Computational and Structural Biotechnology Journal 23 (2024) 3759–3770

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Determining key residues of engineered scFv antibody variants with improved MMP-9 binding using deep sequencing and machine learning

Masoud Kalantar [a], Ifthichar Kalanther [b], Sachin Kumar [a], Elham Khorasani Buxton [b], Maryam Raeeszadeh-Sarmazdeh [a,*]

[a] Department of Chemical and Materials Engineering, University of Nevada, Reno, NV 89557, USA
[b] Department of Computer Science, University of Illinois, Springfield, USA

ARTICLE INFO

Keywords:
Metalloproteinase
Antibody engineering
MMP-9
Single-chain antibody fragment
Yeast surface display
Protein complex structural modeling
Machine learning

ABSTRACT

Given the crucial role of specific matrix metalloproteinases (MMPs) in the extracellular matrix, an imbalance in the regulation of activation of matrix metalloproteinase-9 (MMP-9) zymogen and inhibition of the enzyme can result in various diseases, such as cancer, neurodegenerative, and gynecological diseases. Thus, developing novel therapeutics that target MMP-9 with single-chain antibody fragments (scFvs) is a promising approach. We used fluorescent-activated cell sorting (FACS) to screen a synthetic scFv antibody library displayed on yeast for enhanced binding to MMP-9. The screened scFv mutants demonstrated improved binding to MMP-9 compared to the natural inhibitor of MMPs, tissue inhibitor of metalloproteinases (TIMPs). To identify the molecular determinants of these engineered scFv variants that affect binding to MMP-9, we used next-generation DNA sequencing and computational protein structure analysis. Additionally, a deep-learning language model was trained on the screened scFv library of variants to predict the binding affinities of scFv variants based on their CDR-H3 sequences.

1. Introduction

Matrix metalloproteinases (MMPs) are a group of proteases responsible for remodeling of the extracellular matrix [1,2]. When not properly regulated by their inhibitors and activators, specific MMPs are associated with various diseases, making them promising targets for novel protein-based therapeutics [1,3]. Monoclonal antibodies (mAbs) are well-established therapeutic proteins, generally well-tolerated with minimal risk of side effects [4]. Various antibodies have been previously used [5–7] and engineered [8–10] to target proteases, especially MMPs. The large antigen-binding interface of mAbs, combined with multiple flexible binding loops and complementarity-determining regions (CDRs), allows for high binding affinity and selectivity [11]. This can be further enhanced through protein engineering techniques like directed evolution and yeast surface display [12–14]. Additionally, established platforms for antibody production and purification make mAbs great protein alternatives for developing MMP binders [12,15,16].

Among MMPs, the overexpression of MMP-9 is strongly associated with poor prognosis in various cancers [17,18], neurodegenerative [1,4] and female health-related diseases [19–21]. Further, advances in generating humanized antibodies, and various antibody fragments, such as antigen-binding fragments (Fab), single chain antibody fragments (scFv), and camelid antibodies, as recombinant proteins [22] has made antibodies one of the leading classes of biological binding molecules. Antibody variants have been easily cloned, produced, and genetically manipulated for diverse purposes. Antibody discovery efforts to find potential therapeutics targeting MMPs, specifically MMP-9 due to its pathological contribution in several diseases, led to several hit antibodies [23,24].

The Complementarity-Determining Region 3 of the Heavy Chain (CDR-H3) in monoclonal antibodies (mAbs) is essential for antigen recognition, playing a significant role in improving antigen binding affinity and selectivity [25,26]. However, developing effective antibodies targeting enzymes like MMPs can be challenging, as the active sites of

Abbreviations: MMP, matrix metalloproteinase; FACS, fluorescent-activated cell sorting; scFv, single-chain antibody fragment; TIMP, tissue inhibitors of metalloproteinases; DL, deep learning; NLP, natural language processing; LPLM, large protein language model; LSTM, long-short-term memory; SHAP, Shapley Additive exPlanations; ML, machine learning; ESM, evolutionary scale modeling; PCA, principal component analysis; t-SNE, t-distributed Stochastic neighbor embedding; MLM, Masked Language Modeling.

