Current Trends in Biotechnology and Pharmacy
Vol. 18(3) 1866-1880, July 2024, ISSN 0973-8916 (Print), 2230-7303 (Online)
10.5530/ctbp.2024.3.33

Design and Development of Machine Learning Model for Antibody Design: Rituximab a Case Study

Harit Kasana[1,2], Harish Chander[1], and Ashwani Mathur[2*]

[1]National Institute of Biologicals (Ministry of Health and Family Welfare, Government of India) A-32, Sector-62, NOIDA-201 309, India
[2]Department of Biotechnology, Jaypee Institute of Information Technology Noida, A-10, Sector-62, Noida, 201309, Uttar Pradesh, India
*Corresponding Author: ashwani.mathur@mail.jiit.ac.in

Abstract

The design of antibodies using machine learning has emerged as a cutting-edge approach in the field of medical sciences and therapeutics. This study delves into the principles, methods, and deployment of machine learning approach for antibody design. The present study showed use of extensive antibody databases to train computational models, facilitating the prediction of antibody-antigen interactions. Seven different encoders were used for the vectorization of the antibody and antigen sequences. Conjoint triad showed unsurpassed performance in the machine-learning algorithm with 0.78 correlation and outperformed other encoding methods. The case study of Rituximab was used to demonstrate the practical application of the machine-learning model that was developed. The affinity score predicted by the model was utilised to select the most promising rituximab sequence. Structural investigation employed molecular dynamics (MD) simulation to authenticate the novel sequence (variant) of rituximab. Rituximab variant showed -45.44 kcal/mole compared to the wild type that had -37.66 kcal/mole as a binding free energy for the antigen-antibody complex. Free energy landscape (FEL) was calculated on the first two principal components (PC1 and PC2). The wild type has three minimum energy basins, whereas this variant exhibited only one. This showed that complex produced by antibody variant has a greater capacity to attain global minima. This study sheds light on the innovative application of machine learning in antibody design and also provides compelling evidence of its efficacy through the case study of rituximab. Present study opened new avenues for the development of antibodies with enhanced binding capabilities.

Keywords: Rituximab, Machine Learning, Immunogenicity, Molecular Dynamic Simulation.

Introduction

Monoclonal antibodies (MAbs) also known as therapeutic antibodies (TAbs)(1). Therapeutic antibodies are classified as protein-based therapeutic agents that have undergone engineering to recognise and bind to specific molecules (e.g., proteins, cells, and disease-associated molecules) within the organism. The creation of these antibodies is facilitated using specialized techniques that enable the generation of substantial quantities of identical antibodies capable of recognizing and engaging with a specific target (1). A range of therapeutic antibodies were introduced in the mid-1990s and have since been implemented in clinical settings (2). Currently, the quantity of licenced therapeutic antibodies (TAbs) stands

at approximately one hundred, with a greater number in the process of development this includes various designed and modified antibody formats. TAbs are commonly employed in the management of diverse severe diseases, including cancer, rheumatoid arthritis, multiple sclerosis, Crohn's disease and leukaemia(3). The therapeutic efficacy of these TAbs can be attributed to one or more mechanisms, including target binding and neutralization, direct cytotoxicity, antibody-dependent complement-dependent cytotoxicity, antibody-dependent cellular cytotoxicity, or other unidentified mechanisms (4). The range of disease areas that can be targeted by therapeutic antibodies has expanded significantly. Antibodies are now being used as medicines in various fields, including cancer, inflammatory illness, organ transplantation, cardiovascular disease, infection, respiratory disease, ophthalmologic disease, and others (2).

The essential role of immune-mediated effector functions in therapeutic antibodies targeting tumour cells is their specific binding to targets via the antibody-binding fragment (Fab) region(5), followed by the induction of immune responses via the fragment crystallizable (Fc) region. Cross-linking of cell-bound antibodies with Fc gamma receptors (FcRs) is responsible for activating immune effector cells. This interaction results in natural killer (NK) cells mediating antibody-dependent cellular cytotoxicity (ADCC). Similarly, antibody-dependent cellular phagocytosis (ADCP) is initiated by the interaction between FcRs and macrophages. This process involve the internalisation of antibody-opsonized target cells into phagosomes, where they are degraded via phagosome maturation and acidification(6).

The therapeutic antibody interacts with the antigen at a specific epitope region. The epitope refers to a spatial arrangement acknowledged by the complementary paratype of an antibody (7). The binding affinity between antibody and antigen can be determined by various traditional methods, including surface plasmon resonance (SPR) and enzyme-linked immunosorbent assay (ELISA). Although these are effective methods, there are still some limitations, such as time, cost, and potential variation in results (8).

$$Kd = \frac{koff}{kon}$$

The equilibrium dissociation constant, denoted as 'kd', represents the ratio of the dissociation rate constant ('koff') to the association rate constant ('kon') for the binding interaction between an antibody and its corresponding antigen (9). An inverse relationship exists between 'kd' and affinity. The 'kd' value is a measure of the antibody concentration required for a specific experiment. A lower kd value indicates a stronger affinity of the antibody, as it corresponds to a lower concentration needed (10). Various traditional techniques were used to determine kd during antibody –antigen interactions, including flow cytometry and ligand tracer (11). Traditional methods for determining the dissociation constant (kd) in antibody-antigen interactions face difficulties such as the assumption of 1:1 binding, affinity heterogeneity, omission of kinetic information, inability to account for conformational changes, reliance on equilibrium conditions, and potential interference from immobilization and sample matrix effects. Moreover, the other factors that affect the antigen-antibody interaction are temperature, pH, ionic strength, concentration antigen-antibody(12). With all of these factors, it may be difficult to determine an accurate kd value.

In the light of the challenges for determining the binding affinity between Antibody and Antigen, Therefore, in this study, a machine learning (ML) model to determine kd was proposed. The ML model was trained on antigen and antibody protein sequences with known kd values. It presents a novel machine learning (ML) approach for the efficient determination of kd, utilising existing kd values of antigen and antibody protein sequences. Determining the binding affinity using the sequence made the calculation faster. This could allow modifying the antibody sequence to improve the sensitivity and speci-

ficity against the given antigen. Later, this study demonstrated a case study of application of this method on the Rituximab. Engineered Rituximab was designed and validated using molecular dynamics simulation.

