0
➢ SODA	➢ Solubis	➢ PASTA 2.0	➢ MUpro
➢ SOLpro	➢ GAP	➢ TANGO	➢ PEARS
➢ SOLart	➢ Aggrescan 3D	➢ MusiteDeep	➢ NetMHCIIpan
➢ ANTIGENpro	➢ DiscoTope	➢ SVMTriP	➢ RANKPEP
➢ COBEpro	➢ ElliPro	➢ AbAdapt	➢ T20 Score

Table 14: List of biopharmaceutical informatics tools and features evaluated in this project

CamSol is a tool for solubility screening of protein libraries that uses features related to proximity of the amino acids in the three-dimensional structure and for their solvent exposure. SODA is another tool for protein solubility based on disorder and aggregation. It introduces new features related to disorder, aggregation, helix, and strand propensity differences. SOLpro, SOLart, and Solubis are other well-known solubility predictors available as publicly available biopharmaceutical informatics resources for solubility assessment. We have also utilized tools for aggregation prediction. AggScore is a tool to identify aggregation-prone regions based on features like intensity and relative orientation of hydrophobic and electrostatic surface patches. Another similar tool is Aggrescan 3D that uses solvent-exposed aggregation-prone regions as the main feature for selection. We have especially looked at additional tools for prediction of immunogenicity risk in this comparative analysis. ANTIGENpro is a sequence-based web tool that predicts protein antigenicity using protein microarray data features. COBEpro is another SVM-based tool that has been used in computational immunology for predicting continuous B-cell epitopes. The COBEpro tool uses epitopic propensity scores as the main feature for peptide fragments and residues within a sequence. These tools have been reviewed previously in detail in Chapter Two and listed in Table 2 for additional information.

Tool Name	Main biophysical property	Phase -1	Phase -2	Phase -3
CamSol[7]	Solubility Score	−0.531 ± 0.124	−0.582 ± 0.358	−0.447 ± 0.286
SODA[8]	Solubility Score	62 ± 35	58 ± 47	74 ± 39
SOLpro	Solubility Score	N/A	N/A	N/A
SOLart[9]	Solubility Score	71% ± 28%	79% ± 16%	75% ± 18%
ANTIGENpro[10]	Protein Antigenicity	0.04 ± 0.01	0.04 ± 0.015	0.04 ± 0.01
COBEpro[11]	Epitopic propensity scores	25 ± 8	37 ± 14	28 ± 12
SAP	Aggregation propensity score	0.12 ± 0.29	0.18 ± 0.35	0.14 ± 0.22
Solubis[12]	Aggregation propensity score	61.73 ± 34.59	59.16 ± 42.03	64.82 ± 31.58
GAP[13]	β-strand aggregation	0.851 ± 0.416	0.912 ± 0.394	0.837 ± 0.208
Aggrescan 3D[14]	Aggregation (A3D) score	−1.64 ± 0.62	−1.81 ± 0.54	−1.69 ± 0.85
DiscoTope[15]	discontinuous epitopes score	−11.67 ± 7.16	−14.81 ± 8.62	−13.44 ± 11.09
ElliPro[16]	discontinuous epitopes score	0.574 ± 0.215	0.519 ± 0.202	0.538 ± 0.281
Aggscore[17]	Aggregation score	82.9 ± 63.2	102.3 ± 78.5	87.2 ± 54.6
PASTA 2.0[18]	PASTA Energy Unit (PEU)	2.8 ± 1.5	2.3 ± 1.1	2.9 ± 1.5
TANGO	aggregation tendency	N/A	N/A	N/A
MusiteDeep[19]	No. of PTM sites	2 ± 2	3 ± 3	3 ± 3
SVMTriP	linear antigenic epitopes	N/A	N/A	N/A
AbAdapt	antibody-antigen docking	N/A	N/A	N/A

SCWRL4.0	side-chain conformations	79.1% ± 11.5%	74.1% ± 12.4%	78.0% ± 13.9%
MUpro	Single Site Mutations	N/A	N/A	N/A
PEARS	Side chain prediction	N/A	N/A	N/A
NetMHCIIpan[20]	% MHC II binding score	9.47 ± 5.83	10.32 ± 7.02	10.16 ± 7.54
RANKPEP	% MHC II binding score	8.86 ± 6.91	10.67 ± 8.48	9.34 ± 6.70
T20 Score[21]	Humanness score	77.46 ± 16.16	81.67 ± 15.13	84.64 ± 14.88

Table 15: Evaluation of average ± standard deviation for multiple tools across clinical stages.

5.5.2 Humanness Score: A reliable estimate of clinical trial progression

Our comparative analysis in Table 15 revealed that the 'T20 Humanness Score' was the most reliable feature among all tools for estimating the clinical trial progression of mAbs. The T20 score quantifies the humanness of the variable region of monoclonal antibodies which is derived by comparing the sequence identity of an input sequence to a large database of ~38,700 human antibody variable region sequences.

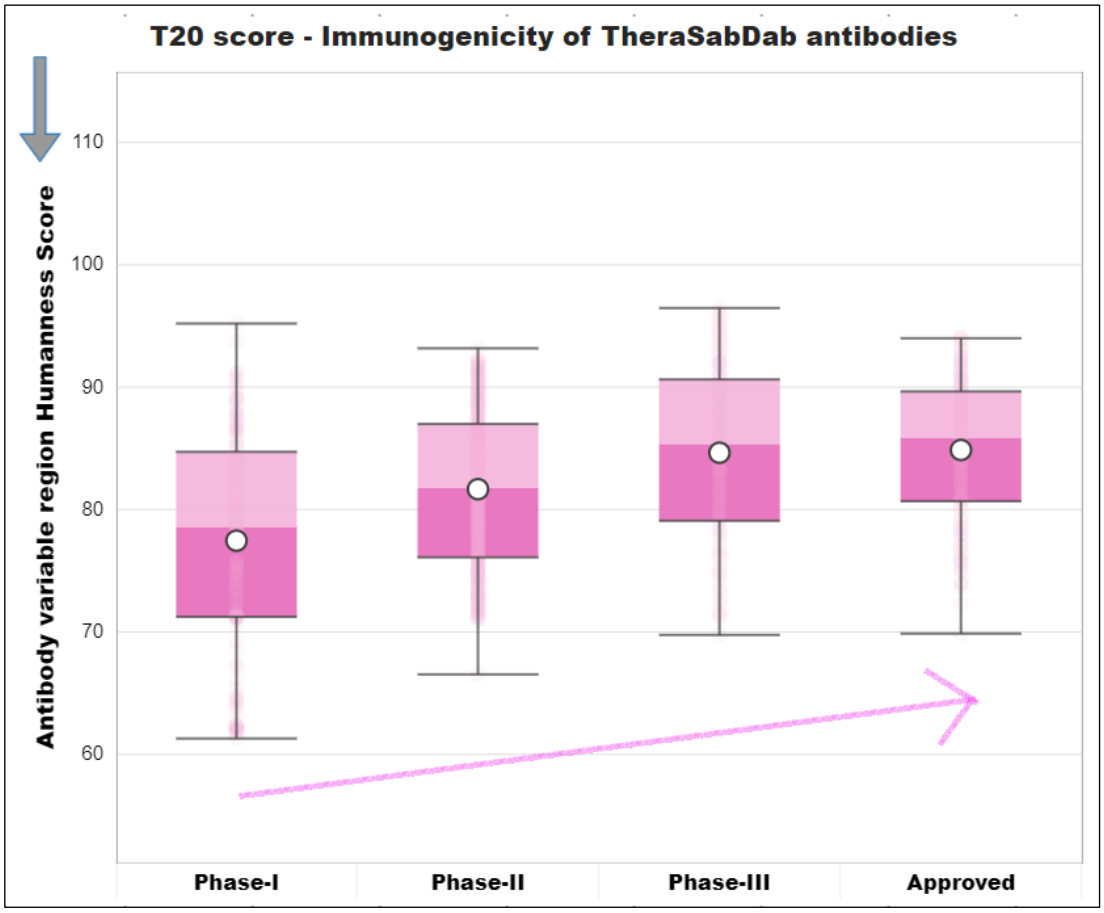


Figure 45: T20 score box and whisker plots for antibodies in different clinical trial stages.

The Protein BLAST methods are used to determine the percent identities of the top 20 matched sequences which are then averaged to obtain the T20 score. The T20 score is scaled from 0 to 100, where a higher score is a more human-like antibody. Gao *et al.*

have previously demonstrated that a high T20 score is correlated with decreased immunogenicity.[21] The rationale for using the T20 humanness score is that a more human-like antibody with a corresponding high T20 score is likely to have few B-cell epitopes and T-cell epitopes that trigger an inflammatory immune response and cause immunogenicity upon antibody administration. Also, non-human sequences would likely be identified as foreign proteins by our immune system and patches of unusual residues on the surface may lead to immune response. So, high T20 scores are desired.

We analyzed the T20 humanness score for all categories of clinical stages. The results are summarized in Figure 45. We observe enrichment of human regions as antibodies progress through clinical trials with an increase in T20 score from Phase 1 to approved therapeutic antibodies. Phase 1 antibodies have a mean score of 77.46 with the lower whisker at 61.30. Phase 2 antibodies have a mean score of 81.67 with the lower whisker now increased to 66.54. Phase 3 antibodies have a mean score of 84.64 and a lower whisker at 69.76. Approved antibodies have a mean T20 score of 84.87 with the highest T20 scores among all categories. Immunogenic antibodies face attrition as they progress through clinical trials. Therefore, among all biophysical features the T20 humanness score is the most appropriate to get a fair estimate of which stage the target antibody would at least achieve in clinical trials. However, as discussed before more subjective features beyond biophysical properties would ultimately decide the clinical trial progression or the final clinical outcome of therapeutic monoclonal antibodies.

5.6 **Failed antibodies dataset of withdrawn and discontinued mAbs**

The machine learning classification algorithms were not able to distinguish between different stages of clinical trials. So, in this final section, we have tested the ability of a custom binary classification algorithm based on our computational developability assessment criteria from Chapter 3 Table 6 to flag the mAbs that failed clinical trials.

We have manually created a novel dataset of the withdrawn and discontinued mAbs in clinical trials. This data was obtained from different sources including PubMed, the Clinical trial database (www.clinicaltrials.gov), patents, regulatory filings, company websites, media news and the ImMunoGeneTics information system (www.imgt.org). Table 16 shows this new dataset of withdrawn or discontinued therapeutic antibodies. Many drugs are withdrawn, at various stages, with several potential reasons for limited success in clinical trials, including insufficient therapeutic effect or, in some cases the

antibody being poorly tolerated. Here, we have selected antibodies that are part of this natural attrition pipeline. These antibodies faced attrition from clinical trials either due to safety reasons, low therapeutic efficacy, financial, commercial, or strategic reasons.

INN (International Nonproprietary Name)	Common name / Proprietary name	Company	Clinical indication	Highest Clinical Trial	Additional Information
Aprutumab	BAY-1179470; FGFR2-TTC	Bayer AG (Germany)	Solid tumours	Phase-I	NCT02368951 800038337
Ascrinvacumab	PF-03446962	Pfizer (NY USA)	Colorectal cancer	Phase-II	NCT01911273 27329247/
Azintuxizumab	ABBV-838	AbbVie Inc. (IL USA)	Multiple myeloma	Phase-I	NCT02462525
Bapineuzumab	AAB-001: Bapi	Pfizer (NY USA) Johnson & Johnson (PA USA)	Alzheimer's disease (AD)	Phase-II	reuters.com/article/
Bivatuzumab	BIWA4	Boehringer Ingelheim Pharmaceuticals (CT USA)	Cancers, squamous	Phase-I	800016679
Bococizumab	PF-04950615; RN316	Pfizer (NY USA)	Cardiovascular diseases	Phase-III	NCT01975376 fiercebiotech.com/
Briakinumab	ABT-874	Abbott GmbH & Co. KG	Psoriasis; Rheumatoid arthritis	Phase-III	800010080 doi
Carlumab	CNTO-888	Centocor Biotech, Inc. MorphoSys AG	Idiopathic pulmonary fibrosis; Ovarian cancer; Solid tumours	Phase-II	800026524 doi
Cixutumumab	IMC-A12	ImClone Systems Eli Lilly and Company (IN USA)	Lung Cancer, Malignant Neoplasm, Adenocarcinoma	Phase-II	NCT01182883 doi
Clivatuzumab	Clivatuzumab 90Y-Hpam4	Immunomedics (NJ, USA)	Pancreatic cancer	Phase-III	800018609 NCT01956812
Dacetuzumab	SGN 14, SGN-40, huS2C6	Seattle Genetics (WA USA)	Chronic lymphocytic leukaemia	Phase-II	800014913 articles/
Daclizumab	ZENAPAX® ZINBRYTA™	AbbVie Inc. (IL USA) Roche (EU)	Multiple sclerosis (MS)	Approved	reuters.com/article
Demcizumab	OMP-18M21; OMP-21M18	Celgene Corporation; OncoMed Pharmaceuticals	Fallopian tube cancer; Ovarian cancer	Phase-I	800023162 doi
Depatuxizumab	ABT-414; Depatux-M	AbbVie Inc. (IL USA)	Glioblastoma; Gliosarcoma	Phase-III	800035129 news.abbvie
Duligotuzumab	MEHD-7945A; RO-5541078	Genentech Inc. (CA USA)	Colorectal cancer; Head & neck cancer	Phase-II	800033001 doi
Duvortuxizumab	MGD 011; CD3xCD19 DART	MacroGenics; Janssen Biotech	B-cell Malignancies; Haematological malignancies	Phase-I	NCT02454270 800042606
Ecromeximab	KW-2871	Kyowa Kirin, Inc.	Malignant Melanoma	Phase-II	NCT00199342
Efalizumab	hu1124 RAPTIVA®	Merck (EU) Xoma (CA USA)	Psoriasis (moderate to severe infection)	Approved	fda.gov/raptiva 800007491
Efungumab	HSP90mab, Mycograb®	NeuTec Pharma plc (Manchester UK)	Candidiasis (yeast infection)	Phase-III	ema 800018318
Enokizumab	MEDI-528	Genaera Corporation (PA USA) MedImmune (MD USA)	Asthma	Phase-II	NCT00590720 800011810
Fasinumab	REGN 475; SAR 164877	Mitsubishi Pharma; Teva Pharmaceutical	Osteoarthritis	Phase-III	NCT00944892 reuters.com/article/
Figitumumab	CP-751871	Pfizer (NY USA)	Non-small cell lung cancer; Cancers	Phase-III	800020634 reuters.com/article
Garivulimab	BGB-A333	BeiGene (China)	Advanced Solid Tumors	Phase-II	800050989 NCT03379259

Lexatumumab	HGS-ETR2	Cambridge Antibody Technology (UK)	Solid tumor and Lymphoma	Phase-I	800017268 NCT00428272
Lucatumumab	CHIR-12,12, HCD122	Xoma (Berkeley CA USA)	Follicular lymphoma, Multiple Myeloma	Phase-I	800013077 NCT00108108
Matuzumab	EMD 72000	Merck (Geneva Switzerland)	Cervical cancer	Phase-II	800007164
Muromonab CD3	OKT3; OKT-3;	Johnson & Johnson (PA USA)	Renal cancer; Organ transplant	Approved	800005151 Muromonab-CD3
N/A	CDX-3379	Celldex Therapeutics (NJ USA)	Malignant melanoma Head and Neck cancer	Phase-II	ir.celldex.com /news-releases/
N/A	PF-04605412	Pfizer (NY USA)	Solid tumours	Phase-I	NCT00915278 doi
N/A	Hu3S193	Recepta Biopharma (SP Brasil)	Breast cancer Colorectal cancer	Phase-II	NCT01370239 doi
Ocaratuzumab	AME-133; AME-133v; LY 2469298	Mentrik Biotech; Applied Molecular Evolution (CA USA)	Follicular lymphoma; Non-Hodgkin's lymphoma	Phase-II	800018236 doi
Ozanezumab	1223249; GSK1223249	GlaxoSmithKline (Brentford UK)	Amyotrophic lateral sclerosis	Phase-II	28139349/ 800033680
Pinatuzumab	DCDT-2980S; FCU-2703; RO-5541072;	Roche (Basel Switzerland), Genentech Inc. (CA USA)	Chronic lymphocytic leukaemia; Non-Hodgkin's lymphoma	Phase-II	PIIS2352-3026 800027651
Rilotumumab	AMG-102	Amgen (CA USA)	Colorectal cancer; Gastric cancer	Phase-II	Rilotumumab NCT02137343
Ruplizumab	BG 9588; 5c8; hu5c8	Biogen (MA USA)	Systemic lupus erythematosus	Phase-II	12632425/ 800010330
Solitomab	MT110; AMG-110	Amgen (CA USA)	Solid tumours	Phase-I	PMC6136859/ 800028176
Theralizumab	TGN1412, CD28-SuperMAB	TeGenero Immuno Therapeutics (Würzburg, Germany)	B cell chronic lymphocytic leukemia	Phase-I	wiki/ ncbi.nlm.nih.gov/
Vadastuximab	SGN-CD33A; 33A	Seattle Genetics (WA USA)	Acute Myeloid Leukemia (AML)	Phase-III	NCT02785900 genengnews.com/
Vesencumab	MNRP1685A; R 7347; RG7347	Roche (Basel Switzerland), Genentech Inc. (CA USA)	Advanced solid tumors	Phase-I	24604265/ 800027393
Visilizumab	HuM-291; Nuvion	PDL BioPharma Inc.	Crohn's disease; Ulcerative colitis	Phase-III	clinicaltrials.gov/ pdl-hazard
Zalutumumab	HuMax-EGFr	Genmab (Copenhagen, Denmark)	Head and neck cancer; Other Cancers	Phase-II	24714973/ 800015831

Table 16: Dataset of withdrawn and discontinued antibodies. These antibodies faced attrition either due to safety reasons, low therapeutic efficacy, commercial, or strategic reasons.

A previous review has outlined the important considerations and types of nonclinical safety evaluation for therapeutic antibodies.[22] An IND-enabling safety package for a mAb will most likely include a human tissue cross-reactivity study and a general toxicity study in at least one relevant species, most likely a non-human primate (NHP). Safety pharmacology and immunotoxicity studies are also usually included in this data package to support the clinical development and regulatory filing documents.

Our computational developability assessment criteria evaluate the corresponding *in silico* features to predict clinical trial attrition of therapeutic antibodies. We evaluated these criteria for both the clinical-stage and failed antibodies dataset. The comparative performance of our computational developability criteria is shown in Figure 46. We

have checked the count (shown as percentage of total dataset) of the antibodies that failed the 10% threshold criteria for each dataset. Most of the clinical antibodies had zero flags as 67.3% of TheraSabDab dataset had optimal scores in each of the twelve assays. The rest of the clinical-stage dataset had one or two flags with a count of 14.9% and 12.2% respectively. An outlier case in the clinical-stage dataset with six flags was Murlentamab, a humanized anti-Müllerian hormone receptor antibody that is currently being evaluated for colorectal cancer in phase 2 trials and other cancer types in phase 1 clinical trials. So, only 17.8% (100% – 67.3% – 14.9%) of the clinical-stage dataset had two or more flags from our computational developability assessment criteria.

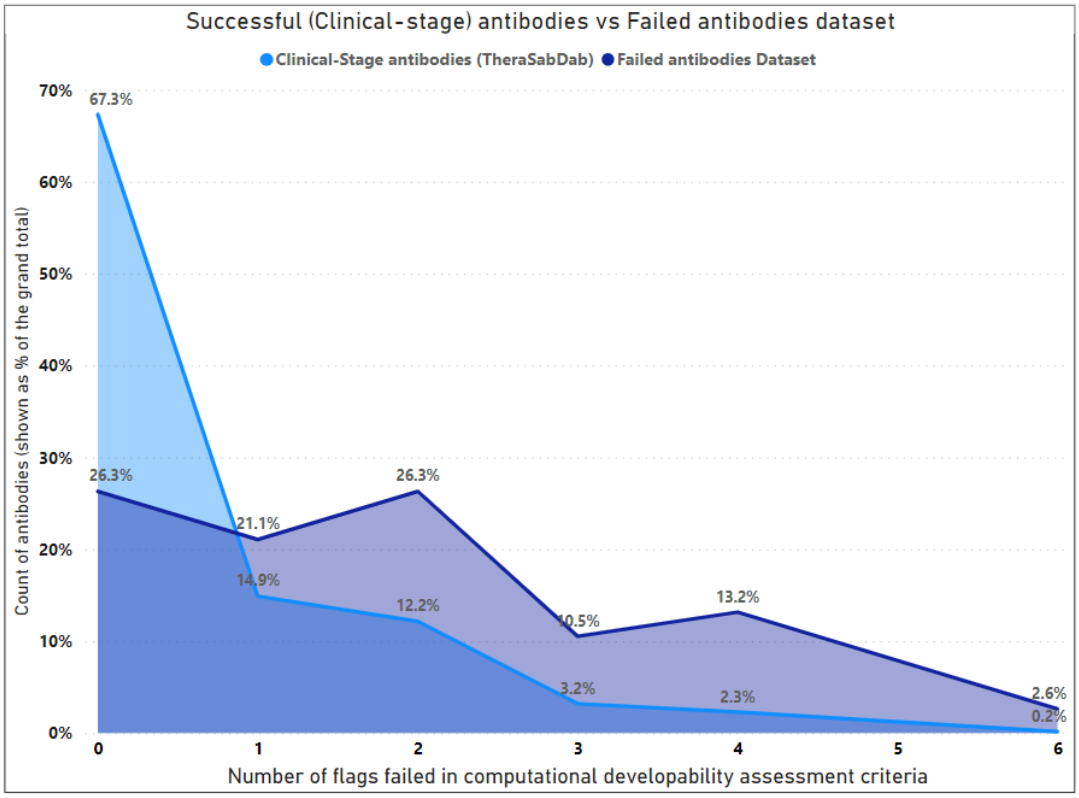


Figure 46: Computational Developability Assessment criteria performance in flagging mAbs for clinical-stage antibodies dataset vs failed antibodies dataset. No. of flags shown on x-axis. 52.6% of the failed antibodies were flagged at least twice by our developability criteria while only 17.8% of clinical-stage antibodies were flagged at least twice by our criteria.

However, for the failed antibodies dataset, we had majority of antibodies being flagged at least once with only 26.3% of the failed antibodies having zero flags. 21.1% of the failed antibodies were flagged by one assay score, 26.3% of the failed antibodies were flagged by two assay scores, 10.5% of the failed mAbs were flagged by three assay scores and 13.2% of the failed antibodies were flagged by four assay scores. Also, a failed antibody Bococizumab had a total of six flags. Therefore, 52.6% of the failed

antibodies were flagged at least twice by our developability criteria while only 17.8% of clinical-stage antibodies were flagged at least twice by our developability criteria. So, overall, we infer that failing two or more flags can be a reliable criterion that has a satisfactory separation and predictability between datasets (% flagged at least twice).

We tested all these mAbs in the failed antibodies dataset against our developability criteria developed based on TheraSabDab clinical-stage antibodies. These criteria are detailed previously in Table 6. We observe that our criteria successfully flagged failed antibodies with a high accuracy as 28 mAbs out of 38 mAbs in the dataset had at least one assay score above the 10% threshold value. Figure 47 below shows an example of the successful flagging of failed antibodies that lie beyond the 10% threshold value (SGAC < 234.11) or even the 5% threshold (SGAC < 78.01) for the SGAC assay.

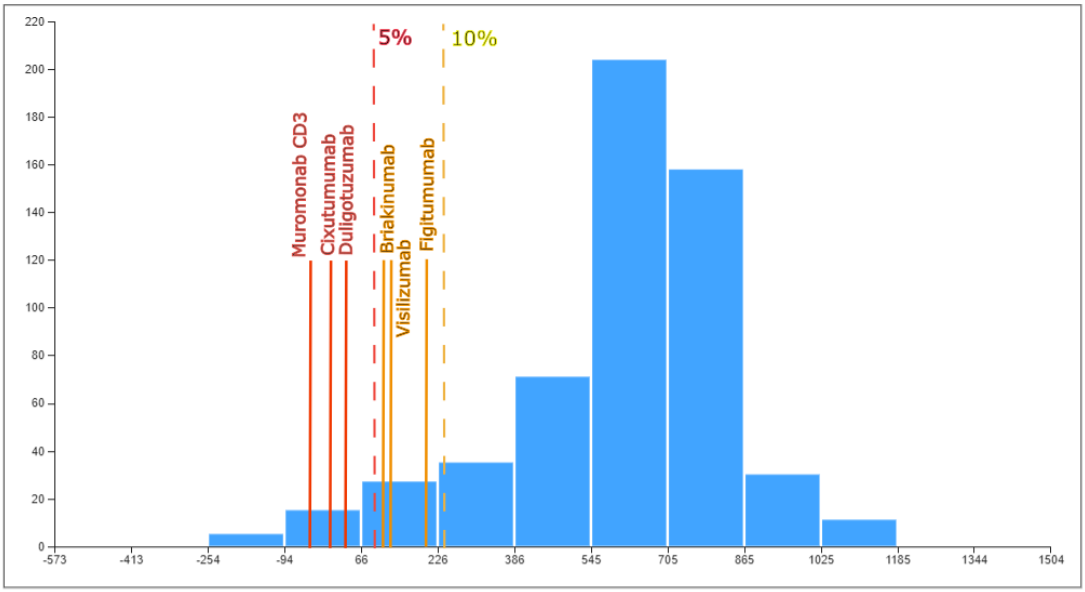


Figure 47: Histogram distribution of SGAC assay score for TheraSabDab clinical-stage antibodies and visualization of the SGAC score for discontinued and withdrawn antibodies.

Our criteria were successful in predicting clinical trial attrition by flagging the outlier scores of failed antibodies in developability assay metrics. For example, Bococizumab a humanized antibody developed by Pfizer for the treatment of high LDL cholesterol levels was terminated from Phase III trials in November 2016.[23] It is an inhibitor of the proprotein convertase subtilisin/kexin type 9 (PCSK9) target. PCSK9 inhibitors work by blocking the protein that degrades LDL receptors on the liver that remove the LDL cholesterol from the blood. The clinical trials reported a higher degree of immunogenicity and a greater incidence of injection-site reactions with bococizumab

compared with other drugs in the class. These two key trials, known as SPIRE-1 and SPIRE-2, had enrolled over 32,000 patients but unfortunately witnessed major adverse cardiovascular events.

Overall, this failure can be attributed to off-target binding to other regions apart from PCSK9 leading to immunogenicity and injection-site reactions. Our developability criteria have successfully flagged the corresponding poor biophysical performances for Bococizumab in six assays with the worst performances in the PSR, ELISA, and BVP assays. Bococizumab had a score of 0.5033 in the PSR assay which is well above the 10% threshold criteria (PSR > 0.4204) and 5% threshold criteria (PSR > 0.4396).

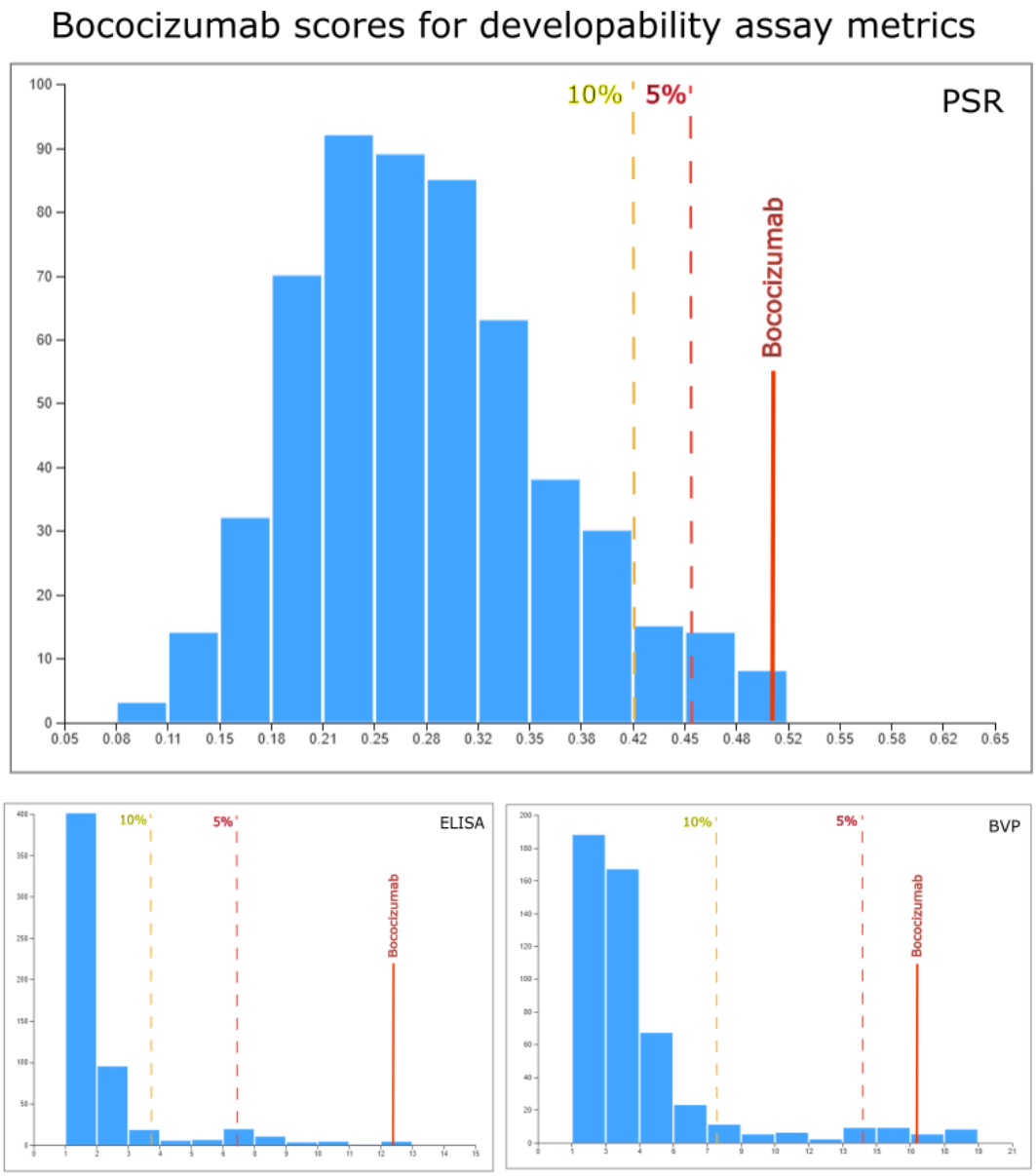


Figure 48: Assay scores in PSR, ELISA, and BVP for Bococizumab compared to the 10% and 5% threshold cutoffs from the developability criteria derived from clinical-stage antibodies.

A similar outlier was observed for Bococizumab for the ELISA assay. This failed antibody had a very high ELISA assay score of 12.345 which exceeded the 10% threshold (ELISA > 3.8382) and 5% threshold values (ELISA > 6.8259). Also, Bococizumab had an extremely high value of 16.249 for the BVP assay, much higher than the 10% threshold (BVP > 7.4012) and 5% threshold (BVP > 14.1383). A very high value in the PSR assay, BVP assay, and ELISA assay is a very clear signal of multiantigen nonspecificity since these assays involve measuring the binding profile of a target antibody to a panel of antigens. So, our developability criteria successfully predicted the off-target binding for this failed antibody as shown in Figure 48.

We have proved the successful flagging of other liabilities beyond polyspecificity as well. For example, Duvortuxizumab, a bispecific antibody by Macrogenics and J&J was terminated in the Phase 1 trial due to toxicity concerns.[24] Duvortuxizumab, was a humanized CD19 x CD3 Dual-Affinity Re-Targeting (DART®) bispecific that was being evaluated in Phase 1, first-in-human, open-label, dose-escalation study for major B-cell cancers including acute lymphoblastic leukaemia, chronic lymphocytic leukaemia, and diffuse large B cell lymphoma. The most common adverse events were infusion-related reactions (80%), fever, chills, pyrexia, constitutional symptoms, and reversible neurological events. The neurotoxicity seen in this Phase 1 clinical trial of duvortuxizumab led to termination of license deal for this asset in B-cell malignancies.

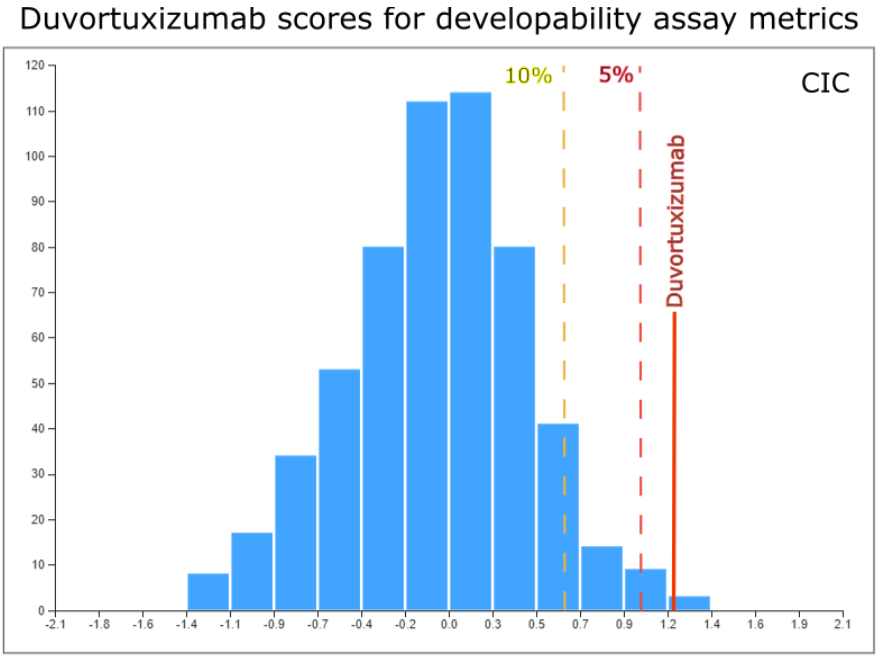


Figure 49: CIC assay score for Duvortuxizumab compared to the 10% and 5% thresholds.

Our developability criteria have successfully flagged the corresponding poor biophysical performances for Duvortuxizumab in CIC, AC SINS, and PSR assays. Duvortuxizumab had a very high value of 1.2124 in the CIC assay which is well above the 10% threshold (CIC > 0.6425) and 5% threshold values (CIC > 1.0554) shown in Figure 49. It suggests high potential cross-interactions with polyclonal human serum antibodies which may explain the adverse events observed in the Phase 1 clinical trials.

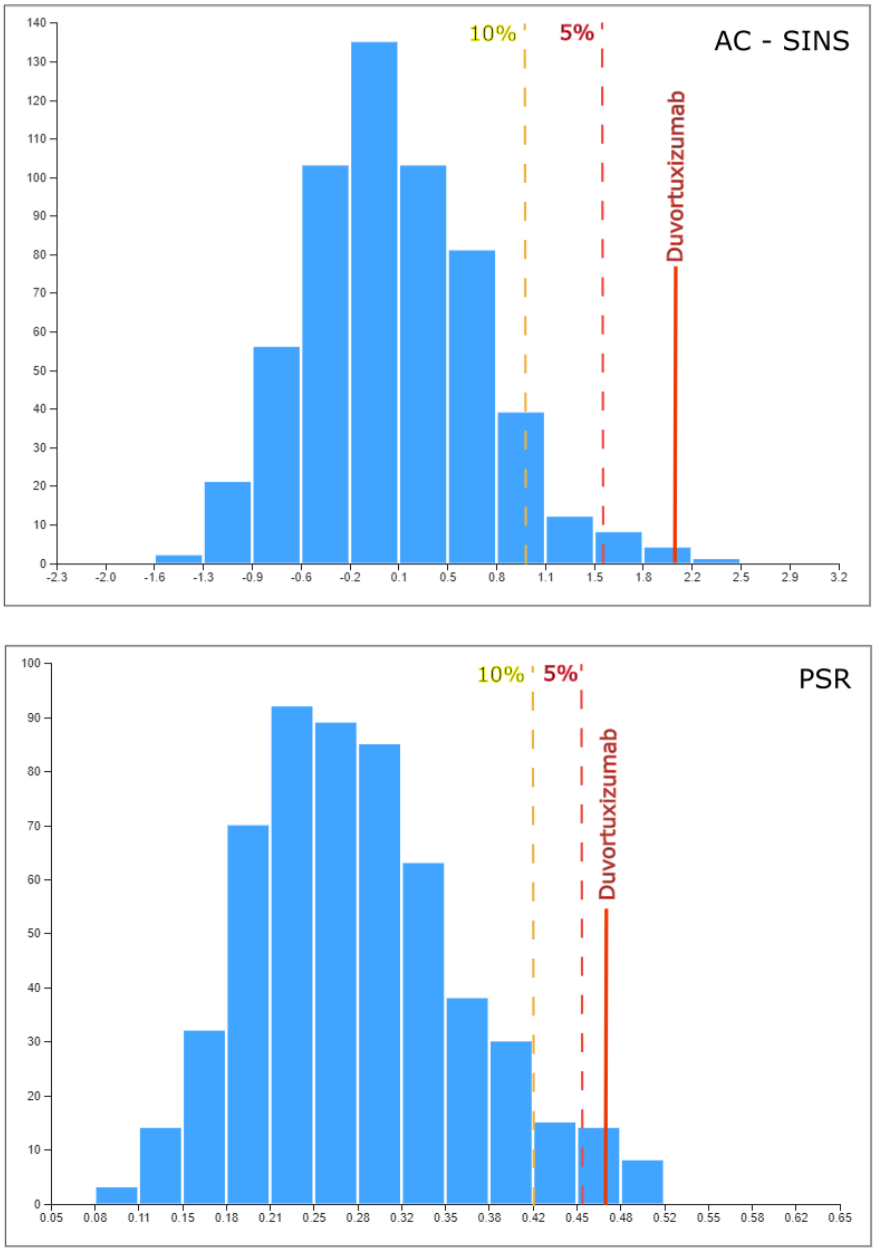


Figure 50: AC-SINS and PSR scores for Duvortuxizumab compared to clinical-stage mAbs.

A similar outlier is seen for the AC-SINS assay where duvortuxizumab had a very high score of 2.0923 which is much higher than the 10% threshold (0.985) and 5% threshold values (1.572). Also, for the PSR assay duvortuxizumab had a score of

0.4660 which exceeded the 10% threshold value (PSR > 0.4204) and 5% threshold value (PSR > 0.4396) shown in Figure 50. These results suggest that Duvortuxizumab exhibits problematic self-interaction and cross-reactivity profile that may have caused the reported adverse events in the Phase 1 clinical trial.

The confusion matrix from a new binary classification algorithm in MATLAB shown in Figure 51 and the corresponding measures shown in Table 17 confirm and validate that our binary classification algorithm derived from the computational developability assessment criteria has been fairly successful in classifying failed antibodies from the clinical-stage antibodies with a high sensitivity (0.5263) and specificity (0.8222). We achieved an overall model accuracy of 80.6% with the binary classification criteria being that the target antibody is flagged at least twice as per the 10% threshold values.

TARGET / OUTPUT	Failed	Clinical-Stage	SUM
Failed	20 2.87%	117 16.81%	137 14.60% 85.40%
Clinical-Stage	18 2.59%	541 77.73%	559 96.78% 3.22%
SUM	38 52.63% 47.37%	658 82.22% 17.78%	561 / 696 80.60% 19.40%

Figure 51: Confusion matrix of the binary classification algorithm from our computational developability assessment criteria for combined failed and clinical-stage antibodies dataset.

In Figure 51, we have 20 out of the 38 failed antibodies that are correctly classified as failed (True Positives: TP); 18 out of 38 failed antibodies that are wrongly classified as clinical-stage (False Negatives: FN); 117 out of 658 antibodies that are wrongly classified as failed (False Positives: FP); and 541 out of 658 antibodies are correctly

classified as clinical-stage (True Negatives: TN). Here, green boxes represent the true or correct predictions, while the red boxes represent the false or wrong predictions made by our binary classification algorithm. Also, the overall sensitivity, specificity, precision, negative predictive value and accuracy values are shown in Figure 51.

Overall, we infer that our developability criteria are partially successful in flagging antibodies that are expected to face clinical trial attrition. A target therapeutic mAb with assay scores exceeding multiple threshold criteria is likely to fail due to adverse events or high immunogenicity caused by the underlying developability liabilities. However, these criteria are not suitable to estimate a clinical trial progression stage.

Measure	Value	Derivations
Sensitivity	0.5263	TPR = TP / (TP + FN)
Specificity	0.8222	SPC = TN / (FP + TN)
Precision	0.1460	PPV = TP / (TP + FP)
Negative Predictive Value	0.9678	NPV = TN / (TN + FN)
False Positive Rate	0.1778	FPR = FP / (FP + TN)
False Discovery Rate	0.8540	FDR = FP / (FP + TP)
False Negative Rate	0.4737	FNR = FN / (FN + TP)
Accuracy	0.8060	ACC = (TP + TN) / (P + N)
F1 Score	0.2286	F1 = 2TP / (2TP + FP + FN)

Table 17: Major binary classification measures for our developability assessment criteria.

5.7 Conclusion

Clinical trials are expensive long-term projects and any insights and predictions from the available biopharmaceutical informatics resources are highly valuable. Efficient management and strategic solutions to clinical trial failures can be a good way to raise the R&D productivity and inspire new therapeutic innovation. In this chapter, we have employed data science and machine learning classification approaches to predict the clinical trial progression stages of antibody therapeutics. An important focus has been on validating the accuracy and applicability of our new computational developability assessment criteria in discriminating the ‘good’ (clinical-stage) antibodies from the ‘bad’ (withdrawn or discontinued antibodies) antibodies.

We started by exploring over 50+ features extracted from the tools used in previous work namely – ProteinSol, AbPred, and Therapeutic Antibody Profiler (TAP). We

found overlap among different stages of clinical-stage antibodies for all these scores. This work has demonstrated that none of these 50+ biophysical features could predict the clinical trial stage outcome for a target antibody. Our analysis then explored new feature engineering techniques and machine learning classification approaches in MATLAB to test clinical trial progression. We observed that coarse KNN, linear discriminant, SVM and ensemble methods had the best model performance among all machine learning algorithms while the neural network algorithms had low accuracy due to high model complexity. The model performance metrics like AUC-ROC curves and the corresponding confusion matrices concluded that our classification models have no discrimination capacity to distinguish between different clinical trial stages.

The results suggest that other additional factors such as location, clinical trial design, therapeutic area, and primary outcome measure also impact the clinical trial progression of therapeutic antibodies. Other external elements beyond developability associated with biology, mechanism, risk vs benefit profile of a biotherapeutic drug, patients, and indication can also significantly affect mAb clinical trial progression.

We analyzed several additional biophysical features from new tools to find other biophysical properties that are most relevant in deciding the clinical trial success. We concluded that only T20 Humanness Score can be potentially used to estimate the clinical trial progression of antibody therapeutics as we observed enrichment in the human regions (higher T20 humanness score) for mAbs in the advanced stages.

Finally, we have checked if our developability criteria can flag mAbs that have failed clinical trials by introducing a new dataset of failed antibodies that were withdrawn or discontinued. We concluded that our criteria were partially successful in flagging failed antibodies with a high model accuracy (80.6%). Evidence from our case studies on multiantigen nonspecificity in Bococizumab and then non-optimal interactions in Duvortuxizumab further support the satisfactory predictability of our developability criteria in flagging the failed mAbs in at least two assays. Our detailed insights and proposed developability assessment criteria can guide antibody discovery or screening by successfully flagging antibodies that are likely to fail while informing any future approaches in characterizing the clinical trial progression of therapeutic antibodies.