* Correspondence to: Department of Chemical and Materials Engineering, University of Nevada, 1664 N. Virginia St, Reno, NV 89557, USA.
E-mail address: maryamr@unr.edu (M. Raeeszadeh-Sarmazdeh).

https://doi.org/10.1016/j.csbj.2024.10.005
Received 16 July 2024; Received in revised form 1 October 2024; Accepted 1 October 2024
Available online 10 October 2024

these enzymes are often deeply recessed within a major cleft or concave structure, which can be difficult for the antigen-binding sites of native human or murine antibodies to access. Native antibodies typically have relatively flat surfaces, with CDR-H3 regions averaging 9 to 12 amino acids in length. To address this issue, several synthetic human antibody libraries have been developed by incorporating antigen-binding regions (paratopes) from camelid antibodies, which are known for their longer CDR-H3 regions. These longer CDR-H3 loops are better suited to penetrate the concave active sites of MMPs, improving the chances of effective binding [27]. To date, some MMP-9 antibodies have been developed with inhibitory function [28–30]. For example, REGA-3G12, an MMP-9 monoclonal antibody, binds to the enzyme's catalytic site but does not directly interact with the catalytic zinc ion or its adjacent residues [23]. These results suggest that particularly for active site inhibitors, the optimal distribution of CDR-H3 lengths and amino acid compositions is crucial for achieving paratope conformations compatible with the structural conformation of the targeted MMPs' active sites.

The advancement of machine learning (ML), particularly, deep learning (DL) and natural language processing (NLP) technologies, along with increased computing power, has further enhanced biotechnological applications, including protein design and engineering [31–35]. These developments have led to the creation of Large Protein Language Models (LPLMs), which assist in discovering the evolutionary, structural, and functional properties across protein space by encoding amino-acid sequences into numeric vector representations [36]. In this study, we leveraged pretrained LPLMs to extract features from CDR-H3 sequencing data to train a downstream Long-Short-Term Memory (LSTM) model to predict the binding affinity between CDR-H3 and MMP-9cd. To understand the predictive influence of each amino acid in CDR-H3 on the binding affinity, we used Shapley Additive exPlanations (SHAP). This technique applies game theory to explain the contribution of each input feature to the prediction made by an ML model. By analyzing the file generated by this technique, we can gain insights into feature importance, understand the distribution of feature impacts, detect interactions between features, and interpret the overall behavior of the model.

This study uses directed evolution and yeast display to screen a synthetic scFv antibody library previously engineered for decreasing non-specific binding [25] to improve binding to the MMP-9cd (Fig. 1). We achieved this by combining the fluorescent-activated cell sorting (FACS) screening approach with next-generation sequencing (NGS) analysis. FACS allows for high throughput screening of high-affinity binders, while NGS offers an unparalleled level of DNA sequencing details compared to traditional methods. By providing both the sequence and frequency information for each scFv antibody in the library, NGS allows us to pinpoint key amino acid residues crucial for binding to MMP-9cd. The amino acid frequency of CDRH3 of scFv variants, obtained from deep sequencing of negative (non-binders) and positive (binders) after FACS sorts toward MMP-9, revealed some residues were consistently present in high-affinity binders, suggesting their crucial role in binding interactions. These results were also used to fine-tune protein language models. These ML-developed models focused on CDR-H3 predicted not only the scFv variants kept as testing population with high precision, but also other negative and positive MMP-9 binders that were unrelated to this study or library. This comprehensive approach goes beyond traditional screening techniques, enabling the identification of promising candidates that might otherwise be missed.