Materials and Methods

Data collection

Collection

The study focuses on the structural characteristics and binding affinity of antigen-antibody complexes, using data obtained from the SAbDab database(13, 14). (SAbDab is an in-depth database of antibody-antigen pairs. Every structure is accompanied with several information, including experimental details, antibody nomenclature (e.g., pairings of heavy and light chains), curated affinity data, and sequence annotations. As of July 25, 2023, 15,065 antigen-antibody complexes were accessible. After eliminating blanks in affinity and removing incorrect entries, 916 antigen-antibody pair remained that had affinity values (kd). These antibody sequences have heavy chain and light chain both. In the creating the data for machine learning model the heavy and light chain sequences were concatenated. Antigen sequences were also fed separately to the model with the sequence of the antibody and the kd value.

Machine learning model

DeepPurpose framework(15) was used for building the machine learning model. This framework is essentially constructed for drug-target interaction (DTI) used in drug repurposing and virtual screening. However, it also has model for protein-protein interaction. The source codes were sourced from github repository (kexinhuang12345/DeepPurpose: A Deep Learning Toolkit for DTI, Drug Property, PPI, DDI, Protein Function Prediction (Bioinformatics) (github.com)). DeepPurpose was preferred due to its user-friendly design which facilitated the use of multiple encoders for encoding the sequence of proteins and multiple evaluation metrics such as MSE, R-squared and Concordance Index. It also enabled the support for custom model architecture with utilities for training and evaluating the models. Here, 8 different encoders were used for encoding the protein sequence and result was evaluated: 'CNN', 'AAC', 'PseudoAAC', 'Conjoint_triad', 'Quasi-seq', 'ESPF', 'Transformer', 'CNN_RNN'. ‘CNN’ and 'Conjoint_triad' encoder outperformed the others and thus used as preferred encoding technique. Though DeepPurpose library enabled the utilisation of some pretrained models, a custom deep learning model was developed combining two different neural network architectures. The combined neural network architectures were, Convolutional Neural Network (CNN) and Message Passing Neural Network (MPNN). The configuration of the model used was, 100 epochs, 0.001 learning rate and 128 batch sizes, these parameters along with other essential parameters are represented in Table 1. The interaction prediction for antibodies was complex due to their large sequence length. Thus, such custom model was utilised with testing on multiple encoders to ensure highly accurate predictions. The model was trained on 70% of the known dataset of antibody-antigen pairs with their corresponding Kd values in log. The example data is shown below:

Antibody

EVQLVESGGGLVQPGGSLRLS-
CAASGYTFTNYGMNWVRQAPGKGLEWVG-
WINTYTGEPTYAADFKRRFTFSLDTSK-
STAYLQMNSLRAEDTAVYYCAKY-
PHYYGSSHWYFDVWGQGTLVTVSSAST-
KGPSVFPLAPSSKSTSGGTAALGCLVKDYF-
PEPVTVSWNSGALTSGVHTFPAVLQSSGLYS-
LSSVVTVPSSSLGTQTYICNVNHKPSNTK-
VDKKVEPKSCDKTHTXXXDIQMTQSPSSL-
SASVGDRVTITCSASQDISNYLNWYQQKP-
GKAPKVLIYFTSSLHSGVPSRFSGSGSGTD-
FTLTISSLQPEDFATYYCQQYSTVPWTFGQ-
GTKVEIKRTVAAPSVFIFPPSDEQLKSGT-
ASVVCLLNNFYPREAKVQWKVDNALQSGN-
SQESVTEQDSKDSTYSLSSTLTLSKADYEKH-
KVYACEVTHQGLSSPVTKSFNRGEC

Antigen

GQNHHEVVKFMDVYQRSYCHPIETL-VDIFQEYPDEIEYIFKPSCVPLMRCGGCCN-DEGLECVPTEESNITMQIMRIKPHQGQHI-GEMSFLQHNKCECRPKKD

kd:1.22

The model used 10% of the data for validation while rest 20% for testing (coefficient of determination), making the split ratio of the dataset to be 70:10:20, where each ratio part corresponds to train, validation and test respectively. The regression score (r^2) was used as evaluation metric to calculate the performance.

Table 1. Parameters associated with the machine learning model.

Experimental Parameter	Value
Learning rate	0.001
Batch size	128
Epoch	100
Hidden layer dimension	1024, 1024, 512
Size of hidden layers in MPNN	128
Number of layers in the MPNN	3
Number of filters in convolutional layer	32, 64, 96
Size of the filters in each convolutional layer	4, 8, 12

CDR Identification

The recognition of Complementarity Determining Regions (CDR) sequences plays a crucial role in augmenting the properties of antibodies, such as their binding affinity, specificity, and overall therapeutic effectiveness. AbRSA tool(16) was used to predict the CDR regions, which is a tool specifically stands for Antibody Numbering and CDR Delimiting. This tool was employed to predict the Complementarity Determining Regions (CDR) for both the VH (Heavy chain) and VL (Light chain) of each antibody.

Rituximab structure

The structure of rituximab was sourced from protein data bank (PDB)(17), numerous structures of rituximab were submitted in the PDB. Here, the structure with PDB ID: 2OSL(18, 19) was downloaded for determining the known interaction of CDR region and epitope of antigen.

Sequence engineering

In this study, the Complementarity Determining Region (CDR) of rituximab was focused on, being a pivotal component guiding the binding specificity and affinity. A comprehensive computational analysis was undertaken to explore all possible amino acid combinations within the CDR.

Structure model building

Novel sequence of the rituximab that predicted the maximum affinity with the antigen was modelled using a machine learning (ML)-based technology provided by the SAbPred server(20). The ABodyBuilder-ML module, which is a component of the SAbPred server, was utilised for the purpose of novel rituximab structure modelling using sequence data. The ABodyBuilder-ML programme was applied to predict the structure of the variable domains of the antibody, namely VH (heavy chain) and VL (light chain), that are commonly referred to as the Fv (fragment variable).