5.8 **References**

1. Lynch C, Grewal I. Preclinical safety evaluation of monoclonal antibodies. Therapeutic Antibodies. 2008:19-44.
2. Hebditch M, Warwicker J. Charge and hydrophobicity are key features in sequence-trained machine learning models for predicting the biophysical properties of clinical-stage antibodies. PeerJ. 2019;7:e8199. doi:10.7717/peerj.8199.
3. Wang M-D. Applications of probability of study success in clinical drug development. Applied Statistics in Biomedicine and Clinical Trials Design: Selected Papers from 2013 ICSA/ISBS Joint Statistical Meetings; 2015.
4. Andrade C. The primary outcome measure and its importance in clinical trials. The Journal of clinical psychiatry. 2015;76(10):15598.
5. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. Biostatistics. 2019;20(2):273-86.
6. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. Contemporary clinical trials communications. 2018;11:156-64.
7. Sormanni P, Aprile FA, Vendruscolo M. The CamSol method of rational design of protein mutants with enhanced solubility. Journal of molecular biology. 2015;427(2):478-90. doi:10.1016/j.jmb.2014.09.026.
8. Paladin L, Piovesan D, Tosatto SC. SODA: prediction of protein solubility from disorder and aggregation propensity. Nucleic acids research. 2017;45(W1):W236-W40.
9. Hou Q, Kwasigroch JM, Rooman M, Pucci F. SOLart: a structure-based method to predict protein solubility and aggregation. Bioinformatics. 2020;36(5):1445-52. doi:10.1093/bioinformatics/btz773.
10. Magnan CN, Zeller M, Kayala MA, Vigil A, Randall A, Felgner PL, Baldi P. High-throughput prediction of protein antigenicity using protein microarray data. Bioinformatics. 2010;26(23):2936-43.
11. Sweredoski MJ, Baldi P. COBEpro: a novel system for predicting continuous B-cell epitopes. Protein Engineering, Design & Selection. 2009;22(3):113-20.
12. Van Durme J, De Baets G, Van Der Kant R, Ramakers M, Ganesan A, Wilkinson H, Gallardo R, Rousseau F, Schymkowitz J. Solubis: a webserver to reduce protein aggregation through mutation. Protein Engineering, Design and Selection. 2016;29(8):285-9. doi:10.1093/protein/gzw019.
13. Thangakani AM, Kumar S, Nagarajan R, Velmurugan D, Gromiha MM. GAP: towards almost 100 percent prediction for β-strand-mediated aggregating peptides with distinct morphologies. Bioinformatics. 2014;30(14):1983-90.
14. Conchillo-Solé O, de Groot NS, Avilés FX, Vendrell J, Daura X, Ventura S. AGGRESCAN: a server for the prediction and evaluation of" hot spots" of aggregation in polypeptides. BMC bioinformatics. 2007;8:1-17.
15. Haste Andersen P, Nielsen M, Lund O. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Science. 2006;15(11):2558-67.
16. Ponomarenko J, Bui H-H, Li W, Fusseder N, Bourne PE, Sette A, Peters B. ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC bioinformatics. 2008;9:1-8.
17. Sankar K, Krystek Jr SR, Carl SM, Day T, Maier JK. AggScore: Prediction of aggregation-prone regions in proteins based on the distribution of surface patches. Proteins: Structure, Function, and Bioinformatics. 2018;86(11):1147-56. doi:10.1002/prot.25594.
18. Walsh I, Seno F, Tosatto SC, Trovato A. PASTA 2.0: an improved server for protein aggregation prediction. Nucleic acids research. 2014;42(W1):W301-W7.

19. Wang D, Liu D, Yuchi J, He F, Jiang Y, Cai S, Li J, Xu D. MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. Nucleic acids research. 2020;48(W1):W140-W6. doi:10.1093/nar/gkaa275.
20. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic acids research. 2020;48(W1):W449-W54. doi:10.1093/nar/gkaa379.
21. Gao SH, Huang K, Tu H, Adler AS. Monoclonal antibody humanness score and its applications. BMC biotechnology. 2013;13(1):1-12. doi:10.1186/1472-6750-13-55.
22. Lynch CM, Hart BW, Grewal IS. Practical considerations for nonclinical safety evaluation of therapeutic monoclonal antibodies. MAbs; 2009.
23. Yokote K, Kanada S, Matsuoka O, Sekino H, Imai K, Tabira J, Matsuoka N, Chaudhuri S, Teramoto T. Efficacy and Safety of Bococizumab (RN316/PF-04950615), a Monoclonal Antibody Against Proprotein Convertase Subtilisin/Kexin Type 9, in Hypercholesterolemic Japanese Subjects Receiving a Stable Dose of Atorvastatin or Treatment-Naive—Results From a Randomized, Placebo-Controlled, Dose-Ranging Study—. Circulation Journal. 2017;81(10):1496-505.
24. Izhak L, Cullen DE, Elgawly M, Luistro L, Johnson S, Bald J, Sasser AK, Balasubramanian S. Potent antitumor activity of duvortuxizumab, a CD19 x CD3 DART® molecule, in lymphoma models. Cancer Research. 2017;77(13_Supplement):3636-.

CHAPTER 6

6 Concluding Remarks and Future Work

In this chapter, we summarize the key findings and conclusions from our work. We also provide an outlook towards potential future work to address the additional gaps that were not explored in this project. Finally, we comment on what general trends are expected to rise in biopharmaceutical informatics for the development of biologics.

6.1 Summary and Conclusion

In this work, we have applied computational methods to understand, model, and predict the developability characteristics of therapeutic monoclonal antibodies. We did this by performing computational developability assessment on several antibody datasets, and by studying the biophysical features of these datasets. The model focused on developing a better understanding of the sequence or structure-derived features that determine the antibody developability profile and biophysical performance. Our final developability assessment criteria were based on AbPred score cutoffs on twelve standard biophysical assays and five TAP metrics. We have deeply explored how the developability profile can be predicted and modelled using *in silico* tools to accelerate antibody therapeutic development and predict any sequence or structural liabilities.

In the **introduction** to this thesis, we highlight the scope, background, and challenges faced in application of biopharmaceutical informatics approaches for computational developability assessment. The current molecular modelling and simulation tools used to study antibody therapeutics were discussed. Much of the work conducted for this thesis continued the investigation into the sequence-structural context of stability. We have captured the previous work done by our and other academic groups in the next section. The concept of developability to de-risk antibody development in early stages was introduced with subsequent discussion on the role of computational methods to aid this developability assessment. In particular, we have documented typical features and *in silico* descriptors used in the available developability assessment tools. Our research hypothesis and objectives are then stated to provide context to the readers. The main research hypothesis is that a target antibody with assay scores exceeding multiple developability criteria thresholds is likely to fail due to adverse events caused by the underlying developability liabilities. Our objective is to propose and validate new computational developability assessment criteria derived from clinical-stage antibodies to estimate the clinical trial success or failure of antibody therapeutics.

Continuing this work, in **Chapter 1**, we performed a comprehensive literature review of the current status and utility of biopharmaceutical informatics databases and tools for developability assessment. Table 1 and Table 2 have tabulated all major resources in biopharmaceutical informatics relevant to antibody-based drugs. Such metadata can be a valuable reference for current and future projects in this field. We found that the current tools are not solely reliable for estimating the overall developability profiles. Therefore, we concluded that an orthogonal combination of conceptually different tools and algorithms should be used in the developability assessment protocols to reduce method-specific biases. In particular, we have proposed a new combinatorial triage approach in computational developability assessment workflow to combine scores and rankings from multiple tools together in this chapter.

In **Chapter 2**, we have provided full details of the methods and procedures used in this work. It provides the methodology of the dataset curation, antibody informatics tools and data processing used throughout this project. In particular, this chapter provides in depth information about the machine learning algorithms used in AbPred tool for estimating the biophysical performance in twelve assays. This chapter also has the information on MATLAB codes and procedures used for the machine learning classification work for estimating the clinical trial progression of antibodies.

In **Chapter 3**, we have analyzed the biophysical property distributions to design the developability criteria for clinical-stage antibodies. We concluded that the biophysical property distributions are asymmetrically long-tailed in the unfavourable direction for clinical-stage mAbs. Furthermore, developability criteria were derived from the worst 5% and 10% cut-off values in the histogram distributions for each assay score that were tabulated in Table 6. These criteria for TheraSAbDab clinical-stage antibodies were consistent with the experimental reality of the Jain clinical dataset which was statistically validated by the Kolmogorov-Smirnov test.

We found that our established developability criteria were successfully flagging therapeutic antibodies that caused serious adverse events or failure in clinical trials such as Teclistamab (Tecvayli), Otilimab (GSK3196165), and Brolucizumab (Beovu) while good examples like Trastuzumab (Herceptin) were within threshold limits. Next, we compared the human immune repertoire dataset to clinical-stage therapeutics. We found that natural human antibodies have better performance in binding and reactivity

assays such as CIC, CSI BLI, AC SINS, and PSR but lower performance in assays measuring hydrophobicity and long-term stability such as HIC and SMAC. So, from this analysis, we concluded that future engineering towards deploying natural human antibodies as therapeutics should be on optimizing the hydrophobicity and stability. A case study on True Human™ antibody therapeutics validated that naturally occurring human antibodies can be successfully engineered as commercial antibody therapeutics that display optimal biophysical properties.

In **Chapter 4**, we performed computational developability assessment on engineered antibodies to guide the selection of platform technologies for creating next-generation biotherapeutics. We found that the DVD-Ig™ platform has the lowest non-specific interactions and cross-reactivity among all bispecific engineered platforms. This might be because scFv-based constructs have constraints imparted by the linker sequences and a tendency to form aggregates due to domain exchange of the variable regions. A favourable developability profile was predicted for an Azymetric™ antibody case study. The CAT phage display platform had the best overall developability profile among all categories of phage display platforms. However, we also concluded that future CAT technology phage display optimization and engineering approaches should be directed towards increasing the degree of expression to improve the ease of manufacturability based on HEK assay results. Also, we found that the Therapeutic Antibody Profiler (TAP) is not accurate for phage display antibody datasets.

The developability assessment results also suggested that the VelocImmune® mouse technology antibodies have the best developability profile among the transgenic mice platforms owing to their lowest polyspecificity and non-specific interactions in PSR, CIC, and SGAC assays. This can be attributed to normal maturation and Fc-mediated effector functions retained in the VelocImmune® edits by design.

Chapter 5 presented machine learning classification approaches to predict the clinical trial progression of antibody therapeutics. We found overlap among different stages of clinical-stage antibodies for TAP, Protein-Sol, and AbPred scores. None of these 50+ features used in our work could predict the clinical trial stage outcome for a target antibody. We conclude that other additional factors such as location, clinical trial design, therapeutic area, primary outcome measure beyond developability associated with biology, mechanism of action, risk vs benefit profile of a biotherapeutic drug,

and patient indication can also significantly affect mAb clinical trial progression. We tested several additional biophysical features to find that only T20 Humanness Score can be potentially used to estimate the clinical trial stage progression as we observed enrichment in the human regions represented by a higher T20 score for phase 3 or approved antibodies compared to phase 1 antibodies.

In summary, sincere efforts have been made in this thesis to contribute to the field of biopharmaceutical informatics which can be divided into three general directions: (i) Computational developability assessment that estimates clinical success or attrition, (ii) Antibody informatics to compare discovery platforms and candidate libraries, and (iii) Validation of tools to predict biophysical performance and other assay scores.

6.2 Limitations of computational developability assessments:

The computational developability assessment approaches reported in this work have several limitations. Firstly, the number of antibody therapeutics available in the market are very few in comparison to the small molecule drug products. Also, there are very few publicly available experimental datasets on these marketed antibody therapeutics. Sequences with information gaps and missing residues are usually discarded, which significantly diminishes available data. Therefore, there is an inherent dataset problem in using biopharmaceutical informatics for antibody therapeutics. We can tackle these limitations by using novel predictive tools to fill the incomplete dataset information. For instance, Olsen *et al*. have demonstrated the use of the AbLang tool to restore missing residues in antibody sequences.[1] It is a language model trained on the antibody sequences in the OAS database to restore residues lost due to sequencing errors.

Next, generalization of data outside of those used for training remains a challenge for many predictive models. Biopharmaceutical informatics tools usually employ a set of labeled training data. The model learns to make predictions based on this training data and the model's parameters are tuned until the model predictions and known outcomes align. However, if a model has been trained too well on training data, it will be unable to generalize for new data. We suggest using data augmentation techniques to prevent overfitting and teach the model to make accurate predictions for out-of-sample data.

Finally, current antibody-based biotherapeutics come in several different molecular formats (IgGs, Fabs, ScFvs, and Fvs), formulations (lyophilized powders, liquid), and different routes of administration (intravenous, subcutaneous, intramuscular). Our

work does not consider these different characteristics specific to individual antibody biotherapeutic product classes for each assay. It is possible that a high-concentration liquid mAb formulation suitable for subcutaneous administration differs from low-concentration mAb formulations suitable for intravenous administration in its overall developability profile. However, our results in this work show that dividing the dataset into such classes does not significantly change the average values of these descriptors.

Jain *et al.* have recently performed a detailed review of the ability of *in silico* and *in vitro* flagging rules to identify the clinical progression of antibodies.[2] They have demonstrated problems in the reproducibility of assessments due to differences in homology modelling, complex *in vitro* assessments, and curation of experimental data. For instance, the HIC prediction model on a set of 152 mAbs with a claimed R^2 of 0.6, had an unsatisfactory R^2 of 0.21 for the predictions on a distinct subset of 64 clinical mAbs in this study. The potential concerns on assay reproducibility can possibly be addressed by the inclusion of multiple controls with known sequences that span a range of measurement values and scenarios.

It is also important to note that biotherapeutic products span a broad spectrum of disease indications, molecular targets, patient populations, mechanisms of action, and sequence—structural characteristics. Therefore, data analysis studies involving them are inherently subjective. Our study has attempted to mitigate this subjectivity by focusing on variable regions of antibody therapeutics and also using manually refined comprehensive datasets that have information from multiple data sources. Despite the limitations discussed above, our work has important implications for devising rational biopharmaceutical informatics approaches toward biologics.

6.3 **Contribution to scientific knowledge**

This thesis work done over the past four years is evidence of original contribution that adds to existing scientific knowledge. The key contributions are summarized below:

1. In this thesis work we have proposed completely original new developability criteria thresholds for performance in biophysical assays. Our proposed developability criteria is successfully estimating clinical trial attrition or success with a good model accuracy (80.6%), high sensitivity (52.6%) and specificity (82.2%) that can accelerate antibody drug development and help filter antibody candidates saving valuable time and resources. This validation is also demonstrated with several real-world case studies.

2. Our work informs the scientific community about the unique developability profile of the human immune repertoire. Due to the huge size of the natural human immune repertoire, experimental characterization is not possible, but our work is the first study ever to provide biophysical assay estimates for the natural human antibodies.

3. Insights into the developability profiles of several antibody platforms from Chapter 4 will be a huge help in the experimental design and planning. This chapter guides the reader towards selection of antibody platform technology for their desired application.

4. Next, in terms of published work, the literature review article published in mAbs journal has been having a good traction with 24 citations and over 12,000 views since the formal publication in January 2022. This is a direct quantitative evidence that this research impacted scientific community positively.

5. Also, this thesis work has undertaken the largest ever machine learning study for benchmarking developability from clinical-stage antibodies. All previous scientific knowledge on clinical-stage antibodies were never scrutinized from a machine learning perspective in a large study. In Chapter 5, we have used biophysical estimates of clinical-stage antibodies to train machine-learning classification algorithms for estimating clinical trial progression that has generated key insights into the role of key biophysical features in clinical trial outcomes.

6. Finally, this a comprehensive thesis work that proposes key databases, tools and new computational developability assessment framework which serve as a reference resource for academic researchers and industrial teams that are commercializing antibody therapeutics and undertaking computational assessment projects. Our work would serve as an excellent reference resource for obtaining refined and curated datasets that would facilitate further scientific research. For instance, the manually created failed antibodies dataset would serve as negative control dataset for future computational studies evaluating biophysical performance and clinical trial success.

6.4 **Future work**

The task of establishing developability guidelines is becoming more approachable with more sequence information and biophysical data becoming publicly available. We expect computational developability assessment will play an increasingly larger role early in the antibody development process. However, using *in silico* models may

come with inherent challenges such as skewed, incomplete, or non-representative datasets, inaccurate or sub-optimal informatics tools, and finally data availability. Therefore, it will be important to re-evaluate algorithms, build and curate new datasets.

Another major hurdle in the implementation of biopharmaceutical informatics is the lack of comprehensive and reproducible experimental data obtained under reliable conditions and protocols. The ongoing innovation in biophysical characterization and digital transformation of the biopharmaceutical industry are expected to ameliorate this limitation in the near future. An industry-wide initiative and a collaborative consortium between leading institutions to archive and disseminate verified datasets can be another great step forward in computational developability assessment goals.

The structural aspects of CDR and Fv regions such as the three-dimensional locations of charged and hydrophobic residues and other valid structural patterns such as the aggregation-prone regions, immune epitopes, and antigen-binding interface properties can be valuable inputs for antibody informatics tools. The currently used tools have none or a very limited use of such structural features which may sometimes contain much more valuable insights beyond what can be captured by sequence information.

Ahmed *et al.* have previously demonstrated the use of new five structural features for lead identification and optimization of biotherapeutic candidates in their work.[3] The five structural features introduced in this study are namely - Variable Domain Interface Stability (BSA V_L: V_H); Structure-Based pI of Fv Region (pI Fv3D); Ratio of Dipole Moment to Hydrophobic Moment (RM); Ratio of Surface Areas of Charged Patches to Hydrophobic Patches (RP) and Average Hydrophobic Imbalance (Avg HI). These five physicochemical descriptors capture several biophysical properties ranging from conformational stability to anisotropy of hydrophobic residues and were derived from homology-based models of the Fv regions of 77 marketed antibody therapeutics. So, we suggest future work in biopharmaceutical informatics towards incorporating these novel structural features and insights into prediction algorithms for better precision.

Future work after this thesis should attempt to perform experimental confirmation of the proposed conclusions and insights from our computational developability results. For instance, we have suggested in Chapter 4 that the scFv-based constructs have a tendency to form aggregates due to domain exchange of the V regions. It would be good to validate this result experimentally with hydrogen-deuterium exchange mass

spectrometry (HDX-MS) experiments in the future. Furthermore, other experimental techniques like X-ray crystallography for epitope mapping of antibody-antigen interactions and surface plasmon resonance (SPR) assay to study the binding kinetics of antibodies can be valuable to confirm our conclusions in this thesis.[4-6]

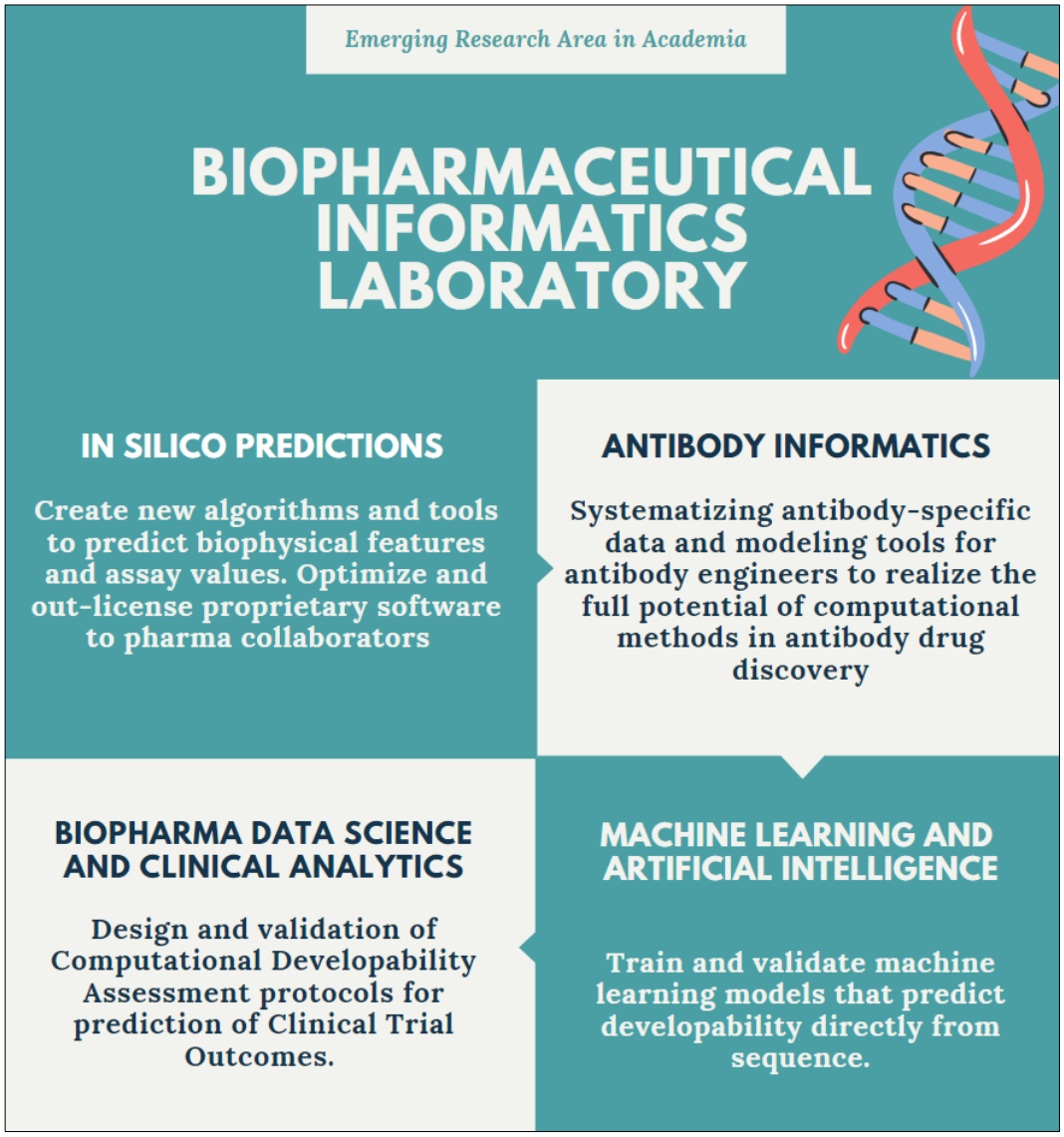


Figure 52: Research themes for Biopharmaceutical Informatics lab in academic institutions.

Also, we hope to openly publish the software and developability criteria for use by the scientific community. We aim to set up a dedicated GitHub Resource for collaboration and knowledge sharing on biopharmaceutical informatics. This can serve as a platform to share comprehensive resources like experimental datasets, tools, and algorithms for use in computational developability assessments. We also hope to introduce new tools focused on the developability score at the Protein-sol website. Our team is currently working on adding structural-based calculations to the webserver in AbPred 2.0. We

intend to keep updating our resources with the recent data and further the use of our web server and GitHub by introducing new features and calculations over time. Finally, in future we propose to set up the Biopharmaceutical Informatics Labs at the partner academic institutions. This is an inevitable transition to address the expanding research in this newly emerging field. The potential major research themes in a Biopharmaceutical Informatics lab in academic institutions are shown in Figure 52.

In summary, the work outlined in this thesis has aimed to investigate the scope and role of biopharmaceutical informatics approaches for computational developability assessment of antibody therapeutics. Using a combination of biophysical property distribution analysis for clinical-stage antibodies, developability comparisons of novel engineered antibody formats, and machine learning classification algorithms, we have studied how computational resources can be used for developability profiling of input antibody therapeutics. Our results suggest that it may be possible to predict some key sequence or structural liabilities in antibody candidates under consideration and also triage appropriate antibody discovery platform or format for the intended application.

Accurate estimation of clinical trial progression is a complicated process determined by additional factors beyond available *in silico* descriptors. As a result, there are many physicochemical properties still to study, and a comprehensive investigation of all of these features is beyond the scope of a single doctoral thesis. We hope, however, that the investigation outlined within this work can contribute to a greater understanding of the scope of antibody informatics resources toward developability assessments.

6.5 References

1. Olsen TH, Moal IH, Deane CM. AbLang: an antibody language model for completing antibody sequences. Bioinformatics Advances. 2022;2(1):vbac046.
2. Jain T, Boland T, Vásquez M. Identifying developability risks for clinical progression of antibodies using high-throughput in vitro and in silico approaches. Mabs; 2023.
3. Ahmed L, Gupta P, Martin KP, Scheer JM, Nixon AE, Kumar S. Intrinsic physicochemical profile of marketed antibody-based biotherapeutics. Proceedings of the National Academy of Sciences. 2021;118(37):e2020577118.
4. Toride King M, Brooks CL. Epitope mapping of antibody-antigen interactions with X-ray crystallography. Epitope Mapping Protocols. 2018:13-27.
5. Adamczyk M, Mattingly PG, Shreder K, Yu Z. Surface plasmon resonance (SPR) as a tool for antibody conjugate analysis. Bioconjugate chemistry. 1999;10(6):1032-7.
6. Malmqvist M. Surface plasmon resonance for detection and measurement of antibody-antigen affinity and kinetics. Current opinion in immunology. 1993;5(2):282-6.

APPENDIX

7 Supplementary Information

7.1 Approved therapeutic monoclonal antibodies in the market

S. No	International Non-Proprietary Name	Brand Name	Therapeutic Area	Clinical Indication	First approval (country, year)
1	Abciximab	Reopro	Hematological Disorders	Prevention of blood clots in angioplasty	US, 1994
2	Adalimumab	Humira	Immunology	Rheumatoid arthritis	US, 2002
3	Ado-trastuzumab	Kadcyla	Oncology	Breast cancer	US, 2012
4	Aducanumab	ADUHELM	Neurological disorders	Alzheimer's disease	US, 2021
5	Alemtuzumab	Lemtrada	Oncology	Chronic myeloid leukemia	US, 2001
6	Alirocumab	Praluent	Other	High cholesterol	US, 2015
7	Amivantamab	RYBREVANT	Oncology	NSCLC w/ EGFR exon 20 mutations	US, 2021
8	Amubarvimab	(Pending)	Infectious disease	SARS-CoV-2 infection	China, 2021
9	Anifrolumab	Saphnelo	Immunology	Systemic lupus erythematosus	US, 2021
10	Ansuvimab	Ebanga	Infectious Diseases	Ebola virus infection	US, 2020
11	Atezolizumab	Tecentriq	Oncology	Bladder cancer	US, 2016
12	Atoltivimab	Inmazeb	Infectious Diseases	Ebola virus infection	US, 2020
13	Avelumab	Bavencio	Oncology	Merkel cell carcinoma	US, 2017
14	Basiliximab	Simulect	Immunology	kidney transplant rejection	US, 1998
15	Belantamab mafodotin	BLENREP	Oncology	Multiple myeloma	US, 2020
16	Belimumab	Benlysta	Immunology	Systemic lupus	US, 2011
17	Benralizumab	Fasenra	Immunology	Asthma	US, 2017
18	Bevacizumab	Avastin	Oncology	Colorectal cancer	US, 2004
19	Bezlotoxumab	Zinplava	Infectious Diseases	Clostridium difficile infection	US, 2016
20	Bimekizumab	Bimzelx	Immunology	Psoriasis	EU, 2021
21	Blinatumomab	Blincyto	Oncology	Acute lymphoblastic leukemia	US, 2014
22	Brentuximab vedotin	Adcetris	Oncology	Hodgkin lymphoma	US, 2011
23	Brodalumab	Siliq	Immunology	Plaque psoriasis	Japan, 2016
24	Brolucizumab	Beovu	Ophthalmology	Age-related macular degeneration	US, 2019
25	Burosumab	Crysvita	Genetic Diseases	X-linked hypophosphatemia	EU, 2018
26	Cadonilimab	(Pending)	Oncology	Cervical cancer	China, 2022
27	Camrelizumab	AiRuiKa	Oncology	Hodgkin's lymphoma	China, 2019
28	Canakinumab	Ilaris	Genetic Diseases	Muckle-Wells syndrome	US, 2009
29	Caplacizumab	Cablivi	Immunology	Thrombocytopenic purpura	EU, 2018
30	Casirivimab	REGEN-COV	Infectious	COVID-19	Japan, 2021
31	Catumaxomab	Removab	Oncology	Malignant ascites	EU, 2009
32	Cemiplimab	Libtayo	Oncology	Cutaneous cell carcinoma	US, 2018
33	Certolizumab pegol	Cimzia	Immunology	Crohn disease	US, 2008
34	Cetuximab	Erbitux	Oncology	Colorectal cancer	EU, 2004
35	Cetuximab saratolacan	Akalux®	Oncology	Head and neck cancer	Japan, 2020
36	Crizanlizumab	Adakveo	Hematological	Sickle cell disease	US, 2019
37	Daclizumab	Zinbryta	Immunology	Multiple sclerosis	US, 1997
38	Daratumumab	Darzalex	Oncology	Multiple myeloma	US, 2015

39	Denosumab	Prolia	Musculoskeletal Disorders	Bone Loss	EU, 2010
40	Dinutuximab	Unituxin	Oncology	Neuroblastoma	US, 2015
41	Disitamab vedotin	Aidixi	Oncology	Gastric cancer	China, 2021
42	Docaravimab	Twinrab	Infectious disease	Rabies exposure	India, 2019
43	Donanemab	(Pending)	Neurological disorders	Alzheimer's disease	In review
44	Dostarlimab	Jemerli	Oncology	Endometrial cancer	EU, 2021
45	Dupilumab	Dupixent	Immunology	Atopic dermatitis	US, 2017
46	Durvalumab	IMFINZI	Oncology	Bladder cancer	US, 2017
47	Eculizumab	Soliris	Hematological Disorders	Paroxysmal nocturnal hemoglobinuria	US, 2007
48	Edrecolomab	Panorex	Oncology	Colon cancer	EU, 1995
49	Efalizumab	Raptiva	Immunology	Psoriasis	US, 2003
50	Elotuzumab	Empliciti	Oncology	Multiple myeloma	US, 2015
51	Emapalumab	Gamifant	Oncology	Hemophagocytic lymphohistiocytosis	US, 2018
52	Emicizumab	Hemlibra	Hematological Disorders	Hemophilia A	US, 2017
53	Enfortumab vedotin	Padcev	Oncology	Urothelial cancer	US, 2019
54	Envafolimab	ENWEIDA	Oncology	Advanced solid tumors	China, 2021
55	Eptinezumab	VYEPTI	Neurological	Migraine prevention	US, 2020
56	Erenumab	Aimovig	Neurological	Migraine prevention	US, 2018
57	Evinacumab	Evkeeza	Genetic Diseases	Homozygous familial hypercholesterolemia	US, 2021
58	Evolocumab	Repatha	Other	High cholesterol	EU, 2015
59	Faricimab	Vabysmo	Ophthalmology	AMD	US, 2022
60	Fremanezumab	Ajovy	Neurological	Migraine prevention	US, 2018
61	Galcanezumab	Emgality	Neurological	Migraine prevention	US, 2018
62	Gemtuzumab	Mylotarg	Oncology	Acute myeloid leukemia	US, 2000
63	Geptanolimab	(Pending)	Oncology	Peripheral T cell lymphoma	In review
64	Glofitamab	(Pending)	Oncology	Diffuse large B-cell lymphoma	In review
65	Golimumab	Simponi	Immunology	Rheumatoid arthritis	US, 2009
66	Guselkumab	TREMFYA	Immunology	Plaque psoriasis	US, 2017
67	Ibalizumab	Trogarzo	Infectious Diseases	HIV infection	US, 2018
68	Ibritumomab tiuxetan	Zevalin	Oncology	Non-Hodgkin lymphoma	US, 2002
69	Idarucizumab	Praxbind	Hematological Disorders	Dabigatran-induced anticoagulation	US, 2015
70	Inebilizumab	Uplizna	Immunology	Neuromyelitis optica spectrum disorders	US, 2020
71	Inetetamab	Cipterbin	Oncology	HER2-positive metastatic breast cancer	China, 2020
72	Infliximab	Remicade	Immunology	Crohn disease	US, 1998
73	Inolimomb	(Pending)	Immunology	Graft vs. host disease	In review
74	Inotuzumab	BESPONSA	Oncology	Acute lymphoblastic leukemia	US, 2017
75	Ipilimumab	Yervoy	Oncology	Metastatic melanoma	US, 2011
76	Isatuximab	Sarclisa	Oncology	Multiple myeloma	US, 2020
77	Itolizumab	Alzumab	Immunology	Psoriasis	India, 2013
78	Ixekizumab	Taltz	Immunology	Psoriasis	US, 2016
79	Lanadelumab	Takhzyro	Genetic Diseases	Hereditary angioedema	US, 2018
80	Lecanemab	(Pending)	Neurological	Alzheimer's disease	In review
81	Levilimab	Ilsira	Immunology	Inflammation	Russia, 2020

82	Loncastuximab	Zynlonta	Oncology	Diffuse B-cell lymphoma	US, 2021
83	Margetuximab	MARGENZA	Oncology	Breast cancer	US, 2020
84	Mepolizumab	Nucala	Immunology	Severe eosinophilic asthma	US, 2015
85	Mirikizumab	(Pending)	Immunology	Ulcerative colitis	In review
86	Mirvetuximab	(Pending)	Oncology	Ovarian cancer	In review
87	Mogamulizumab	Poteligeo	Oncology	Mycosis fungoides or Sézary syndrome	Japan, 2012
88	Mosunetuzumab	Lunsumio	Oncology	Follicular lymphoma	EU, 2022
89	Moxetumomab	Lumoxiti	Oncology	Hairy cell leukemia	US, 2018
90	Muromonab-CD3	Orthoclone Okt3	Immunology	Reversal of kidney transplant rejection	US, 1986
91	Narsoplimab	(Pending)	Hematological Disorders	Thrombotic microangiopathies	In review
92	Natalizumab	Tysabri	Immunology	Multiple sclerosis	US, 2004
93	Naxitamab	DANYELZA	Oncology	Refractory osteomedullary disease	US, 2020
94	Nebacumab	Centoxin	Infectious Diseases	Gram-negative sepsis	EU, 1991
95	Necitumumab	Portrazza	Oncology	Non-small cell lung cancer	US, 2015
96	Nemolizumab	Mitchga.	Immunology	Pruritus with atopic dermatitis	Japan, 2022
97	Netakimab	Efleira	Immunology	Plaque psoriasis	Russia, 2019
98	Nimotuzumab	TheraCIM	Oncology	Head and neck cancer	Cuba, 2002
99	Nirsevimab	(Pending)	Infectious diseases	Respiratory syncytial virus infection	In review
100	Nivolumab	Opdivo	Oncology	Melanoma, non-small cell lung cancer	US, 2014
101	Obiltoxaximab	Anthim	Infectious Diseases	Prevention of inhalational anthrax	US, 2016
102	Obinutuzumab	Gazyva	Oncology	Chronic lymphocytic leukemia	US, 2013
103	Ocrelizumab	OCREVUS	Immunology	Multiple sclerosis	US, 2017
104	Ofatumumab	Arzerra	Oncology	Chronic lymphocytic leukemia	US, 2009
105	Olaratumab	Lartruvo	Oncology	Soft tissue sarcoma	US, 2016
106	Olokizumab	ARTLEGIA	Immunology	Rheumatoid arthritis	Russia, 2020
107	Omalizumab	Xolair	Immunology	Asthma	US, 2003
108	Omburtamab	(Pending)	Oncology	CNS metastasis	In review
109	Oportuzumab	(Pending)	Oncology	Bladder cancer	In review
110	Ormutivimab	(Pending)	Infectious disease	Post-exposure prophylaxis of rabies	China, 2022
111	Ozoralizumab	(Pending)	Immunology	Rheumatoid arthritis	In review
112	Pabinafusp alfa	IZCARGO	Metabolic disorders	Mucopolysaccharidosis type II	Japan, 2021
113	Palivizumab	Synagis	Infectious Diseases	Respiratory syncytial virus infection	US, 1998
114	Panitumumab	Vectibix	Oncology	Colorectal cancer	US, 2006
115	Pembrolizumab	Keytruda	Oncology	Melanoma	US, 2014
116	Penpulimab	(Pending)	Oncology	Metastatic nasopharyngeal carcinoma	In review
117	Pertuzumab	Perjeta	Oncology	Breast Cancer	US, 2012
118	Polatuzumab vedotin	Polivy	Oncology	Diffuse large B-cell lymphoma	US, 2019
119	Prolgolimab	Forteca	Oncology	Melanoma	Russia, 2020
120	Racotumomab	Vaxira®	Oncology	Non-small cell lung cancer	Cuba, 2013
121	Ramucirumab	Cyramza	Oncology	Gastric cancer	US, 2014
122	Ranibizumab	Lucentis	Ophthalmology	Macular degeneration	US, 2006
123	Ravulizumab	Ultomiris	Hematological Disorders	Paroxysmal nocturnal hemoglobinuria	US, 2018

124	Raxibacumab	(Pending)	Infectious	Anthrax infection	US, 2012
125	Regdanvimab	Regkirona	Infectious	COVID-19	Korea, 2021
126	Relatlimab	Opdualag	Oncology	Melanoma	US, 2022
127	Reslizumab	Cinqaero	Immunology	Asthma	US, 2016
128	Retifanlimab	(Pending)	Oncology	Carcinoma of the anal canal	In review
129	Ripertamab	(Pending)	Oncology	Non-Hodgkin's lymphoma	In review
130	Risankizumab	Skyrizi	Immunology	Plaque psoriasis	Japan, 2019
131	Rituximab	Rituxan	Oncology	Non-Hodgkin lymphoma	US, 1997
132	Romosozumab	Evenity	Musculoskeletal	Osteoporosis	Japan, 2019
133	Sacituzumab govitecan	TRODELVY	Oncology	Triple-neg. breast cancer	US, 2020
134	Sarilumab	Kevzara	Immunology	Rheumatoid arthritis	Canada, 2017
135	Satralizumab	Enspryng	Immunology	Neuromyelitis optica spectrum disorder	Canada, 2020
136	Secukinumab	Cosentyx	Immunology	Psoriasis	Japan, 2014
137	Serplulimab	(Pending)	Oncology	Solid tumors	China, 2022
138	Siltuximab	Sylvant	Immunology	Castleman disease	US, 2014
139	Sintilimab	Tyvyt	Oncology	Non-small cell lung cancer	In review
140	Socazolimab	(Pending)	Oncology	Cervical cancer	In review
141	Sotrovimab	Xevudy	Infectious diseases	COVID-19	Australia, 2021
142	Spesolimab	(Pending)	Immunology	Generalized psoriasis	In review
143	Sugemalimab	Cejemly®	Oncology	Non-small cell lung cancer	China, 2021
144	Sutimlimab	Enjaymo	Immunology	Cold agglutinin disease	US, 2022
145	Tafasitamab	Monjuvi	Oncology	Diffuse B-cell lymphoma	US, 2020
146	Tebentafusp	KIMMTRAK	Oncology	Metastatic uveal melanoma	US, 2022
147	Teclistamab	(Pending)	Oncology	Multiple myeloma	In review
148	Teplizumab	(Pending)	Immunology	Type 1 diabetes	In review
149	Teprotumumab	Tepezza	Ophthalmology	Thyroid eye disease	US, 2020
150	Tezepelumab	Tezspire	Immunology	Severe asthma	US, 2021
151	Tildrakizumab	Ilumya	Immunology	Plaque psoriasis	US, 2018
152	Tislelizumab	(Pending)	Oncology	Squamous cell carcinoma	In review
153	Tisotumab	TIVDAK	Oncology	Cervical cancer	US, 2021
154	Tixagevimab	Evusheld	Infectious	COVID-19	EU, 2022
155	Tocilizumab	RoActemra	Immunology	Rheumatoid arthritis	Japan, 2005
156	Toripalimab	Tuoyi	Oncology	Nasopharyngeal carcinoma	In review
157	Tositumomab-I131	Bexxar	Oncology	Non-Hodgkin Lymphoma	US, 2003
158	Tralokinumab	Adtralza	Immunology	Atopic dermatitis	EU, 2021
159	Trastuzumab	Herceptin	Oncology	Breast cancer	US, 1998
160	Trastuzumab deruxtecan	Enhertu	Oncology	Breast cancer	US, 2019
161	Trastuzumab duocarmazine	(Pending)	Oncology	Breast cancer	In review
162	Tremelimumab	(Pending)	Oncology	Antineoplastic; liver cancer	In review
163	Ublituximab	(Pending)	Immunology	Multiple sclerosis	In review
164	Ustekinumab	Stelara	Immunology	Psoriasis	EU, 2009
165	Vedolizumab	Entyvio	Immunology	Ulcerative colitis	US, 2014
166	Zimberelimab	(Pending)	Oncology	Hodgkin’s lymphoma	China, 2021

Table 18: Therapeutic monoclonal antibodies in approved or review stages (2022). Adapted from https://www.antibodysociety.org/resources/approved-antibodies/ and www.fda.gov/.

7.2 **Full scatterplot matrix for AbPred and ProteinSol features**

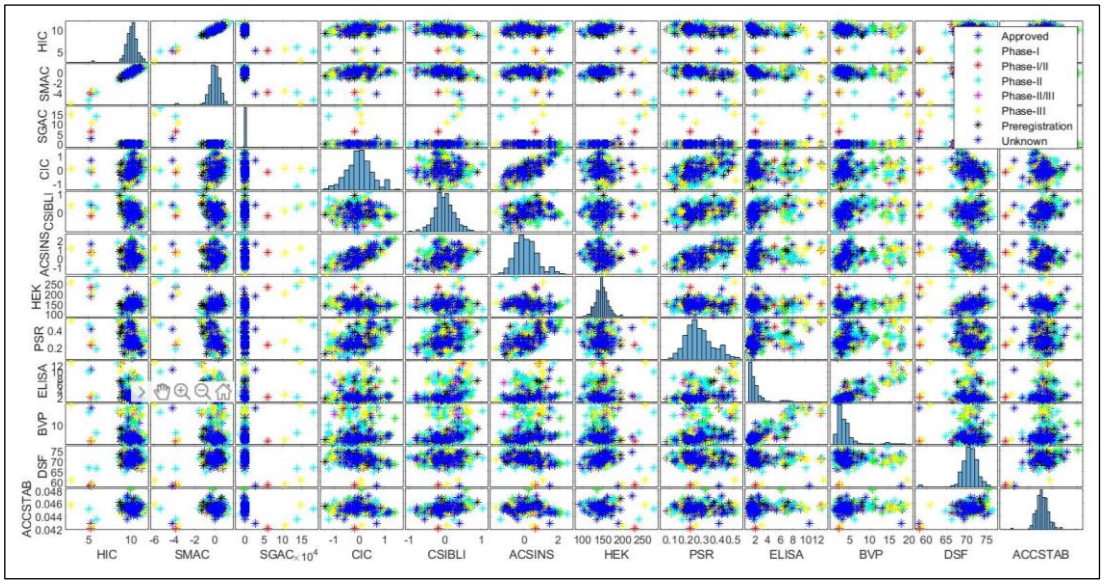


Figure: Scatterplot matrix of 12 AbPred biophysical features as per clinical trial status.

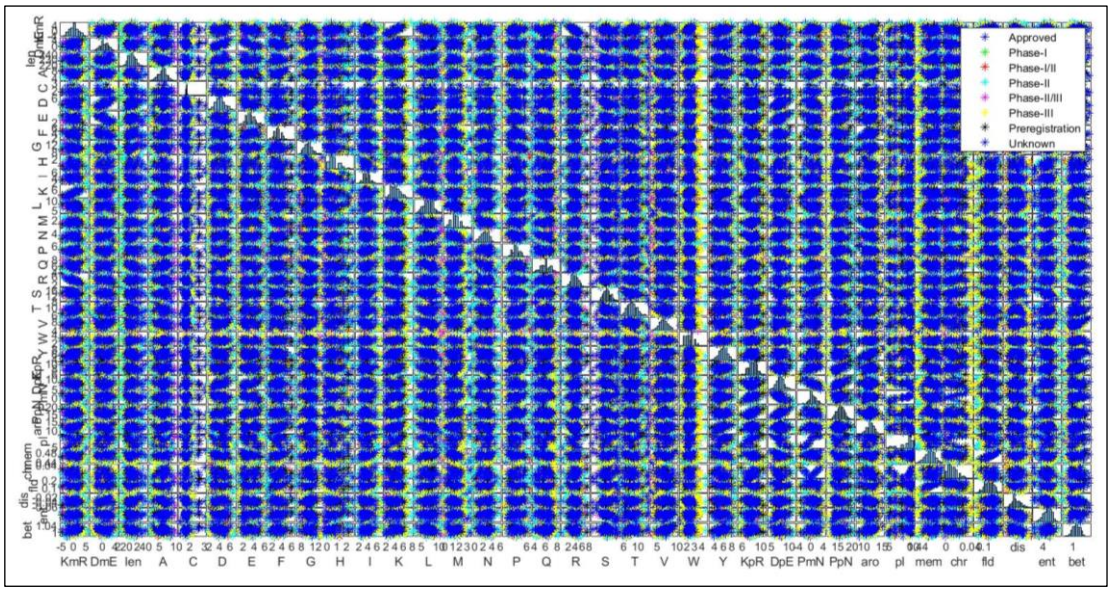


Figure: Scatterplot matrix of 35 ProteinSol sequence features as per clinical trial status.