2. Materials and methods

2.1. Strain and plasmids

Saccharomyces cerevisiae, strain EBY100 (aGAL1-AGA1:: URA3 ura3 52trp1leu2Δ200pep4::HIS2prb1Δ1.6Rcan1 GAL) and RJY100 strain [25] were used for yeast surface display of the N-terminal domain of TIMP-1 (N-TIMP-1) as the control, and scFv variants in the naïve library, respectively. The N-TIMP-1 is expressed at the N-terminus of the Aga2p protein and subsequently integrated into the pCHA vector backbone, resulting in a free N-terminus capable of binding to MMP-9cd. As for scFv variants, the pCTCON2 vector was employed to express the scFv at the C-terminus of Aga2p [25].

2.2. Yeast surface display of N-TIMP-1 and scFvs

The yeast cells, which had been electroporated with plasmids, were incubated overnight in minimal SDCAA, pH 6. This media composition included 20 g/L dextrose, 6.7 g/L yeast nitrogen base, 5 g/L Bacto casamino acids, 10.19 g/L sodium phosphate dibasic ($Na_2HPO_4{\cdot}7H_2O$),

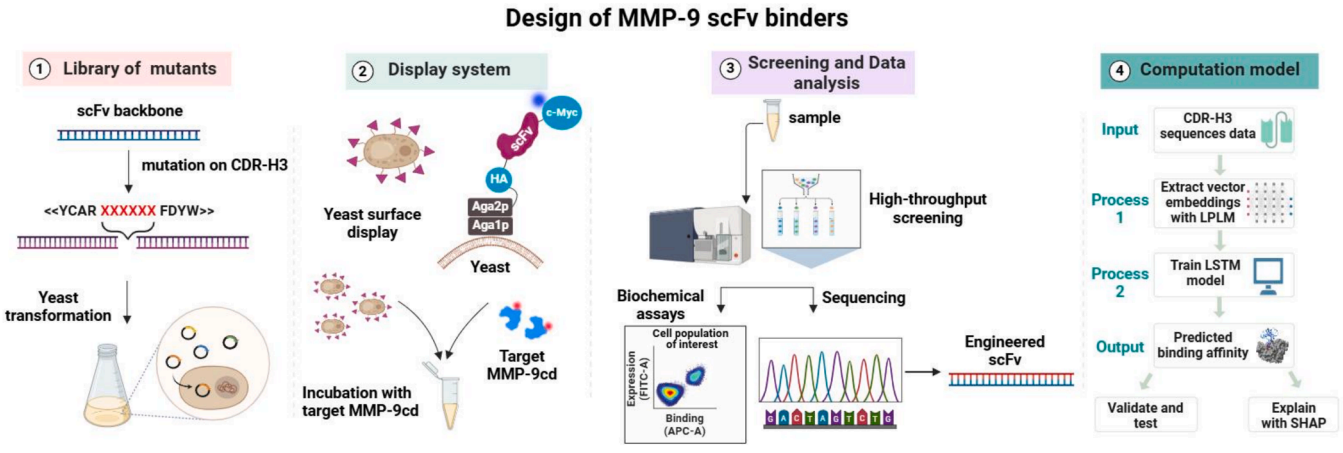


Fig. 1. The general approach for protein engineering and design of antibody scFv variants using directed evolution, yeast surface display to target MMP-9. 1) *Library generation:* a library of scFv variants with mutations in the CDR-H3 region to introduce diversity in both amino acid composition and length was used. These scFv variants were then electrotransformed into a yeast strain for display, labeling, and screening. 2) *Expression and display:* yeast cells carrying expression plasmid vectors encoding different scFvs were grown and induced for display of the scFvs variants genetically fused to the C-terminus of Aga2p on the yeast surface. Cells expressing scFv variants were incubated with MMP-9cd enzyme and further with proper fluorescent conjugated ligands that label scFv variants for further binding analysis. 3) *Screening and sequencing*: the scFv library of mutants was screened for binding to MMP-9cd into binder (positive) and non-binder (negative) populations using FACS. The DNA isolated and amplified from the sorted scFv libraries were sequenced via Sanger sequencing and/or next-generation sequencing. 4) *Data analysis and model training*: The sequencing data extracted from NGS was used as an input for a machine learning model determining key residues of CDR-H3 on MMP-9cd binding. This data was used to train and validate a protein language model to predict the binding affinity of specific CDR-H3 regions to MMP-9cd.