MD Molecular Dynamics Simulation was performed for the top-selected antigen-antibody complexes. GROMACS2022 package was the software that was used to carry out the molecular dynamic simulations of the top-selected antigen-antibody complexes for a time period of 100 ns(21–23). The CHARMM force field parameters (24)were used to define the topological parameters, which were then allocated to both the antibodies and the antigen. The Ewald Particle Mesh approach was used with the purpose to perform the calculations necessary to determine the distance electrostatic force (25). In order to establish a neutral system, Na+ and Cl- ions were introduced, and the TIP3P water cube model was employed with the purpose to

solvate the system. The steepest descent (SD) method, which consists of 50,000 minimization steps, was used on the system with the aim of removing any steric conflicts. In addition, the temperature of the entire system was raised to 310 K, and the timestep was set to 2 fs for 100 ps of simulation time (NVT). After that, the system underwent additional equilibration at a pressure of 1 bar for a duration of 1 ns during the simulation time (NPT). GROMACS was able to simulate at a constant temperature by employing the velocity-rescaling strategy in conjunction with a temperature coupling scheme (26). Temperature control was linked with pressure coupling, and the result was the NPT ensemble. The pressure coupling employed was the Parrinello-Rahman pressure coupling(27). Using the SHAKE approach(28), all hydrogen bonds were constrained, and after that, the coordinates of the structure were recorded after every 10 ps during the entire 100 ns (100,000 ps) production run. The conformational stability and variation of the antigen-antibody complexes were determined by post MD analysis. The results of the 100 ns MD simulation were analysed with the RMSD (root mean square deviation) tool in GROMACS. Following the MD simulation, clustering was performed on the complexes by applying the gromos cluster method using the g_cluster packages of GROMACS(29, 30) with a threshold RMS of 0.3 nm. Later, these clusters were used for further analysis and interpreting the stability and flexibility of the antigen-antibody complexes. Furthermore, the binding free energy of the antigen-antibody was calculated using the MM/GBSA approach as mentioned in the section 1.6. Here, the middle cluster of the most popular cluster was used for determination of binding free energies. All the clusters generated were used for SASA determination, which were compared with the epitopes of the spike protein.Results and Discussion

Data collection

Data was collected from the SAbDab database, eventually 916 entries with their PDB ids and Kd values that had units in nanomolar. These Kd values were further converted to log values under the normalization process. The logarithmic values of the Kd are shown in Figure1, most data points were ranged in between 10^{-6} to 10^{-11}. There were few outliers in the datasheet for Kd values. Corresponding sequence of the antibodies and antigens were extracted from the protein data bank using the PDB Ids sourced from the SAbDab database. These sequences were for the Fab region of the antibody with their heavy and light chains sections. Both heavy and light chains were merged together as a single component of antibody. Similarly, antigen sequences were also extracted as a separate input for the machine learning model.

Building machine learning model

Encoders, ‘CNN’, ‘AAC’, ‘PseudoAAC’, ‘Conjoint_triad’, ‘Quasi seq’, ‘ESPF’, ‘Transformer’, ‘CNN_RNN’ were used for encoding the protein sequences of antibodies and antigens. All these encoders were used serially to encode the sequence of the target protein sequence and used in the DeepPurpose machine-learning model. Table 2 shows the performance of each encoder for predicting the Kd values. The original values of Kd were in nM and the magnitude was converted into log scale after multiplying with 10^9. Figure 1(a) shows the original Kd values and converted logarithmic values. As observed in the plot that original Kd values are skewed towards 0 as most of the datapoints had low Kd values in nano molar. This skewedness can affect the performance of the prediction and thus they were converted into log scale after multiplying with 10^9. Figure 1(b) showed the normalized values of CNN and ConjointTriad performed similarly in the pearson correlation with the coefficient value of 0.78. However, mean square error for CNN was 4.08 while for ConjointTriad it was 4.14. Other encoders, AAC, ESPF and Transformer also showed correlation coefficient greater than 0.7.

Table 2. Performance of the different encoder on the data set to predict the Kd values.

Encoder	MSE	Pearson Correlation Coefficient	p-value	Concordance Index
CNN	4.08	0.78	7.32E-37	0.82
AAC	5.28	0.71	1.04E-27	0.78
PseudoAAC	8.82	0.39	1.18E-07	0.66
ConjointTriad	4.14	0.78	1.43E-36	0.83
Quasi-Seq	8.79	0.42	1.12E-08	0.63
ESPF	4.96	0.73	6.68E-30	0.80
Transformer	5.60	0.71	4.88E-28	0.81
CNN_RNN	7.90	0.50	3.51E-12	0.67

ConjointTriad model showed the best performance for correlation coefficient and mean square error and thus selected for the further processing. ConjointTriad is widely used for dealing with the sequential data including protein sequence. The architecture comprises multiple convolutional layers and encodes the sequence within a 1D encoder, thereby capturing the features and patterns inherent in the sequence. As the sequence of the protein was of different length thus the padding was also used during the ConjointTriad encoder. Moreover, p-values in the Table 1 shows the relevance score of the prediction for their correlation co-efficient. As shown in the table that all the p-values are lower than 0.05 and thus the confidence was greater than 90% for the calculated correlation coefficients. The concordance index evaluates how well the model's predicted probabilities of survival align with the actual outcomes. The ConjointTriad model has the highest concordance index value of 0.83, while CNN also has a close value of 0.82. Considering all the evaluation metrics, Conjoint Triad encoder moder was selected as the best performing model. Moreover, various studies used machine learning models based deep convolutional neural networks, that used in both bio-medical and sensor-based study(31–34)deep convolutional neural networks have demonstrated dominant performance in human activity recognition (HAR.

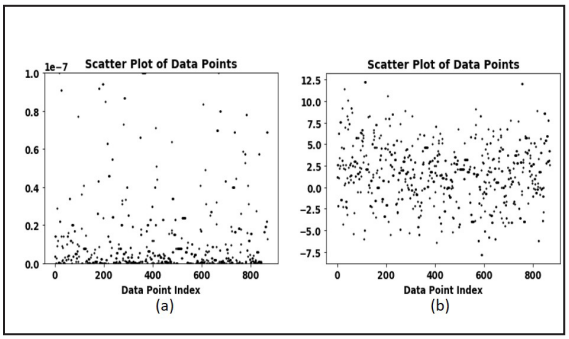

Figure 1. Plot of Kd values with respect to each 916 entries of the antibodies (a) original Kd (b) logarithmic values of the kd.