7.3 Biopharma licensing and Merger and Acquisition (M&A) trends in the 21st-century landscape

This chapter was published in the Journal of Commercial Biotechnology as Khetan, Rahul. "Biopharma licensing and M&A trends in the 21st-century landscape." Journal of Commercial Biotechnology 25.3 (2020). doi: https://doi.org/10.5912/jcb943

7.3.1 Abstract:

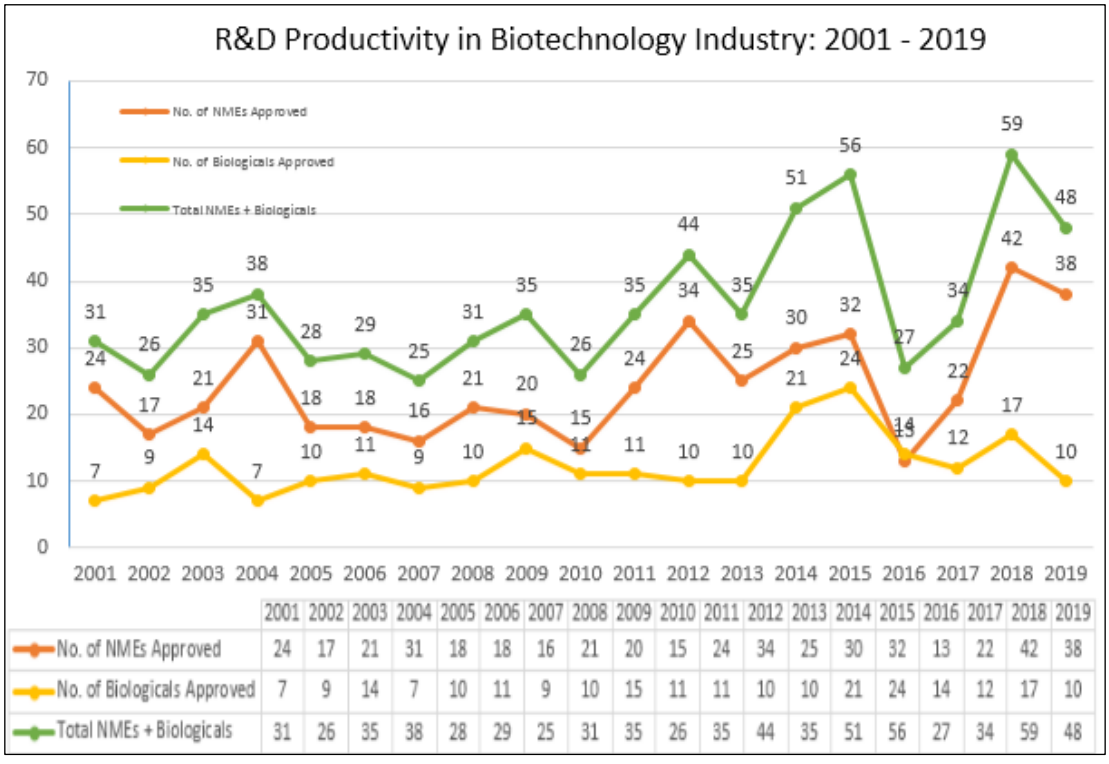
The declining in-house R&D productivity has compelled the biopharmaceutical firms to supplement their innovation pipelines with well-managed licensing or acquisition deals. The reliance on new licensed products and acquired technologies continues to increase across the biotech industry to support pipeline expansion and technological diversification. In this article, we present another application of biopharmaceutical informatics to study licensing and merger and acquisition (M&A) trends across the biopharmaceutical industry by examining the partnering activities of major biopharmaceutical companies and analyzing the trends in the business development transactions of biotechnology companies worldwide. The information has been extracted from various public and proprietary sources such as company annual reports, Industry reports, and press releases from biopharmaceutical companies. The transaction data and partnership information has been analyzed over a contiguous twenty-year time period of 2000–2020 to understand the change in strategic focus in dealmaking and trends over the past two decades.

7.3.2 Introduction:

The global pharmaceutical market is estimated to be worth nearly 1,430 billion USD($) by 2020.[1] In the last decade, the R&D expenditure of biopharmaceutical companies exceeded half a trillion USD resulting in advances and discoveries that will make a huge difference in millions of patients' lives.[2] In this new era of medicine, many diseases previously regarded as deadly are now manageable and potentially curable. Novartis (10.5), Roche (9.1), Pfizer (7.5), Merck & Co. (7.1), J&J (6.7), Sanofi (6.1), AstraZeneca (5.6) and Glaxo-SmithKline (5.4) are expected to invest more than 5 billion USD on R&D in 2020 with an industry-wide forecasted total R&D spend of USD 160 billion by 2020.[3] In the United States, VC investments in the biopharmaceutical industry have doubled between 2010 and 2015 from $3.7 billion to $8.2 billion.[4] Industry analysts predict that 80% of the revenues for

biopharmaceuticals and diagnostics in 2030 will be driven by advances in biologic drugs that were not on the market by 2010.[5]

Pavlou and Belsey have reviewed biopharma licensing and M&A trends in 2005 where they have discussed the reliance of the leading US and European pharma players on licensing and M&A, types of M&A deals in the industry and their contribution to total M&A value.[6] Gautam et al. have also previously reviewed the key trends in R&D portfolio mix, revenue distribution and operational model over the 1995–2015 period that have impacted and transformed the top 12 big-pharma companies.[7] They concluded that the pharmaceutical companies are now adapting their strategic focus towards their areas of strength, consolidating R&D towards hotspots, shifting towards specialty drugs and recognize the emerging markets as major revenue drivers. Third-party collaborations are now an essential part of biopharmaceutical companies' strategy to supplement product pipelines and to maximise revenues using commercial deals.[8] For instance, Boehringer Ingelheim data from 2003 depicted that over two-thirds of sales of three pharmaceutical companies among top 15 were from in-licensed products. This reliance on licensed products and acquired technologies continues to increase across the industry to support expansion and technological diversification.



	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
No. of NMEs Approved	24	17	21	31	18	18	16	21	20	15	24	34	25	30	32	13	22	42	38
No. of Biologicals Approved	7	9	14	7	10	11	9	10	15	11	11	10	10	21	24	14	12	17	10
Total NMEs + Biologicals	31	26	35	38	28	29	25	31	35	26	35	44	35	51	56	27	34	59	48

Figure 53: NMEs and Biologicals approved by the FDA over the last two decades. Source: US Food and Drug Administration (FDA) reports and Evaluate Pharma database search.

7.3.2.1 Overview of Licensing and M&A in the Biotechnology Industry

Most early-stage biotechnology companies lack sufficient funds and experience to sustain their discoveries through complex and expensive clinical testing and subsequent regulatory approval hurdles. Further, these companies don't have the sales and marketing competence needed to bring their approved drugs swiftly into the market. So, they mostly rely on much larger pharmaceutical companies to finance and conduct clinical testing and to market the drugs once they have received regulatory approval. This is achieved sometimes through a license agreement and sometimes through outright acquisition of the biotechnology pioneer by a larger and established pharma company.[9]

Intellectual Property (IP) can be effectively commercialized and exploited through licensing the IP by the owner (the licensor) to another company (the licensee) which would carry out the marketing, distribution, and sales activities. 'Licensing Out' an IP means that the licensor retains the ownership of the IP and allows the licensee to use the IP. On the contrary, 'Licensing In' refers to the act of securing rights to use the IP from the licensor by a licensee. Selling IP to a third party is technically known as an 'assignment' where ownership is transferred to a new party with outright disposal of the IP by the owner. In the biopharmaceutical industry, spin-out companies are actively involved in licensing to commercialize research and innovation. This is usually in the form of an exclusive License or an assignment as any company investors may insist that the company should own the IP rights. However, in some cases where the assignment is not feasible, the spin-out company may instead take a sole or non-exclusive license of the IP. The value, commercialization viability, company strategy and licensing/assignment costs determine the approach adopted in transferring an IP in the biotechnology sector.

The traditional operational model in big pharmaceutical companies has been that of a fully integrated company. Every operation in R&D, sales and marketing were carried out within the company. Big pharma now lays far greater emphasis on external collaborations to procure and develop new medicines and therapies. In 2010, for example, big pharma GlaxoSmithKline's Chief Executive Officer (CEO) Andrew Witty announced further cuts to the company's in-house R&D organisation, focusing the company strategy towards a "more virtual, more partner-orientated" model. Witty's remarks echoed with his counterparts across the industry, all now disinvesting

from the traditional research model in favour of in-licensing drug candidates while outsourcing development work.[10] Also, the current GlaxoSmithKline's CEO Emma Walmsley has also been focused on creating strategic and operational synergies to build broad industry portfolio while maintaining a leading position in therapeutic areas such as HIV, respiratory and pain relief. The leading big pharma companies such as GSK are likely to continue pursuing external collaborations and in-licensing of technologies with their BD&L strategy in the next few decades.

Morgan Stanley economic value analysis report from 2010 also supports divestment of in-house R&D. The report suggests that every $1 invested in licensed drugs will on average deliver three times as much more value as $1 invested in R&D within the company.[11] Major pharma companies have echoed this suggestion and have declared disinvestment in early-stage research in several disease areas in the previous two decades. They are being more reliant on external sources for maintaining their drug innovation pipeline. For instance, in 2010, the same year when Morgan Stanley report was released, AstraZeneca closed down discovery research in 10 therapeutic categories affecting nearly 3500 R&D jobs in the UK, Sweden and the US.

7.3.2.2 Scope of Business Development deals in Biotechnology industry

Business development deals in the biotechnology industry fall into one of two broad categories—asset-based and non-asset-based. The asset-based partnerships include acquisitions and licensing of drugs, technology, and patented innovations whereas the non-asset-based partnerships include joint ventures, consortia, and collaborations where two parties pool resources to achieve a common goal. The types of deals can range from simple patent licensing deals to complex co-development deals, co-promotion deals and commercialisation deals.[12] These collaborative R&D deals (discovery or preclinical-stage) are considered in the licensing section of our study. The motive behind these deals is to develop external collaborations to obtain products to supplement the internal R&D pipeline.

In our study, M&A is defined as outright acquisitions that result in the exit of the target firm. However, it must be considered that outright acquisition is one extreme variant of the range of pharmaceutical-biotech and biotech-biotech relationships, including the purchase of a major equity stake (e.g. Roche-Genentech), co-development

alliances and co-commercialization or marketing agreements. This continuum of activity makes the definition of merger /acquisition somewhat arbitrary.

Danzon et al. have analyzed the scope, determinants and effects of significant M&A transactions over the period 1988–2000 using a multinomial logit model to test several competing hypotheses to explain the M&A activity across the entire pharma–biotechnology industry.[13] They have concluded that pharmaceutical acquisitions of biotechnology companies are consistently driven by an asset-specific motive, such as cross-national acquisitions, assuming that it is a cheaper, quicker and more effective way to buy a local company with established connections rather than building a foreign subsidiary.

The 'valley of death' between drug discovery and its ability to attract formal venture capital has been widening. In particular, the venture capital for early-stage biopharmaceutical companies must compete with alternative low-risk profile opportunities that consistently offer high returns in the near-term. Many bioscience venture capitalists are increasingly focusing their investments in emerging life science companies only once their drug candidates enter clinical trials.[14] Finally, it has been evident from previous studies that solely trusting the valuations from the hired investment banks for due diligence can be misleading. So, all the large companies now have business development & licensing teams to search, evaluate and negotiate deals. The implementation of a business development strategy depends on the availability and ability to identify opportunities and execute them at acceptable costs.

7.3.3 Key licensing trends:

7.3.3.1 Reliance on licensed life-science products and technologies

Post-patent-expiration price competition has become more intense, compelling the pharmaceutical companies to either innovate or indulge in licensing deals to replace innovations in the R&D pipeline. Table 19 provides an overview of the key licensing deals in the biopharmaceutical industry from the 21st century with an overall value of over 500 million USD. We have only included the deals involving a preclinical compound or drugs in advanced clinical trials. So, we have excluded discovery stage collaboration deals or any commercial rights deals for proper representation of the biopharmaceutical dealmaking landscape.

Table 19: Key licensing deals from the 21st century.

S.No	Year	Licensee	Licensor	Value	Description
1.	2017	Merck	AstraZeneca	8500 million USD	Strategic oncology collaboration with MSD to co-develop and co-commercialise AstraZeneca's Lynparza for multiple cancer types.
2.	2019	AstraZeneca	Daiichi Sankyo	6900 million USD	License of HER2-targeted antibody-drug conjugate Trastuzumab deruxtecan for breast cancer
3.	2018	Merck	Eisai	5755 million USD	Strategic collaboration for LENVIMA® (lenvatinib mesylate), an orally available tyrosine kinase inhibitor.
4.	2018	Roche	Affirmed	5096 million USD	License to commercialize novel NK cell engager-based immunotherapeutics to treat multiple cancers.
5.	2019	Gilead Sciences	Galapagos NV	5050 million USD	License of phase 3 candidate for idiopathic pulmonary fibrosis known as GLPG1690
6.	2019	GlaxoSmithKline	Merck KGaA	4200 million USD	License of M7824 (bintrafusp alfa) a bifunctional fusion protein-based cancer immunotherapy for solid tumours
7.	2020	AbbVie	Genmab	3800 million USD	License of bispecific drugs led by CD3xCD20 bispecific antibody epcoritamab
8.	2018	BMS	Nektar Therapeutics	3630 million USD	Worldwide license and collaboration for immuno-oncology program, NKTR-214
9.	2018	Gilead (Kite Pharma)	Sangamo Therapeutics	3160 million USD	Exclusive license for cell therapies using zinc finger technology.
10.	2014	Pfizer	Cellectis	2885 million USD	Partnership to develop Chimeric Antigen Receptor T-cell (CAR-T) cancer immunotherapies
11.	2014	Pfizer	Merck KGaA	2850 million USD	Partnership to co-develop and co-commercialize MSB0010718C, an investigational anti-PD-L1 antibody
12.	2019	Roche (Genentech)	Sarepta Therapeutics	2850 million USD	License of Duchenne muscular dystrophy gene therapy SRP-9001
13.	2018	Allogene	Cellectis	2800 million USD	Exclusive license for UCART Cell therapies
14.	2017	Sanofi	Ablynx	2700 million USD	Exclusive worldwide license of Nanobody®-based therapeutics
15.	2019	Gilead Sciences	Nurix	2350 million USD	License of Nurix Protein degradation technology for multiple therapeutic categories
16.	2015	BMS	uniQure	2307 million USD	Global license and commercialization rights for gene therapies against 10 cardiovascular targets
17.	2014	AstraZeneca	Almirall	2095 million USD	Divestment of Almirall's respiratory assets including the marketed drug Eklira plus pipeline candidates.
18.	2015	Sanofi	Regeneron	2000 million USD	License of clinical-stage bispecific antibodies for cancer immunotherapy.

19.	2015	Amgen	Xencor	1745 million USD	License for Xencor's Preclinical CD38 Bispecific T Cell Engager for Multiple Myeloma
20.	2014	BMS	Five Prime Therapeutics	1740 million USD	Partnership to co-commercialize phase I cancer/immunology compound FPA008 and other CSF1R compounds.
21.	2015	Sanofi	Lexicon	1700 million USD	Exclusive license for Sotagliflozin, an oral treatment for Diabetes.
22.	2019	Neurocrine Biosciences	Xenon Pharmaceuticals	1700 million USD	Exclusive license to Nav1.6 sodium channel inhibitor candidate, XEN901 for epilepsy treatment.
23.	2015	Eli Lilly	Innovent Biologics	1456 million USD	Multiple drug development collaborations to enter Chinese oncology market
24.	2012	Johnson & Johnson	Genmab	1100 million USD	Global license and development agreement for daratumumab (HuMax®-CD38), a human CD38 monoclonal antibody
25.	2016	BMS	Nitto Denko	998 million USD	Exclusive worldwide license agreement for siRNA molecules targeting HSP47
26.	2007	Novartis	Antisoma Plc.	990 million USD	License deal of vascular disrupting agent AS1404 a promising oncology drug
27.	2020	Sanofi	Kiadis Pharma	986 million USD	License deal of K-NK004, modified NK cells to prevent the expression of CD38
28.	2016	BMS	PsiOxus Therapeutics	936 million USD	Exclusive worldwide license of NG-348, a Tumour-Specific Immuno-gene Therapy (T-SIGn)
29.	2016	Takeda Pharmaceuticals	Crescendo Biologics	790 million USD	License for discovery, development, and commercialisation of Humabody®-based therapeutics
30.	2017	Sanofi	Principia Biopharma	765 million USD	Exclusive worldwide license of PRN2246
31.	2016	J&J	MacroGenics	740 million USD	Global license to MGD015, a preclinical DART® (dual-affinity re-targeting) molecule for various hematological malignancies and solid tumours
32.	2015	AstraZeneca	Inovio Pharmaceuticals	728 million USD	License agreement for clinical-stage INO-3112 HPV cancer vaccine
33.	2017	BMS	CytomX	723 million USD	Exclusive worldwide license to develop and commercialize Probody therapeutics for eight additional targets.
34.	2017	Biogen	BMS	710 million USD	License of Phase 2 anti-eTau compound for Progressive Supranuclear Palsy.
35.	2007	Sanofi-Aventis	Oxford BioMedica	690 million USD	License of cancer immunotherapeutic TroVax (vaccinia-delivered tumour-associated antigen 5T4)
36.	2007	Schering-Plough	Anacor Pharmaceuticals	625 million USD	License of its phase 2 antifungal ANA2690 retaining the rights to copromote it in the US.

Note: Only Licensing deals with an overall value of over 500 million USD considered. Deals without financial terms have been excluded. Only the deals involving drugs in advanced clinical trials are included in this Table.

7.3.3.2 Objectives and nature of the Licensing deals

Companies often employ a licensing strategy for therapeutic areas of challenging scientific nature such as oncology and infectious diseases to hedge against clinical failure. In fact, the proportion of biopharma revenue generated by in-licensed or acquired compounds rose from 41% in 2005 to 50% in 2014.[15] Generic makers are signing distribution and marketing contracts to reach foreign regulated and developing markets such as the recent out-licensing deals between Pfizer & Aurobindo and GlaxoSmithKline & Dr Reddy's Labs to expand in emerging markets such as India. This trend is expected to increase even further due to a large number of drugs with pending patent expiration in the next few years.

AstraZeneca has been the most prolific pharmaceutical dealmaker in terms of the completed number of deals. AstraZeneca signed a record 169 agreements in total between 2014-2018, 66 of which were out-licensing deals. Such a leading dealmaking rate is also demonstrated when AstraZeneca completed five late-stage deals in 2010, including a 1.24 billion USD deal with Targacept and the 350 million USD acquisition of Novexel, to develop two late-stage antibiotics in partnership with Forest Laboratories. The mean transaction value of licensing deals in 2019 was 455 million USD, a 41% increase from the mean value of 322 million USD in 2018. Also, there was a staggering rise in the value of the mean Upfront payment of the licensing deals, changing 48% from the value of 32.6 million USD in 2018 to 48.3 million USD in 2019.[16] Big Pharma companies signed two-to-three times as many in-licensing agreements as out-licensing deals annually between 2011 and 2015. The majority of the licensing deals throughout the 21st century were in the Discovery and Pre-clinical stages which represent the interest of Licensors in capturing the early-stage assets at a lower price and utilizing the in-house expertise in later developmental stages.

In particular, the early-stage deals that offer access to novel technology platforms and next-generation biologics are very popular amongst large pharmaceutical companies. Last year, in 2019, Gilead was the leading dealmaker with three deals collectively worth almost $10 billion. These three deals are Gilead Sciences-Galapagos, Gilead Sciences-Nurix and Gilead Sciences-Goldfinch Bio. The Gilead Sciences-Galapagos deal for late-stage idiopathic pulmonary fibrosis (IPF) drugs, in particular, has the biggest upfront payment at $3.95 billion. Gilead has emerged actively in the licensing

market with such high-value deals ever since the company recently appointed Gilead's Chief executive Daniel O' Day.

In April 2014, Novartis and GlaxoSmithKline agreed to swap a series of assets where Novartis acquired GlaxoSmithKline's marketed oncology portfolio for $16 billion and sold its vaccines business to GlaxoSmithKline for $7.1 billion, a deal that reshaped two of the world's biggest drugmakers. Similarly, in 2017, Sanofi acquired Boehringer Ingelheim's consumer healthcare (CHC) business in exchange for Sanofi's Animal Health business (Merial). This strategic asset swap was valued at a combined total of 24 billion USD. Such 'exchange deals' seem to be an attractive alternative for business development transactions in biotechnology in upcoming years since it helps the firms focus on their key business areas. The big pharma has started to embrace a focused strategic approach on their key therapeutic areas while divesting non-core assets.

7.3.3.3 Statistics on Licensing deals by sector/focus area

Oncology, and the field of immuno-oncology in particular, has continued to dominate the dealmaking landscape, while some noteworthy early-stage deal activity for novel biological programs across a variety of therapy areas was observed throughout the decade. The rise of immuno-oncology as a therapeutic strategy is reflected in the number of licensing deals in the biotechnology industry such as the global strategic Oncology collaboration between Merck and Eisai for LENVIMA© in 2018. Analysts predict that Oncology will continue to lead all the therapy areas and would represent 17.5% of all prescription/OTC drug sales by 2022, more than the next three highest therapy areas combined.[17]

The largest CAGR growth in the top 15 therapy categories except oncology is predicted to be from immunosuppressants, dermatologicals and anti-coagulants. CNS diseases, infectious diseases, endocrine diseases, and cardiovascular diseases were the next prevalent therapy areas after oncology for dealmaking. Gene therapy has also emerged as a top priority focus area in licensing deals. The global personalized medicine market is forecasted to reach $2.4 trillion with projected sales of $118.15 billion in 2022 at a CAGR of 11.8%, double the projected 5.2% annual growth rate for the overall health care sector. Also, worldwide Medtech sales are forecasted to grow at an annual compound growth rate of 5.1%, reaching US$521.9 billion by 2022

where In-vitro diagnostics is estimated to be the largest Medtech segment with annual sales of more than USD $70 billion by 2022.

Table 20: Therapy Area and Projected sales in 2022 for pharma assets (USD billion)

S.No	Therapeutic focus area	Average Number of Deals per year	No. of R&D products in 2019	Projected sales in 2022
1.	Oncology	1040	2731	192.2 billion USD
2.	Antidiabetics	430	571	57.9 billion USD
3.	Anti-inflammatory	390*	473	55.4 billion USD
4.	Anti-virals	410*	439	42.8 billion USD
5.	Vaccines	440*	364	35.3 billion USD
6.	Bronchodilators	170*	480	30.1 billion USD
7.	Sensory organs	220	459	28.3 billion USD
8.	Immunosuppressants	370	511	26.3 billion USD
9.	Anti-hypertensives	290	412	24.4 billion USD
10.	Anti-coagulants	210	410	23.2 billion USD
11.	Musculoskeletal	200	461	21.7 billion USD
12.	Dermatologicals	250*	200*	19.9 billion USD
13.	Anti-fibrinolytics	230*	210*	17.1 billion USD
14.	Anti-hyperlipidemics	240*	200*	13.4 billion USD
15.	Anti-bacteria	140	270*	12.8 billion USD
	Top 15	5030	8191	601 billion USD
	Total	5800*	9500*	1100 billion USD

*Source: EvaluatePharma, 2017 for projected sales. IQVIA™ Pharma Deals Half-Year Review of 2018/2019 and Author's calculations for Licensing statistics.*represent estimated projections of global deal count. Note: Deals covering more than one therapeutic area are counted more than once (in each relevant therapeutic area).*

7.3.3.4 Challenges to successful execution of Licensing deals

The expected benefits of the licensing transactions may never be fully realized or may even take longer to realize than expected due to 10-15 year development timelines, extensive R&D costs, and high rates of scientific & regulatory uncertainty. Also, competition from possible generic or biosimilar alternatives has to be taken into account. When a drug expires from patent protection, the owner loses some market share through generics. For instance, Pfizer lost the US patent protection for their top-selling drug Lipitor in November 2011 which dwindled Lipitor sales from 5 billion USD per year to only 0.93 million USD the year after the patent expired. So, any unforeseen delay such as the COVID-19 pandemic in 2020 can jeopardize the drug development/clinical programs while the nearing patent expiry date would continue to decrease the revenue generated after product launch.

The COVID-19 pandemic has disrupted several other industries such as in the hospitality sector, however, we have witnessed active dealmaking in the pharmaceutical healthcare sector even during this global pandemic. For instance,

Gilead acquired cancer drugmaker Forty Seven for $4.9 billion in April 2020 adding Forty Seven's investigational lead product Magrolimab to their immuno-oncology portfolio. The statistics have shown that the number of deals has been unchanged but the overall deal values and upfront payment values have declined in the second quarter of 2020. Big pharma companies have been resilient in this crisis by redirecting resources towards developing drugs and vaccines against the SARS-CoV-2 virus.

Licensing deals can often involve extensive clinical development programs across multiple indications which may involve co-development and co-commercialisation roles between the licensor and the licensee. This presents a unique challenge to the licensing dealmaking since adequate involvement of both parties is required for the success of the target product. Furthermore, in recent years, increased access to capital for early-stage companies continues to slow down the licensing activity. The investor sentiment towards biotech companies has been increasingly bullish overall owing to the huge return on investment provided by several blockbuster drugs. Early-stage biotechnology companies now have a variety of funding options available to them to fund their pipeline programs for longer. This allows them to retain the rights to their pipelines in the development phase in the hope of achieving higher returns in the clinical stage. Finally, several intangible liabilities such as lawsuits and binding long-term contracts can hamper licensing or acquisition deals.

7.3.3.5 Licensing litigation activity and trends

The licensing deals are often disrupted by various lawsuits, claims, government investigations and other legal proceedings that arise in the business development transactions. Such legal proceedings can involve various types of parties such as governments, competitors, customers, suppliers, service providers, licensees, employees, or shareholders. These legal disputes usually involve patent infringement, antitrust, securities, pricing, sales and marketing practices, environmental, commercial, contractual rights, licensing obligations, health and safety matters, consumer fraud, employment matters, product liability and insurance coverage. Moreover, failure to enforce the patent rights likely results in substantial decreases in the respective product revenues from generic competition.

Last year in 2019, Sanofi terminated a 1.7 billion USD licensing deal with Lexicon Pharmaceuticals due to unsatisfactory results in Phase III trials of Sotagliflozin.

Lexicon accused Sanofi of 'breach of contract' by ending the partnership. Sanofi has contractual obligations to fund ongoing clinical trials for a specified period following termination as per documents with US SEC. This incident has shown the importance of properly discussing and agreeing upon the restrictions and obligations involved in any licensing deal in case of deal termination. The same year, Amgen and Novartis entered a legal dispute regarding the collaboration agreements of the migraine drug Aimovig (FDA approved drug). Amgen terminated the partnership alleging that Novartis was in breach of the collaboration agreements for entering into a new joint development agreement with Alder BioPharmaceuticals, innovating a potential rival to Aimovig. However, Novartis accused Amgen of attempting to back out of their partnership and keep all the profits from the drug's sales and considered the notice of termination "unjustified and without legal merit".[18] Here the conflict arose due to Novartis' involvement with Alder's development of a similar drug to Aimovig, so the licensing and M&A deals should avoid any overlapping projects to avoid litigations.

Every deal structure carries different tax and accounting implications. M&A, for example, may add tax benefits if the target company bears net operating losses. Also, R&D incentives can be utilized to reduce cash taxes by the acquiring company. However, M&A can negatively impact the financial statements because it requires consolidation of assets, liabilities, and other financial items of two or more entities into one. Proper due diligence should unearth any issues which can cause legal disputes after the completion of any deal. Furthermore, risk management through the use of IP insurance can be very helpful for firms involved in licensing deals.

7.3.3.6 Strategies for successful licensing deals

Initial stage assignment deals usually provide the least return in the longer term, as IP tends to become more valuable as it is developed further and commercialized. Besides, the valuation of early-stage innovations is very challenging which increases the risk involved in selling an IP at an undervalue or overpaying for an economically unrewarding IP. However, licensing has shown a satisfactory track record for early-stage patented innovations. Licensing out of an IP minimizes the capital investment and maximizes the return on IP for the owner by creating an additional income stream while retaining the ownership. Out-licensing has also emerged as a viable option to offload non-core assets and share development risks.

An established firm already has its own marketing team, salesforce, distribution channels and a well-respected brand and reputation. These will enable it to access the market for an IP product very effectively; by contrast, commercialising the IP through a start-up company will require the IP owner to create his access to the marketplace from scratch. In such situations, an optimal licensing deal is a win-win for both parties. Financial rewards from successful licensing are usually not immediate but can build up to respectable levels over the years. Therefore, a business development executive can seek to reduce the cash at risk by using deal structures that make payment contingent upon hitting specific milestones.

Another important strategy is to negotiate licenses with the technology transfer offices within the research institutions. This would enable companies to invest in potential early-stage inventions at a much lower capital investment which can be financially very rewarding. So, for advancing innovative biopharmaceutical solutions, academic collaborations should be an integral part of the business development strategy. For instance, Merck & Co. have launched academic partnerships with universities and academic institutes, such as the California Institute for Biomedical Research (Calibr), to accelerate the commercialization of academic research.

Current practices in due diligence are varied across the biotechnology industry. For example, Bristol-Myers Squibb ranking of potential licensing deals is based on three measures: Probability of technical and regulatory success, Expected NPV and Risk-adjusted internal rate of return calculated for each asset. Novo Nordisk and Celgene were ranked in top with the best positive partnering attributes in BCG survey of Biotechnology CEOs and Licensing Executives, 2012 reflecting the trend towards inclination of big-pharma to partner with these two companies. Also, the survey indicated that GSK, Merck, and Roche were the preferred buy-side companies for licensors. Thus, the partnering characteristics of a company also influence the business development deals in the biotechnology industry.

7.3.4 Key M&A trends:

7.3.4.1 Reliance on M&A for Biotechnology and Pharma companies

Big pharmaceutical companies use M&A deals to access strategically important intellectual property (IP), enter new therapeutic areas and fill R&D pipeline gaps in the company. Major pharmaceutical companies have broadened their R&D portfolio

by accessing research projects and drug candidates from mergers and acquisitions of external sources. M&A activity in the pharmaceutical–biotechnology industry during the last decade of the 20th century (1988-2000) had exceeded 500 billion USD. Whereas, the aggregate value of all M&A deals in 2010-2020 has exceeded 1200 billion USD.[19] Table 21 below summarizes the key M&A deals from the 21st century.

Table 21: Key M&A deals from the 21st century.

S.No	Year	Parent Company	Target Company	Value (billion $)	Description
1	2000	Pfizer	Warner-Lambert	111.8	Pfizer acquired Warner-Lambert and gained product lines such as Parke-Davis branded pharmaceuticals
2	2000	Glaxo Wellcome Plc	SmithKline Beecham	76.0	Merger of two UK-based drugmakers to form the new company known as GlaxoSmithKline
3	2019	Bristol-Myers Squibb	Celgene	74.0	Definitive merger agreement expected to achieve $2.5 Billion Run-Rate Cost Synergies by 2022
4	2004	Sanofi	Aventis	73.5	Birth of Sanofi-Aventis by merger of France's largest drugmaker.
5	2015	Actavis	Allergan	70.5	The merger provided dominant position in segments like Eyecare, Neurosciences, Dermatology, Gastroenterology and Urology
6	2009	Pfizer	Wyeth	68.0	Merger for Diversification of the in-line and pipeline patent-protected portfolio of biopharmaceuticals
7	2003	Pfizer	Pharmacia	64.3	Pfizer-Pharmacia merger was fueled by the Arthritis drugs Celebrex and Bextra, which were expected to have $3.75 billion in sales per annum
8	2018	Takeda Pharmaceutical	Shire	62.0	Acquisition focused on four therapeutic areas - Oncology, Neuroscience, Rare diseases, and Plasma-derived therapies
9	2016	Bayer	Monsanto	54.5	Acquisition to create the world's biggest agro-chemical and seed company
10	2010	Novartis	Alcon	52.5	Novartis expands reach in eye-care business by buying Alcon shares from Nestlé
11	2009	Merck & Co.	Schering-Plough	47.1	A reverse merger to obtain market rights for Infliximab (Remicade)
12	2009	Roche	Genentech	46.8	Consolidated 1990 acquisition of Genentech
13	2014	Medtronic	Covidien	42.3	Mergers of two giants in medical device community – Spinal Implants, Heart devices and Insulin pumps
14	2015	Teva Pharmaceutical Industries	Actavis	40.5	Increased scale and pricing power in the generics market was the deal driver.

15	2016	Shire	Baxalta	32.0	Merger focused on rare disease products – HAE, Endocrine diseases, and lysosomal storage diseases.
16	2017	Johnson & Johnson	Actelion	30.0	Four focused therapeutic areas: Cardiovascular disorders, CNS disorders, Immunological disorders, and Orphan diseases.
17	2006	Boston Scientific Abbott Laboratories	Guidant	27.2	Merger for medical devices portfolio especially cardiovascular devices
18	2000	Pharmacia & Upjohn	Monsanto	25.2	The company retained Monsanto's pharmaceutical division (Searle) and spun off the remaining interests
19	2017	Abbott Laboratories	St Jude Medical	25.0	Merger to create a diverse portfolio of devices, diagnostics, Nutritionals and branded generic pharmaceuticals
20	2015	AbbVie	Pharmacyclics	21.0	Focus on Imbruvica® (ibrutinib), a Bruton's tyrosine kinase inhibitor approved for the treatment of certain B-cell malignancies
21	2014	Actavis	Forest Laboratories	20.7	Merger to Strengthen Actavis' Specialty Brands Business
22	2011	Sanofi	Genzyme Corporation	20.1	France's pharmaceutical company Sanofi acquisition of Genzyme is symbolic of the Pharma 'shift' into Biotechnology
23	2012	Johnson & Johnson	Synthes	19.7	Synthes integrated with DePuy franchise to establish the DePuy Synthes Companies of Johnson & Johnson.
24	2006	Bayer	Schering	18.4	Created Bayer-Schering Pharmaceuticals headquartered in Berlin
25	2016	Quintiles	IMS Health	17.6	Created IQVIA, one of the world's largest contract research organizations
26	2015	Pfizer	Hospira	17.0	Expanded business in Injectable drugs, Biosimilars and Infusion technologies market
27	2015	Merck Group	Sigma-Aldrich	17.0	New entity to enhance product range, capabilities, and geographic reach
28	2001	Amgen	Immunex	16.8	Immunex's key product Enbrel, a rheumatoid arthritis drug was a key driver
29	2006	Johnson & Johnson	Pfizer Consumer Health	16.6	All-cash transaction which provided a boost to global personal care and OTC medicines business
30	2014	Novartis	GlaxoSmithKline Oncology	16.0	Newly-acquired therapies such as Tafinlar®, Votrient® and Promacta®
31	2015	Valeant	Salix Pharmaceuticals	15.8	Created a new speciality platform for growth in U.S. Gastrointestinal Market
32	2007	AstraZeneca	MedImmune	15.6	Acquisition of U.S. biotechnology company MedImmune to expand

					towards vaccines and biologicals
33	2007	Schering Plough	Organon International	14.5	The acquisition added five drugs in Phase III development
34	2014	Bayer	Merck & Co Consumer Health	14.2	The acquisition significantly enhanced Bayer's OTC business across multiple therapeutic categories and geographies
35	2016	Pfizer	Medivation	14.0	Acquired promising late-stage oncology pipeline to accelerate position in Oncology
36	2015	Zimmer Inc.	Biomet Inc.	13.4	Created the Zimmer Biomet Holdings, a leader in musculoskeletal healthcare market
37	2019	Amgen	Otezla (drug program)	13.4	Otezla® (apremilast) acquired from Celgene in connection with its merger with Bristol-Myers Squibb
38	2006	Merck Group	Serono	13.2	Merck's Pharma Ethicals division combined with Serono to create Merck-Serono Biopharmaceuticals
39	2018	GlaxoSmithKline	Novartis Consumer Healthcare	13.0	Buyout of Novartis' 36.5% stake in the Consumer Healthcare Joint Venture
40	2017	Gilead Sciences	Kite Pharma	11.9	Acquisition aimed to position Gilead as a Leader in Cell Therapy
41	2018	Sanofi	Bioverativ	11.6	Creating a Leading Hemophilia Portfolio by acquiring therapies in rare blood disorders.
42	2019	Pfizer	Array BioPharma	11.4	Acquisition to bolster cancer treatment portfolios
43	2011	Gilead Sciences	Pharmasset	11.2	Acquisition directed towards the promising Hepatitis C treatment portfolio
44	2013	Amgen	Onyx Pharmaceuticals	10.4	Oncology Portfolio and Pipeline such as Multiple Myeloma drug Kyprolis® were the key deal driver
45	2019	Novartis	The Medicines Company	9.7	Inclisiran was the target drug to expand cardiovascular disease R&D portfolio

Source: Mergers and Innovation and Media release event study by the author. Note: We have NOT considered Net Present value (NPV) of the deals and the actual deal figures are shown. Only deals with an overall value above 9 billion USD have been considered.

7.3.4.2 Objectives and nature of the M&A deals

The mean transaction value of M&A deals was 2690 million USD in 2019 compared to 1613 million USD in 2018 indicating a positive trend in dealmaking activity across the industry. Pfizer has been the most active in BD&L transactions in the first decade of the 21st century with some high-value M&A deals. Pfizer acquired three large companies — Warner-Lambert (in 2000), Pharmacia (in 2003) and Wyeth (in 2009) — and multiple smaller companies, such as Vicuron, Rinat and Esperion to meet its business objectives. However, in the second decade, acquisitions were largely driven

by the strategic rationale to build complementary capabilities rather than a desire to be massive. For instance, AbbVie acquired Pharmacyclics to enhance AbbVie's scientific and commercial presence in oncology with the addition of Imbruvica®, a blockbuster drug approved in multiple indications for blood cancers.

In March 2009, Roche announced a $46.8 billion deal to acquire full ownership of Genentech which has been a key pharma industry merger. This acquisition was strategically directed towards Genentech's three best-selling products — the cancer drugs Avastin, Herceptin and Rituxan. Their Swiss rival Novartis AG. announced $39 billion takeover of U.S. eye care company Alcon the same year. Also, in the 'merger wave' of 2009, Pfizer acquired Wyeth for $68 billion, while Merck paid $41 billion to acquire Schering-Plough to diversify their pipeline with the addition of Remicade and Simponi.

Mergers can intensify the research performance of the transacting firms by creating large knowledge synergies, optimizing R&D expenditure, and improving the research productivity. The M&A dealmaking trends indicate that the big pharma has transitioned into a leaner and focused model by divesting non-core assets and focusing on their speciality therapeutic areas. In 2019, Bristol-Myers Squibb acquired Celgene for a massive 74 billion USD because of enhanced margins, highly complementary portfolios, strong combined cashflows and revenue potential of more than 15 billion USD of six near-term product launches. Therefore, in such cases, the resulting synergies in R&D, administrative and market from an M&A deal usually make the resulting combined company greater than the 'sum of the parts'.

Mean upfront payments for clinical-stage assets have also increased markedly over the 2011-2015 time period. However, discovery and preclinical projects continue to be popular among dealmakers. The level of M&A and licensing activity for preclinical assets has been dominant over the deal volume for clinical or approved products. Also, the dealmaking activity for Phase I and Phase II assets was lower than Phase III and pre-registration assets which demonstrates the reluctance of investors to cashing-in on the riskier Phase I and II projects. Some may argue that the trend of sole interest in late-stage innovations is detrimental to drug discovery in biotechnology industry because then fewer funds are available for early-stage research projects. However, owing to the very low success probability (<5%) of drug development projects, such

early-stage projects should be funded primarily by government and philanthropic organizations. This ensures that a single company does not bear any loss for undertaking drug discovery initiatives and the underlying risk is shared by the use of public funds. On the contrary, the pharmaceutical companies should focus their investments and resources in accelerating late-stage projects by adopting rigorous licensing and acquisition strategies. The logic of comparative advantage strongly favours big pharma companies in acquiring the late-stage projects. So, this current financial landscape at various stages of drug development does facilitate innovation with more emphasis towards bringing the medicine to market.

7.3.4.3 Statistics on M&A deals by sector/focus area

Oncology remained a priority but other areas of research that gained momentum in both licensing and M&A were neuroscience, infectious diseases, cardiovascular and gene therapies. This inclination towards oncology is reflected in a recent acquisition of Array Biopharma by Pfizer for 11.4 billion USD to enrich Pfizer's R&D pipeline with high-potential targeted investigational cancer therapies such as BRAFTOVI® and MEKTOVI® for metastatic colorectal cancer. In 2018, a wave of M&A deals emerged in oncology such as 9 billion USD acquisition of Juno Therapeutics by Celgene and 5 billion USD acquisition of Tesaro by GlaxoSmithKline.

The first decade of the 21st century had deals directed towards diversification into new therapeutic areas and were majorly driven by key blockbuster drugs that could provide entry into a new therapeutic area. For instance, Merck had succeeded with the transformational acquisition of Serono in 2007 driven by blockbuster drugs such as Rebif®, a treatment for relapsing-remitting multiple sclerosis. Similarly, Merck & Co. had a reverse merger deal with Schering-Plough in 2009 which doubled their number of late-stage drugs in development. New innovative therapies emerged in the second decade such as cell-based therapies in 2011 with Provenge and gene-based therapies in 2012 with Glybera. In the recent years, the total value of Medtech venture financing deals has increased drastically, since exponential advances in machine learning and Artificial intelligence (AI) technology has converged digital health technologies with Medtech which has attracted more venture capital investment. M&A deals relating to biomarkers, biosensors and companion diagnostics were also very popular. New alliances with Artificial intelligence (AI) technology developers to accelerate drug discovery and improve R&D productivity and efficiency have become more common.

For example, AstraZeneca collaborated with BenevolentAI in 2019 to use AI and machine learning for the discovery and development of new treatments for chronic kidney disease and idiopathic pulmonary fibrosis.

7.3.4.4 Challenges to successful M&A deals

Big mergers reshape the R&D and growth of the therapeutic areas targeted in M&A strategy. They are likely to raise anticompetitive concerns and may provide fewer incentives to innovate in the long-run. Licensing deal involves working with a licensor who is committed to the continued success of the asset. Such a structure creates more accountability for both the licensee and licensor to hit key milestones in the development and launch of the asset. However, In M&A, if the strategic focus of the acquiring company changes, the assets could linger in development pipelines without being progressed or terminated, especially in phase I or II.

In 2015, Pfizer attempted to acquire Allergan Biologics Ltd, the maker of Botox for 160 billion USD which would have been the largest pharmaceutical deal ever. The plan was to move Pfizer to where Allergan was located in Ireland so that the company could pay the Irish corporate tax rate of 12.5% instead of America's 35% corporate rate. The deal was contingent on several factors including shareholder agreement, US, and EU approval. This deal was structured as a reverse merger so that the smaller Allergan was technically acquiring the much larger Pfizer. However, the deal ultimately fell through because of new laws that were introduced by U.S. President Barack Obama to limit corporate tax inversions.

Furthermore, key talent or capabilities could be lost in M&A transactions, potentially disrupting R&D with a substantial negative impact on the momentum of research programs. The integration demands of acquisition must not be underestimated, and a thoughtful post-merger integration planning should be implemented for the success of the acquired assets. Finally, novel and highly sought-after assets are usually tied up in licensing agreements with other companies early on in development, which causes acquisition deals to be overpriced to gain majority equity of the Intellectual property.