and 8.56 g/L sodium dihydrogen phosphate monohydrate ($NaH_2PO_4{\cdot}H_2O$) and in a shaker incubator at 30 °C and 250 rpm. Subsequently, the yeast cells were induced in SGCAA media (similar to SDCAA but with 20 g/L galactose instead of dextrose, pH 6) for 20 h at 30 °C, starting with an initial OD600 of 1.00. Yeast cells displaying N-TIMP-1 or scFv variants were harvested at an OD600 of 0.2 and washed twice in 750 μL of ice-cold PBSA (8 g/L NaCl, 0.2 g/L KCl, 1.44 g/L Na_2HPO_4, 0.24 g/L KH_2PO_4, pH 7.4 %, and 0.1 % BSA). Subsequently, these cells were incubated with 6xHis-MMP-9cd from Enzo Life Sciences (Farmingdale, NY), a final concentration of 400 nM, for 60 min at room temperature, following previously reported methods [37]. After incubation, the samples underwent two washes with 750 μL of ice-cold PBSA, and the cells were kept on ice. For the primary antibody labeling process, mouse anti-c-Myc Antibody (GenScript, NJ) solution at a 1:100 ratio in PBSA buffer (0.25 mg/ML) was used to incubate the cell for 30 min on ice. Subsequently, for the secondary antibody labeling, the cells washed three times in ice-cold PBSA were incubated with goat anti-mouse Alexa Fluor 488 antibody (Invitrogen, 2 mg/ML) and anti-6xHis monoclonal antibody conjugated with Alexa Fluor 647 (Invitrogen, 1 mg/ML) at 1:100 dilution in PBSA for either one for 30 min on ice, shielded from light. Following the completion of the final washes, the labeled cells were resuspended in 750 μL of PBSA for analysis using a BD Accuri™ C6 Plus flow cytometer (BD Biosciences, NJ). The data obtained from the flow cytometer were further analyzed using FlowJo software (FlowJo, LLC, OR).

2.3. Protein expression and purification

Recombinant active human MMP-9cd (auto-cleavage resistance) [38] with an N-terminal 6xHis tag was expressed in Rosetta™(DE3) pLysS (Millipore Sigma) cells using a pET-28a(+)−6xHis-MMP-9cd vector as previously described [37,39,40]. Briefly, MMP-9cd protein expression was induced with 0.5 mM IPTG for 21 h at 25 °C. Insoluble MMP-9cd protein was extracted from E. coli cells via sonication and solubilized under denaturing conditions with urea. Bacterial pellets were mixed with lysis buffer and kept shaking overnight at 4 °C. Lysate was then incubated with sodium deoxycholate and DNase I for 1 h at room temperature. The lysate was then separated by centrifugation, and aliquots were collected for subsequent SDS-PAGE analysis. This cycle of sonication, solubilization, lysis, and centrifugation was repeated until no insoluble material remained in the pellet fraction. The pooled supernatants containing solubilized MMP-9cd were purified by immobilized metal affinity chromatography (IMAC) using Ni resin. Finally, the purified MMP-9cd was refolded through gradient dialysis. For the original sorts, purchased 6xHis-MMP-9cd protein from Enzo Life Sciences (Farmingdale, NY) was used. Moreover, human MMP-3cd with a C-terminal 6xHis tag was expressed using pET-29b(+) and purified as previously discussed [39].