CDR and structure of rituximab

Once the ML model was selected with the best performing encoder technique (ConjointTriad), the Fab sequence of the Rituximab was collected. PDB ID: 6VJA was sourced from protein databank for extracting the sequence of heavy and light chains of the Fab segment of the Antibody. Sequence of heavy and light chains were submitted to the AbRSA tool for detecting the CDRs. Complementarity-determining regions (CDRs) are immunoglobulin (Ig) hypervariable domains that regulate particular antibody (Ab) binding(35)fungi, protozoa and viruses. The possibility that isolated CDRs, represented by short synthetic peptides, may display antimicrobial, antiviral and antitumor activities irrespective of Ab specificity for a given antigen

is addressed here.\nMETHODOLOGY/PRINCIPAL FINDINGS: CDR-based synthetic peptides of murine and human monoclonal Abs directed to: a.

The chains are shown below with the coloured CDR region detected by AbRSA tool:

> RITUXIMAB FAB HEAVY CHAIN

1 QVQLQQPGAELVKPGASVKMSCKASGYTFTSYNMHWVKQTPGRGLEWIGAIYPGNGDTSY 60
61 NQKFKGKATLTADKSSSTAYMQLSSLTSEDSAVYYCARSTYYGGDWYFNVWGAGTTVTVS 120
121 AASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQS 180
181 SGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKVEPKSC

> RITUXIMAB FAB LIGHT CHAIN

1 QIVLSQSPAILSASPGEKVTMTCRASSSVSYIHWFQQKPGSSPKPWIYATSNLASGVPVR 60
61 FSGSGSGTSYSLTISRVEAEDAATYYCQQWTSNPPTFGGGTKLEIKRTVAAPSVFIFPPS 120
121 DEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTL 180
181 SKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC

Heavy chain showed 3 CDR segments, CDR1 coloured as red, CDR2 coloured as orange while CDR3 coloured as green. In heavy chain, CDR3 is the longest with 12 residues while CDR2 showed 8 residues and CDR1 with 7 residues, respectively. In contrast, CDR1 has the maximum length of 10 in light chain, while CDR2 has the minimum of 7 amino acids. Structure of Rituximab is submitted in the PDB with PDB ID: 6VJA, this entry has CD20 B-lymphocyte antigen bound with the antibody structure. Figure 2, shows the Antigen and Antibody structures. Only the Fab segment that interacts with Antigen is submitted in the protein databank, heavy chain has 224 amino acids while light chain has 213 amino acids. CD20 antigen has 278 amino acids in the PDB entry.

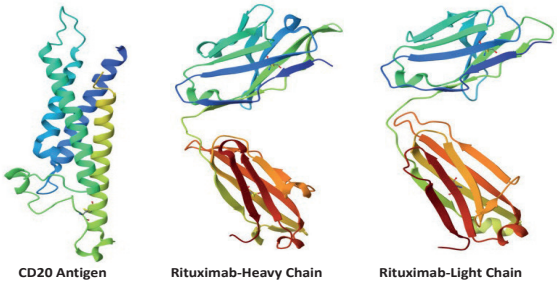

Figure 2. Structure of CD antigen that complexed with Rituximab using heavy and light chain of Fab segment, deposited in protein data bank with PDB ID: 6VJA.

Antigen in the structure is deposited as dimer in this structure with chain C and D while H and L are the heavy and light chains of the antibody. SER95, TRP100, HIS35, ASN33, SER58, TYR52, and ASP56 from the heavy chain was interacting with the chain C of the antigen while SER28, SER29, SER31, TRP91, and ASN94 from the light chain was involved in the antigen antibody interaction. It is noteworthy that SER28, SER29, and SER31 interacting residues from the light chain were from CDR1, while TRP91, and ASN94 were from CDR3. Similarly, SER58, TYR52, and ASP56 were from CDR2 of the heavy chain, while SER95, and TRP100, also seen as interactive residues, were from CDR3. In overall, the outcome provides a comprehensive analysis of the molecular composition and binding properties of the Rituximab antibody, with specific emphasis on its interaction with the CD20 antigen, which is crucial for its therapeutic efficacy.

Modification of antibody sequence

Later, the CDR segments of antibody were targeted for the substitution of single amino acids serially. This resulted in generation of 969 sequences that constitute of all possible combination of amino acid for a given substitution.

Essentially, this created a vast library of Rituximab variants. These new sequences of heavy/light chains were constitutively fed to the trained machine-learning model along with the original sequence of CD20 antigen. The model predicted log (Kd) values which further converted into antilog to deduce the original Kd value in nano Molar. The Kd value is a standard measure of the binding affinity between an antibody and an antigen; lower Kd values indicate higher affinity(36). Figure 3, shows the top three sequences of the Rituximab that predicted the minimum Kd values in the ML model. The Top sequence showed 0.000024 nM Kd and it was considered and best modified Antibody sequence at the CDR that can bind with CD20 antigen with the highest affinity. In this sequence, SER31 was converted to LEU31. This specific mutation significantly increased the binding affinity of the antibody to the CD20 antigen.

Figure 3. Top three sequences of the Rituximab with minimum Kd values from the ML model.

Structure modelling

Protein structure of Rituximab and CD20 antigen peptide was collected from the protein data bank (PDB). The required modification in the rituximab was imposed in the complex structure using SWISS Pdb viewer tool. Swiss-PdbViewer is an application that provides a user friendly interface allowing to evaluate many proteins at the same time(37). The structure was further simulated under physiological condition for 100 ns to achieve the most stable conformation. Both wild and new variant of rituximab was simulated for 100 ns. Various evaluation metrics were further computed on the simulation trajectory.Root mean square deviation (RMSD)

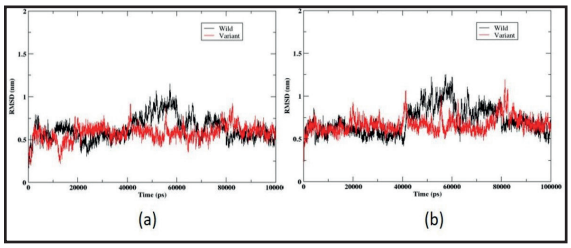

Figure 4. Root mean square deviation (RMSD) for the (a) antibody (rituximab) and (b) antigen (CD20) for wild type and variant type in the complex.