7.3.4.5 Corporate Strategies for successful M&A deals

The percentage of the profit received by both parties in a merger situation will be less than if the initial owner were to commercialize the IP solely, as the financial rewards will be shared between the partners. So, if the IP owner is confident with the success

of the IP and has the economic resources to commercialize the IP by itself then diluting the profits by a merger with another company should be avoided. However, only if the merger adds extra value to the commercialization of IP such as market penetration into new geographical locations/access to new customer segments which compensates for the ownership share loss, the owner should proceed towards a merger deal. The deals where multiple assets are involved, such as megamergers are complicated to evaluate but offer a balanced R&D portfolio.

Often a life-sciences IP will require extensive R&D and a large Infrastructure to be developed and enhanced, which is very expensive. Also, commercialization of this IP may require complementary IP, products and services which are present and owned by established firms. In these circumstances, it makes sense to seek to place the IP in that context through M&A, rather than try to raise capital via a spin-out company and ultimately to compete with established players. Therefore, M&A is an optimal exit strategy for small firms in such situations.

Biopharma companies should also consider thorough due diligence and integration planning in advance of the transaction to help increase the success of assets sourced through M&A. M&A should be strategically used to expand the number of projects in R&D portfolio to compensate for individual project failures and maximize ROI expected by investors. Currently, the corporate R&D pipelines of the top companies include more than 150 drug projects in development phases, with GSK (261), Roche (248), Novartis (223), and Pfizer (205) having more than 200 projects in portfolio.

Table 22: Key products from M&A and Licensing deals for top 20 biopharma companies

S.No	Company	Products / Technologies	Net Sales (2019)
1.	Roche	Ocrevus®, Hemlibra®, Alecensa®, RoActemra®	53.36 billion USD
2.	Pfizer	Eliquis®, Enbrel®, XTANDI®, Celebrex®	51.75 billion USD
3.	Novartis	Promacta®, Jakavi®, Lucentis®, Gilenya®	47.44 billion USD
4.	Merck & Co.	KEYTRUDA®, BRIDION®, SIMPONI®	46.80 billion USD
5.	GlaxoSmithKline	Zejula®, BREO™ ELLIPTA™	44.17 billion USD
6.	Sanofi	Lemtrada®,Libtayo®,Eloxatin®, Aubagio®	42.78 billion USD
7.	Johnson & Johnson	IMBRUVICA®, DARZALEX®, INVOKANA®	42.19 billion USD

8.	AbbVie	Humira®, Mavyret®, Imbruvica®	33.26 billion USD
9.	Takeda	VELCADE®, ADYNOVATE®, TRINTELLIX®	30.87 billion USD
10.	Bristol-Myers Squibb	OPDIVO®, Eliquis®, YERVOY®, EMPLICITI®	26.14 billion USD
11.	AstraZeneca	CRESTOR®, Lumoxiti™, FARXIGA®, ONGLYZA®	24.38 billion USD
12.	Amgen	KANJINTI™, Otezla®, KYPROLIS®, Aimovig™	23.36 billion USD
13.	Boehringer-Ingelheim	Trajenta®, JARDIANCE®, BASAGLAR®	22.49 billion USD
14.	Gilead	YESCARTA®, HARVONI®, Nurix DELIGASE™	22.45 billion USD
15.	Eli Lilly & Co.	Humalog®,VITRAKVI®,QBREXZA®	22.31 billion USD
16.	Bayer	EYLEA®, NEXAVAR®, BETAFERON®	21.27 billion USD
17.	Novo Nordisk	Macrilen™, INDiGO®, Dicerna GalXC™	18.29 billion USD
18.	Teva	Truxima®, BENDEKA®, Attenukine™	16.88 billion USD
19.	Biogen	TECFIDERA®, Spinraza®, Tysabri®	14.37 billion USD
20.	Otsuka	Visterra HIEROTOPE®, ABILIFY MYCITE®, REXULTI®	13.11 billion USD

Source: Company Annual Reports 2019. Only the Pharmaceutical division is considered for net sales.

7.3.5 Conclusion:

In conclusion, over this decade, there have been numerous suggestions of the radical ways in which pharma industry can re-structure itself. Critics have suggested that big pharma companies should go so far as to divest themselves completely of all R&D functions, and simply become companies which acquire new drugs and then market them. The previous trends have indicated that late-stage licensing deals have been a priority for large pharma over preclinical licensing deals. This shift in focus from in-house research to late-stage deals is also reflected from the current trends in licensing and M&A from Table 19 and Table 21. However, in the past 15 years, the M&A and licensing activity for preclinical assets has been dominant over the deal volume for late-stage or approved products. Therefore, most of the big pharma companies have reshaped their BD&L strategy towards creating strategic and operational synergies to bolster their drug pipeline at the preclinical level.

The pressure from investors to launch new products, imminent blockbuster patent expiries and fewer returns from in-house R&D spending has caused the major pharma companies to remain dependent on licensing and M&A deals for supplementing their innovation pipelines. We expect this reliance on licensed products and technologies will continue to increase in the next few decades because of the increasing complexity of innovations in biotechnology can never be sufficiently addressed without external collaborations. Each pharma company does maintain their excellence and leadership in certain therapeutic areas, but the firms need external innovations to stay competitive in the biopharmaceutical industry. We also observed a continued inclination towards oncology in both licensing and M&A deals which is reflected in the higher deal count and mean deal values of the oncology deals. However, other promising areas in dealmaking have been CNS diseases, infectious diseases, endocrine diseases, and cardiovascular diseases. Digital health technologies and medical devices have also emerged as promising areas for M&A and licensing.

The current forecasts suggest that Novartis, Pfizer, and Roche would dominate the pharmaceutical market with expected sales of $49.8 billion, $49.7 billion, and $49.6 billion respectively by 2022. So, these three companies are expected to be the key players in M&A deals for the next few years. A previous study showed that for deals executed between 2007 and 2012, a greater percentage of assets sourced through licensing (22%) made it to market than assets sourced through M&A (14%). This is a result of higher accountability in a licensing deal and a drive to hit the key milestones to gather the next stage funding inherent to the licensing deal structures. Out-licensing of non-core assets would continue to be significant in the next few decades while we project the in-licensing of new innovative products and technologies to be more prominent in the future. So, commercial biotechnology projects directed towards the development of novel technologies are likely to be preferred for in-licensing or acquisition. For instance, recently, AbbVie entered into a collaboration with Genmab for three of Genmab's next-generation bispecific antibodies, including Epcoritamab.

Scenario planning using the licensing and M&A data from this review could help organizations deal with uncertainty and prepare for the future. The best deals are likely to bring synergies in therapeutic areas and build on a life sciences company’s strengths. Divestitures, in the areas where a life sciences company is weak or where an acquisition is not performing, are likely opportunities for growth. Pharma licensing

and acquisition deals are now far more flexible and creative with opportunities to capture value through co-development/co-marketing rights and retaining geographical rights in the deal.[20] Therefore, the licensors can shift from pure licensing deals to deals involving retention of commercial and geographical rights. A key challenge in M&A and licensing over the coming decade will be external collaborations to expand the sales into the emerging markets which have shown to be a major contributor to the big pharma revenue. Finally, wisely positioned licensing deals by pharma companies that complement their R&D innovation synergistically would be important in deciding their market capitalization growth in the biopharmaceutical industry.

7.3.6 Acknowledgement:

I thank the University of Manchester for their support by Dean's Doctoral Scholar Award and President's Doctoral Scholar Award. I also thank my PhD supervisor, Dr Robin Curtis, for his mentorship and guidance and members of the Curtis lab for their enduring support.

7.3.7 References

1. International Federation of Pharmaceutical Manufacturers & Associations. (2017) The Pharmaceutical Industry and Global Health, https://www.ifpma.org/wp-content/uploads/2017/02/IFPMA-Facts-And-Figures-2017.pdf, accessed 26 April 2020.
2. Pharmaceutical Research and Manufacturers of America (PhRMA). (2019) Pharmaceutical Industry Profile 2019, https://www.phrma.org/en/Report/Industry-Profile-2019, accessed 5 November 2019.
3. Schuhmacher, Alexander, Oliver Gassmann, and Markus Hinder.(2016) Changing R&D models in research-based pharmaceutical companies. Journal of translational medicine 14.1: 105, https://doi.org/10.1186/s12967-016-0838-4
4. TEConomy Partners; for PhRMA. (2017) Closing the Gap: Increasing global competition to attract and grow the Biopharmaceutical sector, https://www.phrma.org/-/media/Project/PhRMA/PhRMA-Org/PhRMA-Org/PDF/P-R/PhRMA-InternationalReport-vfinal.pdf, accessed 26 April 2020.
5. OECD. (2009), The Bioeconomy to 2030: Designing a Policy Agenda, OECD Publishing, Paris, https://doi.org/10.1787/9789264056886-en.
6. Pavlou, Alex K., and Mark J. Belsey. (2005) BioPharma licensing and M&A trends. Nat Rev Drug Discov 4, 273–274. https://doi.org/10.1038/nrd1697.
7. Gautam, Ajay, and Xiaogang Pan. (2016) The changing model of big pharma: impact of key trends. Drug discovery today 21.3: 379-384, https://doi.org/10.1016/j.drudis.2015.10.002.
8. Davies, Roger. (2013) The relevance and importance of business development and licensing in the biopharmaceutical industry. Journal of Commercial Biotechnology 19.3, https://doi.org/10.5912/jcb592.

9. Comanor, William S., and Frederic M. Scherer. (2013) Mergers and innovation in the pharmaceutical industry. Journal of health economics 32.1 (2013): 106-113, https://doi.org/10.1016/j.jhealeco.2012.09.006.
10. Pharmafile. (2010) R&D: a new world order, http://www.pharmafile.com/news/r-and-d-new-world-order, accessed 10 December 2019.
11. Morgan Stanley. (2010) Pharmaceuticals Exit Research and Create Value. New York: Morgan Stanley Research Report.
12. Drug Development and Delivery. (2017) R&D Partnerships, https://drug-dev.com/rd-partnerships-partnering-for-progress-how-collaborations-are-fueling-biomedical-advances/, accessed 28 April 2020. .
13. Danzon, Patricia M., Andrew Epstein, and Sean Nicholson. (2007) Mergers and acquisitions in the pharmaceutical and biotech industries. Managerial and Decision Economics 28.4-5 (2007): 307-328, https://doi.org/10.1002/mde.1343.
14. Pharmaceutical Research and Manufacturers of America (PhRMA). (2017) Driving Innovation and Economic growth for the 21st century, http://phrma-docs.phrma.org/files/dmfile/PhRMA-Driving-Innovation_06_01.2017.pdf, accessed 21 February 2020.
15. Deloitte. (2017) External innovation. How biopharma companies are bolstering R&D pipelines through deal-making, https://www2.deloitte.com/us/en/insights/industry/health-care/biopharma-companies-deals-research-development.html, accessed 26 April 2020.
16. Heather Cartwright, Natasha Piper. (2020) IQVIA PHARMA DEALS: Half-Year Review of 2019. IQVIA.
17. Deloitte. (2018) 2018 Global life sciences outlook, https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Life-Sciences-Health-Care/gx-lshc-ls-outlook-2018.pdf, accessed 26 April 2020.
18. FiercePharma. (2019) As Amgen seeks to can Aimovig collaboration, Novartis sues to save the deal, https://www.fiercepharma.com/pharma/not-quite-a-year-into-launch-amgen-wants-to-can-its-aimovig-collaboration-novartis-lawsuit, accessed April 26 2020.
19. LaMattina, John L.(2011) The impact of mergers on pharmaceutical R&D. Nature Reviews Drug Discovery 10.8 (2011): 559, https://doi.org/10.1038/nrd3514.
20. Moran, Nuala. (2007) Licensing deals morph to acquisitions in seller's market. Nat Biotechnol 25, 609–610, https://doi.org/10.1038/nbt0607-609.

7.4 **Computational Developability Assessment Full Results**

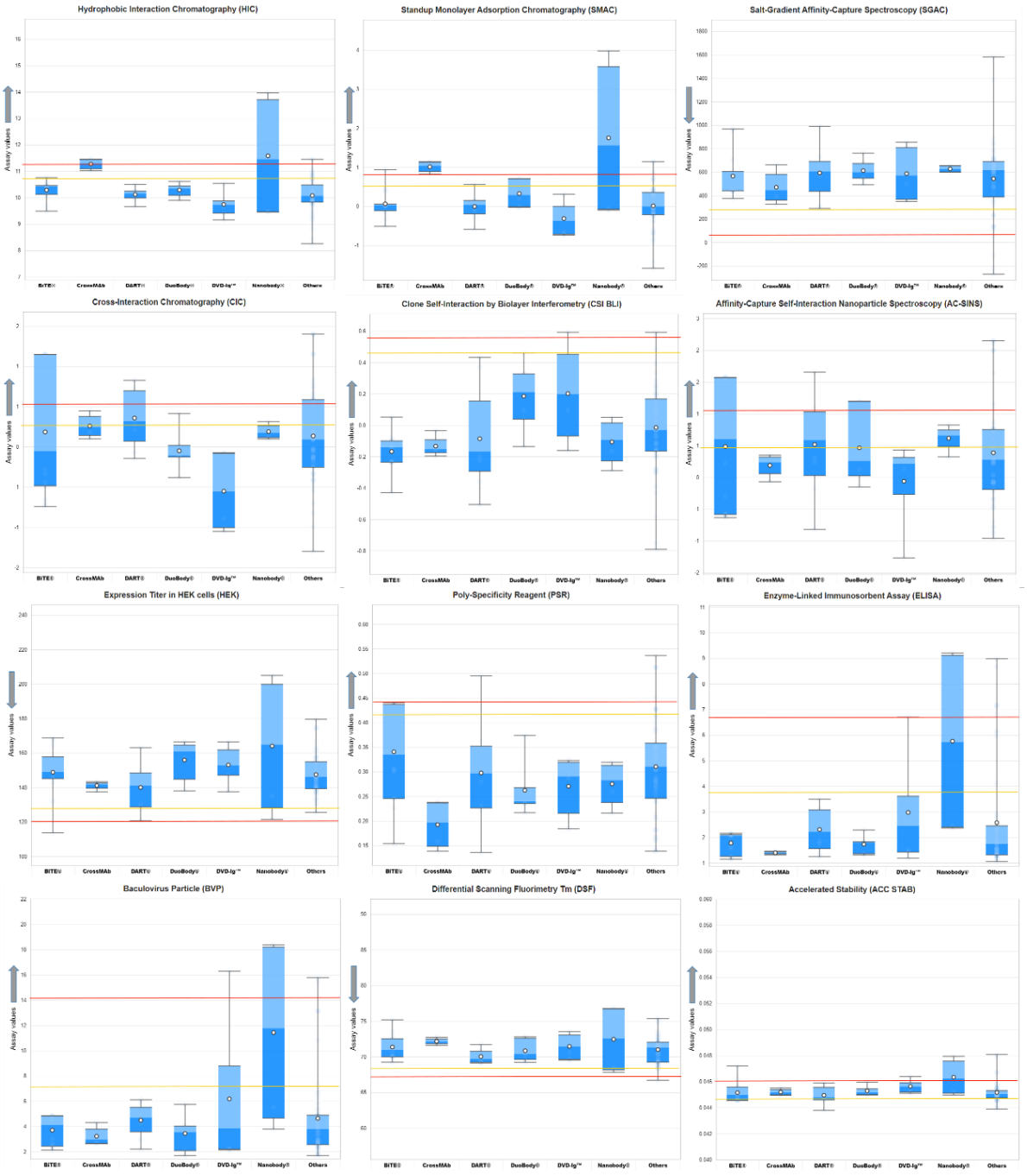


Figure: AbPred scores for 12 developability assays for different categories of engineered antibodies shown on x-axis. The arrow on y-axis indicates the direction of unfavorable values.

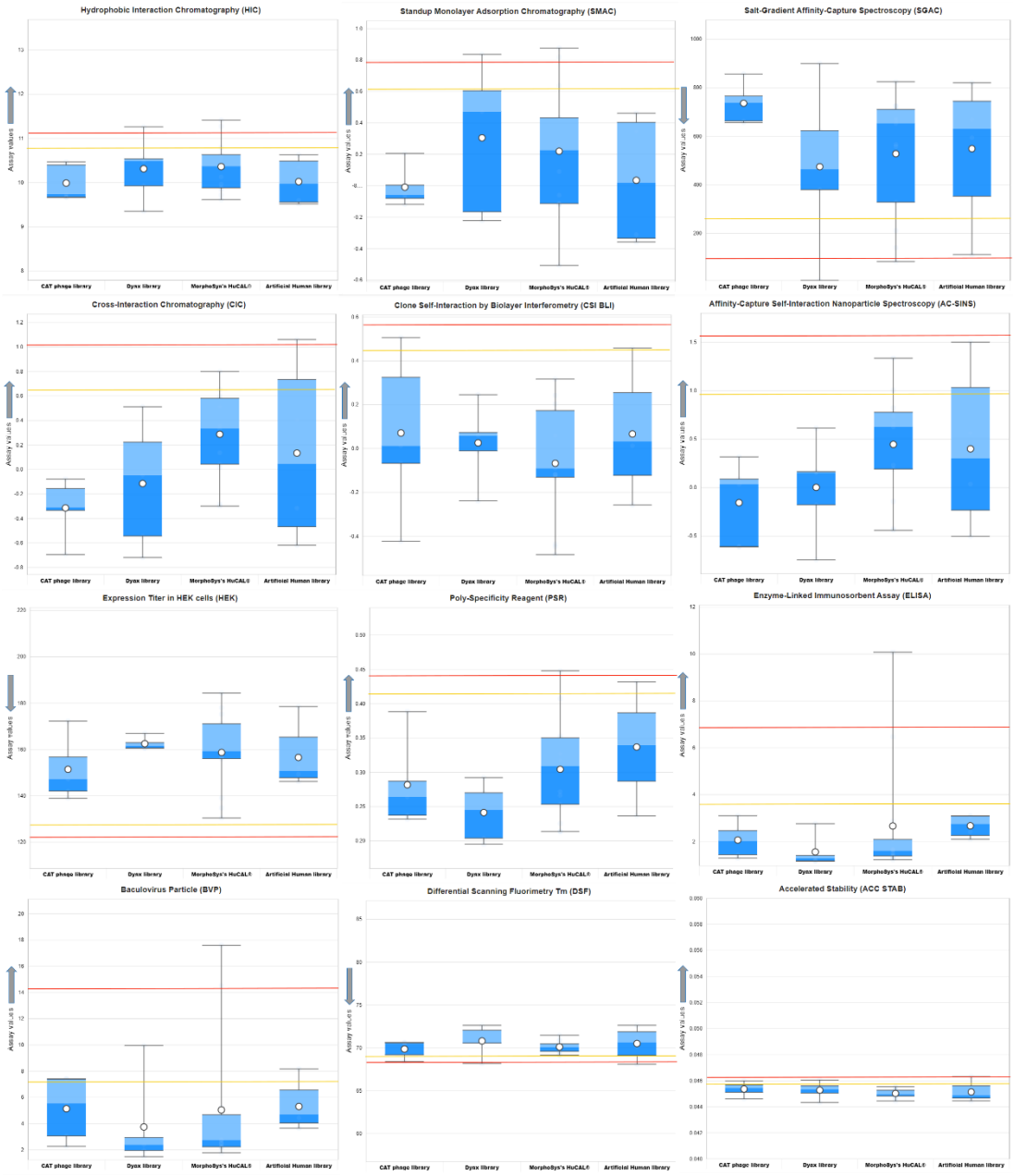
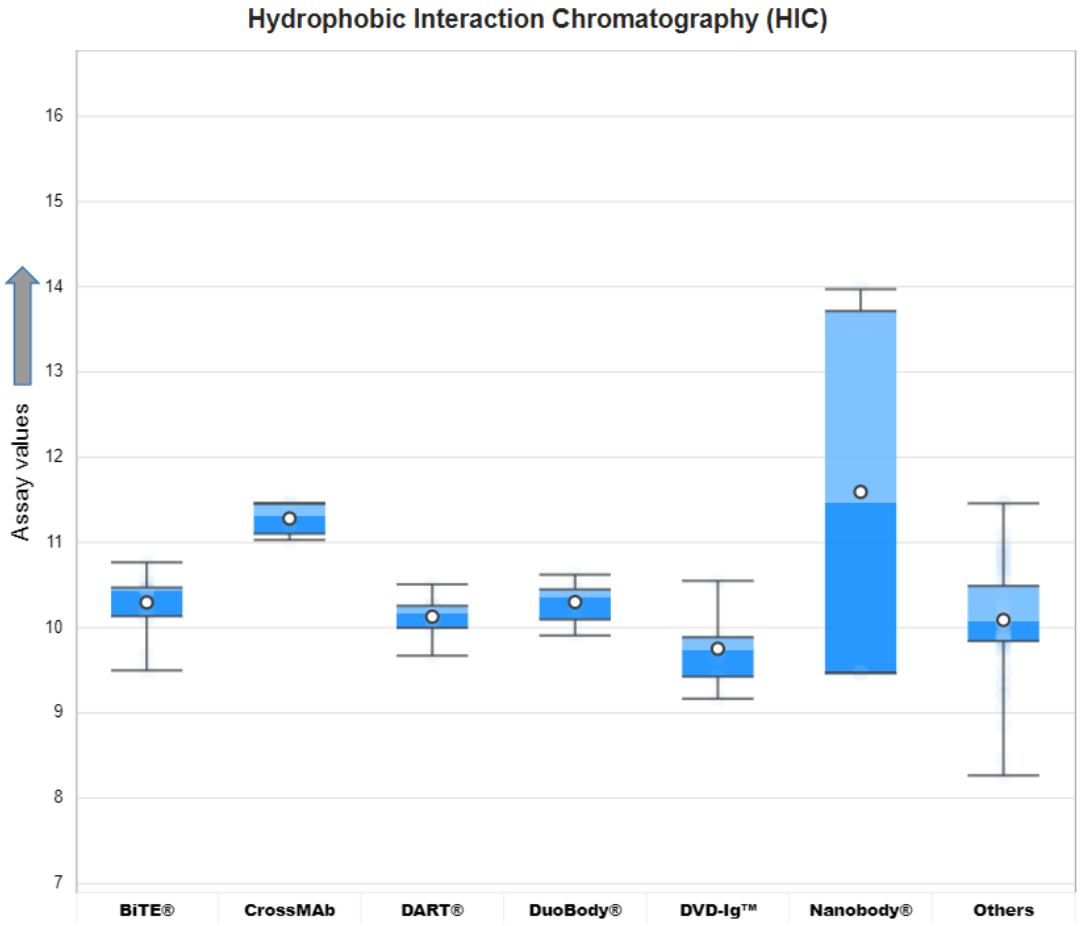


Figure:AbPred scores for 12 developability assays for different categories of phage display antibodies shown on x-axis. The arrow on y-axis indicates the direction of unfavorable values.

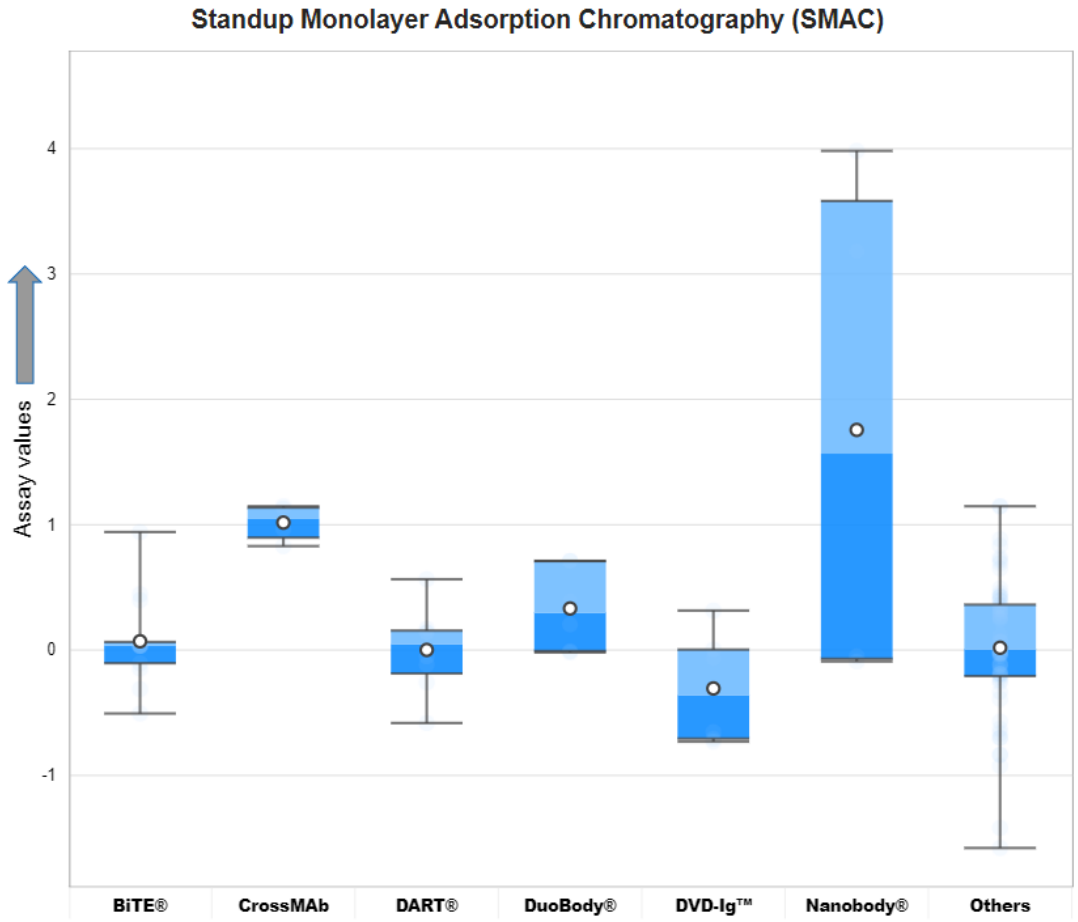
Figure: AbPred scores for 12 developability assays for different transgenic mice antibody platforms shown on x-axis. The arrow on y-axis indicates the direction of unfavorable values.



Hydrophobic Interaction Chromatography (HIC) assay values for different categories of engineered antibodies. The arrow on y-axis indicates the direction of unfavorable values.

	BiTE®	CrossMAb	DART®	DuoBody®	DVD-Ig™	Nanobody®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	10.30	11.28	10.13	10.30	9.75	11.59	10.09
Quartile 1	10.14	11.10	10.00	10.10	9.43	9.47	9.84
Median	10.44	11.32	10.17	10.36	9.74	11.47	10.08
Quartile 3	10.47	11.46	10.26	10.45	9.89	13.71	10.49
Maximum	10.77	11.46	10.51	10.62	10.55	13.97	11.46
Minimum	9.50	11.03	9.67	9.91	9.17	9.47	8.27
IQR	0.34	0.36	0.26	0.35	0.46	4.24	0.65
Upper Whisker	10.77	11.46	10.51	10.62	10.55	13.97	11.46
Lower Whisker	9.50	11.03	9.67	9.91	9.17	9.47	8.27

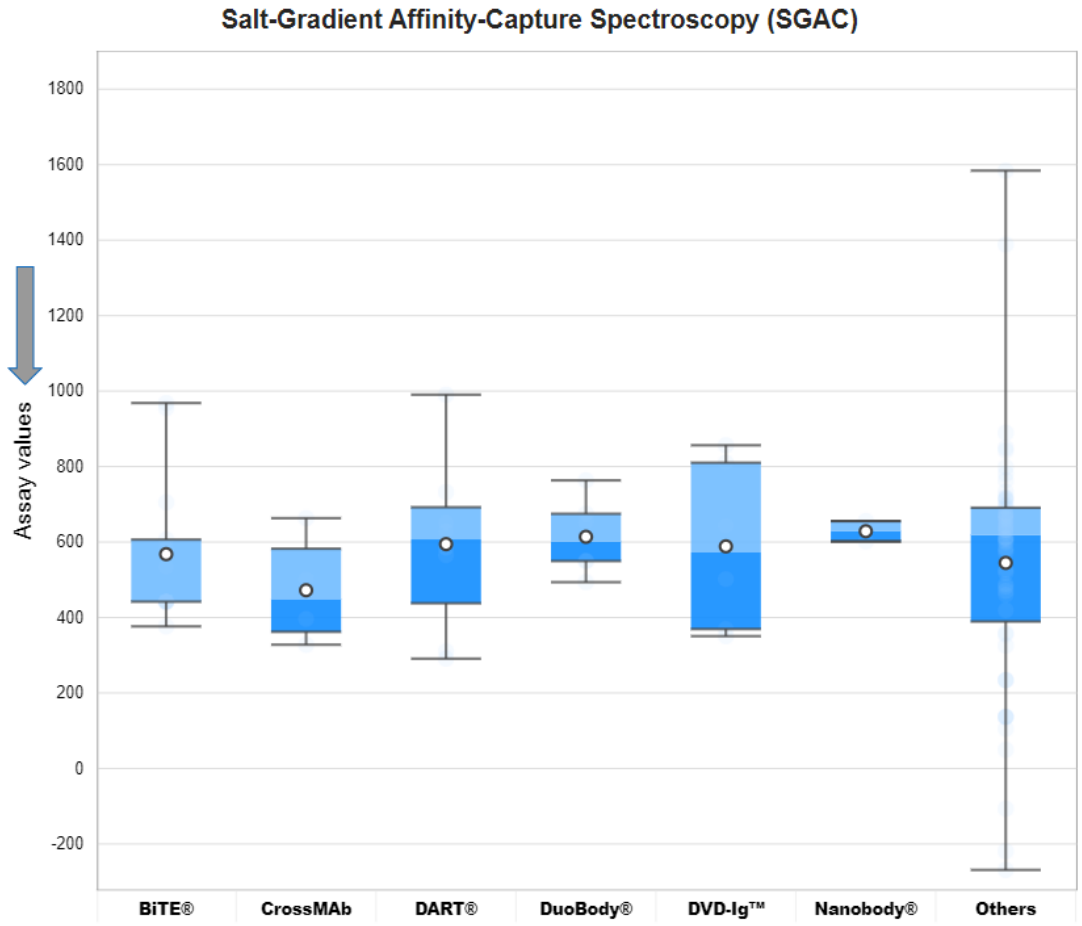
AbPred results for Hydrophobic Interaction Chromatography (HIC).



Standup Monolayer Absorption Chromatography (SMAC) values for different categories of engineered antibodies. The arrow on y-axis indicates the direction of unfavorable values.

	BiTE®	CrossMAb	DART®	DuoBody®	DVD-Ig™	Nanobody®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	0.07	1.02	0.00	0.33	- 0.31	1.76	0.02
Quartile 1	- 0.10	0.90	- 0.19	- 0.01	- 0.71	- 0.07	- 0.21
Median	0.03	1.05	0.04	0.30	- 0.36	1.57	0.01
Quartile 3	0.06	1.14	0.16	0.71	0.00	3.58	0.36
Maximum	0.94	1.15	0.56	0.71	0.32	3.98	1.15
Minimum	- 0.51	0.83	- 0.58	- 0.02	- 0.73	- 0.09	-1.58
IQR	0.17	0.24	0.34	0.72	0.71	3.65	0.57
Upper Whisker	0.94	1.15	0.56	0.71	0.32	3.98	1.15
Lower Whisker	- 0.51	0.83	- 0.58	- 0.02	- 0.73	- 0.09	-1.58

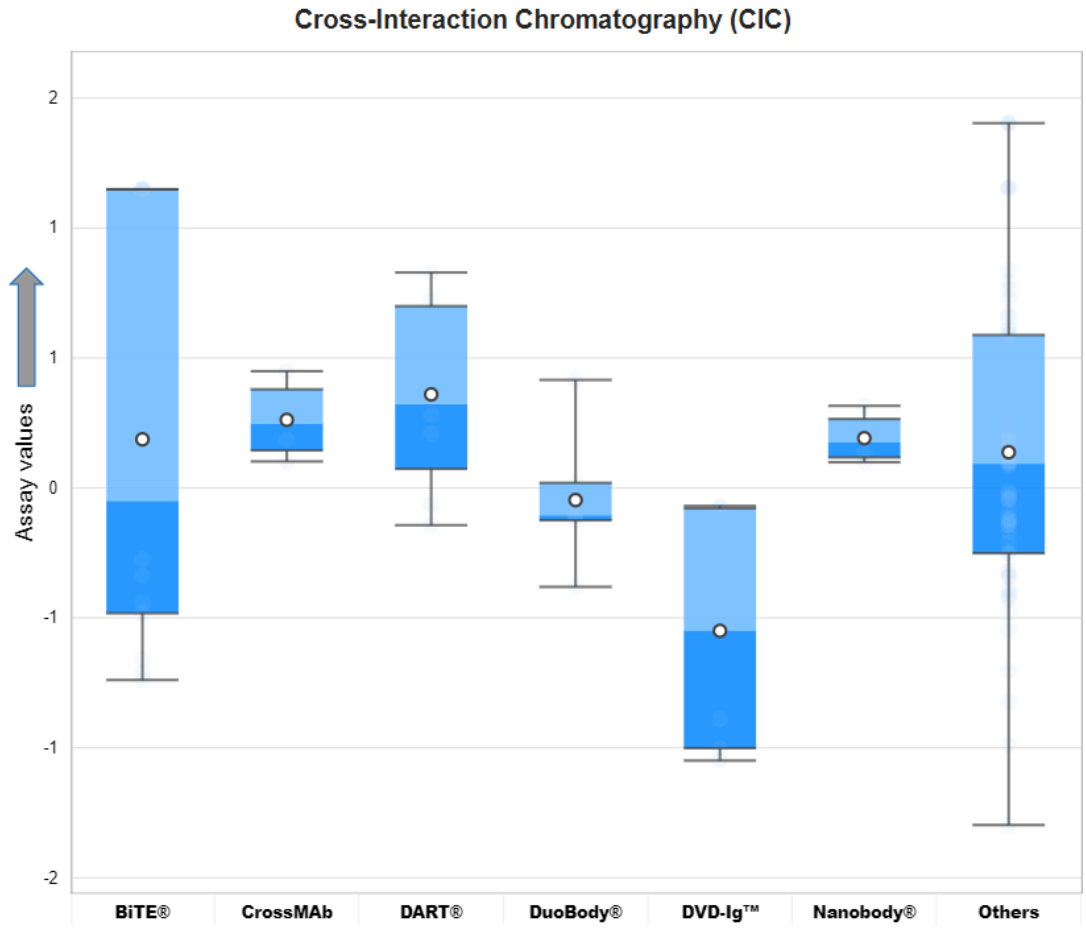
AbPred results for Standup Monolayer Absorption Chromatography (SMAC).



Salt-Gradient Affinity-Capture Spectroscopy (SGAC) assay values for different categories of engineered antibodies. The arrow on y-axis indicates the direction of unfavorable values.

	BiTE®	CrossMAb	DART®	DuoBody®	DVD-Ig™	Nanobody®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	567.89	472.25	594.46	613.87	588.55	628.59	544.78
Quartile 1	441.94	362.10	437.83	549.65	369.40	601.53	389.37
Median	443.51	448.93	607.04	600.81	572.31	628.59	618.50
Quartile 3	606.52	582.39	692.21	674.88	810.19	655.65	690.95
Maximum	968.45	663.22	990.48	763.36	856.48	655.65	1584.17
Minimum	376.40	327.91	290.99	493.70	350.62	601.53	-268.87
IQR	164.57	220.28	254.37	125.24	440.79	54.12	301.58
Upper Whisker	968.45	663.22	990.48	763.36	856.48	655.65	1584.17
Lower Whisker	376.40	327.91	290.99	493.70	350.62	601.53	-268.87

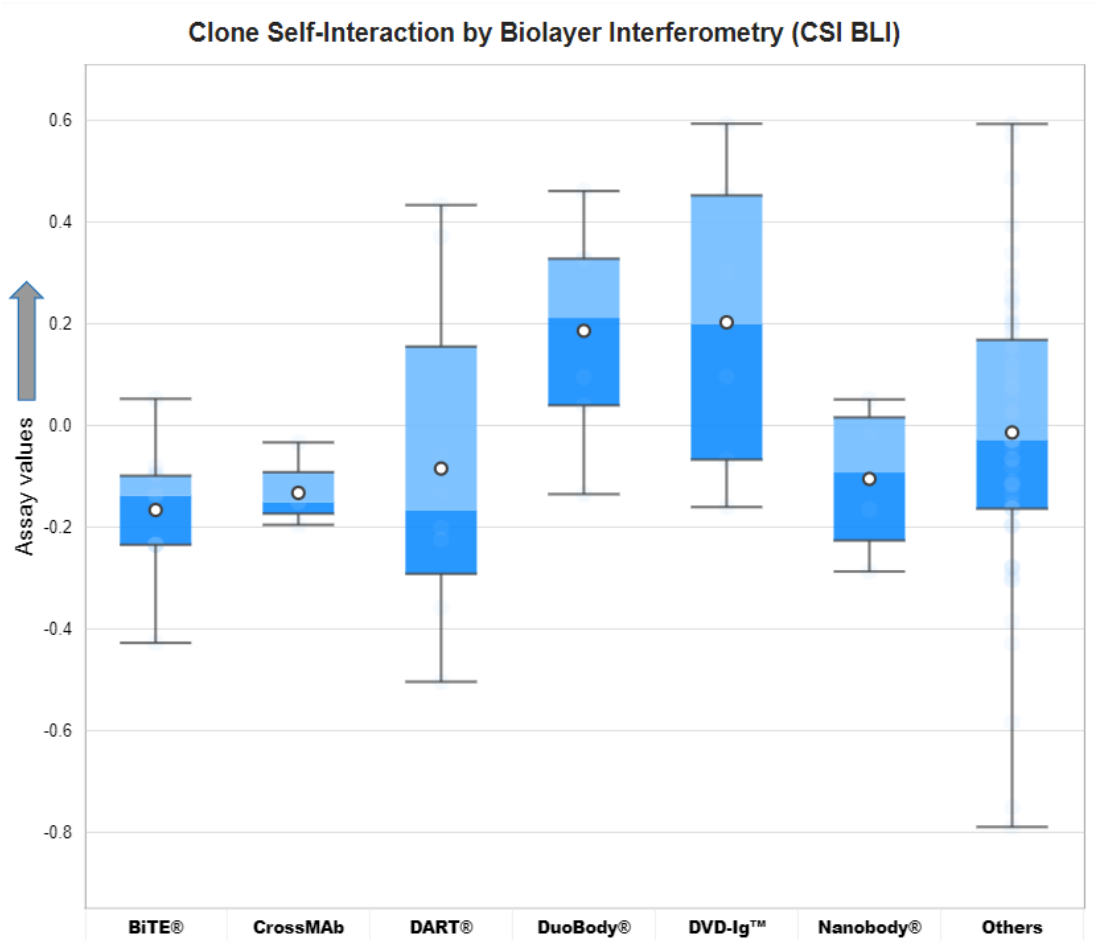
AbPred results for Salt-Gradient Affinity-Capture Spectroscopy (SGAC).



Cross-Interaction Chromatography (CIC) assay values for different categories of engineered antibodies. The arrow on y-axis indicates the direction of unfavorable values.

	BiTE®	CrossMAb	DART®	DuoBody®	DVD-Ig™	Nanobody®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	0.19	0.26	0.36	- 0.05	- 0.55	0.19	0.14
Quartile 1	- 0.48	0.14	0.07	- 0.12	- 1.00	0.12	- 0.25
Median	- 0.05	0.25	0.32	- 0.10	- 0.55	0.18	0.09
Quartile 3	1.15	0.38	0.70	0.02	- 0.08	0.27	0.59
Maximum	1.15	0.45	0.83	0.42	- 0.07	0.32	1.40
Minimum	- 0.74	0.10	- 0.14	- 0.38	- 1.05	0.10	-1.30
IQR	1.63	0.24	0.63	0.14	0.92	0.15	0.84
Upper Whisker	1.15	0.45	0.83	0.42	- 0.07	0.32	1.40
Lower Whisker	-0.74	0.10	- 0.14	- 0.38	- 1.05	0.10	-1.30

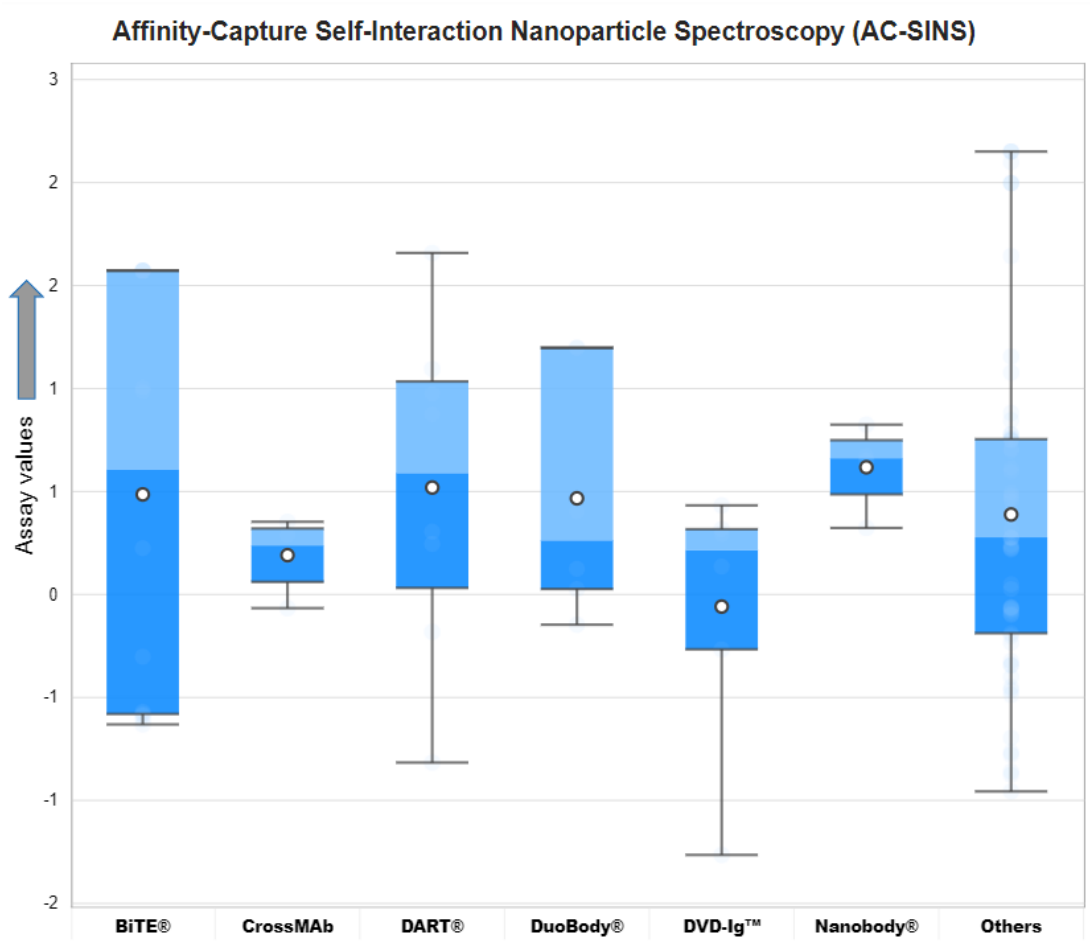
AbPred results for Cross-Interaction Chromatography (CIC).



Clone Self-Interaction by Biolayer Interferometry (CSI BLI) assay values for different categories of engineered antibodies. The arrow on y-axis indicates the direction of unfavorable values.