2.4. Screening the yeast-displayed scFv library using FACS

The non-specific scFv library, a generous gift from the Wittrup lab-MIT-Chemical Engineering, was previously engineered to minimize the non-specific binding to various protein targets [25]. The library's structure consisted of 5 different heavy chain segments (V_H) and 3 variations of light chains (V_L), connected by a glycine linker, all presented in the V_L-V_H format. To expand the diversity of the library, mutations were applied to the CDR-H3 loop in terms of length and amino acid compositions [25]. The library of synthetic scFv library was originally recovered from the glycerol stock in 50 ML SD-CAA media (20 g glucose, 6.7 g yeast nitrogen base without amino acids, 5 g casamino acids, 10.4 g sodium citrate, 7.4 g citric acid monohydrate, pH 4.5) containing 100 μg/ML ampicillin to prevent bacterial growth. The library was diluted to 500–1000 L after overnight growth at 30 °C shaker. Before each round of cell sorting, the yeast cells were induced in SGCAA media. The number of cells was measured to reach a density of OD_{600} of 1.00 ($10^{\wedge}7$ cells/ML). Subsequently, the cells were incubated with purified 6xHis-MMP-9cd protein, followed by primary and secondary antibody labeling. After washing, the cells were resuspended in ice-cold PBSA buffer, following the procedure outlined in the "Yeast surface display of N-TIMP-1 and scFvs" section. These samples were kept on ice and protected from light until loaded into the BD FACS Aria II cell sorter. A pentagon gate was applied to screen variants that exhibited strong signals for Alexa Fluor-488 and Alexa Fluor-647, indicating expression and positive binding to MMP-9cd, respectively. Additionally, another rectangular gate was used to collect clones with a positive signal for Alexa Fluor 488 and a weak signal for Alexa Fluor 647, indicating non-binder scFv variants to MMP-9cd.

After sorting, the cells were recovered by inoculating them in a 50 ML SDCAA medium with a pH of 4.5. They were then allowed to grow overnight at 30 °C. The yeast library was stored in 20 % glycerol stock at −80 °C for longer storage. The library underwent two rounds of screening, each involving staining, sorting, recovery, and regrowth. In the first round, cells were incubated with 300 nM MMP-9cd and sorted based on yield mode to eliminate unwanted variants. Approximately 5 % of the population ($10^{\wedge}6$ cells) was collected for the positive gate (binders), while up to 3 % ($3 \times 10^{\wedge}5$ cells) were collected for the negative gate (weak or non-binders). In the second round, the concentration of MMP-9cd was reduced to 100 nM, and the sorting mode was changed to purity to increase the efficiency of sorted cells, which exhibit both positive expression and binding signals. Up to 2 % of the population ($5 \times 10^{\wedge}5$ cells) was collected in the last sort.

2.5. DNA isolation for Sanger and Next-generation sequencing

After sorting using FACS, isolated yeast clones were grown on selective SDCAA plates at a 30 °C incubator for further binding analysis and DNA sequencing. For the Sanger sequencing of isolated clones, the yeast DNA plasmids were extracted using the Zymoprep Miniprep II kit (Zymoprep), amplified using PCR, and purified using the SV Gel and PCR Clean-Up System (Promega Corporation) and were sequenced at the Eurofins genomics.

For the next-generation sequencing (NGS) analysis, the yeast plasmid DNA from the scFv antibody libraries, either negative or positive sorts, were isolated using the Zymoprep Miniprep II kit. To ensure high-quality sequencing data, the plasmid DNA underwent Lambda-Exo digestion to remove impurities, including yeast cell genomes, from the extracted DNA as previously described [41]. Subsequently, the heavy chain genes were selectively amplified by PCR using Phusion high-fidelity polymerase and corresponding primers that were up and downstream of CDR-H3. This targeted amplification strategy was chosen due to the high mutation rate in the CDR-H3 and further analysis of this region. Finally, the amplified scFv variants (positive or negative) were sequenced by Azenta/Genewiz using an Illumina sequencing platform.