The deviation of antigen-antibody complex from the equilibrated structure achieved after the NVT and NPT ensemble equilibration. This deviation showed the change in the conformational space of the antigen-antibody complex. Root-mean-square deviation (RMSD) was calculated over the 100 ns for the rituximab (antibody) and CD20 (antigen). The RMSD is a metric that quantifies the average distance between the atoms, typically the backbone atoms, of molecules that have been superimposed(38). Both antigen and antibody showed high RMSD during the 100 ns simulation. Here, the wild and variant both started with the similar pattern in the RMSD for antibody as shown in Figure 4 (a) at 0.75 nm from the initial conformation. Until the 40 ns of the simulation trajectory, they behaved similarly. However, post 40 ns, wild type showed a high peak of RMSD compared to variant. Moreover, after 80 ns, they re-showed the similar pattern of RMSD. Similarly, in antigen the RMSD of wild type showed the higher peak between 40 ns to 80 ns. Following an 80-nanosecond interval, they subsequently realign and attain a stable configuration, converging at a distance of 0.75 nanometers, as visually represented in Figure 4 (b). Plots shown in Figure 4, illustrate the higher deviation in the wild conformation of the protein compared to the variant (both for antibody and antigen). Here, the reference structure used for alignment was the

complete complex while the RMSD was calculated for the antigen and antibody separately. The plots showed that RMSD of the antibody and antigen behaved similarly as per their deviation pattern from the initial docked conformation. Here, the complete complex was taken as reference for the alignment and RMSD was calculated for antibody and antigen individually. This suggest that antibody showed translation motion in the given complex. The interaction site has changed to limited extent during the course of the simulation. This happened to find the stronger interaction with the antigen. However, it is noted that variant that designed in this case study has shown lower deviation than the wild type. This specify that if the variant antibody interact with the antigen in native like conformation, then the chances of formation of stable pose is higher compared to the wild type. Both the plots showed a jump in same period (40-70 ns) that indicate the disturbance of the conformation. However, it settled in the later phase to achieve more stability. Overall, this outcome provides significant understanding into the dynamic characteristics of the Rituximab-CD20 complex, highlighting the variations in stability and structural modifications between the original and modified versions of the antibody. These findings hold great importance in understanding the molecular foundation of interactions between antibodies and antigens, and can provide assistance in developing therapeutic antibodies that are more stable and efficient.

Root mean square fluctuation (RMSF)

The root mean square fluctuation (RMSF) quantifies the variation of each atom along the whole trajectory. Later, the root mean square fluctuation (RMSF) was calculated for wild and variant antibody to compare the fluctuation at residue level. The plot shown in Figure 5, shows that both wild type and variant showed the similar trend of RMSF. In heavy chain of the wild type other than the terminal residues LYS-129, SER-130, and SER-132 showed the high RMSF, crossing 0.3 nm cutoff while LYS-129 and SER-130 also showed the similar RMSF pattern in variant type. However, none of these residues were part of the CDR region. In contrast, light chain showed more stability and only terminal residues showed the high RMSF while rest of the residues were settled in their conformation. The graphical representation of the RMSF is shown in the Figure 5. As observed from the plots that both wild and variant type behaved similarly in the simulation and common residues showed the jump in the RMSF.

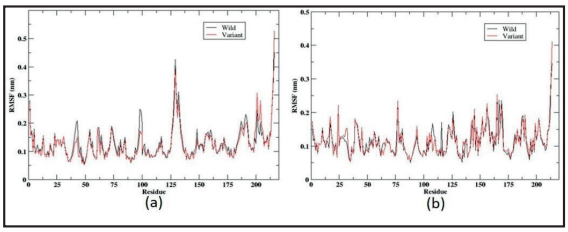

Figure 5. RMSF plots of (a) heavy chain and (b) light chain for both wild and variant types. X axis shows the residue number and Y axis shows the RMSF in nm scale.

Free energy landscape (FEL)

Free energy landscapes are commonly used in computational chemistry and biophysics to represent and simulate the process of protein folding, ligand-receptor interaction, enzyme mechanisms, and other dynamic molecular processes. They offer valuable information regarding the thermodynamics and kinetics of these processes(39). Here, the top two principal components of the conformational motion were calculated to represent the maximum deviation. This presents the movement of the protein along the principal component to achieve an energetically stable state. Moreover, it also presents the variation in the structure of the protein that signifies the overall stability of the molecule. The antibody antigen complexes (wild and variant) was simulated for 100 ns and their principal components were detected, the motion and the energy term were calculated as shown in the 3D energy landscape in Figure 6. The free energy was shown in relative scale where the most stable state or the minimum free energy state was considered as reference. Therefore, the energy

scale started from 0 (minimum) and went to high scale. Moreover, in the funnel shape plot, it is shown that wild type has multiple basins while the variant has single basin. These basins are minimum energy state that was achieved for the given complex. It is noteworthy that going from one basin to another basin there are energy barrier that needs to be crossed. Navigating these energy barriers presents a formidable challenge within the conformational space, and there is a significant likelihood of a given structure becoming trapped in these local energy minima. As shown in the Figure 6, that variant has no energy barrier in their its free energy landscape and this showed that complex has high possibility to reach the global minima without getting trapped in the local minima. On the other hand, wild type has energy barriers and there is more chance that conformation would trap in the local minima and will not be eligible to reach global minima. It is also been seen that basins in the wild type have not the similar energy states conformation as depicted by the colour gradient and the depth of the basin. In conclusion, this investigation shows the significance of studying the energy landscapes of molecular complexes. The outcome is that the mutated version of the antibody-antigen complex is more prone to achieving a stable structure with ease, without the obstruction of energy obstacles, in contrast to the original form. The understanding acquired from this observation is extremely helpful in the realm of drug design and protein engineering, as it pertains to the crucial aspects of molecular interactions' stability and efficiency.

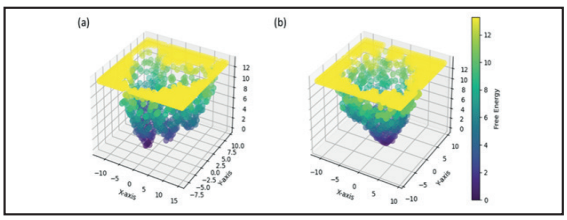

Figure 6. Free energy landscape of the antigen (CD20) and antibody (rituximab) complex for 100 ns simulation calculated over first two principal components for (a) wild type and (b) variant type.