	BiTE®	CrossMAb	DART®	DuoBody®	DVD-Ig™	Nanobody®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	- 0.17	- 0.13	- 0.08	0.19	0.20	- 0.11	- 0.01
Quartile 1	- 0.23	- 0.17	- 0.29	0.04	- 0.07	- 0.23	- 0.16
Median	- 0.14	- 0.15	- 0.17	0.21	0.20	- 0.09	- 0.03
Quartile 3	- 0.10	- 0.09	0.16	0.33	0.45	0.02	0.17
Maximum	0.05	- 0.03	0.43	0.46	0.59	0.05	0.59
Minimum	- 0.43	- 0.20	- 0.50	- 0.14	- 0.16	- 0.29	- 0.79
IQR	0.14	0.08	0.45	0.29	0.52	0.24	0.33
Upper Whisker	0.05	- 0.03	0.43	0.46	0.59	0.05	0.59
Lower Whisker	- 0.43	- 0.20	- 0.50	- 0.14	- 0.16	-0.29	- 0.79

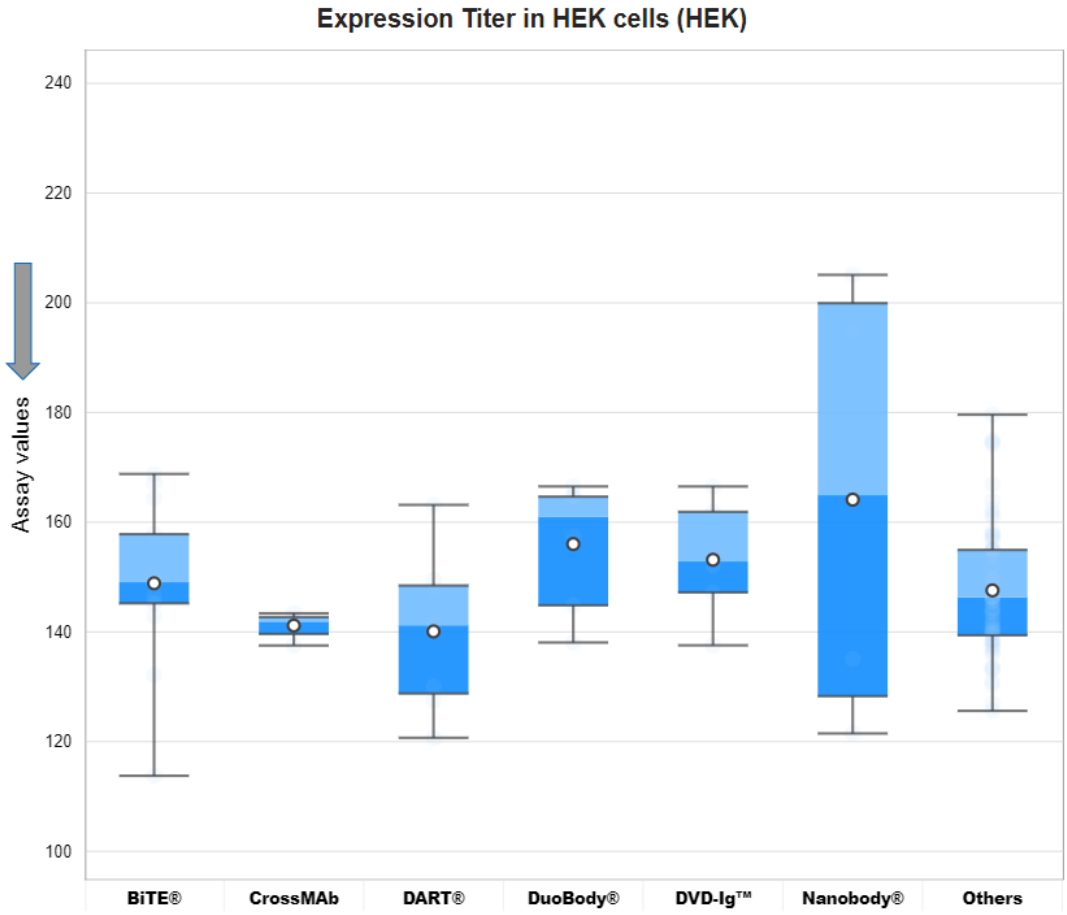
AbPred results for Clone Self-Interaction by Biolayer Interferometry (CSI BLI).



Affinity-Capture Self-Interaction Nanoparticle Spectroscopy (AC-SINS) assay values for different categories of engineered antibodies. The arrow on y-axis indicates the direction of unfavorable values.

	BiTE®	CrossMAb	DART®	DuoBody®	DVD-Ig™	Nanobody®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	0.49	0.19	0.52	0.47	- 0.06	0.62	0.39
Quartile 1	- 0.58	0.06	0.03	0.03	- 0.27	0.49	- 0.19
Median	0.61	0.24	0.59	0.26	0.21	0.66	0.28
Quartile 3	1.57	0.32	1.03	1.20	0.32	0.75	0.75
Maximum	1.57	0.35	1.66	1.20	0.43	0.82	2.15
Minimum	- 0.63	- 0.07	- 0.82	- 0.15	- 1.26	0.32	- 0.96
IQR	2.15	0.26	1.00	1.17	0.58	0.26	0.94
Upper Whisker	1.57	0.35	1.66	1.20	0.43	0.82	2.15
Lower Whisker	- 0.63	- 0.07	- 0.82	- 0.15	- 1.26	0.32	- 0.96

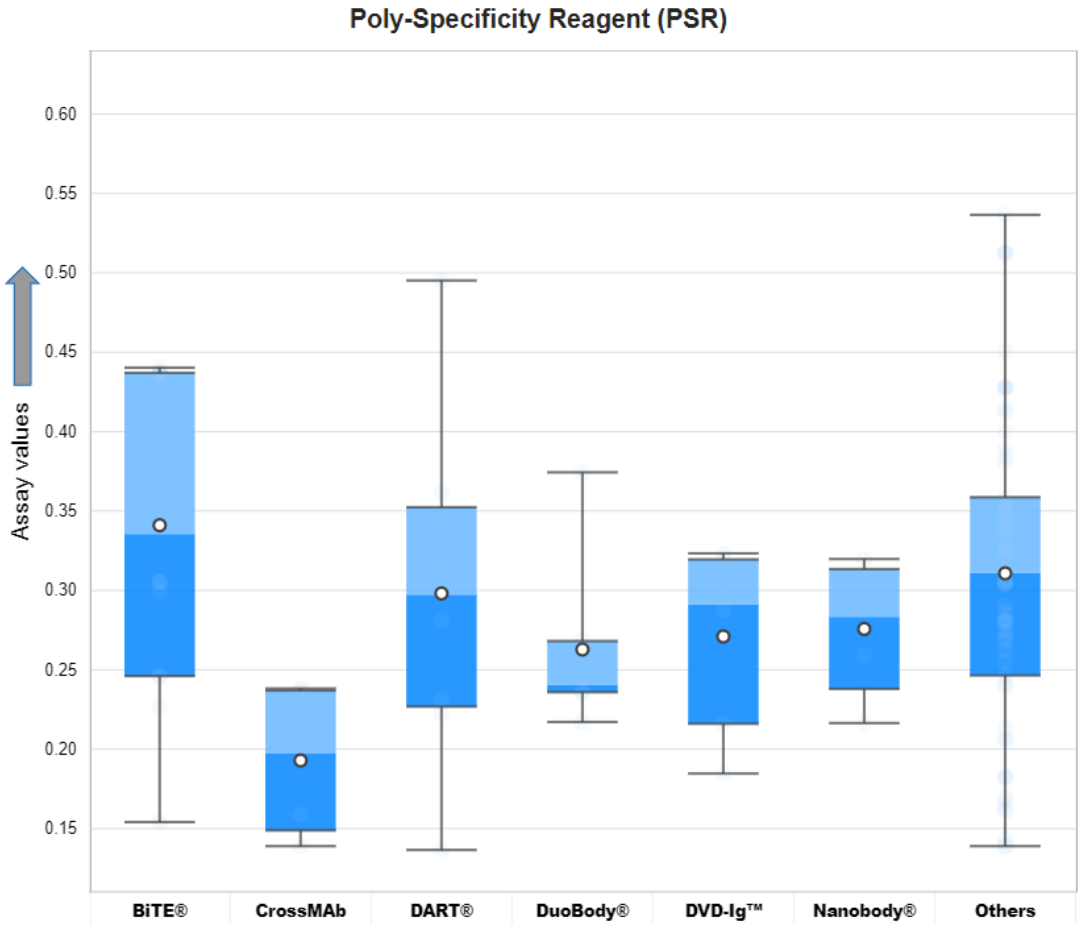
AbPred results for Affinity-Capture Self-Interaction Nanoparticle Spectroscopy (AC-SINS).



Expression Titer in HEK cells (HEK) assay values for different categories of engineered antibodies. The arrow on y-axis indicates the direction of unfavorable values.

	BiTE®	CrossMAb	DART®	DuoBody®	DVD-Ig™	Nanobody®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	148.86	141.17	140.09	156.02	153.15	164.11	147.57
Quartile 1	145.22	139.63	128.77	144.85	147.20	128.29	139.39
Median	149.07	141.88	141.21	161.00	152.85	164.93	146.31
Quartile 3	157.83	142.71	148.46	164.67	161.90	199.94	154.96
Maximum	168.80	143.39	163.16	166.52	166.52	205.10	179.61
Minimum	113.78	137.53	120.70	138.07	137.56	121.49	125.61
IQR	12.61	3.09	19.69	19.82	14.70	71.65	15.57
Upper Whisker	168.80	143.39	163.16	166.52	166.52	205.10	179.61
Lower Whisker	113.78	137.53	120.70	138.07	137.56	121.49	125.61

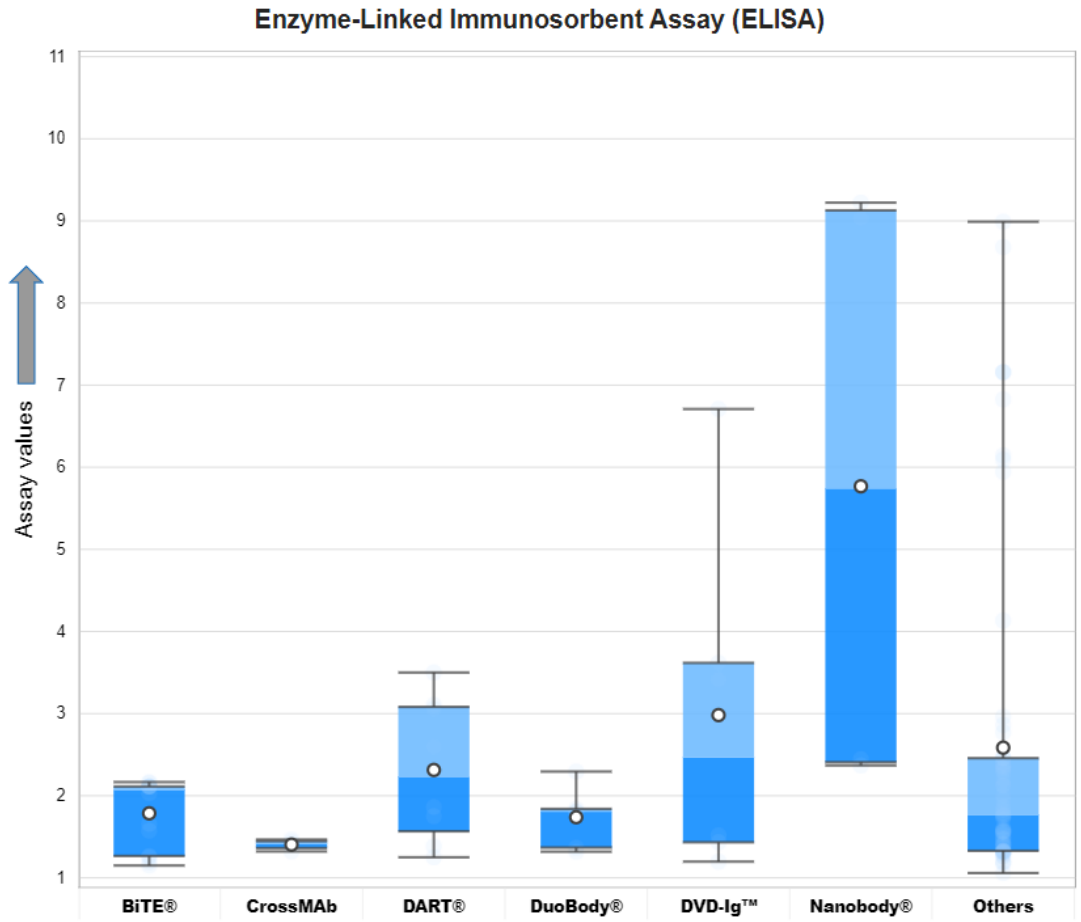
AbPred results for Expression Titer in HEK cells (HEK).



Poly-Specificity Reagent (PSR) assay values for different categories of engineered antibodies. The arrow on y-axis indicates the direction of unfavorable values.

	BiTE®	CrossMAb	DART®	DuoBody®	DVD-Ig™	Nanobody®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	0.34	0.19	0.30	0.26	0.27	0.28	0.31
Quartile 1	0.25	0.15	0.23	0.24	0.22	0.24	0.25
Median	0.34	0.20	0.30	0.24	0.29	0.28	0.31
Quartile 3	0.44	0.24	0.35	0.27	0.32	0.31	0.36
Maximum	0.44	0.24	0.50	0.37	0.32	0.32	0.54
Minimum	0.15	0.14	0.14	0.22	0.18	0.22	0.14
IQR	0.19	0.09	0.13	0.03	0.10	0.08	0.11
Upper Whisker	0.44	0.24	0.50	0.37	0.32	0.32	0.54
Lower Whisker	0.15	0.14	0.14	0.22	0.18	0.22	0.14

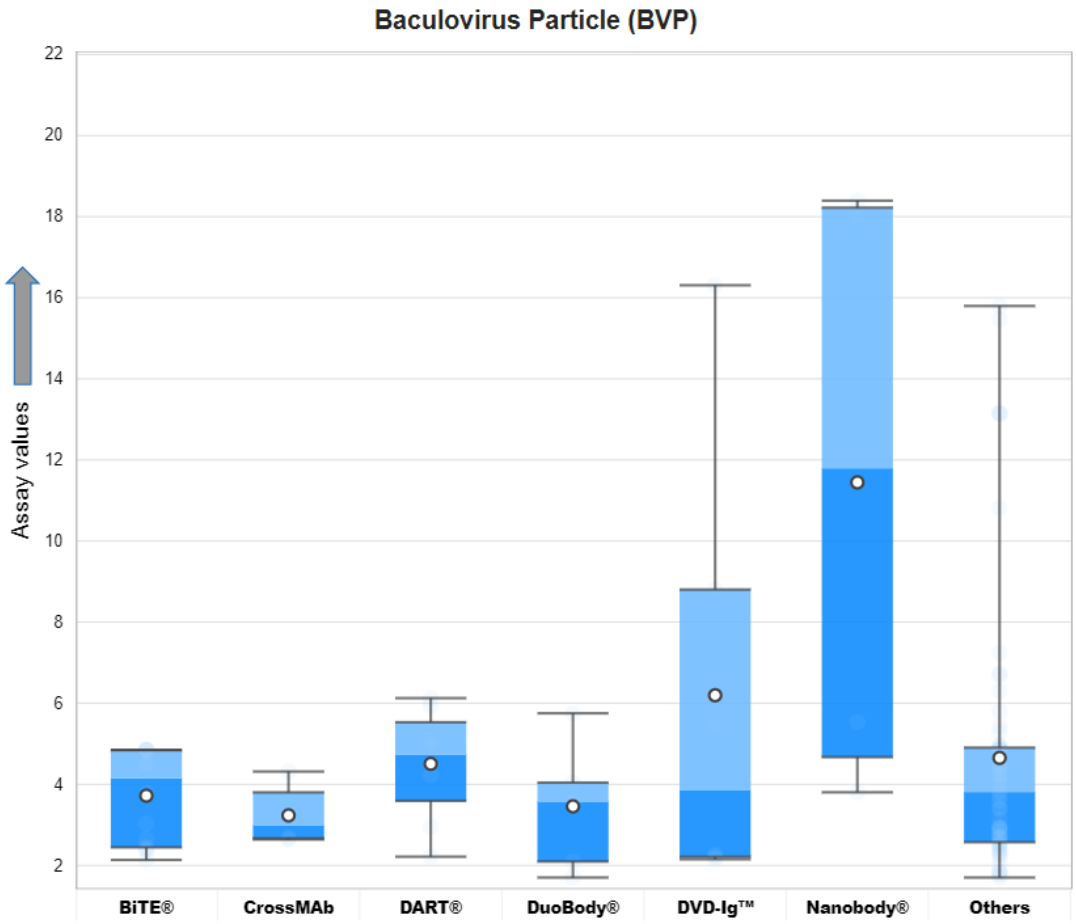
AbPred results for Poly-Specificity Reagent (PSR).



Enzyme-Linked Immunosorbent Assay (ELISA) values for different categories of engineered antibodies. The arrow on y-axis indicates the direction of unfavorable values.

	BiTE®	CrossMAb	DART®	DuoBody®	DVD-Ig™	Nanobody®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	1.79	1.41	2.31	1.74	2.98	5.77	2.59
Quartile 1	1.27	1.36	1.57	1.37	1.43	2.41	1.33
Median	2.07	1.42	2.23	1.81	2.47	5.74	1.77
Quartile 3	2.11	1.45	3.08	1.84	3.62	9.13	2.46
Maximum	2.17	1.47	3.50	2.29	6.71	9.22	8.99
Minimum	1.15	1.32	1.25	1.32	1.20	2.37	1.06
IQR	0.85	0.08	1.52	0.47	2.18	6.72	1.13
Upper Whisker	2.17	1.47	3.50	2.29	6.71	9.22	8.99
Lower Whisker	1.15	1.32	1.25	1.32	1.20	2.37	1.06

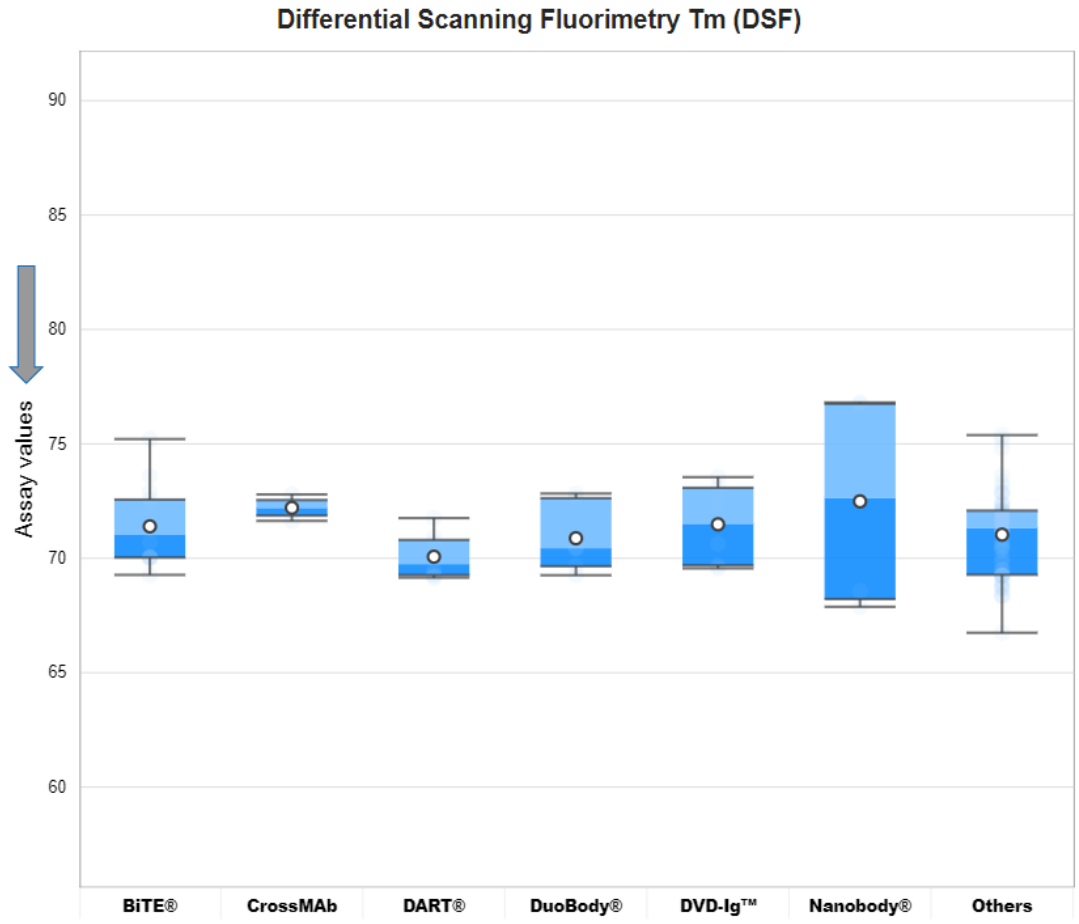
AbPred results for Enzyme-Linked Immunosorbent Assay (ELISA).



Baculovirus Particle (BVP) assay values for different categories of engineered antibodies. The arrow on y-axis indicates the direction of unfavorable values.

	BiTE®	CrossMAb	DART®	DuoBody®	DVD-Ig™	Nanobody®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	3.72	3.23	4.51	3.46	6.20	11.45	4.65
Quartile 1	2.45	2.67	3.59	2.09	2.21	4.67	2.57
Median	4.15	2.99	4.73	3.58	3.85	11.79	3.81
Quartile 3	4.85	3.80	5.53	4.04	8.80	18.22	4.90
Maximum	4.85	4.32	6.13	5.75	16.31	18.39	15.80
Minimum	2.14	2.64	2.21	1.70	2.15	3.80	1.70
IQR	2.40	1.14	1.94	1.95	6.59	13.55	2.33
Upper Whisker	4.85	4.32	6.13	5.75	16.31	18.39	15.80
Lower Whisker	2.14	2.64	2.21	1.70	2.15	3.80	1.70

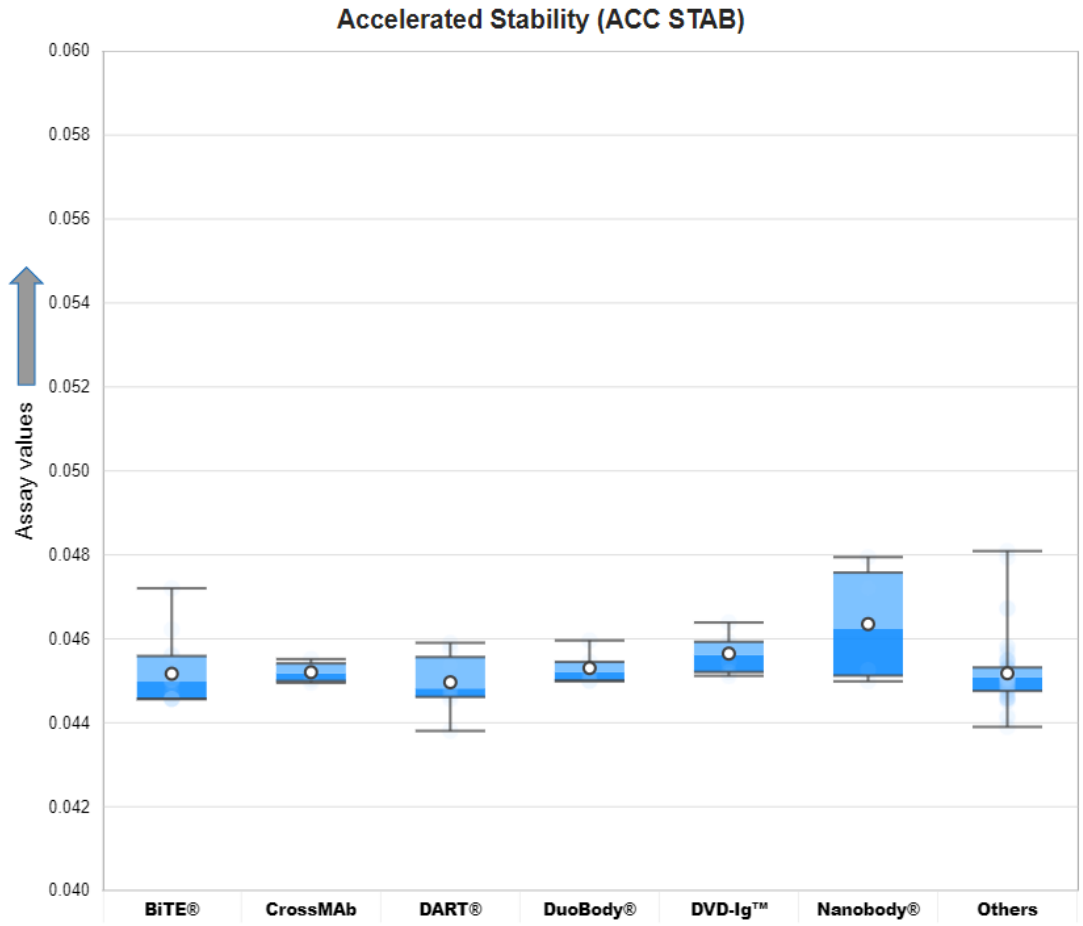
AbPred results for Baculovirus Particle (BVP) assay.



Differential Scanning Fluorimetry (DSF) assay values for different categories of engineered antibodies. The arrow on y-axis indicates the direction of unfavorable values.

	BiTE®	CrossMAb	DART®	DuoBody®	DVD-Ig™	Nanobody®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	71.39	72.20	70.07	70.87	71.48	72.49	71.03
Quartile 1	70.03	71.86	69.26	69.65	69.69	68.22	69.28
Median	71.02	72.20	69.75	70.44	71.51	72.63	71.33
Quartile 3	72.56	72.54	70.80	72.63	73.08	76.75	72.08
Maximum	75.21	72.79	71.76	72.83	73.54	76.81	75.38
Minimum	69.27	71.63	69.15	69.26	69.56	67.88	66.74
IQR	2.53	0.68	1.54	2.97	3.39	8.54	2.80
Upper Whisker	75.21	72.79	71.76	72.83	73.54	76.81	75.38
Lower Whisker	69.27	71.63	69.15	69.26	69.56	67.88	66.74

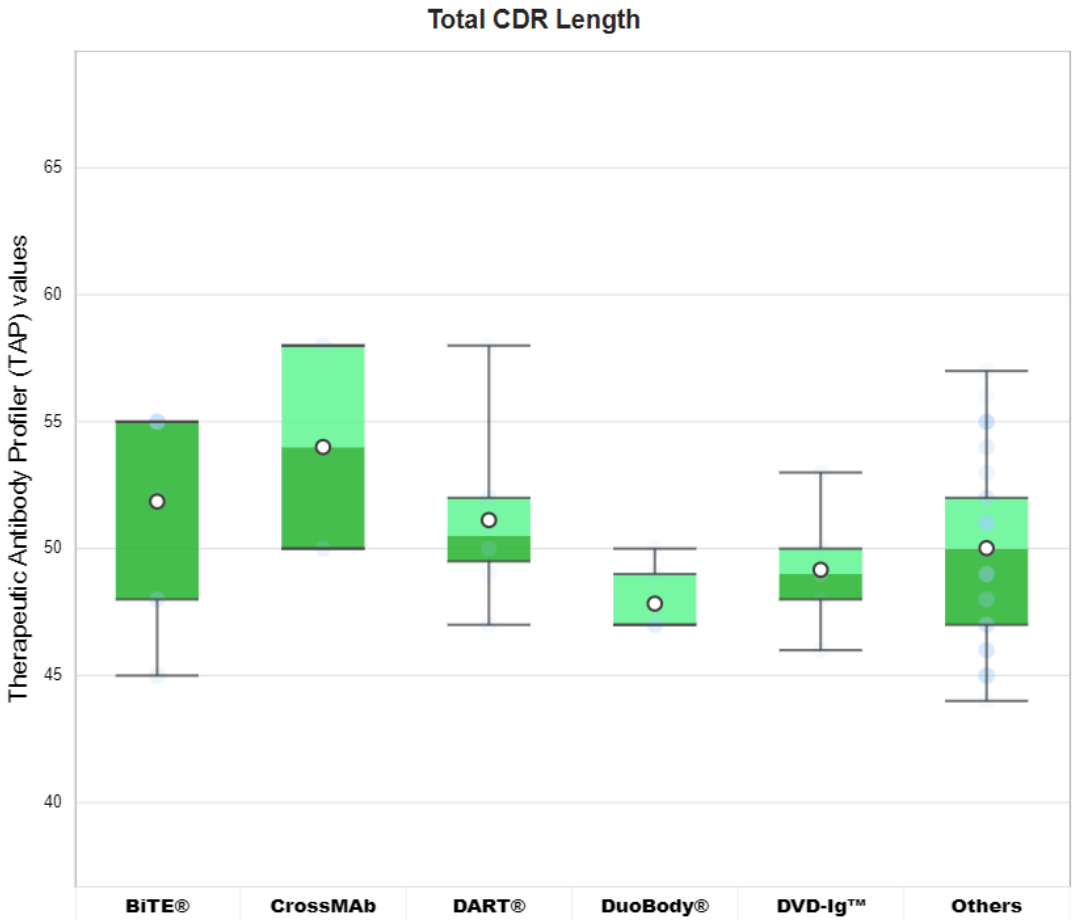
AbPred results for Differential Scanning Fluorimetry (DSF) assay.



Accelerated Stability (ACC STAB) assay values for different categories of engineered antibodies. The arrow on y-axis indicates the direction of unfavorable values.

	BiTE®	CrossMAb	DART®	DuoBody®	DVD-Ig™	Nanobody®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	0.05	0.05	0.04	0.05	0.05	0.05	0.05
Quartile 1	0.04	0.04	0.04	0.04	0.05	0.05	0.04
Median	0.04	0.05	0.04	0.05	0.05	0.05	0.05
Quartile 3	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Maximum	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Minimum	0.04	0.04	0.04	0.04	0.05	0.04	0.04
IQR	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Upper Whisker	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Lower Whisker	0.04	0.04	0.04	0.04	0.05	0.04	0.04

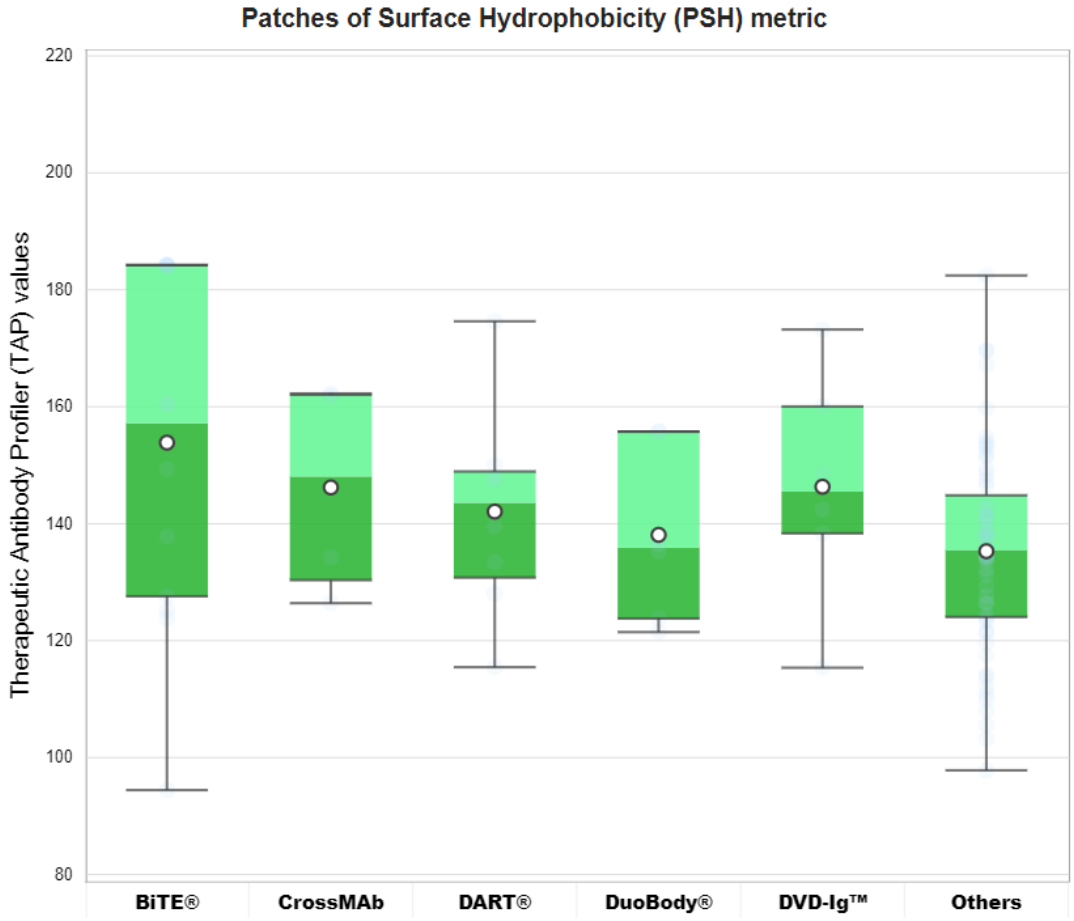
AbPred results for Accelerated Stability (ACC STAB) assay.



Total CDR length metric values for different categories of engineered antibodies.

	BiTE®	CrossMAb	DART®	DuoBody®	DVD-Ig™	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	52	54	51	48	49	50
Quartile 1	48	50	50	47	48	47
Median	55	54	51	47	49	50
Quartile 3	55	58	52	49	50	52
Maximum	55	58	58	50	53	57
Minimum	45	50	47	47	46	44
IQR	7	8	3	2	2	5
Upper Whisker	55	58	58	50	53	57
Lower Whisker	45	50	47	47	46	44

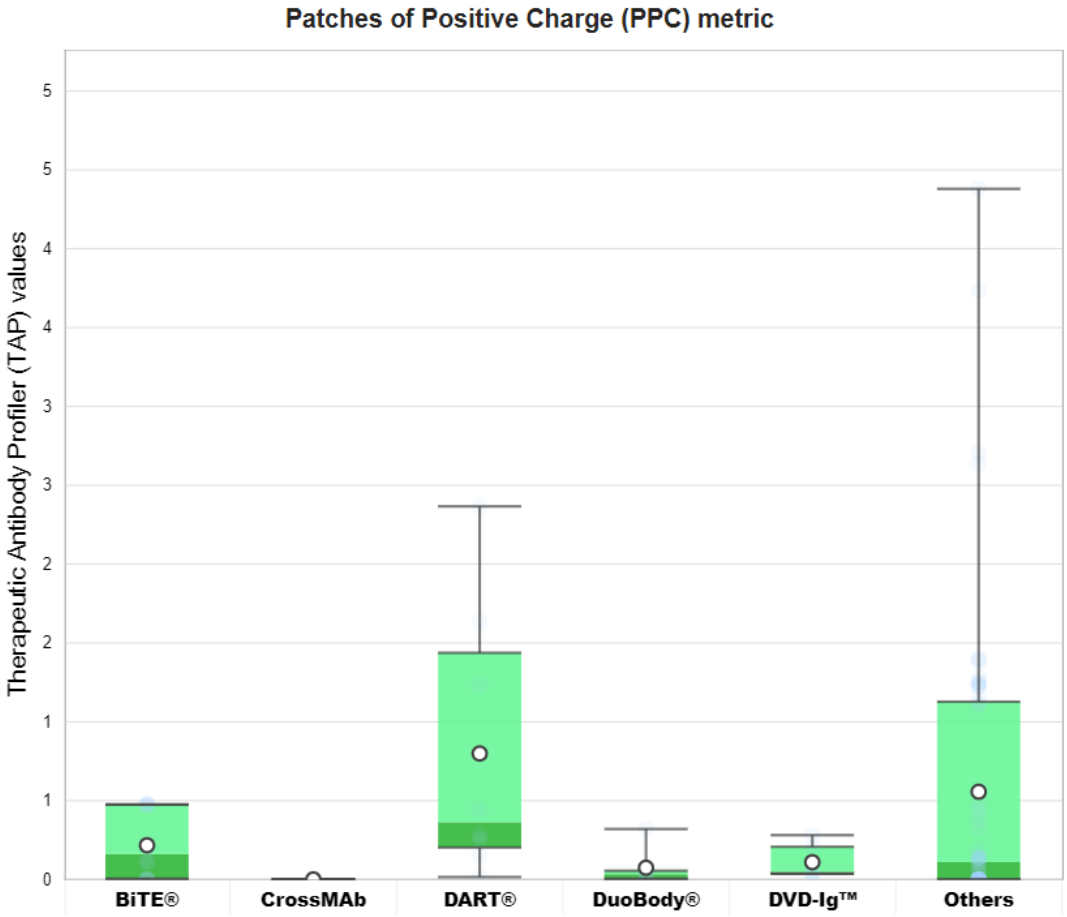
Therapeutic Antibody Profiler (TAP) statistics for Total CDR length.



Patches of Surface Hydrophobicity (PSH) metric values for different categories of engineered antibodies. PSH is calculated across the CDR vicinity.

	BiTE®	CrossMAb	DART®	DuoBody®	DVD-Ig™	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	153.87	146.20	142.08	138.09	146.32	135.29
Quartile 1	127.58	130.35	130.78	123.78	138.34	124.04
Median	157.17	148.07	143.57	135.88	145.48	135.50
Quartile 3	184.23	162.06	148.93	155.76	160.05	144.85
Maximum	184.23	162.27	174.62	155.76	173.21	182.46
Minimum	94.44	126.41	115.46	121.46	115.36	97.82
IQR	56.66	31.71	18.16	31.98	21.70	20.81
Upper Whisker	184.23	162.27	174.62	155.76	173.21	182.46
Lower Whisker	94.44	126.41	115.46	121.46	115.36	97.82

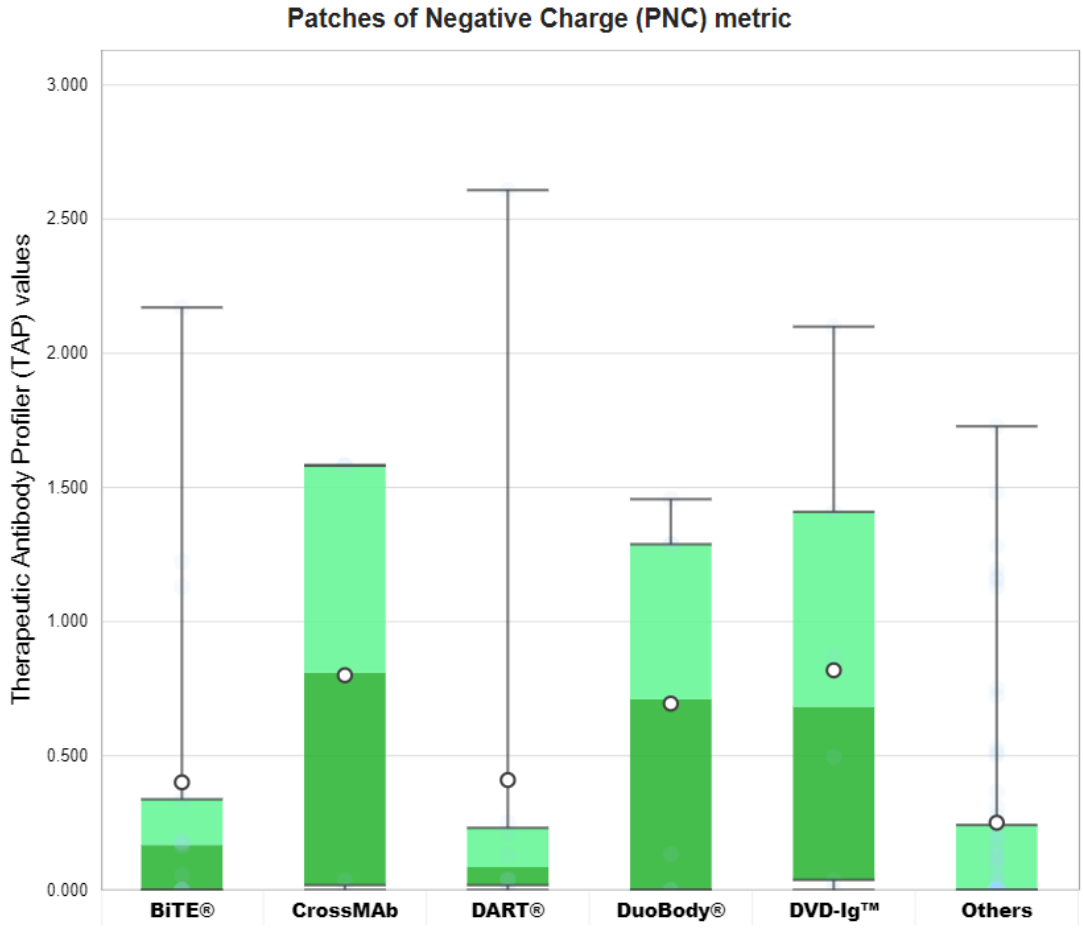
Therapeutic Antibody Profiler (TAP) statistics for Patches of Surface Hydrophobicity (PSH) metric.



Patches of Positive Charge (PPC) metric values for different categories of engineered antibodies. PPC is calculated across the CDR vicinity.

	BiTE®	CrossMAb	DART®	DuoBody®	DVD-Ig™	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	0.22	0.00	0.80	0.07	0.11	0.56
Quartile 1	0.00	0.00	0.20	0.00	0.04	0.00
Median	0.16	0.00	0.36	0.03	0.05	0.11
Quartile 3	0.48	0.00	1.44	0.06	0.21	1.13
Maximum	0.48	0.00	2.37	0.32	0.28	4.38
Minimum	0.00	0.00	0.02	0.00	0.04	0.00
IQR	0.47	0.00	1.23	0.05	0.17	1.13
Upper Whisker	0.48	0.00	2.37	0.32	0.28	4.38
Lower Whisker	0.00	0.00	0.02	0.00	0.04	0.00

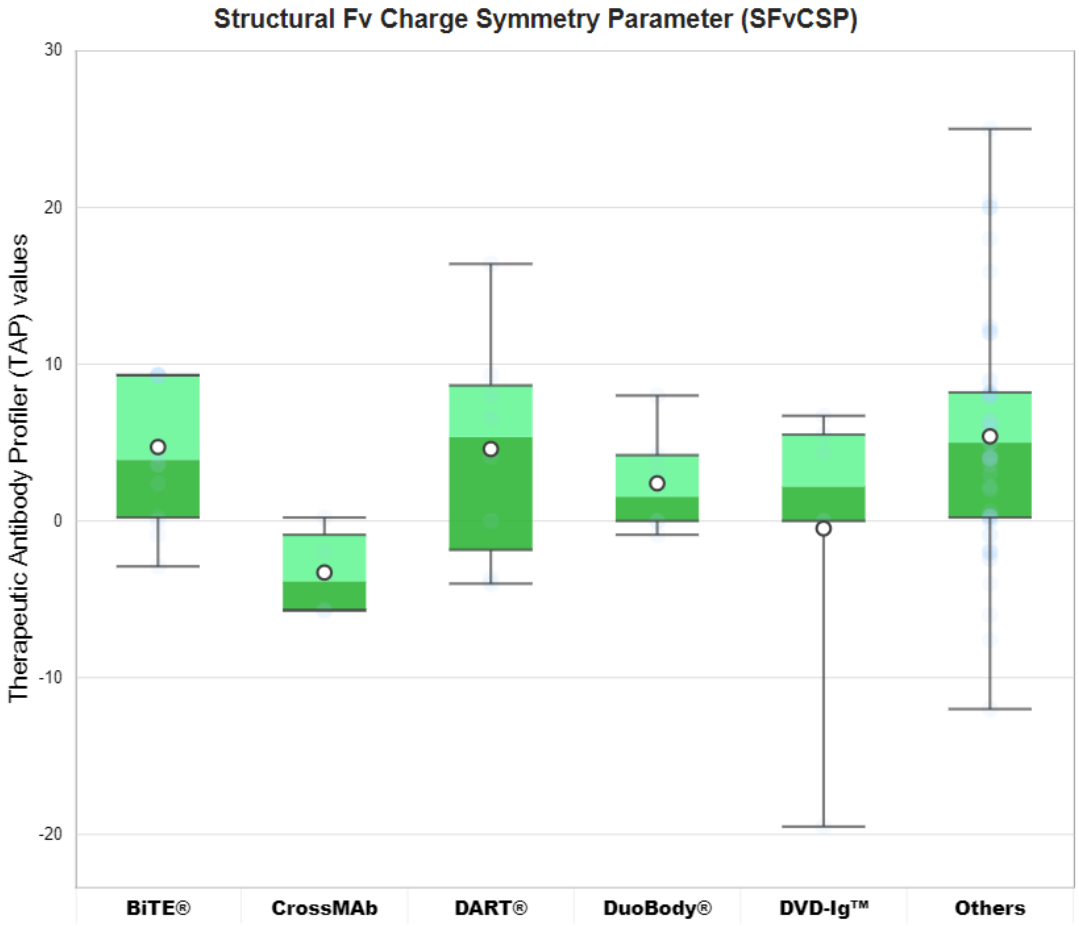
Therapeutic Antibody Profiler (TAP) statistics for Patches of Positive Charge (PPC) metric.