DNA sequences obtained from NGS were analyzed using multiple Python scripts. These scripts were designed to sort each V_H variation and align the CDR-H3 regions. The Python codes included translation of DNA sequences into three forward and three reverse frames of amino acid sequences, identification of amino acid residue mutations in the CDR-H3 region and counting the number of repetitions. Subsequently, the frequency of amino acid residues in the CDR-H3 region and the CDR-H3 length were compiled and graphed using PRISM software packages (GraphPad Software, Inc., CA).

2.6. Training a DL model for predicting binding

DL models were trained using 2380 unique CDR-H3 sequences of scFv variants in the positive gate and 153 unique CDR-H3 sequences in the negative gate. Two large protein language models (LPLM) were used for extracting features from protein sequences: The Evolutionary Scale Modeling (ESM)−2 models (with 650B, 3B, and 15B parameters) [31] and the AntiBERTy model with 26 M parameters [42]. ESM-2, a

state-of-the-art LPLM developed by Meta AI, is trained on sequences from the UniRef protein sequence database using a Masked Language Modeling (MLM) objective. AntiBERTy is similar to the ESM-2 model but has fewer parameters and was trained exclusively on antibody sequences, capturing the diversity and specificity of the immune repertoire. In the MLM, amino acids are randomly masked in a protein sequence, and the model is trained to predict the missing amino acids from their surrounding context. This approach helps the language model learn vector representations (called embeddings) that capture patterns and dependencies in protein sequences, which can then be used in downstream protein prediction tasks. The performance of LPLM embeddings in downstream tasks typically depends on the model's size (number of parameters) and the diversity of its pre-training database [31].

A low-dimensional visualization of features was extracted from each LPLM model. The embeddings produced by the LPLM for each amino acid in a sequence were averaged over CDR-H3 positions (Fig. S1A). Then, the principal components were extracted from the embedding vectors and used for two-dimensional projection using the t-SNE algorithm [43]. The non-binders (orange dots) form a cluster in the embedding space, demonstrating the utility of LPLMs for distinguishing between CDR-H3 regions that are binders and non-binders (Fig. S1B). Embeddings extracted for each amino acid position in the CDR-H3 region were padded to the same length and used to train a downstream LSTM model to predict binding affinity with MMP-9cd. LSTM models are particularly suitable for capturing long-range dependencies in sequence data [44] and can be effectively used to learn interactions between residues that might be far apart in sequence [45]. While more advanced models like transformers can be used to capture these dependencies, LSTMs are computationally more efficient when dealing with relatively short sequences, such as CDR-H3.

The sequences in the dataset were split into 80 % training and 20 % test sets. Each model was trained for 50 epochs with early stopping and the Adam optimizer. The initial learning rate was 0.001, with decay steps of 1000 and a decay rate of 0.9. The predictive performance of the model was evaluated using precision, recall, and F1 metrics in both a cross-validation setting and an independent out-of-sample test set. To understand the predictive influence of each amino acid in CDR-H3 on binding affinity, Shapley Additive exPlanations (SHAP) were employed. SHAP, a technique based on game theory, explains the contribution of each input feature to the prediction made by a machine learning (ML) model [46,47]. The DeepSHAP algorithm [46] was used to compute the predictive significance of each amino acid residue in CDR-H3 binding. DeepSHAP uses reference baselines to compute differences in neuron activations and linearly decomposes the model's output into contributions from each input feature. These contributions were then aggregated over multiple reference baselines to capture the effect of all possible feature combinations, providing a scalable and accurate approximation of Shapley values for deep learning models. Contributions were averaged over the embedding dimension to produce one Shapley value per amino acid residue in a CDR-H3, and SHAP force plots were then created for each sequence. The force plots visualize the impact of each residue on the binding of a single CDR-H3, showing how the residue pushed the prediction for CDR-H3 from the base value to the model output. To understand the overall importance of each position or residue in binding, global Shapley values were calculated by averaging the Shapley values over all CDR-H3 samples in the dataset.