Gibbs binding free energy (∆G)

MM/PBSA and MM/GBSA have been extensively employed in biomolecular studies, particularly in the investigation of protein folding, protein−ligand binding, and protein−protein interactions(40). Here, binding free energy of the antigen-antibody complex was calculated for the last 20 ns of the simulation. MM/GBSA (Molecular mechanics with generalised Born and surface area solvation) method was employed to calculate the binding free energy. This energy has multiple energetic terms that includes gas and solvation term of gibbs energy. Gas energy term has Van der Waal and electrostatic energy components. However, solvation term has both polar and non-polar components. Average values for all the terms were calculated van der Waal showed the most favourable energy (-63.07 kcal/mole) for the wild while for variant electrostatic component dominated (-70.55 kcal/mole). Polar solvation energy was positive in both the cases, wild has 50.08 kcal/mole while for variant it is 129.57 kcal/mole. Overall, the average binding free energy for wild and variant was comparative but variant showed more stability with -45.44 kcal/mole and the wild type had -37.66 kcal/mole. This showed that variation in the antibody provided stronger interaction with the antigen that further resulted in the more stable complex formation. Figure 7, showed the binding free energies for wild and variant types with their various energetic components.

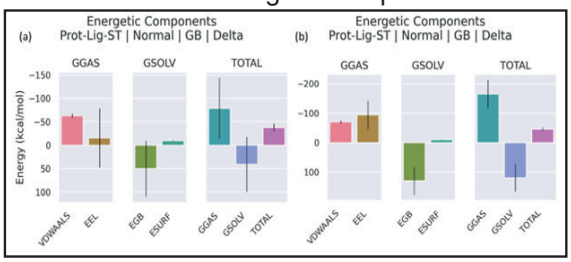

Figure 7. Binding free energy calculated using MM/GBSA techniques for (a) wild type and (b) variant type.

Further, the binding free energy was recorded for each frame, 20 frames were gen-

erated for the last 20 ns. Energy in each of the frame is shown in Figure 8. The energy was stable for the variant the standard deviation was 6.21. In the wild type the energy continuously fallen and reached to minimum at 100 ns.

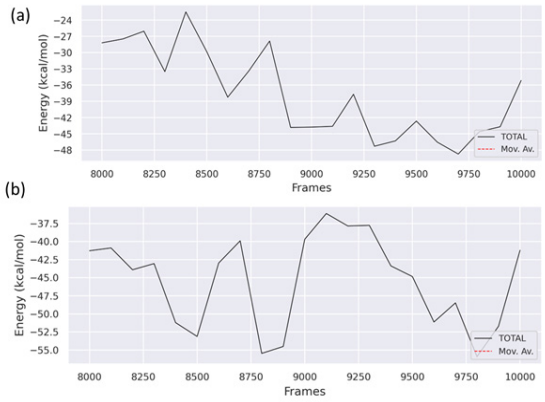

Figure 8. Binding free energy of each frame for (a) wild type and (b) variant type for the rituximab and CD20 complex over last 20 ns simulation trajectory.

RMSD and RMSF shown earlier in the calculation conveyed that the conformation of wild and variant type showed similar trend. However, variant had relatively more stable conformational behaviour during the simulation trajectory. This showed that more stability was attained in the complex after the change imposed in the rituximab sequence. This change was guided by the ML method demonstrated in this study. Later, it was also showed that the variant had similar level of residual fluctuation. The variant had favourable binding free energy calculated from the MM/GBSA. Overall, the variant of rituximab showed stronger binding with the antigen than the wild type.

Discussion

The therapeutic application of monoclonal antibodies (mAbs) is among the most promising areas within the domain of immunotherapy. Immunogenicity is one of the most significant factors that can limit the therapeutic and diagnostic applications of mAbs. The advancement of biological therapeutics has been greatly accelerated by the integration of computational technologies with high-throughput technology(41–43). Numerous cases of in silico vaccine design have been successful, including the vaccine design initiated by Correia et al. through computational protein design(44). They used in-silico methods to design small, thermally and conformationally stable protein scaffolds. Additional testing has shown that these protein scaffolds successfully imitate the structure of the viral epitope and stimulate powerful antibodies that can neutralise the virus.

The present study demonstrated the conceptualization and construction of a resilient machine learning framework for forecasting the antigen-antibody binding affinity. The conjoint triad encoder method demonstrated the highest degree of accuracy, as evidenced by its correlation coefficient of 0.78. An additional application of the method was to illustrate its efficacy on rituximab through the generation of original sequences that targeted the CDR region. Rituximab induces apoptosis, complement activation, and antibody-dependent cell-mediated cytotoxicity by binding to the CD20 marker expressed on B lymphocytes (Miranda-Hernández MP, López-Morales CA, Ramírez-Ibáñez ND, Piña-Lara N, Pérez NO, Molina-Pérez A, et al. Assessment of physicochemical properties of rituximab related to its immunomodulatory activity(45, 46). The study identified the optimal sequence exhibiting the greatest affinity for the CD20 antigen. A molecular dynamics (MD) simulation study was conducted to investigate the binding characteristics and structure of the variant rituximab. The RMSD of this variant was lower than that of the wild type. This indicates that the probability of stable pose formation is greater for variant antibodies that interact with the antigen in a native-like conformation, as opposed to the wild type. The binding free energy of the antigen-antibody complex was -45.44 kcal/mole for the Rituximab variant and -37.66 kcal/mole for the wild type. This showed that variation in the antibody provided stronger interaction with the antigen that further resulted

in the more stable complex formation. A free energy landscape (FEL) was calculated by utilising the initial two principal components (PC1 and PC2). In contrast to the wild type, which possesses three minimum energy basins, this variant displayed only one. Thus, it was demonstrated that the antibody variant-produced complex is more capable of achieving global minima. In the machine learning model, the variant that was generated and justified demonstrated a higher level of stability and affinity with the antigen. Thus, the methodology that was presented in the study has the potential to be utilised in the design of antibodies in order to produce therapeutic antibodies that are more effective.