Patches of Negative Charge (PNC) metric values for different categories of engineered antibodies. PNC is calculated across the CDR vicinity.

	BiTE®	CrossMAb	DART®	DuoBody®	DVD-Ig™	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	0.40	0.80	0.41	0.69	0.82	0.25
Quartile 1	0.00	0.02	0.02	0.00	0.04	0.00
Median	0.17	0.81	0.09	0.71	0.68	0.00
Quartile 3	0.34	1.58	0.23	1.29	1.41	0.24
Maximum	2.17	1.58	2.61	1.46	2.10	1.73
Minimum	0.00	0.00	0.00	0.00	0.00	0.00
IQR	0.34	1.56	0.21	1.29	1.37	0.24
Upper Whisker	2.17	1.58	2.61	1.46	2.10	1.73
Lower Whisker	0.00	0.00	0.00	0.00	0.00	0.00

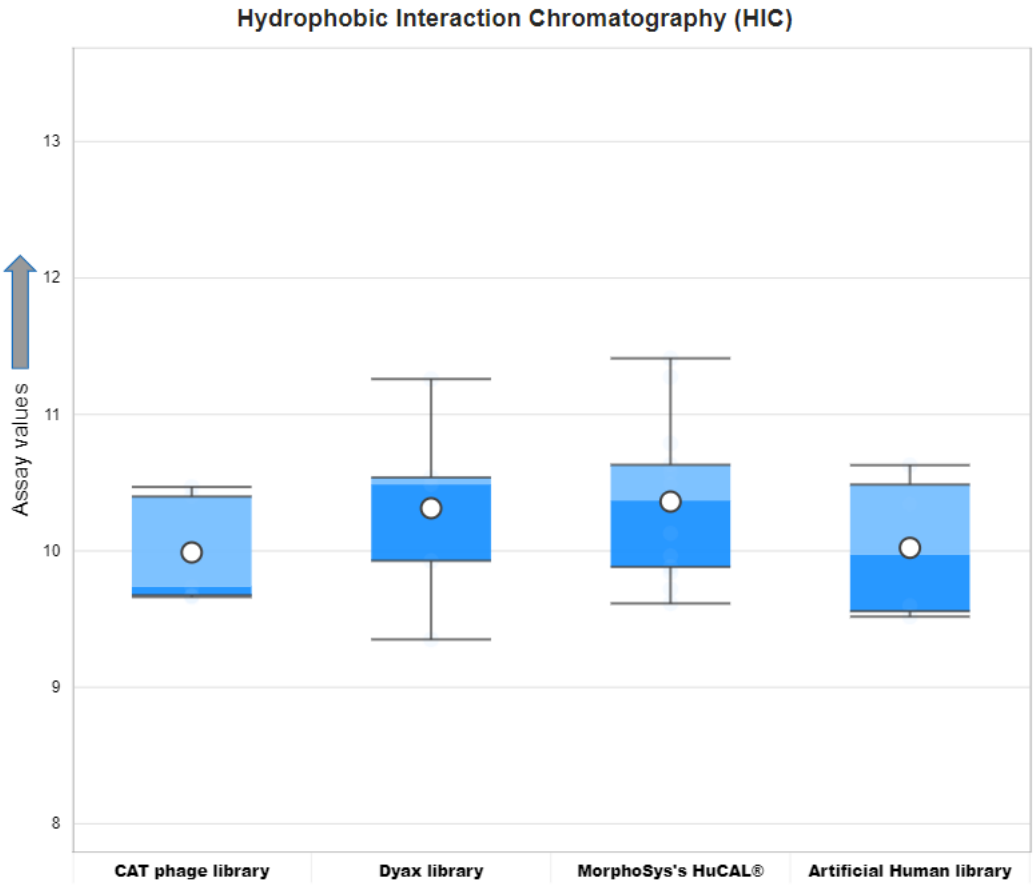
Therapeutic Antibody Profiler (TAP) statistics for Patches of Negative Charge (PNC) metric.



Structural Fv Charge Symmetry Parameter (SFvCSP) values for different categories of engineered antibodies.

	BiTE®	CrossMAb	DART®	DuoBody®	DVD-Ig™	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	4.72	- 3.29	4.59	2.40	- 0.48	5.40
Quartile 1	0.22	- 5.70	- 1.84	0.00	0.00	0.22
Median	3.90	- 3.84	5.35	1.55	2.21	5.00
Quartile 3	9.30	- 0.88	8.65	4.20	5.51	8.20
Maximum	9.30	0.22	16.40	8.00	6.71	25.01
Minimum	- 2.90	- 5.70	- 4.00	- 0.88	- 19.50	- 12.00
IQR	9.08	4.82	10.50	4.20	5.51	7.98
Upper Whisker	9.30	0.22	16.40	8.00	6.71	25.01
Lower Whisker	- 2.90	- 5.70	- 4.00	- 0.88	- 19.50	- 12.00

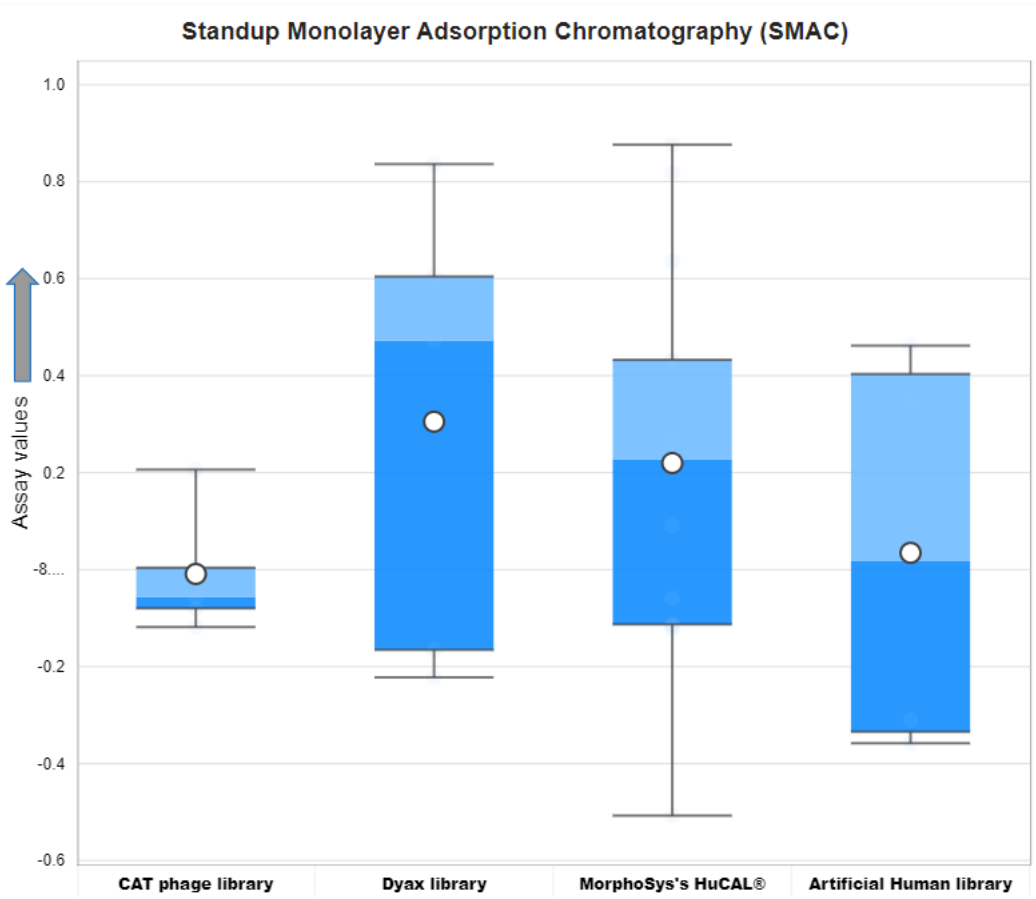
Therapeutic Antibody Profiler (TAP) statistics for Structural Fv Charge Symmetry Parameter (SFvCSP).



Hydrophobic Interaction Chromatography (HIC) values for different categories of antibody phage display platforms. The arrow on y-axis indicates the direction of unfavorable values.

	CAT phage library	Dyax phage library	MorphoSys's HuCAL®	Artificial Human library
Median Type	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max
Mean	9.99	10.31	10.36	10.02
Quartile 1	9.67	9.93	9.88	9.56
Median	9.74	10.49	10.37	9.97
Quartile 3	10.40	10.54	10.63	10.49
Maximum	10.47	11.26	11.41	10.63
Minimum	9.66	9.35	9.61	9.52
IQR	0.72	0.61	0.75	0.93
Upper Whisker	10.47	11.26	11.41	10.63
Lower Whisker	9.66	9.35	9.61	9.52

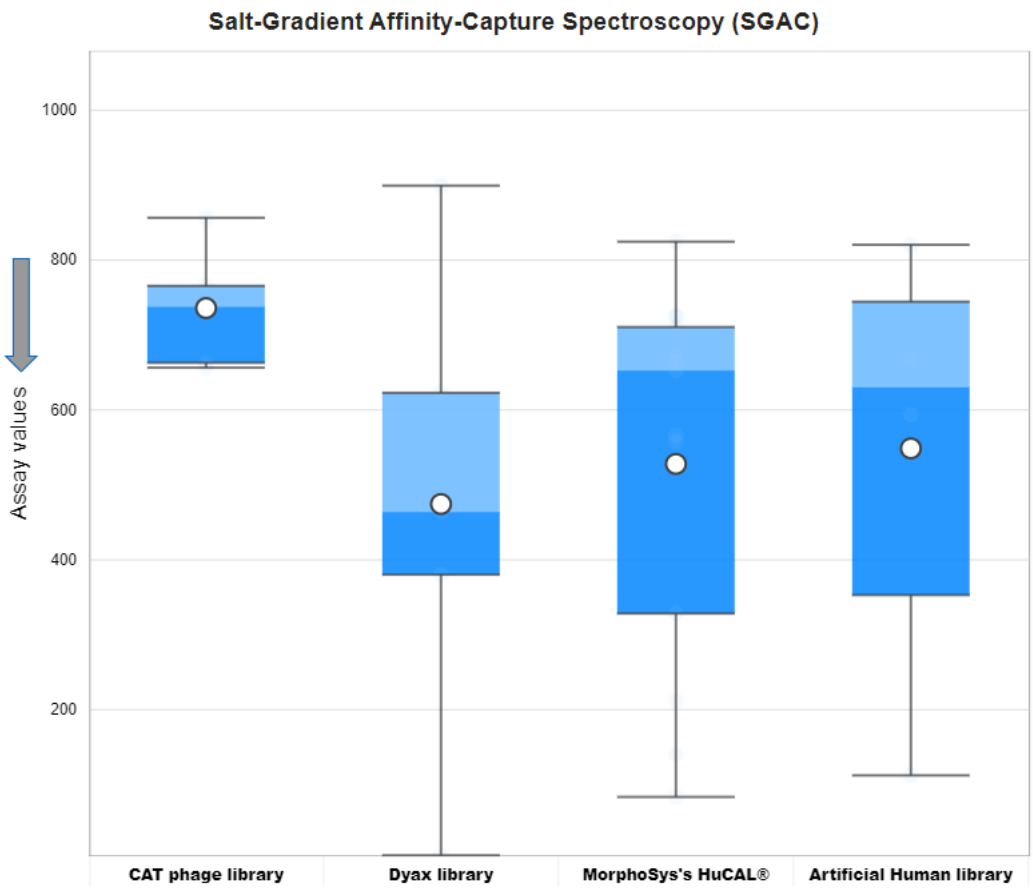
AbPred results for Hydrophobic Interaction Chromatography (HIC).



Standup Monolayer Absorption Chromatography (SMAC) assay values for different categories of antibody phage display platforms. The arrow on y-axis indicates the direction of unfavorable values.

	CAT phage library	Dyax phage library	MorphoSys's HuCAL®	Artificial Human library
Median Type	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max
Mean	- 0.01	0.31	0.22	0.03
Quartile 1	- 0.08	- 0.17	- 0.11	- 0.33
Median	- 0.06	0.47	0.23	0.02
Quartile 3	0.00	0.60	0.43	0.40
Maximum	0.21	0.84	0.88	0.46
Minimum	- 0.12	- 0.22	- 0.51	- 0.36
IQR	0.08	0.77	0.55	0.74
Upper Whisker	0.21	0.84	0.88	0.46
Lower Whisker	- 0.12	- 0.22	- 0.51	- 0.36

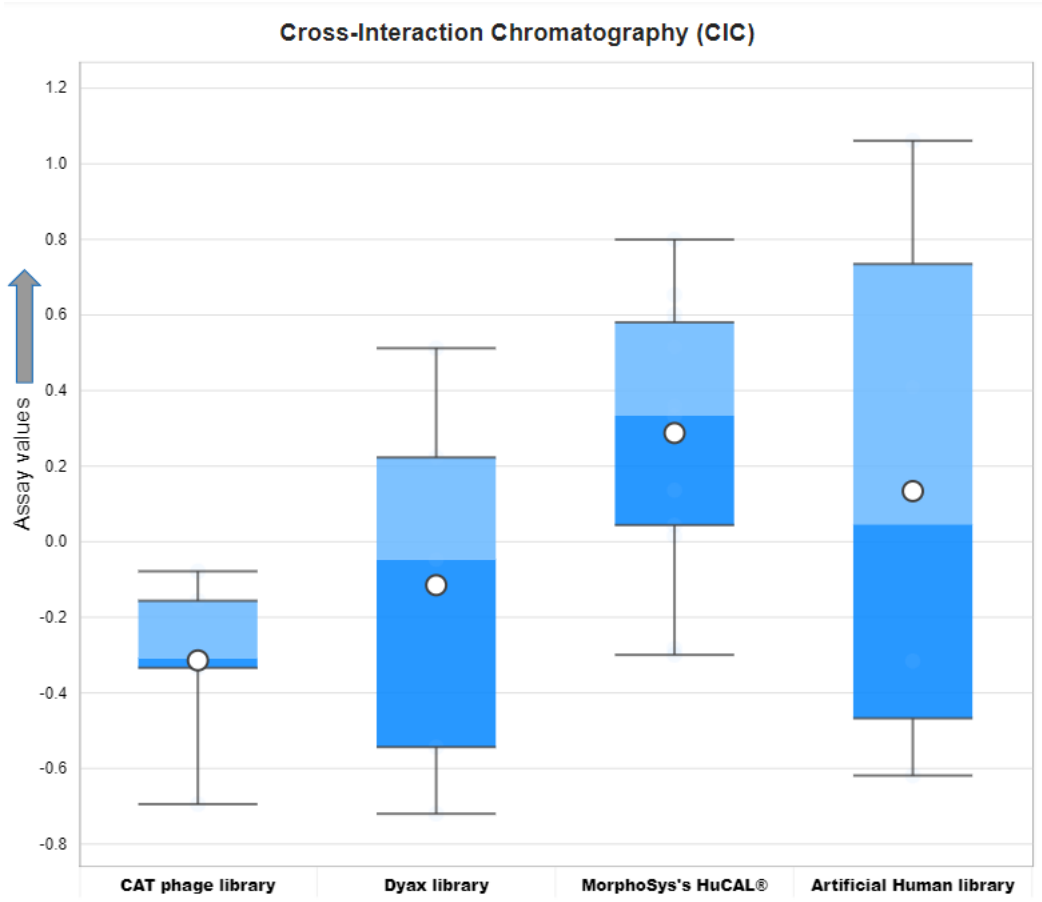
AbPred results for Standup Monolayer Absorption Chromatography (SMAC).



Salt-Gradient Affinity-Capture Spectroscopy (SGAC) values for different categories of antibody phage display platforms. The arrow on y-axis indicates the direction of unfavorable values.

	CAT phage library	Dyax phage library	MorphoSys's HuCAL®	Artificial Human library
Median Type	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max
Mean	735.94	474.44	527.97	548.73
Quartile 1	662.97	380.21	328.35	353.00
Median	738.20	464.00	652.88	631.07
Quartile 3	765.58	622.91	710.66	744.46
Maximum	856.48	899.37	824.59	820.43
Minimum	656.45	5.70	83.61	112.36
IQR	102.62	242.70	382.31	391.46
Upper Whisker	856.48	899.37	824.59	820.43
Lower Whisker	656.45	5.70	83.61	112.36

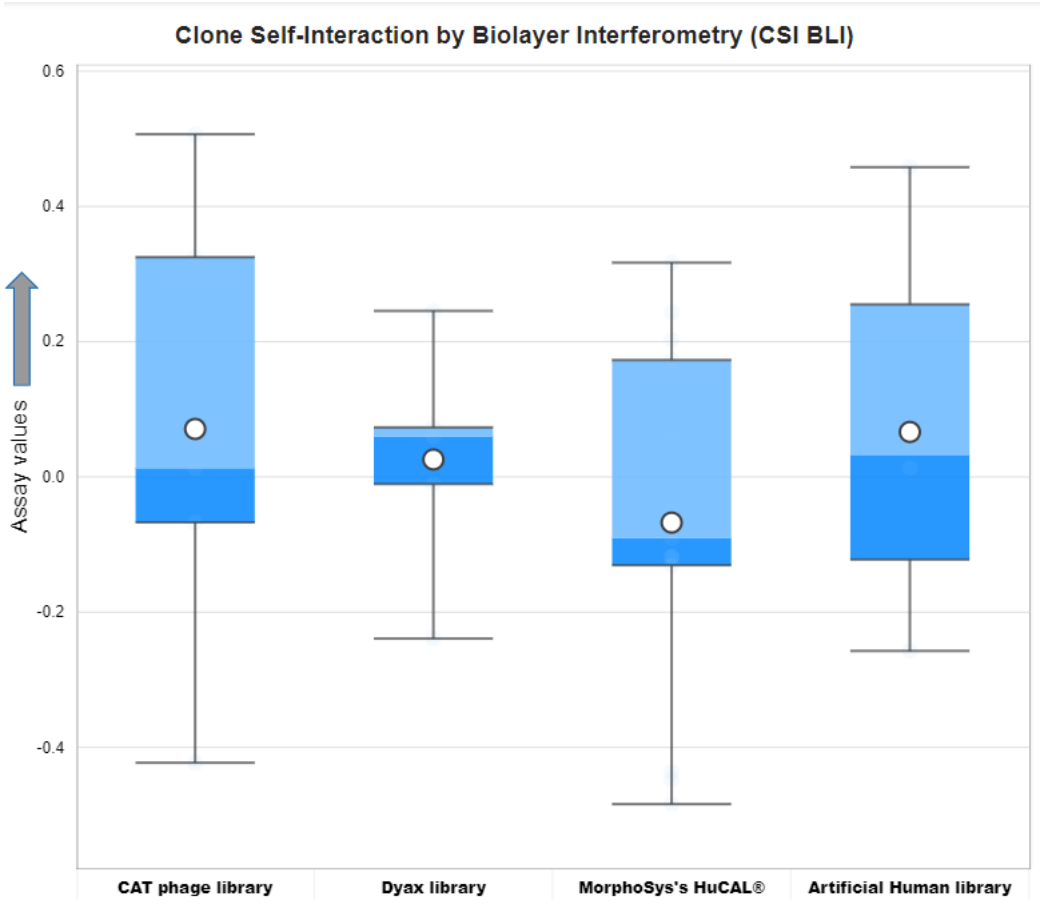
AbPred results for Salt-Gradient Affinity-Capture Spectroscopy (SGAC).



Cross-Interaction Chromatography (CIC) assay values for different categories of antibody phage display platforms. The arrow on y-axis indicates the direction of unfavorable values.

	CAT phage library	Dyax phage library	MorphoSys's HuCAL®	Artificial Human library
Median Type	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max
Mean	- 0.31	- 0.11	0.29	0.13
Quartile 1	- 0.33	- 0.54	0.04	- 0.47
Median	- 0.31	- 0.05	0.33	0.05
Quartile 3	- 0.16	0.22	0.58	0.73
Maximum	- 0.08	0.51	0.80	1.06
Minimum	- 0.69	- 0.72	- 0.30	- 0.62
IQR	0.18	0.77	0.54	1.20
Upper Whisker	- 0.08	0.51	0.80	1.06
Lower Whisker	- 0.69	- 0.72	- 0.30	- 0.62

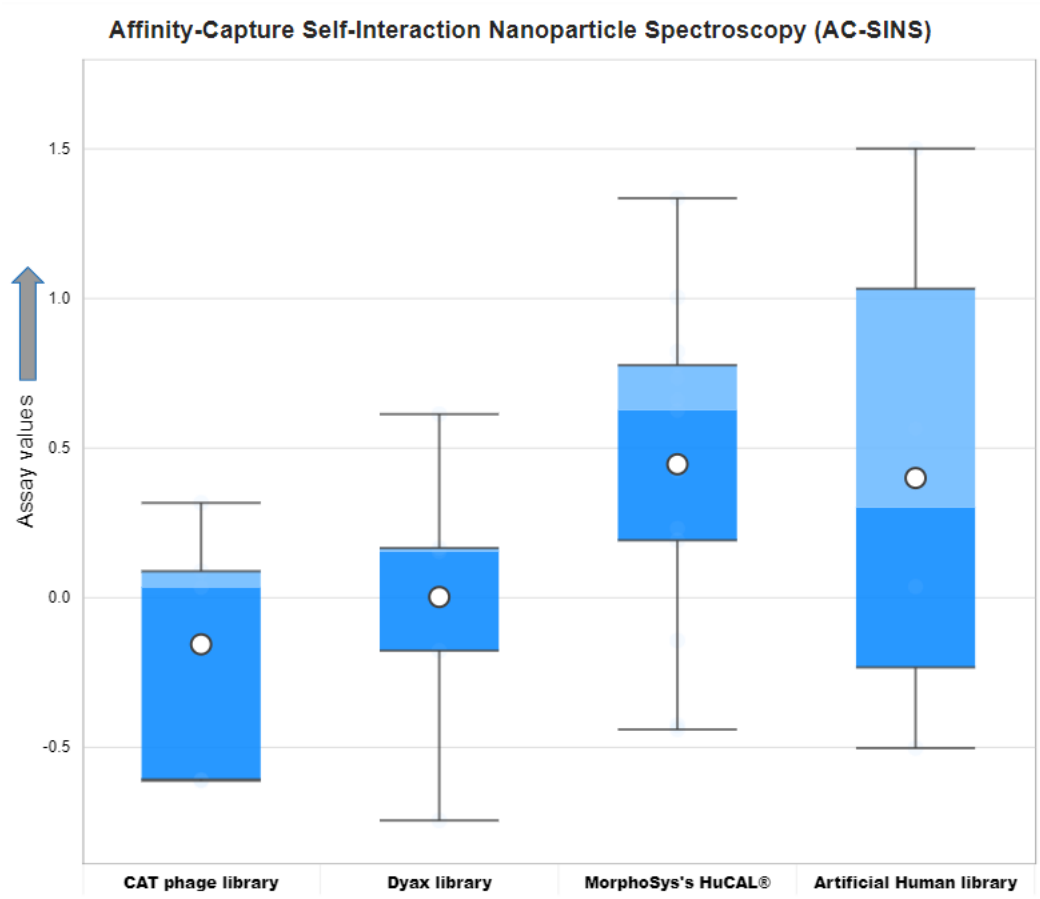
AbPred results for Cross-Interaction Chromatography (CIC).



Clone Self-Interaction by Biolayer Interferometry (CSI BLI) assay values for different categories of antibody phage display platforms. The arrow on y-axis indicates the direction of unfavorable values.

	CAT phage library	Dyax phage library	MorphoSys's HuCAL®	Artificial Human library
Median Type	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max
Mean	0.07	0.03	- 0.07	0.07
Quartile 1	- 0.07	- 0.01	- 0.13	- 0.12
Median	0.01	0.06	- 0.09	0.03
Quartile 3	0.32	0.07	0.17	0.26
Maximum	0.51	0.25	0.32	0.46
Minimum	- 0.42	- 0.24	- 0.48	- 0.26
IQR	0.39	0.08	0.30	0.38
Upper Whisker	0.51	0.25	0.32	0.46
Lower Whisker	- 0.42	- 0.24	- 0.48	- 0.26

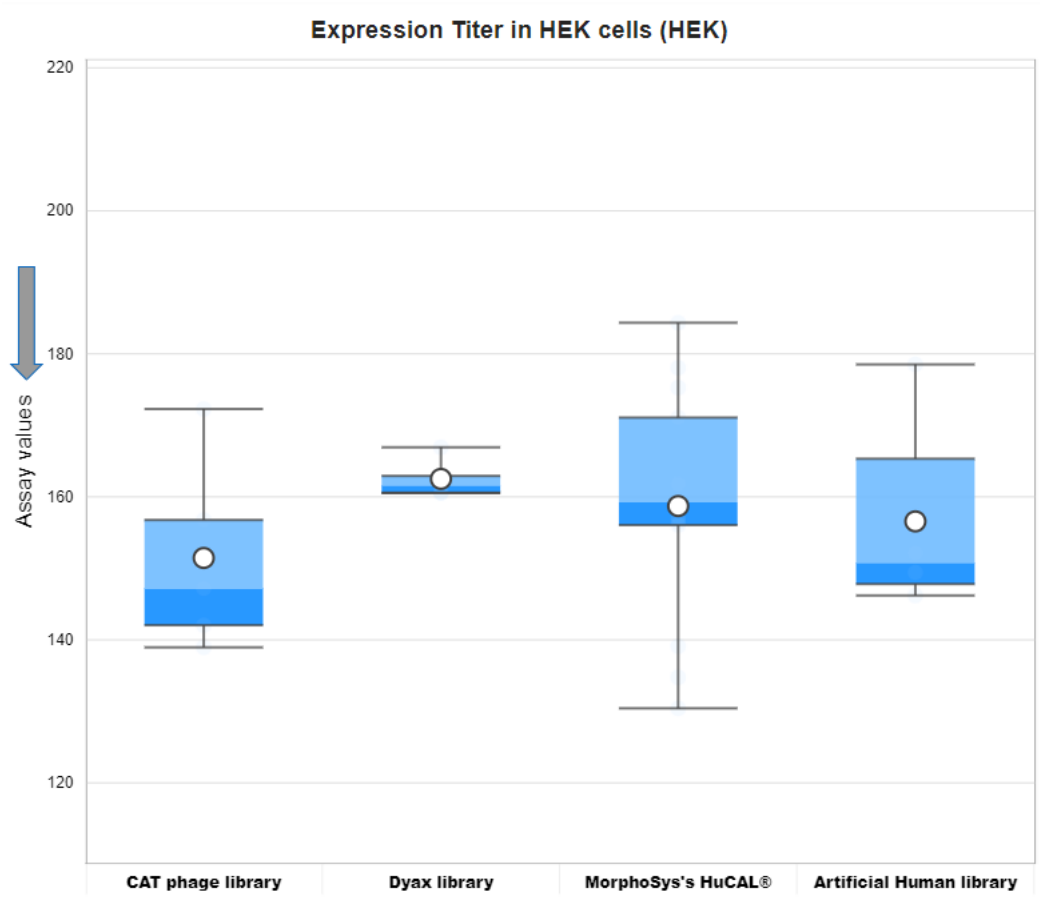
AbPred results for Clone Self-Interaction by Biolayer Interferometry (CSI BLI).



Affinity-Capture Self-Interaction Nanoparticle Spectroscopy (AC-SINS) assay values for different categories of antibody phage display platforms. The arrow on y-axis indicates the direction of unfavorable values.

	CAT phage library	Dyax phage library	MorphoSys's HuCAL®	Artificial Human library
Median Type	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max
Mean	- 0.16	0.00	0.45	0.40
Quartile 1	- 0.61	- 0.18	0.19	- 0.23
Median	0.03	0.15	0.63	0.30
Quartile 3	0.09	0.17	0.78	1.03
Maximum	0.32	0.61	1.34	1.50
Minimum	- 0.61	- 0.74	- 0.44	- 0.50
IQR	0.70	0.34	0.59	1.27
Upper Whisker	0.32	0.61	1.34	1.50
Lower Whisker	- 0.61	- 0.74	- 0.44	- 0.50

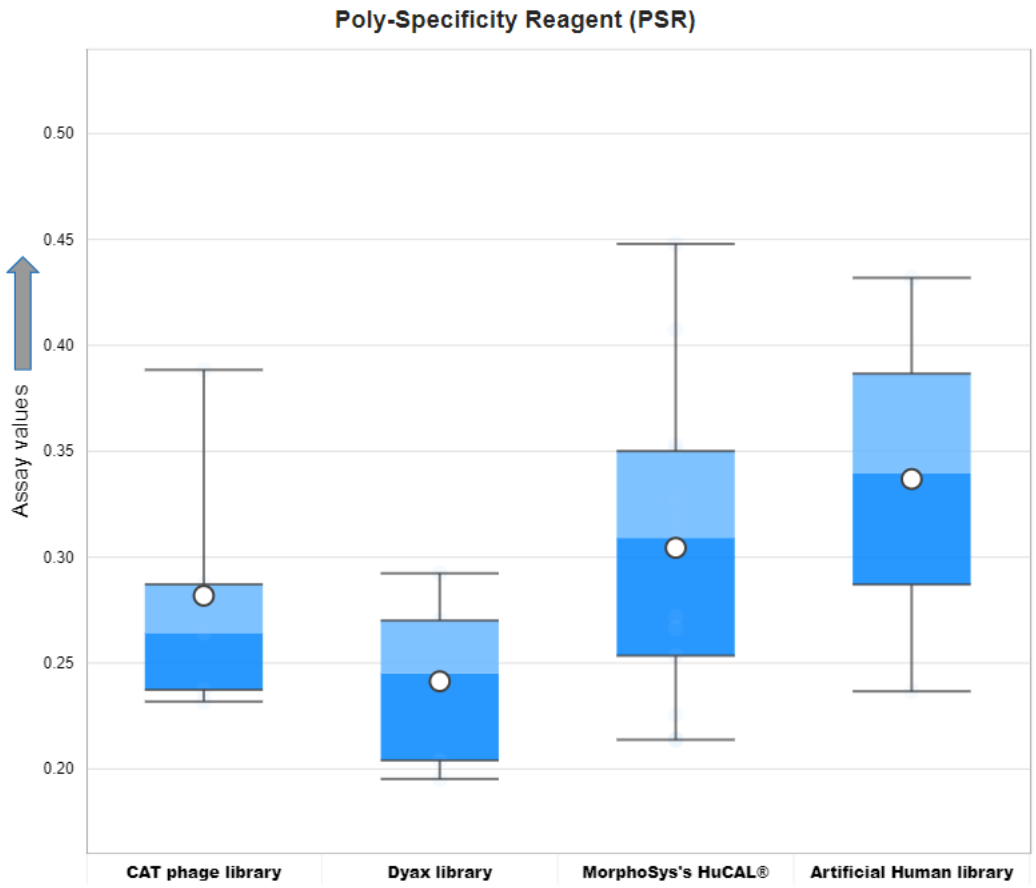
AbPred results for Affinity-Capture Self-Interaction Nanoparticle Spectroscopy (AC-SINS).



Expression Titer in HEK cells (HEK) assay values for different categories of antibody phage display platforms. The arrow on y-axis indicates the direction of unfavorable values.

	CAT phage library	Dyax phage library	MorphoSys's HuCAL®	Artificial Human library
Median Type	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max
Mean	151.45	162.51	158.71	156.57
Quartile 1	142.05	160.56	156.05	147.81
Median	147.20	161.60	159.23	150.78
Quartile 3	156.77	162.93	171.09	165.33
Maximum	172.29	166.93	184.35	178.52
Minimum	138.95	160.55	130.43	146.21
IQR	14.72	2.37	15.03	17.52
Upper Whisker	172.29	166.93	184.35	178.52
Lower Whisker	138.95	160.55	130.43	146.21

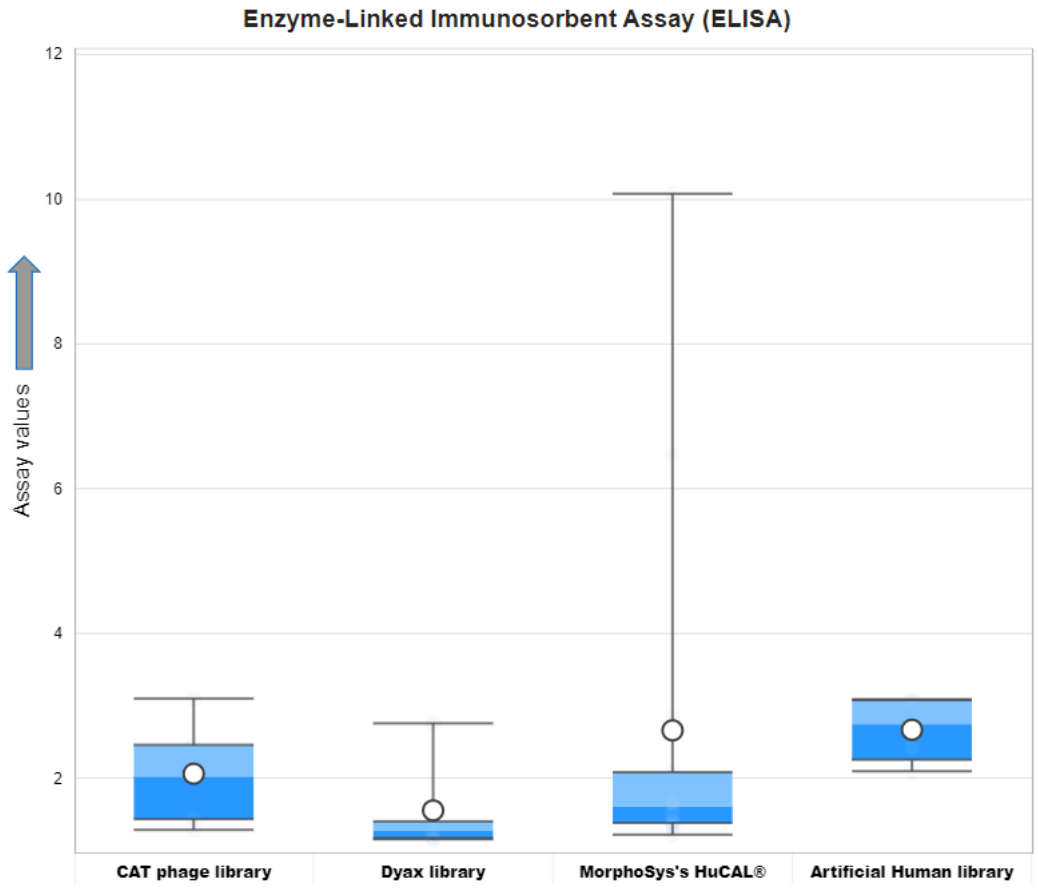
AbPred results for Expression Titer in HEK cells (HEK).



Poly-Specificity Reagent (PSR) assay values for different categories of antibody phage display platforms. The arrow on y-axis indicates the direction of unfavorable values.

	CAT phage library	Dyax phage library	MorphoSys's HuCAL®	Artificial Human library
Median Type	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max
Mean	0.28	0.24	0.30	0.34
Quartile 1	0.24	0.20	0.25	0.29
Median	0.26	0.25	0.31	0.34
Quartile 3	0.29	0.27	0.35	0.39
Maximum	0.39	0.29	0.45	0.43
Minimum	0.23	0.20	0.21	0.24
IQR	0.05	0.07	0.10	0.10
Upper Whisker	0.39	0.29	0.45	0.43
Lower Whisker	0.23	0.20	0.21	0.24

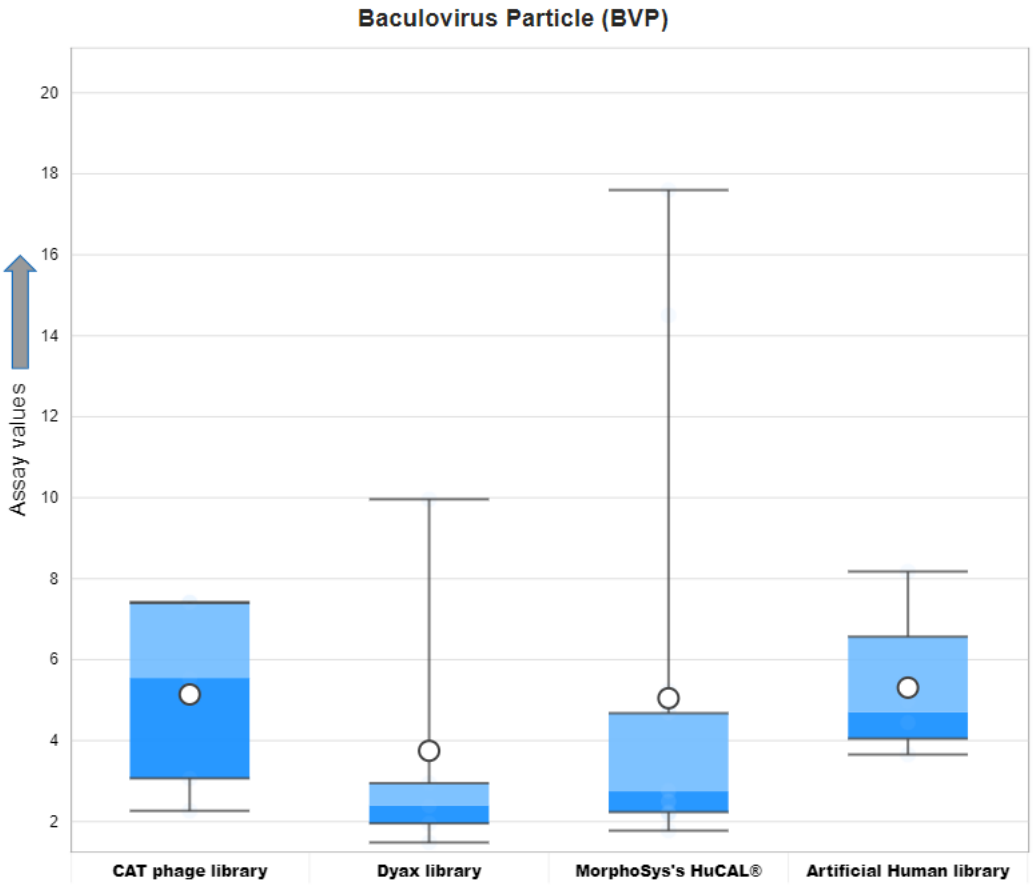
AbPred results for Poly-Specificity Reagent (PSR).



Enzyme-Linked Immunosorbent Assay (ELISA) values for different categories of antibody phage display platforms. The arrow on y-axis indicates the direction of unfavorable values.

	CAT phage library	Dyax phage library	MorphoSys's HuCAL®	Artificial Human library
Median Type	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max
Mean	2.06	1.55	2.66	2.67
Quartile 1	1.43	1.17	1.38	2.26
Median	2.02	1.28	1.61	2.74
Quartile 3	2.46	1.40	2.08	3.08
Maximum	3.10	2.76	10.08	3.09
Minimum	1.29	1.16	1.22	2.10
IQR	1.03	0.23	0.70	0.83
Upper Whisker	3.10	2.76	10.08	3.09
Lower Whisker	1.29	1.16	1.22	2.10

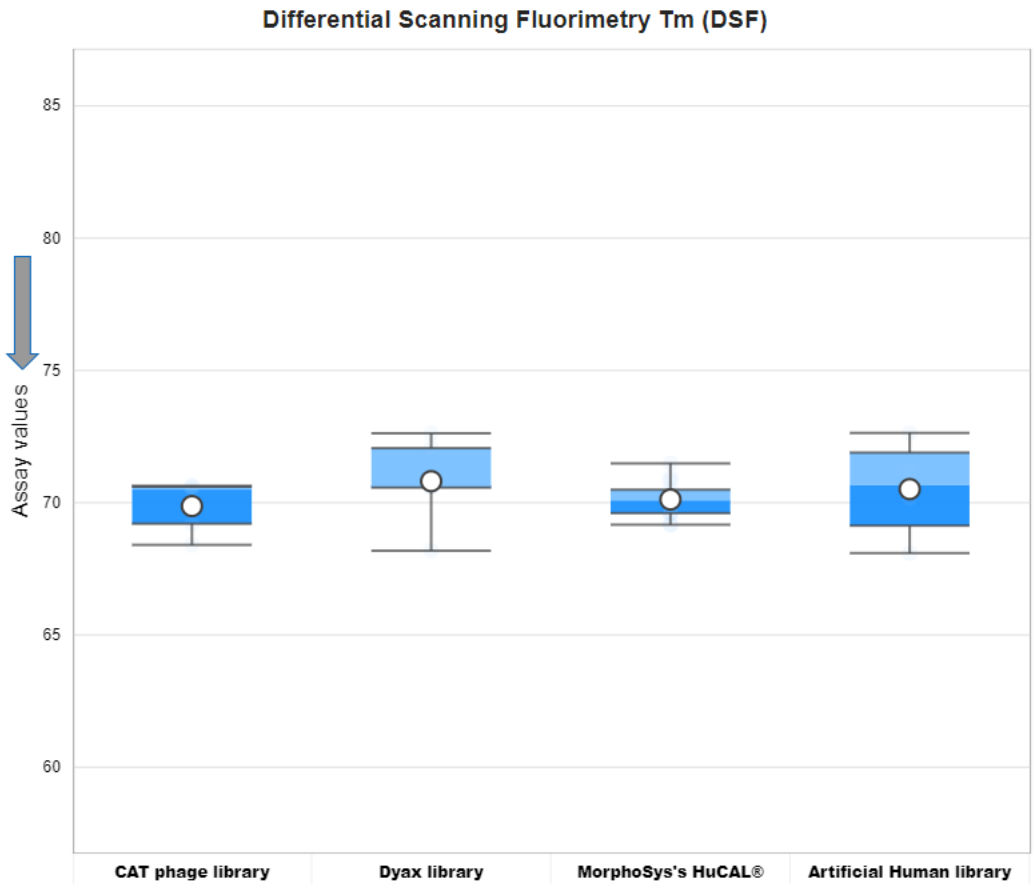
AbPred results for Enzyme-Linked Immunosorbent Assay (ELISA).



Baculovirus Particle (BVP) assay values for different categories of antibody phage display platforms. The arrow on y-axis indicates the direction of unfavorable values.

	CAT phage library	Dyax phage library	MorphoSys's HuCAL®	Artificial Human library
Median Type	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max
Mean	5.14	3.75	5.05	5.31
Quartile 1	3.07	1.96	2.24	4.05
Median	5.55	2.39	2.75	4.70
Quartile 3	7.40	2.95	4.68	6.57
Maximum	7.42	9.96	17.60	8.18
Minimum	2.27	1.48	1.78	3.66
IQR	4.33	0.99	2.44	2.52
Upper Whisker	7.42	9.96	17.60	8.18
Lower Whisker	2.27	1.48	1.78	3.66

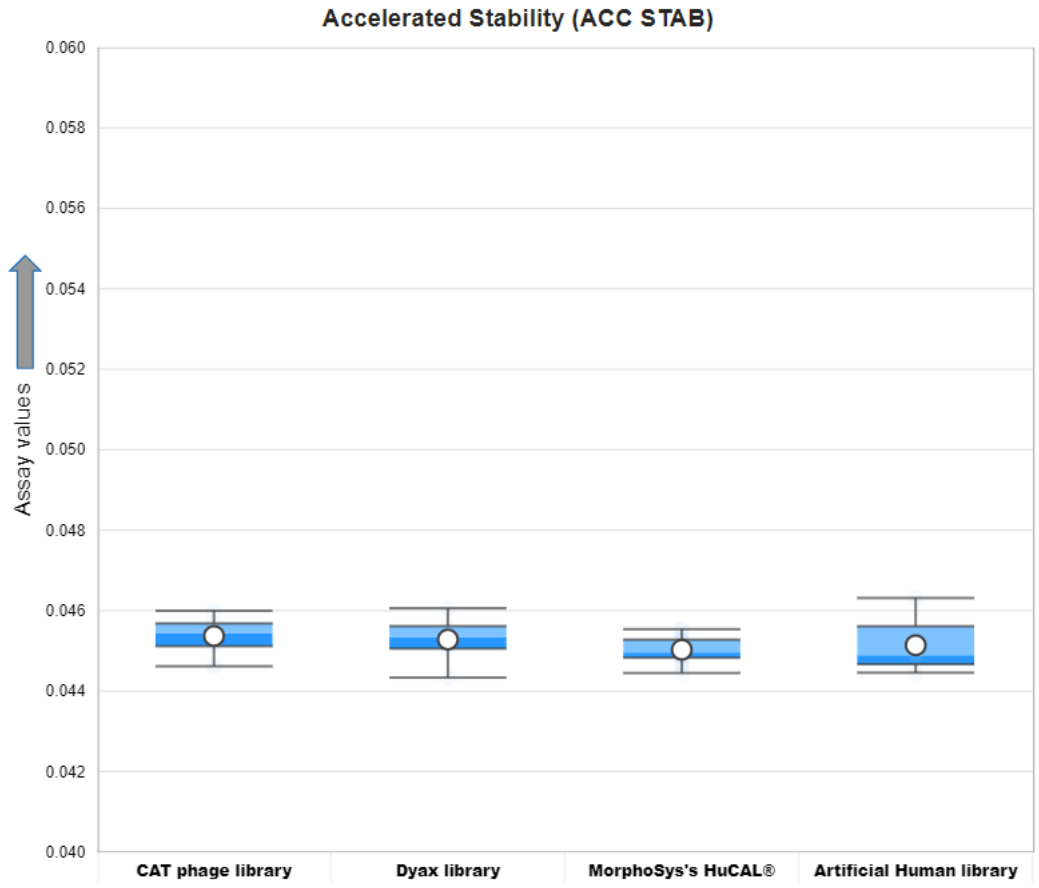
AbPred results for Baculovirus Particle (BVP).



Differential Scanning Fluorimetry (DSF) assay values for different categories of antibody phage display platforms. The arrow on y-axis indicates the direction of unfavorable values.

	CAT phage library	Dyax phage library	MorphoSys's HuCAL®	Artificial Human library
Median Type	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max
Mean	69.87	70.81	70.12	70.51
Quartile 1	69.21	70.57	69.60	69.13
Median	70.50	70.62	70.09	70.66
Quartile 3	70.61	72.06	70.49	71.89
Maximum	70.63	72.62	71.48	72.64
Minimum	68.40	68.18	69.17	68.09
IQR	1.41	1.49	0.89	2.76
Upper Whisker	70.63	72.62	71.48	72.64
Lower Whisker	68.40	68.18	69.17	68.09

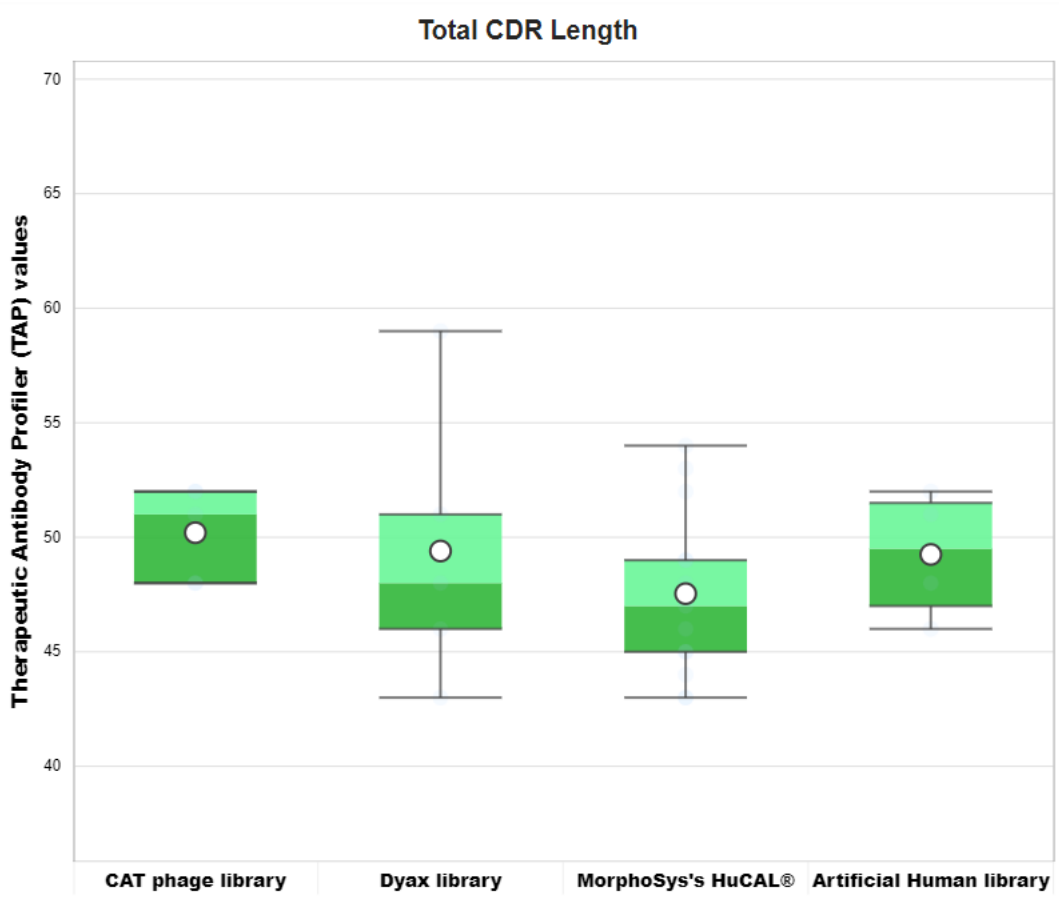
AbPred results for Differential Scanning Fluorimetry (DSF).



Accelerated Stability (ACC STAB) assay values for different categories of antibody phage display platforms. The arrow on y-axis indicates the direction of unfavorable values.

	CAT phage library	Dyax phage library	MorphoSys's HuCAL®	Artificial Human library
Median Type	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max
Mean	0.05	0.05	0.05	0.05
Quartile 1	0.05	0.05	0.04	0.04
Median	0.05	0.05	0.04	0.04
Quartile 3	0.05	0.05	0.05	0.05
Maximum	0.05	0.05	0.05	0.05
Minimum	0.04	0.04	0.04	0.04
IQR	0.00	0.00	0.00	0.00
Upper Whisker	0.05	0.05	0.05	0.05
Lower Whisker	0.04	0.04	0.04	0.04

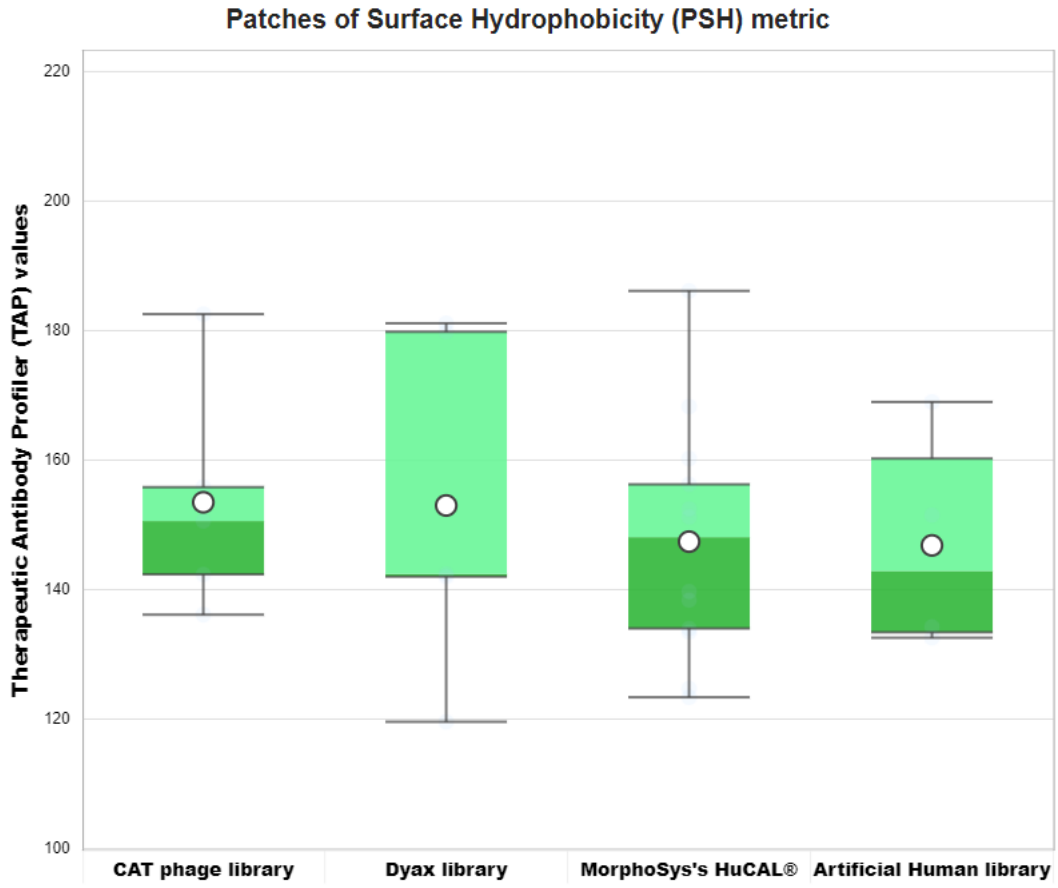
AbPred results for Accelerated Stability (ACC STAB).



Total CDR length metric values for different categories of phage display antibodies.

	CAT phage library	Dyax phage library	MorphoSys's HuCAL®	Artificial Human library
Median Type	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max
Mean	50	49	48	49
Quartile 1	48	46	45	47
Median	51	48	47	50
Quartile 3	52	51	49	52
Maximum	52	59	54	52
Minimum	48	43	43	46
IQR	4	5	4	5
Upper Whisker	52	59	54	52
Lower Whisker	48	43	43	46

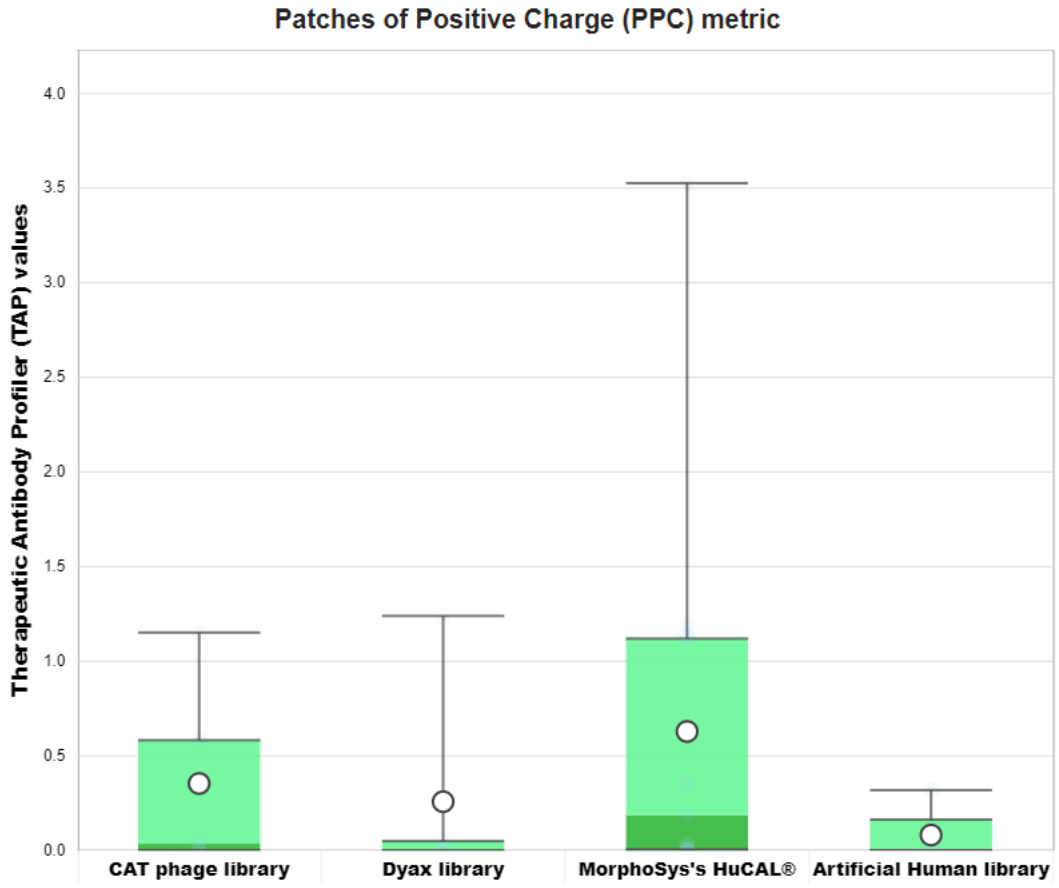
Therapeutic Antibody Profiler (TAP) statistics for Total CDR length metric.



Patches of Surface Hydrophobicity (PSH) metric values for different categories of phage display antibodies. PSH is calculated across the CDR vicinity.

	CAT phage library	Dyax phage library	MorphoSys's HuCAL®	Artificial Human library
Median Type	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max
Mean	153.49	153.00	147.41	146.83
Quartile 1	142.36	142.01	134.01	133.42
Median	150.59	142.43	148.11	142.89
Quartile 3	155.83	179.82	156.26	160.25
Maximum	182.55	181.13	186.14	168.98
Minimum	136.13	119.59	123.38	132.57
IQR	13.47	37.81	22.26	26.83
Upper Whisker	182.55	181.13	186.14	168.98
Lower Whisker	136.13	119.59	123.38	132.57

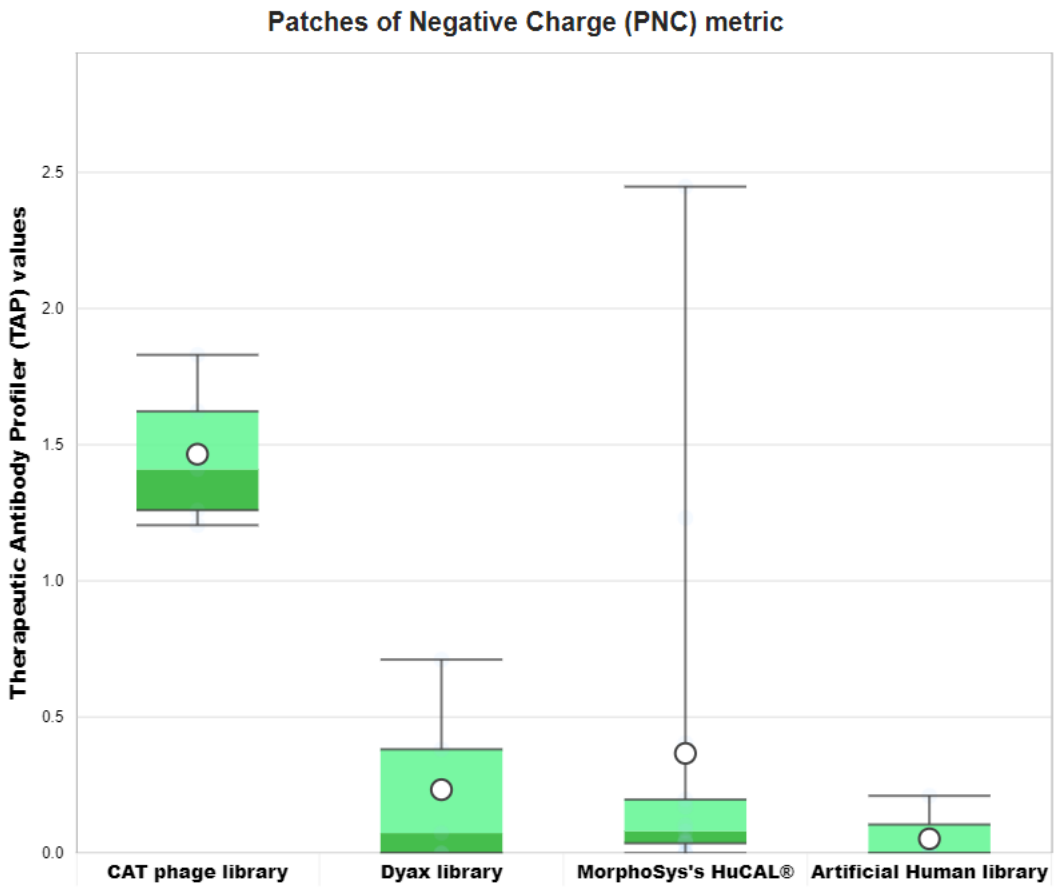
Therapeutic Antibody Profiler (TAP) statistics for Patches of Surface Hydrophobicity (PSH) metric.



Patches of Positive Charge (PPC) metric values for different categories of phage display antibodies. PPC is calculated across the CDR vicinity.

	CAT phage library	Dyax phage library	MorphoSys's HuCAL®	Artificial Human library
Median Type	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max
Mean	0.35	0.26	0.63	0.08
Quartile 1	0.00	0.00	0.01	0.00
Median	0.04	0.00	0.19	0.00
Quartile 3	0.58	0.05	1.12	0.16
Maximum	1.15	1.24	3.53	0.32
Minimum	0.00	0.00	0.00	0.00
IQR	0.58	0.05	1.11	0.16
Upper Whisker	1.15	1.24	3.53	0.32
Lower Whisker	0.00	0.00	0.00	0.00

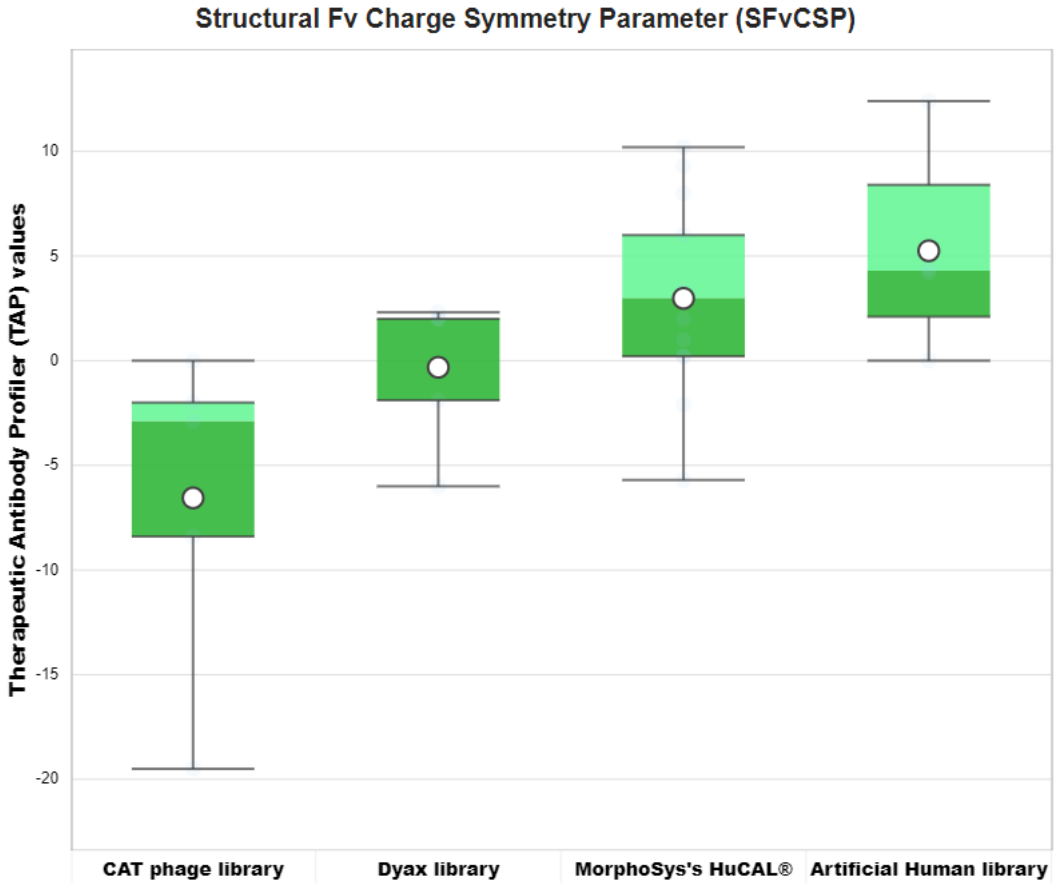
Therapeutic Antibody Profiler (TAP) statistics for Patches of Positive Charge (PPC) metric.



Patches of Negative Charge (PNC) metric values for different categories of phage display antibodies. PNC is calculated across the CDR vicinity.

	CAT phage library	Dyax phage library	MorphoSys's HuCAL®	Artificial Human library
Median Type	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max
Mean	1.47	0.23	0.37	0.05
Quartile 1	1.26	0.00	0.04	0.00
Median	1.41	0.07	0.08	0.00
Quartile 3	1.62	0.38	0.20	0.11
Maximum	1.83	0.71	2.45	0.21
Minimum	1.20	0.00	0.00	0.00
IQR	0.36	0.38	0.16	0.11
Upper Whisker	1.83	0.71	2.45	0.21
Lower Whisker	1.20	0.00	0.00	0.00

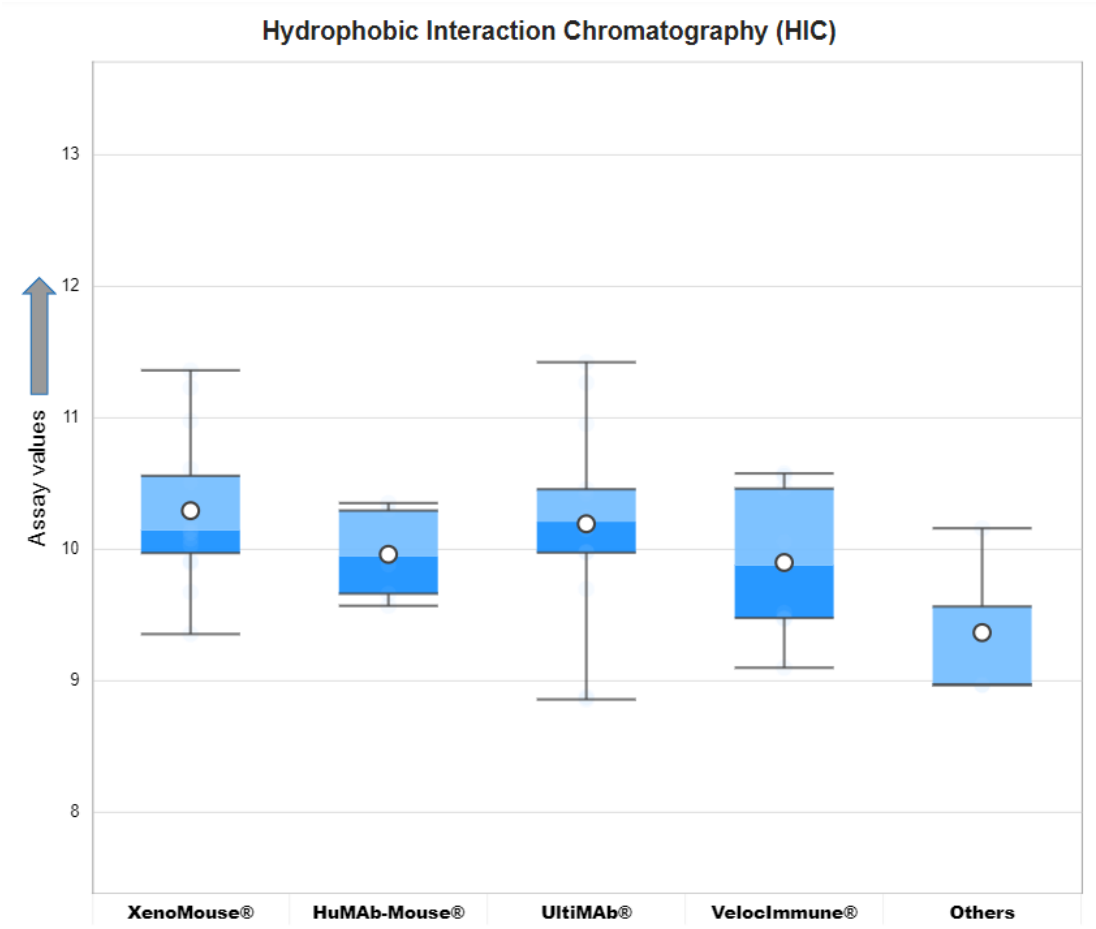
Therapeutic Antibody Profiler (TAP) statistics for Patches of Negative Charge (PNC) metric.



Structural Fv Charge Symmetry Parameter (SFvCSP) values for different categories of phage display antibodies.

	CAT phage library	Dyax phage library	MorphoSys's HuCAL®	Artificial Human library
Median Type	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max
Mean	- 6.56	- 0.32	2.98	5.25
Quartile 1	- 8.40	- 1.89	0.21	2.10
Median	- 2.88	2.00	3.00	4.30
Quartile 3	- 2.00	2.00	6.00	8.40
Maximum	0.00	2.31	10.20	12.40
Minimum	- 19.50	- 6.00	- 5.70	0.00
IQR	6.40	3.89	5.79	6.30
Upper Whisker	0.00	2.31	10.20	12.40
Lower Whisker	- 19.50	- 6.00	- 5.70	0.00

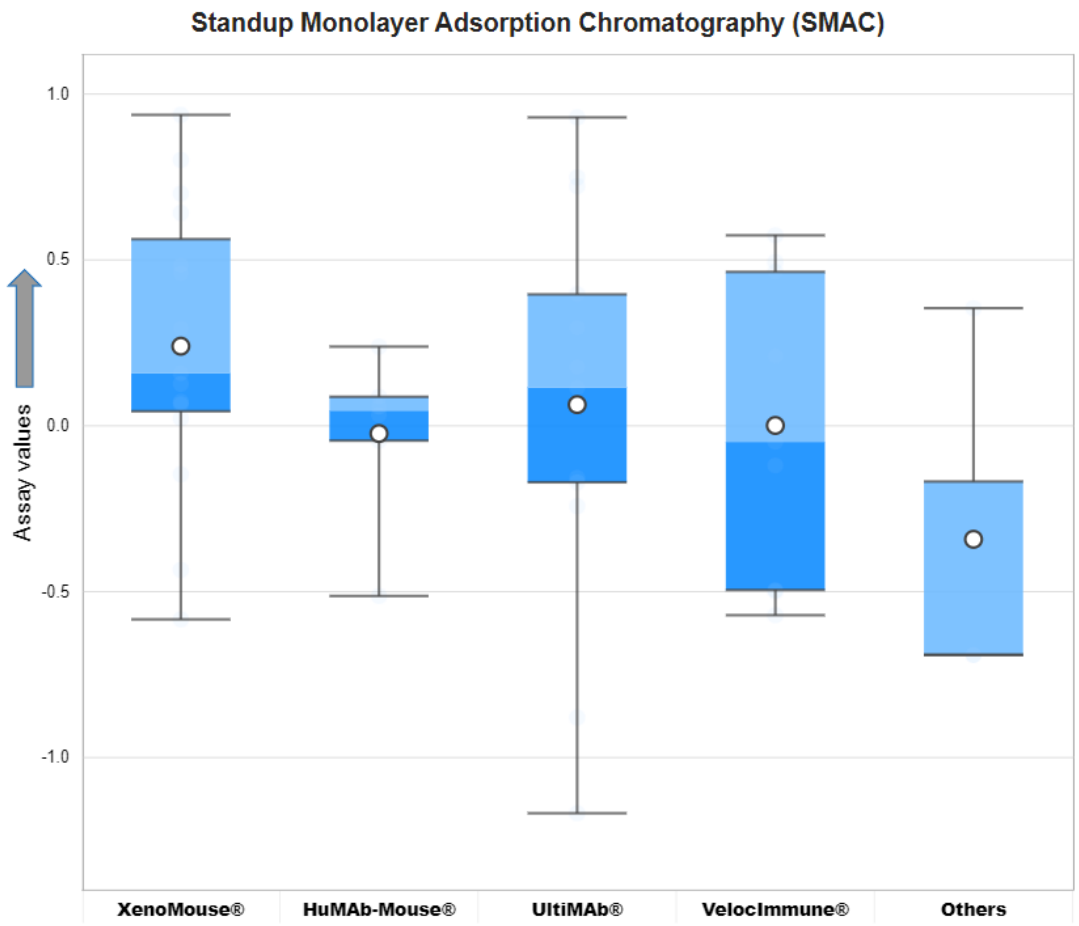
Therapeutic Antibody Profiler (TAP) statistics for Structural Fv Charge Symmetry Parameter (SFvCSP).



Hydrophobic Interaction Chromatography (HIC) assay values for different categories of transgenic mouse platforms. The arrow on y-axis indicates the direction of unfavorable values.

	XenoMouse®	HuMAb-Mouse®	UltiMAb®	VelocImmune®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	10.29	9.96	10.19	9.90	9.37
Quartile 1	9.97	9.66	9.98	9.48	8.97
Median	10.14	9.95	10.21	9.88	8.97
Quartile 3	10.56	10.29	10.46	10.46	9.57
Maximum	11.36	10.35	11.42	10.58	10.16
Minimum	9.36	9.57	8.86	9.10	8.97
IQR	0.59	0.63	0.48	0.98	0.60
Upper Whisker	11.36	10.35	11.42	10.58	10.16
Lower Whisker	9.36	9.57	8.86	9.10	8.97

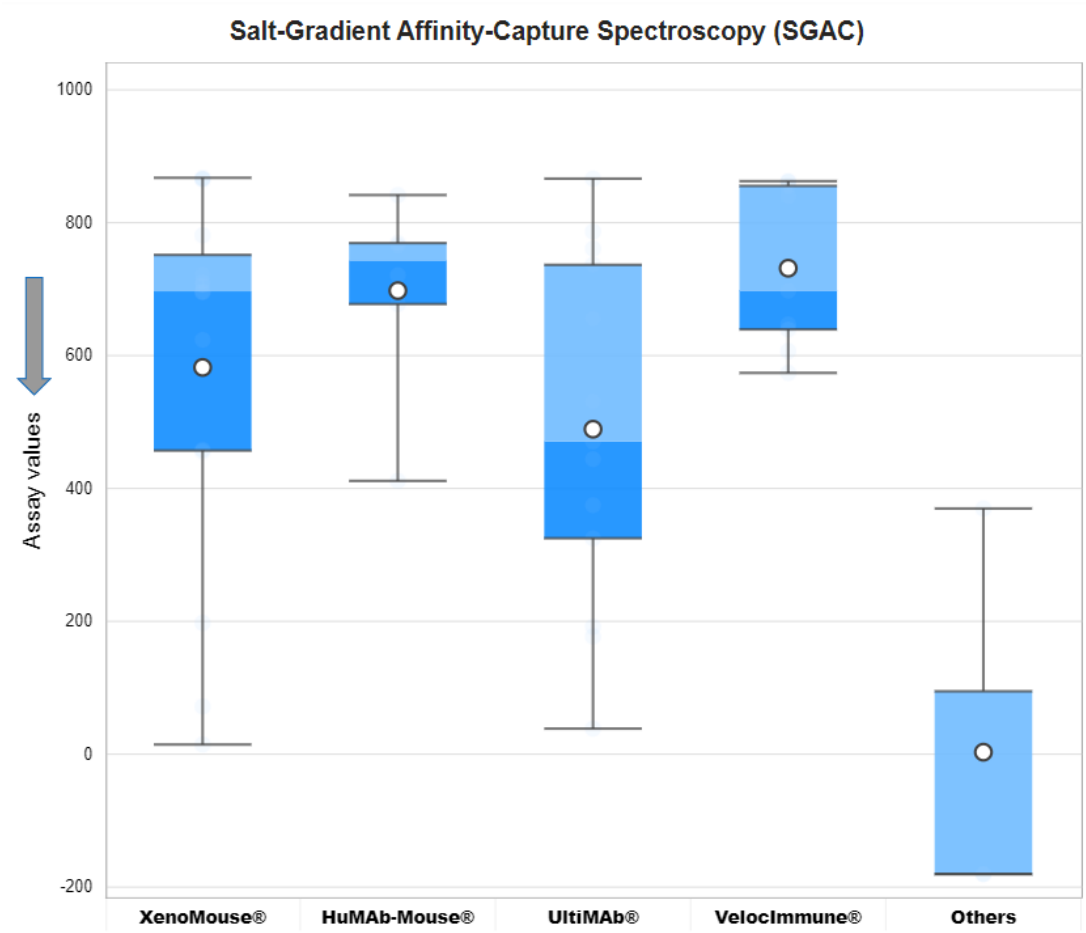
AbPred results for Hydrophobic Interaction Chromatography (HIC).



Standup Monolayer Absorption Chromatography (SMAC) assay values for different categories of transgenic mouse platforms. The arrow on y-axis indicates the direction of unfavorable values.

	XenoMouse®	HuMAb-Mouse®	UltiMAb®	VelocImmune®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	0.24	- 0.02	0.06	0.00	- 0.34
Quartile 1	0.04	- 0.04	- 0.17	- 0.49	- 0.69
Median	0.16	0.05	0.12	- 0.05	- 0.69
Quartile 3	0.56	0.09	0.40	0.46	- 0.17
Maximum	0.94	0.24	0.93	0.57	0.35
Minimum	- 0.58	- 0.51	- 1.17	- 0.57	- 0.69
IQR	0.52	0.13	0.57	0.96	0.52
Upper Whisker	0.94	0.24	0.93	0.57	0.35
Lower Whisker	- 0.58	- 0.51	- 1.17	- 0.57	- 0.69

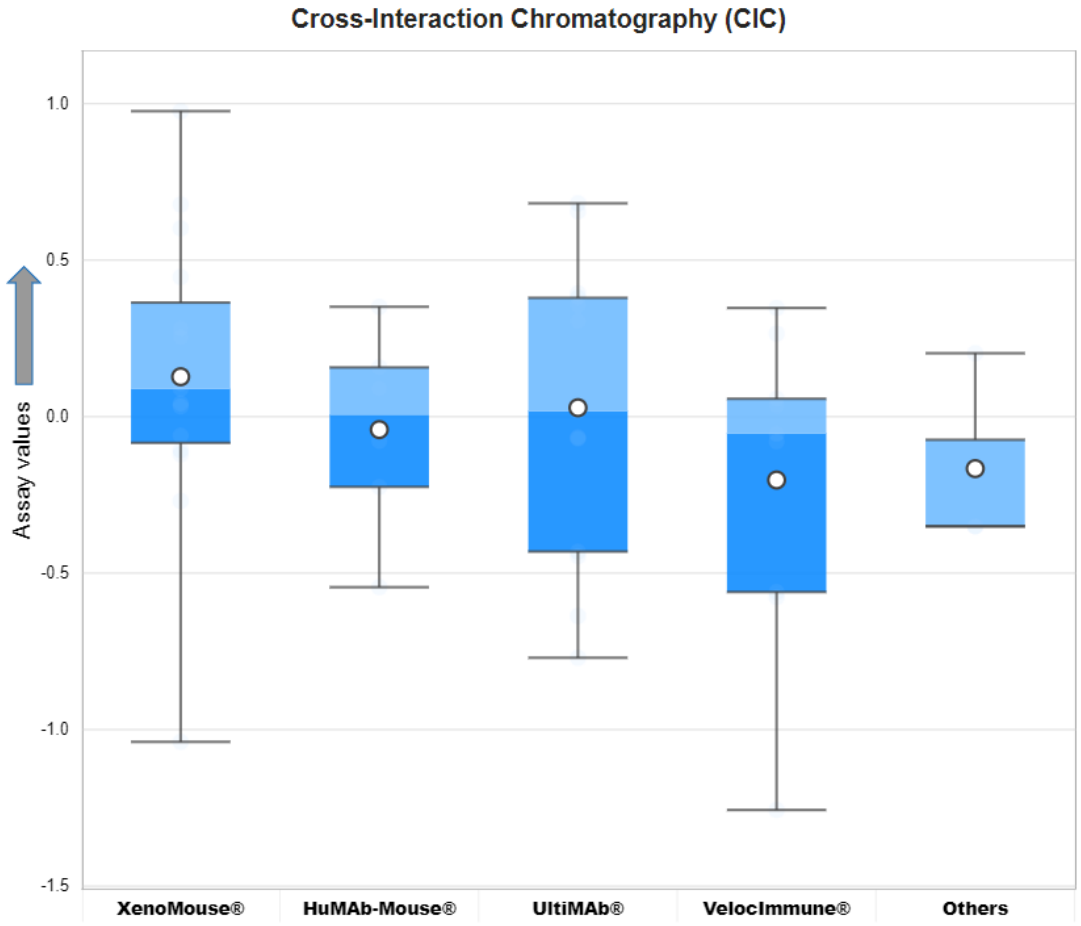
AbPred results for Standup Monolayer Absorption Chromatography (SMAC).



Salt-Gradient Affinity-Capture Spectroscopy (SGAC) assay values for different categories of transgenic mouse platforms. The arrow on y-axis indicates the direction of unfavorable values.

	XenoMouse®	HuMAb-Mouse®	UltiMAb®	VelocImmune®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	582.16	697.74	489.20	731.51	3.06
Quartile 1	457.05	677.62	325.04	639.41	- 180.31
Median	696.70	743.19	470.67	697.74	- 180.31
Quartile 3	751.57	769.39	736.59	855.40	94.75
Maximum	867.59	841.67	866.33	862.54	369.81
Minimum	14.57	411.39	38.53	573.95	- 180.31
IQR	294.52	91.77	411.55	215.99	275.06
Upper Whisker	867.59	841.67	866.33	862.54	369.81
Lower Whisker	14.57	411.39	38.53	573.95	- 180.31

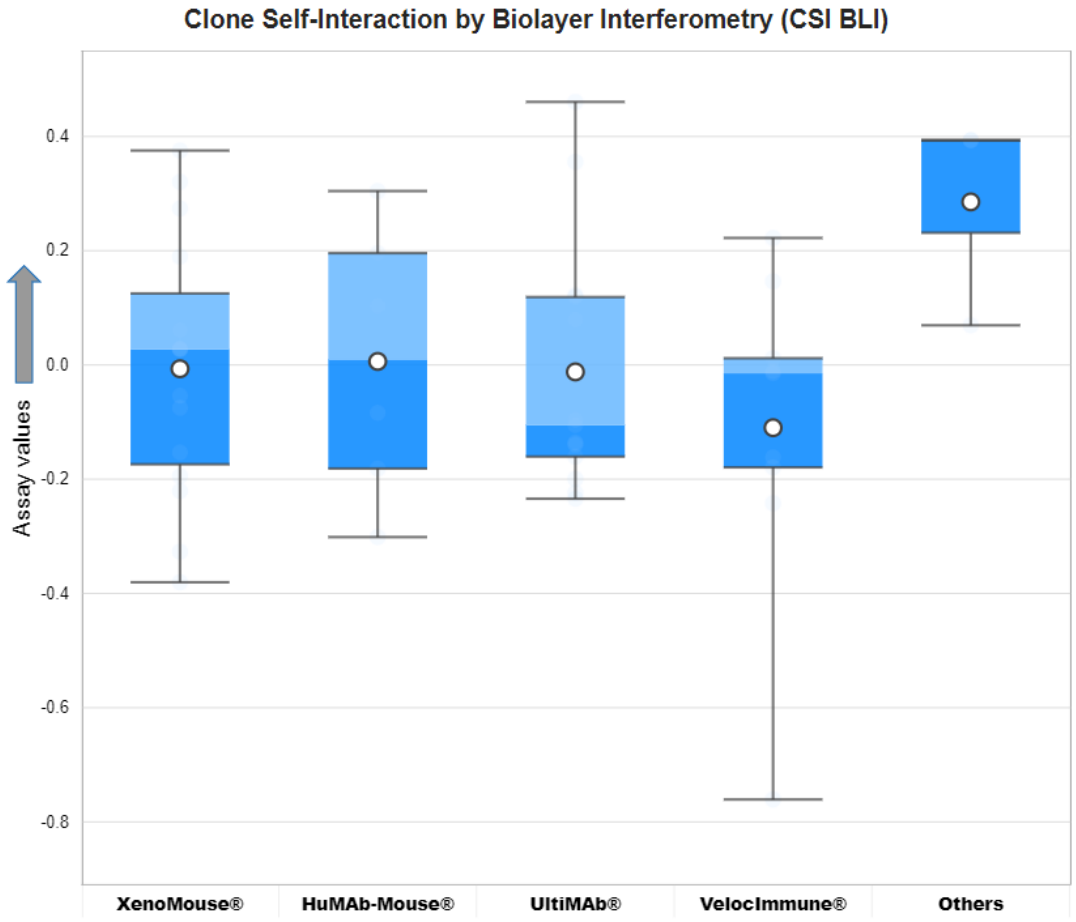
AbPred results for Salt-Gradient Affinity-Capture Spectroscopy (SGAC).



Cross-Interaction Chromatography (CIC) assay values for different categories of transgenic mouse platforms. The arrow on y-axis indicates the direction of unfavorable values.

	XenoMouse®	HuMAb-Mouse®	UltiMAb®	VelocImmune®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	0.13	- 0.04	0.03	- 0.20	- 0.17
Quartile 1	- 0.08	- 0.22	- 0.43	- 0.56	- 0.35
Median	0.09	0.01	0.02	- 0.05	- 0.35
Quartile 3	0.36	0.16	0.38	0.06	- 0.07
Maximum	0.98	0.35	0.68	0.35	0.20
Minimum	- 1.04	- 0.54	- 0.77	- 1.26	- 0.35
IQR	0.45	0.38	0.81	0.62	0.28
Upper Whisker	0.98	0.35	0.68	0.35	0.20
Lower Whisker	- 1.04	- 0.54	- 0.77	- 1.26	- 0.35

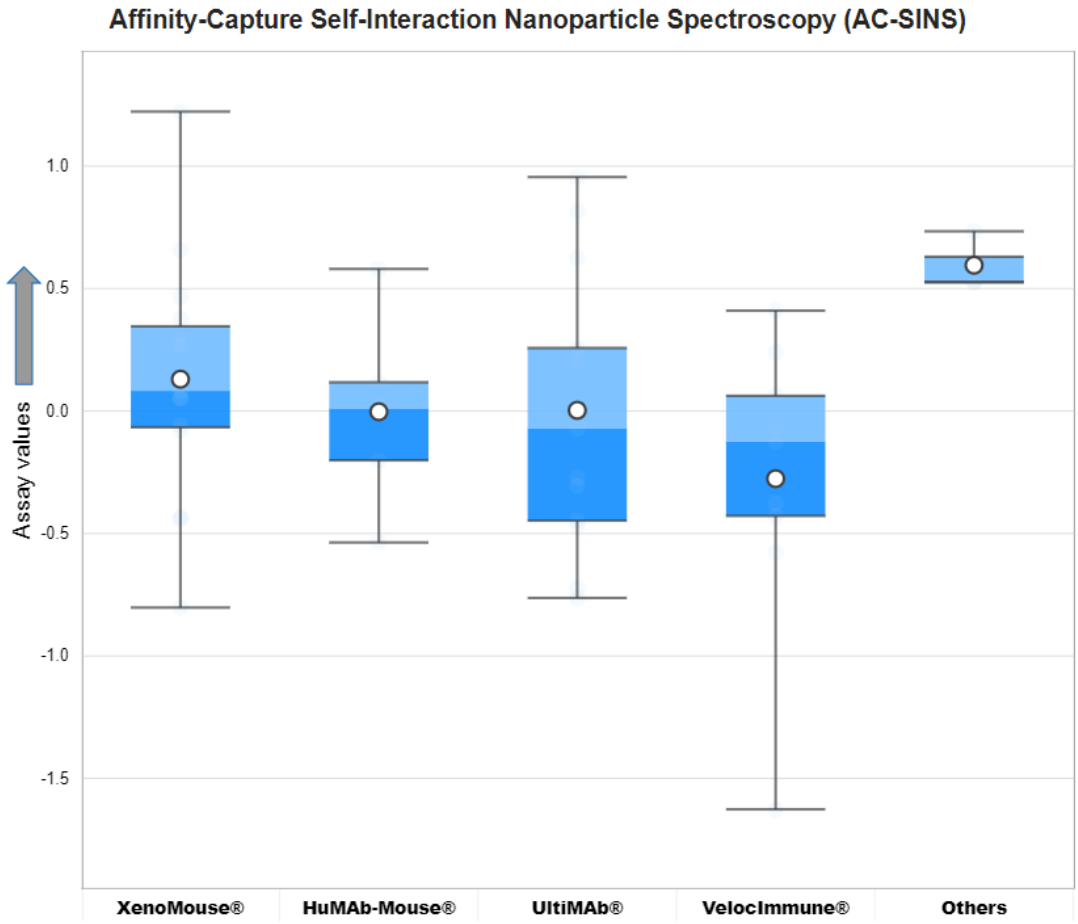
AbPred results for Cross-Interaction Chromatography (CIC).