Conclusion

This study showed the design and development of a robust machine learning model to predict the binding affinity between antigen and antibody. Conjoint triad encoder method showed the highest accuracy with 0.78 correlation coefficient. The method was further used to demonstrate its application on rituximab by generating novel sequences by targeting the CDR region. The study detected the best sequence that showed the highest affinity for the CD20 antigen. The structure of the variant rituximab was modelled and the binding characteristics was explored in MD simulation study. The variant generated and justified in ML model showed the higher stability and affinity with the antigen. The method presented in the study has potential to be used in antibody design to generate more efficacious therapeutic antibody.

Acknowledgements: The authors acknowledge Jaypee Institute of Information Technology, NOIDA, Uttar Pradesh and National Institute of Biologicals, NOIDA for providing infrastructure and support

Conflict of Interest: The authors declare that there are no conflicting interests.

Funding: There is no funding to report

Authors Contribution: H.K. wrote the manuscript. A.M. conceived the idea and methodology. H.C. and A.M. edited the manuscript. All authors read and approve the manuscript

References

1. Berger, M., Shankar, V., & Vafai, A. (2002). Therapeutic Applications of Monoclonal Antibodies. *The American Journal of the Medical Sciences*, *324*(1), 14–30. https://doi.org/10.1097/00000441-200207000-00004
2. Suzuki, M., Kato, C., & Kato, A. (2015). Therapeutic antibodies: their mechanisms of action and the pathological findings they induce in toxicity studies. *Journal of Toxicologic Pathology*, *28*(3), 133–139. https://doi.org/10.1293/tox.2015-0031
3. Quinteros, D. A., Bermúdez, J. M., Ravetti, S., Cid, A., Allemandi, D. A., & Palma, S. D. (2017). Therapeutic use of monoclonal antibodies: general aspects and challenges for drug delivery. In *Nanostructures for Drug Delivery* (pp. 807–833). Elsevier. https://doi.org/10.1016/B978-0-323-46143-6.00025-7
4. Houen, G. (2022). Therapeutic Antibodies: An Overview. In G. Houen (Ed.), *Therapeutic Antibodies* (Vol. 2313, pp. 1–25). New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-1450-1_1
5. Lu, L. L., Suscovich, T. J., Fortune, S. M., & Alter, G. (2018). Beyond binding: antibody effector functions in infectious diseases. *Nature Reviews. Immunology*, *18*(1), 46–61. https://doi.org/10.1038/nri.2017.106
6. Kamen, L., Ordonia, B., Myneni, S., & Chung, S. (2022). Method for Measurement of Antibody-Dependent Cellular Phagocytosis. In G. Houen (Ed.), *Therapeutic Antibodies* (Vol. 2313, pp. 305–312). New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-1450-1_19
7. Akbar, R., Robert, P. A., Pavlović, M., Je-

liazkov, J. R., Snapkov, I., Slabodkin, A., … Greiff, V. (2021). A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Reports*, *34*(11), 108856. https://doi.org/10.1016/j.celrep.2021.108856

8. Beeg, M., Nobili, A., Orsini, B., Rogai, F., Gilardi, D., Fiorino, G., … Gobbi, M. (2019). A Surface Plasmon Resonance-based assay to measure serum concentrations of therapeutic antibodies and anti-drug antibodies. *Scientific Reports*, *9*(1), 2064. https://doi.org/10.1038/s41598-018-37950-4

9. Huber, W., & Mueller, F. (2006). Biomolecular Interaction Analysis in Drug Discovery Using Surface Plasmon Resonance Technology. *Current Pharmaceutical Design*, *12*(31), 3999–4021. https://doi.org/10.2174/138161206778743600

10. Frenzel, A., Roskos, L., Klakamp, S., Liang, M., Arends, R., & Green, L. (2014). Antibody Affinity. In S. Dübel & J. M. Reichert (Eds.), *Handbook of Therapeutic Antibodies* (pp. 115–140). Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA. https://doi.org/10.1002/9783527682423.ch6

11. Spiegelberg, D., Stenberg, J., Richalet, P., & Vanhove, M. (2021). KD determination from time-resolved experiments on live cells with LigandTracer and reconciliation with end-point flow cytometry measurements. *European Biophysics Journal*, *50*(7), 979–991. https://doi.org/10.1007/s00249-021-01560-2

12. Reverberi, R., & Reverberi, L. (2007). Factors affecting the antigen-antibody reaction. *Blood Transfusion*. https://doi.org/10.2450/2007.0047-07

13. Schneider, C., Raybould, M. I. J., & Deane, C. M. (2022). SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker. *Nucleic Acids Research*, *50*(D1), D1368–D1372. https://doi.org/10.1093/nar/gkab1050

14. Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., … Deane, C. M. (2014). SAbDab: the structural antibody database. *Nucleic Acids Research*, *42*(D1), D1140–D1146. https://doi.org/10.1093/nar/gkt1043

15. Huang, K., Fu, T., Glass, L. M., Zitnik, M., Xiao, C., & Sun, J. (2021). DeepPurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics*, *36*(22–23), 5545–5547. https://doi.org/10.1093/bioinformatics/btaa1005

16. Li, L., Chen, S., Miao, Z., Liu, Y., Liu, X., Xiao, Z., & Cao, Y. (2019). AbRSA: A robust tool for antibody numbering. *Protein Science : A Publication of the Protein Society*, *28*(8), 1524–1531. https://doi.org/10.1002/pro.3633

17. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

18. Bank, R. P. D. (n.d.). RCSB PDB - 2OSL: Crystal structure of Rituximab Fab in complex with an epitope peptide. Retrieved October 14, 2023, from https://www.rcsb.org/structure/2OSL

19. Du, J., Wang, H., Zhong, C., Peng, B., Zhang, M., Li, B., … Ding, J. (2007). Structural Basis for Recognition of CD20 by Therapeutic Antibody Rituximab *. *Journal of Biological Chemistry*, *282*(20), 15073–15080. https://doi.org/10.1074/jbc.M701654200

20. Dunbar, J., Krawczyk, K., Leem, J., Marks, C., Nowak, J., Regep, C., … Deane, C. M. (2016). SAbPred: a structure-based antibody prediction server. *Nucleic Acids Research*, *44*(Web Server issue), W474–W478. https://doi.org/10.1093/nar/gkw361