Clone Self-Interaction by Biolayer Interferometry (CSI BLI) assay values for different categories of transgenic mouse platforms. The arrow on y-axis indicates the direction of unfavorable values.

	XenoMouse®	HuMAb-Mouse®	UltiMAb®	VelocImmune®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	- 0.01	0.01	- 0.01	- 0.11	0.29
Quartile 1	- 0.17	- 0.18	- 0.16	- 0.18	0.23
Median	0.03	0.01	- 0.11	- 0.01	0.39
Quartile 3	0.13	0.20	0.12	0.01	0.39
Maximum	0.38	0.30	0.46	0.22	0.39
Minimum	- 0.38	- 0.30	- 0.23	- 0.76	0.07
IQR	0.30	0.38	0.28	0.19	0.16
Upper Whisker	0.38	0.30	0.46	0.22	0.39
Lower Whisker	- 0.38	- 0.30	- 0.23	- 0.76	0.07

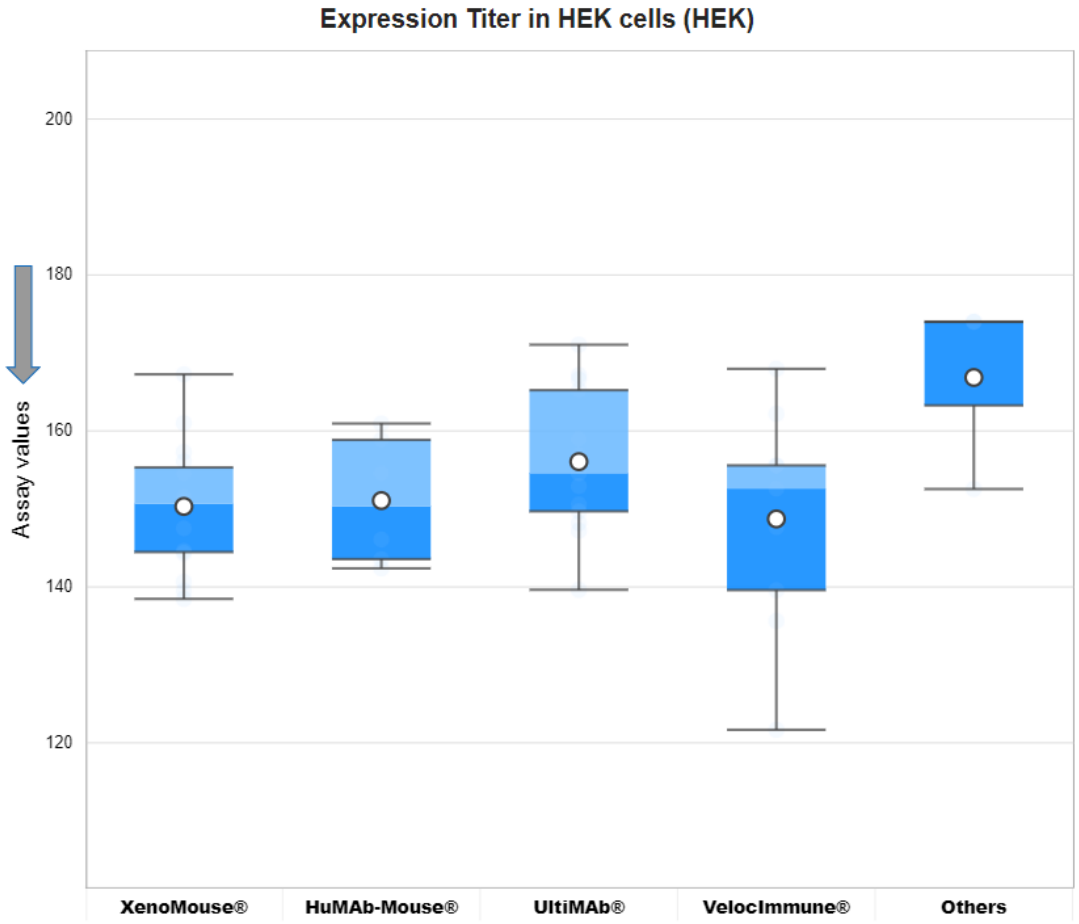
AbPred results for Clone Self-Interaction by Biolayer Interferometry (CSI BLI).



Affinity-Capture Self-Interaction Nanoparticle Spectroscopy (AC-SINS) assay values for different categories of transgenic mouse platforms. The arrow on y-axis indicates the direction of unfavorable values.

	XenoMouse®	HuMAb-Mouse®	UltiMAb®	VelocImmune®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	0.13	0.00	0.00	- 0.28	0.60
Quartile 1	- 0.07	- 0.20	- 0.45	- 0.43	0.53
Median	0.08	0.01	- 0.07	- 0.12	0.53
Quartile 3	0.35	0.12	0.26	0.06	0.63
Maximum	1.22	0.58	0.96	0.41	0.73
Minimum	- 0.80	- 0.54	- 0.76	- 1.63	0.53
IQR	0.41	0.32	0.71	0.49	0.10
Upper Whisker	1.22	0.58	0.96	0.41	0.73
Lower Whisker	- 0.80	- 0.54	- 0.76	- 1.63	0.53

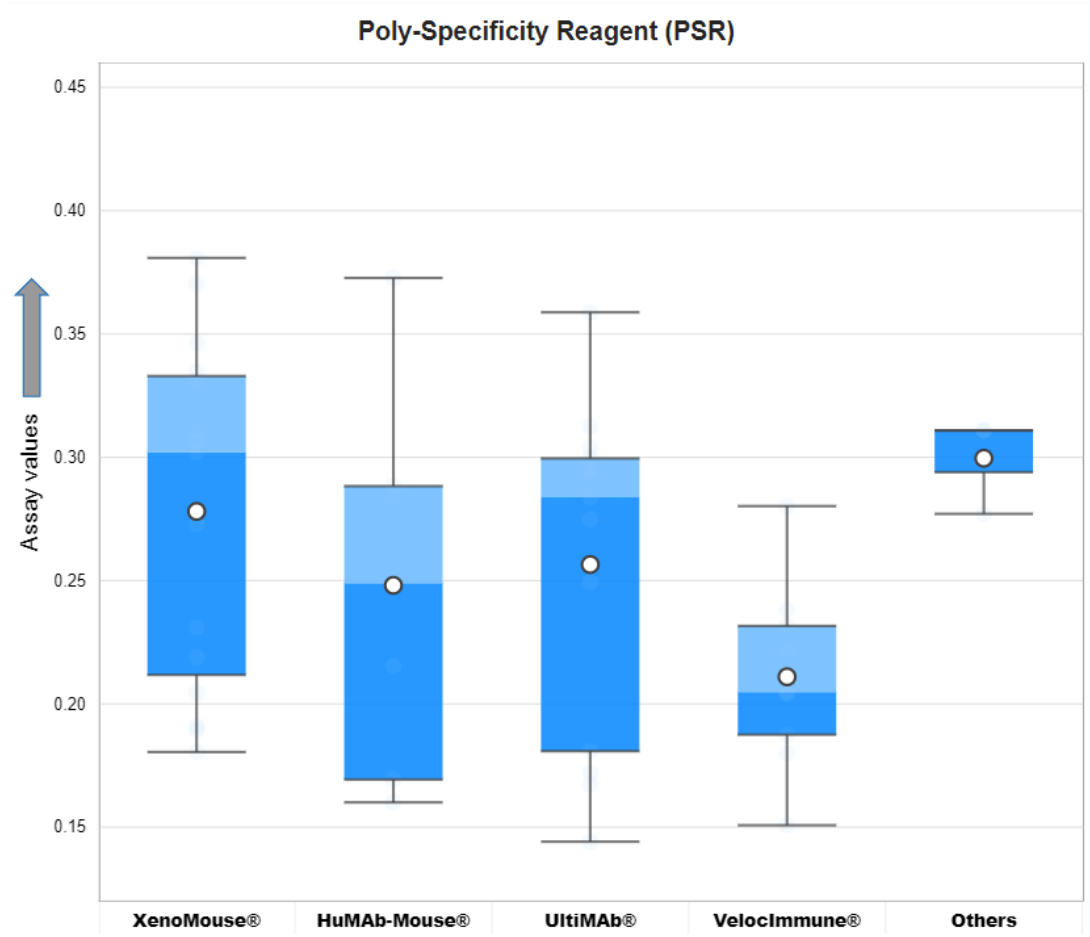
AbPred results for Affinity-Capture Self-Interaction Nanoparticle Spectroscopy (AC-SINS).



Expression Titer in HEK cells (HEK) assay values for different categories of transgenic mouse platforms. The arrow on y-axis indicates the direction of unfavorable values.

	XenoMouse®	HuMAb-Mouse®	UltiMAb®	VelocImmune®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	150.32	151.07	156.06	148.71	166.85
Quartile 1	144.47	143.56	149.69	139.59	163.28
Median	150.65	150.33	154.59	152.61	174.00
Quartile 3	155.32	158.86	165.23	155.58	174.00
Maximum	167.26	160.95	171.06	167.96	174.00
Minimum	138.46	142.38	139.63	121.67	152.55
IQR	10.85	15.30	15.54	15.99	10.73
Upper Whisker	167.26	160.95	171.06	167.96	174.00
Lower Whisker	138.46	142.38	139.63	121.67	152.55

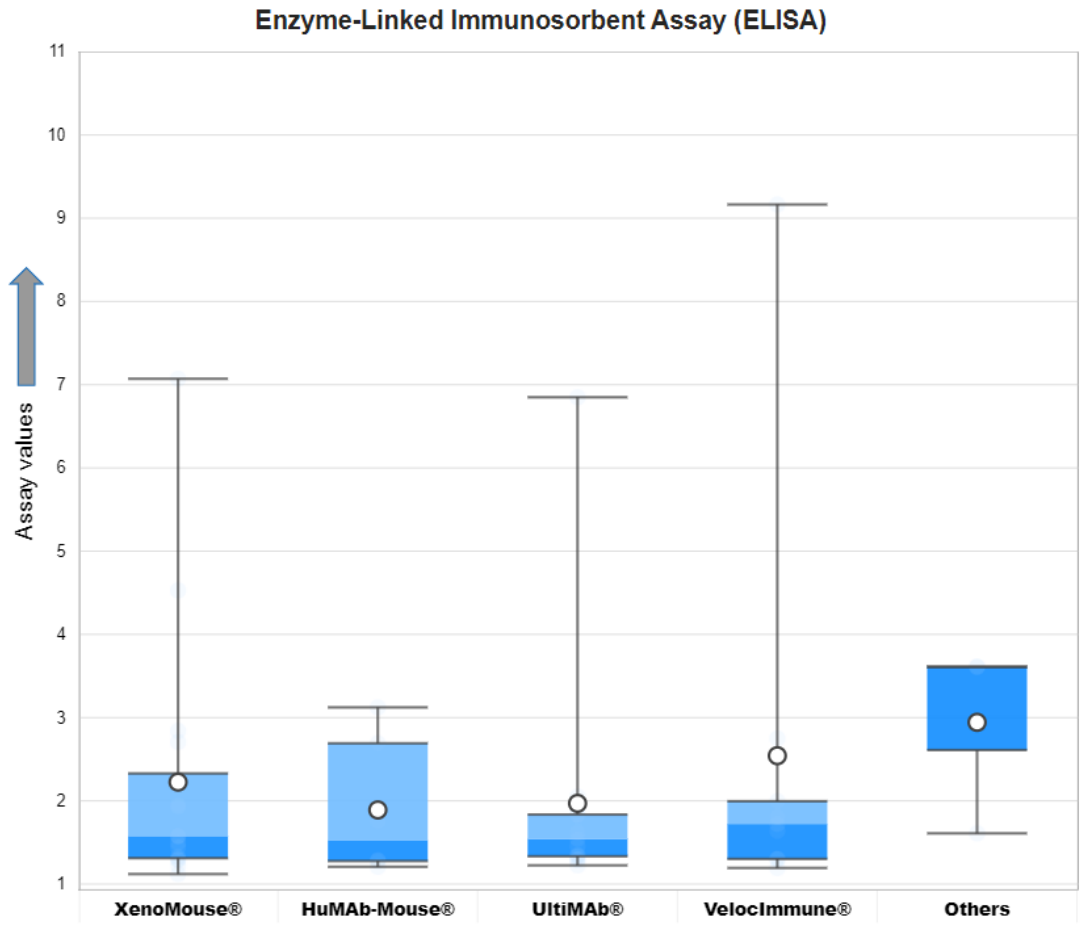
AbPred results for Expression Titer in HEK cells (HEK).



Poly-Specificity Reagent (PSR) assay values for different categories of transgenic mouse platforms. The arrow on y-axis indicates the direction of unfavorable values.

	XenoMouse®	HuMAb-Mouse®	UltiMAb®	VelocImmune®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	0.28	0.25	0.26	0.21	0.30
Quartile 1	0.21	0.17	0.18	0.19	0.29
Median	0.30	0.25	0.28	0.20	0.31
Quartile 3	0.33	0.29	0.30	0.23	0.31
Maximum	0.38	0.37	0.36	0.28	0.31
Minimum	0.18	0.16	0.14	0.15	0.28
IQR	0.12	0.12	0.12	0.04	0.02
Upper Whisker	0.38	0.37	0.36	0.28	0.31
Lower Whisker	0.18	0.16	0.14	0.15	0.28

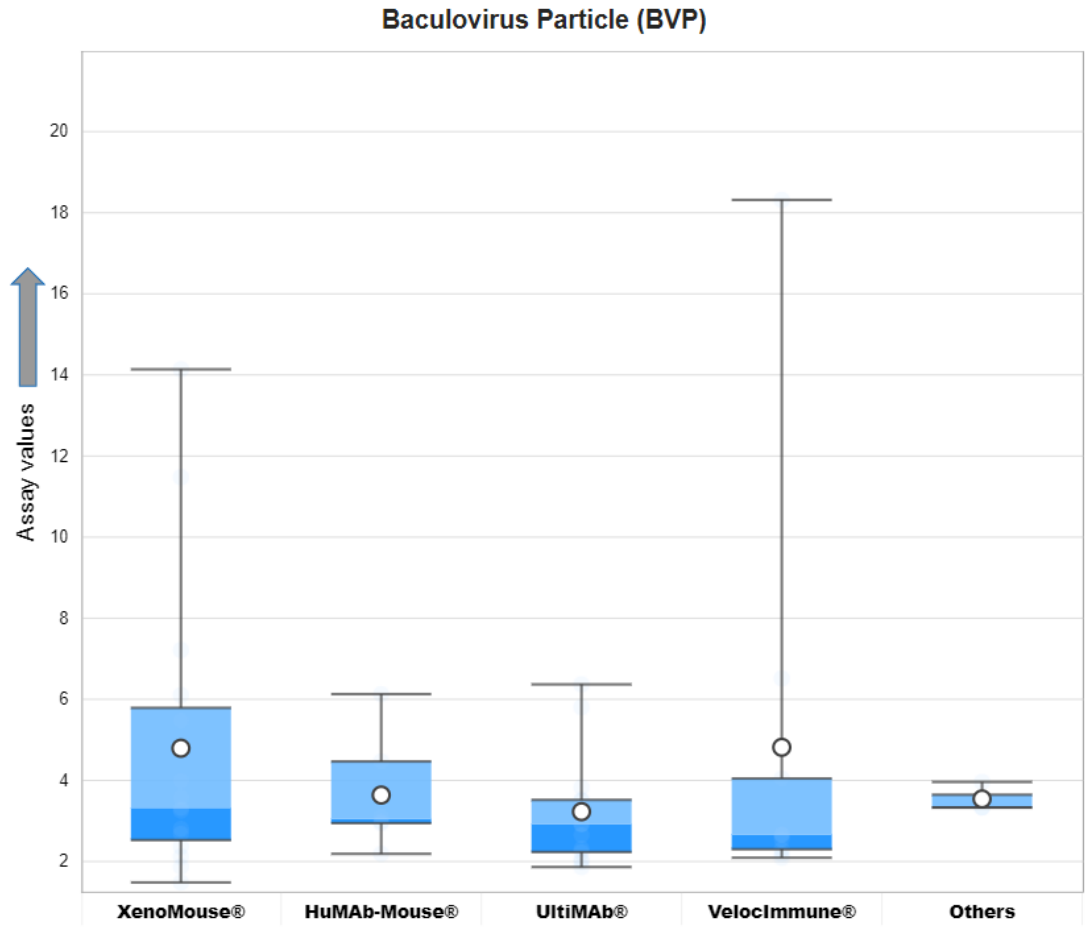
AbPred results for Poly-Specificity Reagent (PSR).



Enzyme-Linked Immunosorbent Assay (ELISA) values for different categories of transgenic mouse platforms. The arrow on y-axis indicates the direction of unfavorable values.

	XenoMouse®	HuMAb-Mouse®	UltiMAb®	VelocImmune®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	2.23	1.89	1.97	2.54	2.94
Quartile 1	1.31	1.28	1.33	1.30	2.61
Median	1.58	1.52	1.54	1.72	3.61
Quartile 3	2.33	2.69	1.84	2.00	3.61
Maximum	7.07	3.12	6.85	9.16	3.61
Minimum	1.12	1.21	1.23	1.19	1.61
IQR	1.02	1.41	0.50	0.69	1.00
Upper Whisker	7.07	3.12	6.85	9.16	3.61
Lower Whisker	1.12	1.21	1.23	1.19	1.61

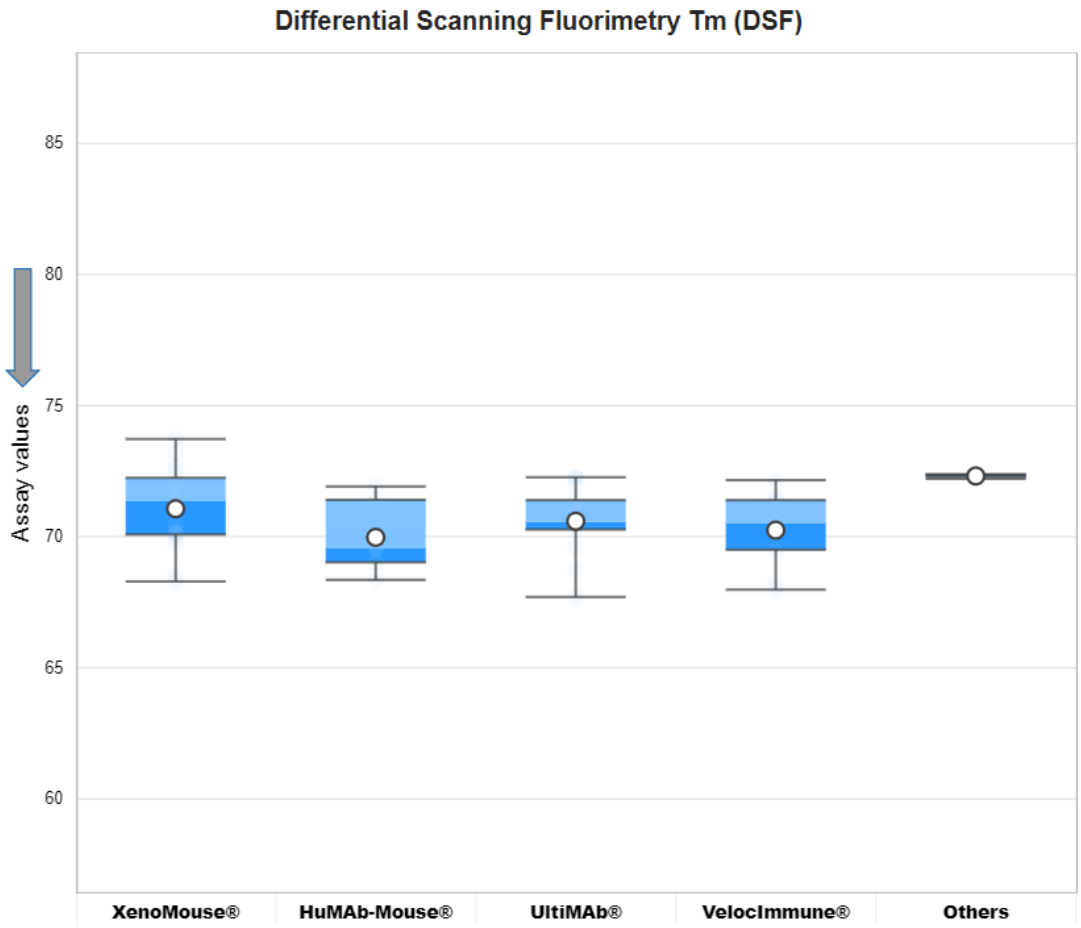
AbPred results for Enzyme-Linked Immunosorbent Assay (ELISA).



Baculovirus Particle (BVP) assay values for different categories of transgenic mouse platforms. The arrow on y-axis indicates the direction of unfavorable values.

	XenoMouse®	HuMAb-Mouse®	UltiMAb®	VelocImmune®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	4.79	3.64	3.23	4.82	3.54
Quartile 1	2.53	2.95	2.23	2.30	3.33
Median	3.32	3.05	2.92	2.66	3.33
Quartile 3	5.79	4.47	3.52	4.05	3.65
Maximum	14.14	6.13	6.37	18.32	3.96
Minimum	1.48	2.19	1.87	2.09	3.33
IQR	3.26	1.52	1.29	1.75	0.32
Upper Whisker	14.14	6.13	6.37	18.32	3.96
Lower Whisker	1.48	2.19	1.87	2.09	3.33

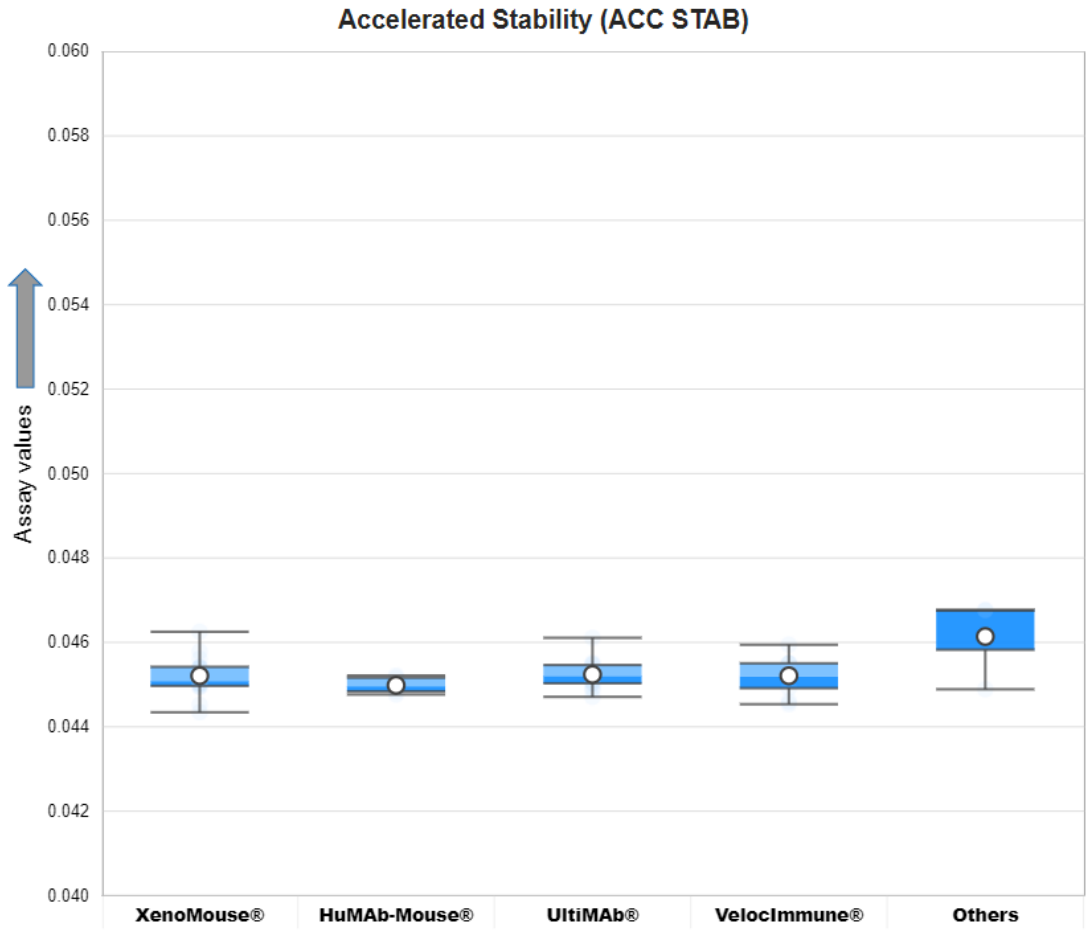
AbPred results for Baculovirus Particle (BVP).



Differential Scanning Fluorimetry (DSF) assay values for different categories of transgenic mouse platforms. The arrow on y-axis indicates the direction of unfavorable values.

	XenoMouse®	HuMAb-Mouse®	UltiMAb®	VelocImmune®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	71.07	69.98	70.59	70.25	72.32
Quartile 1	70.09	69.03	70.28	69.50	72.29
Median	71.36	69.58	70.57	70.53	72.37
Quartile 3	72.25	71.41	71.40	71.40	72.37
Maximum	73.73	71.92	72.27	72.16	72.37
Minimum	68.30	68.35	67.70	67.98	72.21
IQR	2.16	2.38	1.12	1.89	0.08
Upper Whisker	73.73	71.92	72.27	72.16	72.37
Lower Whisker	68.30	68.35	67.70	67.98	72.21

AbPred results for Differential Scanning Fluorimetry (DSF).



Accelerated Stability (ACC STAB) assay values for different categories of transgenic mouse platforms. The arrow on y-axis indicates the direction of unfavorable values.

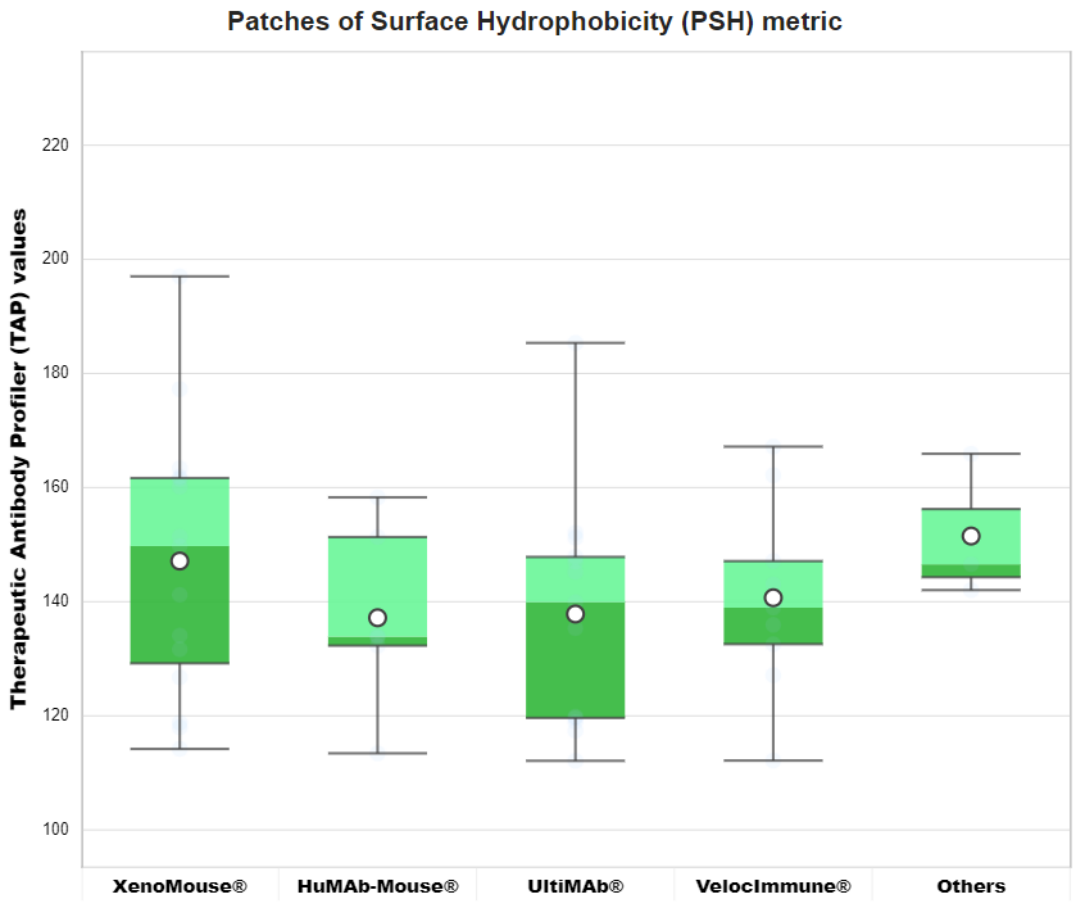
	XenoMouse®	HuMAb-Mouse®	UltiMAb®	VelocImmune®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	0.05	0.04	0.05	0.05	0.05
Quartile 1	0.04	0.04	0.05	0.04	0.05
Median	0.05	0.04	0.05	0.05	0.05
Quartile 3	0.05	0.05	0.05	0.05	0.05
Maximum	0.05	0.05	0.05	0.05	0.05
Minimum	0.04	0.04	0.04	0.04	0.04
IQR	0.00	0.00	0.00	0.00	0.00
Upper Whisker	0.05	0.05	0.05	0.05	0.05
Lower Whisker	0.04	0.04	0.04	0.04	0.04

AbPred results for Accelerated Stability (ACC STAB).

Total CDR length metric values for different categories of transgenic mouse platform antibodies.

	XenoMouse®	HuMAb-Mouse®	UltiMAb®	VelocImmune®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	50	48	48	49	49
Quartile 1	48	47	45	47	47
Median	50	49	46	49	50
Quartile 3	52	49	51	50	51
Maximum	60	52	54	57	52
Minimum	43	42	39	43	44
IQR	5	2	6	3	4
Upper Whisker	60	52	54	57	52
Lower Whisker	43	42	39	43	44

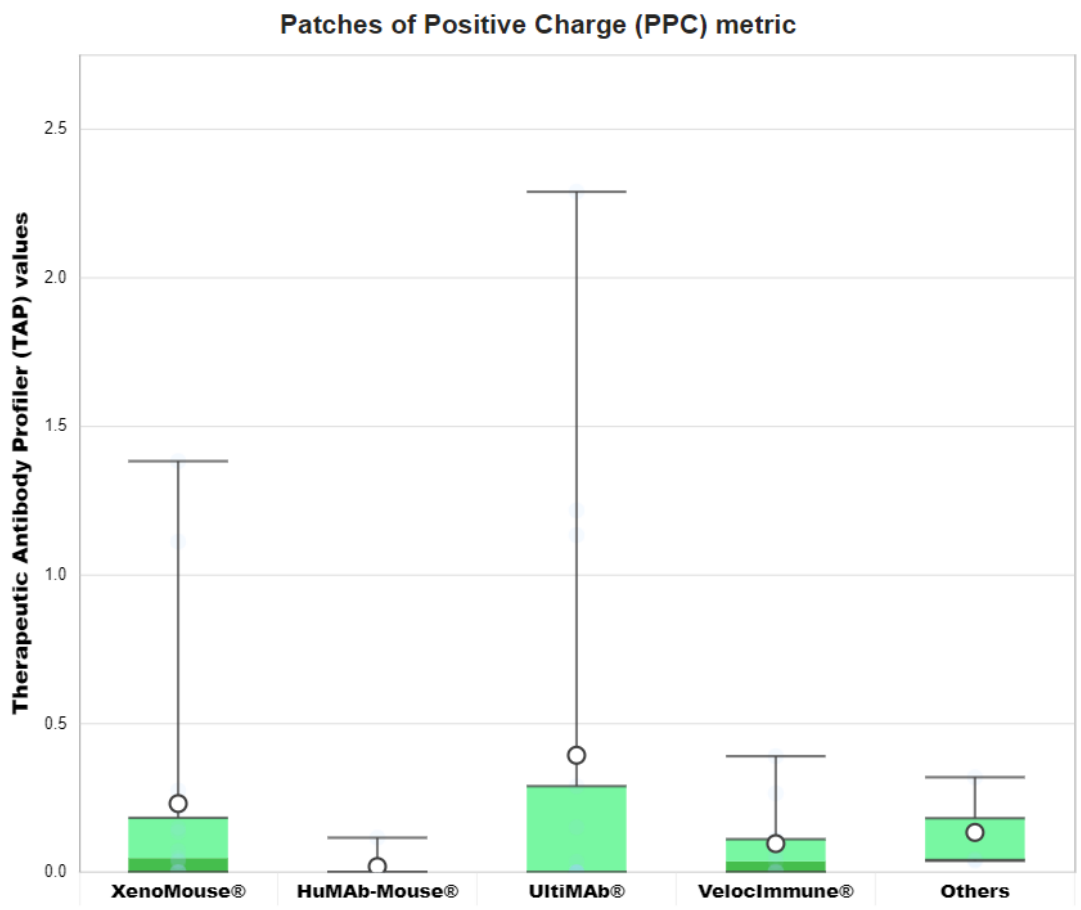
Therapeutic Antibody Profiler (TAP) statistics for Total CDR length metric.



Patches of Surface Hydrophobicity (PSH) metric values for different categories of transgenic mouse platform antibodies. PSH is calculated across the CDR vicinity.

	XenoMouse®	HuMAb-Mouse®	UltiMAb®	VelocImmune®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	147.13	137.18	137.84	140.69	151.51
Quartile 1	129.20	132.30	119.63	132.56	144.29
Median	149.76	133.88	139.92	138.95	146.55
Quartile 3	161.67	151.32	147.84	147.11	156.24
Maximum	197.02	158.29	185.35	167.18	165.93
Minimum	114.20	113.43	112.11	112.15	142.03
IQR	32.47	19.02	28.21	14.55	11.95
Upper Whisker	197.02	158.29	185.35	167.18	165.93
Lower Whisker	114.20	113.43	112.11	112.15	142.03

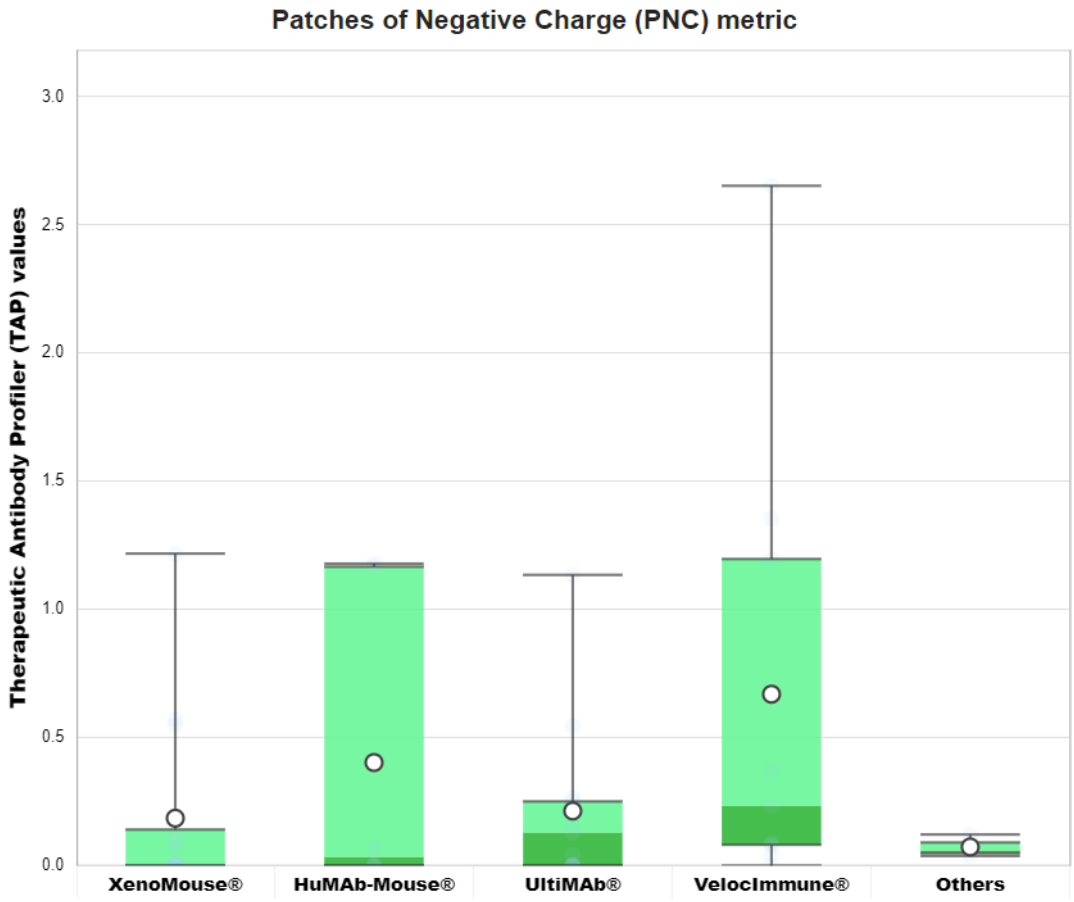
Therapeutic Antibody Profiler (TAP) statistics for Patches of Surface Hydrophobicity (PSH) metric.



Patches of Positive Charge (PPC) metric values for different categories of transgenic mouse platform antibodies. PPC is calculated across the CDR vicinity.

	XenoMouse®	HuMAb-Mouse®	UltiMAb®	VelocImmune®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	0.23	0.02	0.39	0.10	0.13
Quartile 1	0.00	0.00	0.00	0.00	0.04
Median	0.05	0.00	0.00	0.04	0.04
Quartile 3	0.18	0.00	0.29	0.11	0.18
Maximum	1.38	0.12	2.29	0.39	0.32
Minimum	0.00	0.00	0.00	0.00	0.04
IQR	0.18	0.00	0.29	0.11	0.14
Upper Whisker	1.38	0.12	2.29	0.39	0.32
Lower Whisker	0.00	0.00	0.00	0.00	0.04

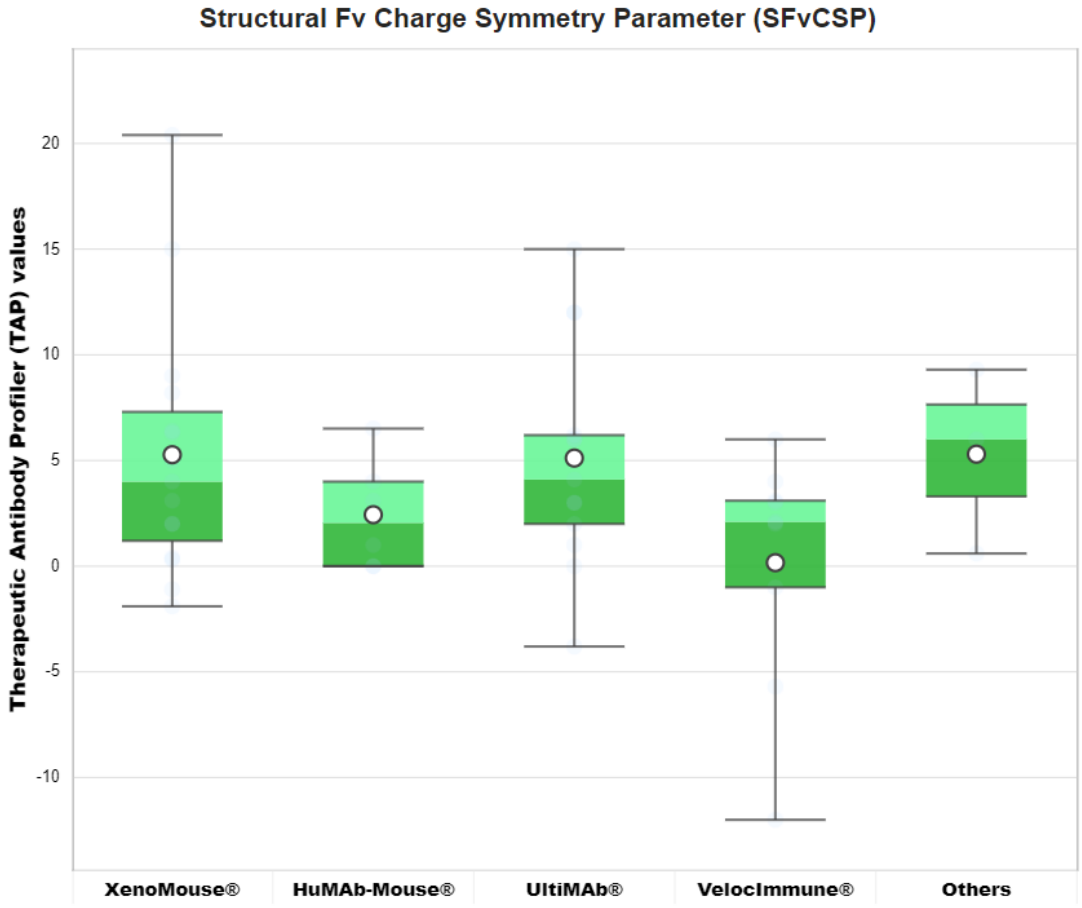
Therapeutic Antibody Profiler (TAP) statistics for Patches of Positive Charge (PPC) metric.



Patches of Negative Charge (PNC) metric values for different categories of transgenic mouse platform antibodies. PNC is calculated across the CDR vicinity.

	XenoMouse®	HuMAb-Mouse®	UltiMAb®	VelocImmune®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	0.18	0.40	0.21	0.67	0.07
Quartile 1	0.00	0.00	0.00	0.08	0.05
Median	0.00	0.03	0.13	0.23	0.06
Quartile 3	0.14	1.17	0.25	1.20	0.09
Maximum	1.22	1.18	1.13	2.65	0.12
Minimum	0.00	0.00	0.00	0.00	0.04
IQR	0.14	1.17	0.25	1.11	0.04
Upper Whisker	1.22	1.18	1.13	2.65	0.12
Lower Whisker	0.00	0.00	0.00	0.00	0.04

Therapeutic Antibody Profiler (TAP) statistics for Patches of Negative Charge (PNC) metric.



Structural Fv Charge Symmetry Parameter (SFvCSP) values for different categories of transgenic mouse platform antibodies.

	XenoMouse®	HuMAb-Mouse®	UltiMAb®	VelocImmune®	Others
Median Type	Inclusive	Inclusive	Inclusive	Inclusive	Inclusive
Whisker Type	Min/Max	Min/Max	Min/Max	Min/Max	Min/Max
Mean	5.27	2.44	5.12	0.17	5.30
Quartile 1	1.20	0.00	2.00	- 1.00	3.30
Median	4.00	2.05	4.10	2.10	6.00
Quartile 3	7.30	4.00	6.20	3.10	7.65
Maximum	20.40	6.51	15.00	6.00	9.30
Minimum	- 1.90	0.00	- 3.80	- 12.00	0.60
IQR	6.10	4.00	4.20	4.10	4.35
Upper Whisker	20.40	6.51	15.00	6.00	9.30
Lower Whisker	- 1.90	0.00	- 3.80	- 12.00	0.60

Therapeutic Antibody Profiler (TAP) statistics for Structural Fv Charge Symmetry Parameter (SFvCSP).