21. Berendsen, H. J. C., van der Spoel, D., & van Drunen, R. (1995). GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, *91*(1), 43–56. https://doi.org/10.1016/0010-4655(95)00042-E

22. Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, *1*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001

23. GROMACS 2022 Manual. (n.d.). https://doi.org/10.5281/zenodo.6103568

24. Huang, J., & MacKerell, A. D. (2013). CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *Journal of Computational Chemistry*, *34*(25), 2135–2145. https://doi.org/10.1002/jcc.23354

25. Darden, T., York, D., & Pedersen, L. (1993). Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, *98*(12), 10089–10092. https://doi.org/10.1063/1.464397

26. Bussi, G., Donadio, D., & Parrinello, M. (2007). Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, *126*(1), 014101. https://doi.org/10.1063/1.2408420

27. Parrinello, M., & Rahman, A. (1981). Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, *52*(12), 7182–7190. https://doi.org/10.1063/1.328693

28. Xu, Y., Gnanasekaran, R., & Leitner, D. M. (2012). Analysis of Water and Hydrogen Bond Dynamics at the Surface of an Antifreeze Protein. *Journal of Atomic and Molecular Physics*, *2012*, e125071. https://doi.org/10.1155/2012/125071

29. Valdés-Tresanco, M. S., Valdés-Tresanco, M. E., Valiente, P. A., & Moreno, E. (2021). gmx_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS. *Journal of Chemical Theory and Computation*, *17*(10), 6281–6291. https://doi.org/10.1021/acs.jctc.1c00645

30. Miller, B. R. I., McGee, T. D. Jr., Swails, J. M., Homeyer, N., Gohlke, H., & Roitberg, A. E. (2012). MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *Journal of Chemical Theory and Computation*, *8*(9), 3314–3321. https://doi.org/10.1021/ct300418h

31. Bu, C., Zhang, L., Cui, H., Yang, G., & Wu, H. (2023). Dynamic Inference via Localizing Semantic Intervals in Sensor Data for Budget-Tunable Activity Recognition. *IEEE Transactions on Industrial Informatics*, *PP*, 1–13. https://doi.org/10.1109/TII.2023.3315773

32. Cheng, D., Zhang, L., Bu, C., Wang, X., Wu, H., & Song, A. (2023). ProtoHAR: Prototype Guided Personalized Federated Learning for Human Activity Recognition. *IEEE Journal of Biomedical and Health Informatics*, *27*(8), 3900–3911. Presented at the IEEE Journal of Biomedical and Health Informatics. https://doi.org/10.1109/JBHI.2023.3275438

33. Mirrashid, M., & Naderpour, H. (2023). Incomprehensible but Intelligible-in-time logics: Theory and optimization algorithm. *Knowledge-Based Systems*, *264*, 110305. https://doi.org/10.1016/j.knosys.2023.110305

34. Xu, S., Zhang, L., Tang, Y., Han, C., Wu, H., & Song, A. (2023). Channel Attention for Sensor-Based Activity Recognition: Embedding Features into all Frequencies in DCT Domain. *IEEE Transactions on Knowledge and Data Engineering*, *35*(12), 12497–12512. Presented at the IEEE Transactions on Knowledge and

Data Engineering. https://doi.org/10.1109/TKDE.2023.3277839

35. Polonelli, L., Pontón, J., Elguezabal, N., Moragues, M. D., Casoli, C., Pilotti, E., … Travassos, L. R. (2008). Antibody complementarity-determining regions (CDRs) can display differential antimicrobial, antiviral and antitumor activities. *PloS One*, *3*(6), e2371. https://doi.org/10.1371/journal.pone.0002371

36. Van Regenmortel, M. H. V., & Azimzadeh, A. (2000). Determination of Antibody Affinity. *Journal of Immunoassay*, *21*(2–3), 211–234. https://doi.org/10.1080/01971520009349534

37. Guex, N., & Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modeling. *ELECTROPHORESIS*, *18*(15), 2714–2723. https://doi.org/10.1002/elps.1150181505

38. Benson, N. C., & Daggett, V. (2012). A Comparison of Multiscale Methods for the Analysis of Molecular Dynamics Simulations. *The Journal of Physical Chemistry B*, *116*(29), 8722–8731. https://doi.org/10.1021/jp302103t

39. Camacho, C. J., Weng, Z., Vajda, S., & DeLisi, C. (1999). Free Energy Landscapes of Encounter Complexes in Protein-Protein Association. *Biophysical Journal*, *76*(3), 1166–1178. https://doi.org/10.1016/S0006-3495(99)77281-4

40. Wang, E., Sun, H., Wang, J., Wang, Z., Liu, H., Zhang, J. Z. H., & Hou, T. (2019). End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chemical Reviews*, *119*(16), 9478–9508. https://doi.org/10.1021/acs.chemrev.9b00055

41. Carter, P. J. (2006). Potent antibody therapeutics by design. *Nature Reviews Immunology*, *6*(5), 343–357. https://doi.org/10.1038/nri1837

42. Reichert, J. (2008). Monoclonal Antibodies as Innovative Therapeutics. *Current Pharmaceutical Biotechnology*, *9*(6), 423–430. https://doi.org/10.2174/138920108786786358

43. Nelson, A. L., & Reichert, J. M. (2009). Development trends for therapeutic antibody fragments. *Nature Biotechnology*, *27*(4), 331–337. https://doi.org/10.1038/nbt0409-331

44. Correia, B. E., Bates, J. T., Loomis, R. J., Baneyx, G., Carrico, C., Jardine, J. G., … Schief, W. R. (2014). Proof of principle for epitope-focused vaccine design. *Nature*, *507*(7491), 201–206. https://doi.org/10.1038/nature12966

45. Weiner, L. M., Dhodapkar, M. V., & Ferrone, S. (2009). Monoclonal antibodies for cancer immunotherapy. *The Lancet*, *373*(9668), 1033–1040. https://doi.org/10.1016/S0140-6736(09)60251-8

46. Tsurushita, N., Hinton, P. R., & Kumar, S. (2005). Design of humanized antibodies: From anti-Tac to Zenapax. *Methods*, *36*(1), 69–83. https://doi.org/10.1016/j.ymeth.2005.01.007