2.7. AlphaFold2 protein complex modeling and analysis

AlphaFold2-Multimer pipeline was used to predict the complete three-dimensional structure of the MMP-9cd-scFv complex [48]. This approach involved incorporating the amino acid sequences of both the antibody heavy and light chains alongside the MMP-9cd sequence within a single FASTA file. Subsequently, the generated structures were subjected to a relaxation process using Amber, a molecular dynamics package, to relieve steric clashes and optimize the geometry for better physical realism. Also, up to 20 template hits were allowed during the modeling process to enhance the prediction accuracy. These templates provide structural references that guide the folding and interaction predictions.

3. Results

3.1. Screening of the synthetic scFv library for binding to MMP-9cd showed improvement in both expression and binding levels compared to the naïve library

A synthetic scFv antibody library previously engineered for nonspecific binding was used [25]. Previous study showed an enrichment of four amino acids (Gly, Val, Trp, and Arg) within the CDR-H3 region [25] of scFv antibody libraries. The Trp residue was shown to have a substantial effect on nonspecificity in binding to various antigen targets. This knowledge was subsequently applied to generate a new library (also known as naïve library in this study), resulting in scFv antibodies exhibiting robust binding to a wide range of antigens with minimal nonspecificity. The overall backbone of the library used a combination of five V_H and three V_L frameworks, with the majority of mutational diversity focused on the CDR-H3 loop [25]. Key highlights of this scFv library included the elimination of Trp and a significant drop in the frequency of Arg and Val. Additionally, allowing CDR-H3 loop length diversity (6 to 17 aa), which mimics the natural repertoire, and ensuring a library size of at least one billion members were among the significant improvements [25]. This synthetic scFv antibody library was subjected to two rounds of FACS targeting MMP-9cd. Two sort gates (positive and negative) were used to collect the cells sorted as positive binders with dual positive expression and binding signals, and negative binders with only positive expression determined by c-Myc detection and low MMP-9cd binding (Fig. 2**A**, Fig. S2).

Although the scFv antibody framework includes a combination of different light chains and heavy chains, the backbone and other CDR regions except for CDR-H3, which was heavily mutated (both length and amino acid composition), exhibit significant similarity to each other with up to 80 % sequence homology across the entire scFv backbone. Additionally, analyzing a large dataset with both bound and unbound antibody structures revealed minimal movement in most CDR regions, except for CDR-H3 from the antigen-binding site [49]. Therefore, the rationale behind this approach was based on previous studies demonstrating that CDR-H3 serves as the primary functional contributor to antigen recognition in most antigen-binding sites [50–52]. Additionally, separate high throughput screenings were conducted to isolate high-affinity binders using the positive gate and to identify weak or non-binders through the negative gate. The number of selection rounds was intentionally limited, aiming not only to isolate the few tightest binders but also to obtain a diverse set of unique binders for comprehensive statistical analysis.

A significant enrichment of positive scFv binders was observed compared to the original or naive library. An increase of more than 10 % in the overall expression of collected cells through the positive gate was noted, along with a similar increase in binding affinity (Fig. S3). The percentage of positive binders collected after two rounds of sorting in the P1 gate increased significantly from 3 % to 26 %. This enrichment suggests that the sorting strategy effectively isolated high-affinity binders from the naïve scFv library (Fig. 2**B**).

3.2. Next-generation sequencing of the screened scFv antibody library revealed enrichment of mainly polar residues after two rounds of FACS

Sequence analysis of scFv binders and non-binders from the positive and negative sort gates revealed amino acids likely to enhance binding affinity towards MMP-9. The next-generation sequencing (NGS) data, which included 2380 unique CDR-H3 sequences from the positive gate

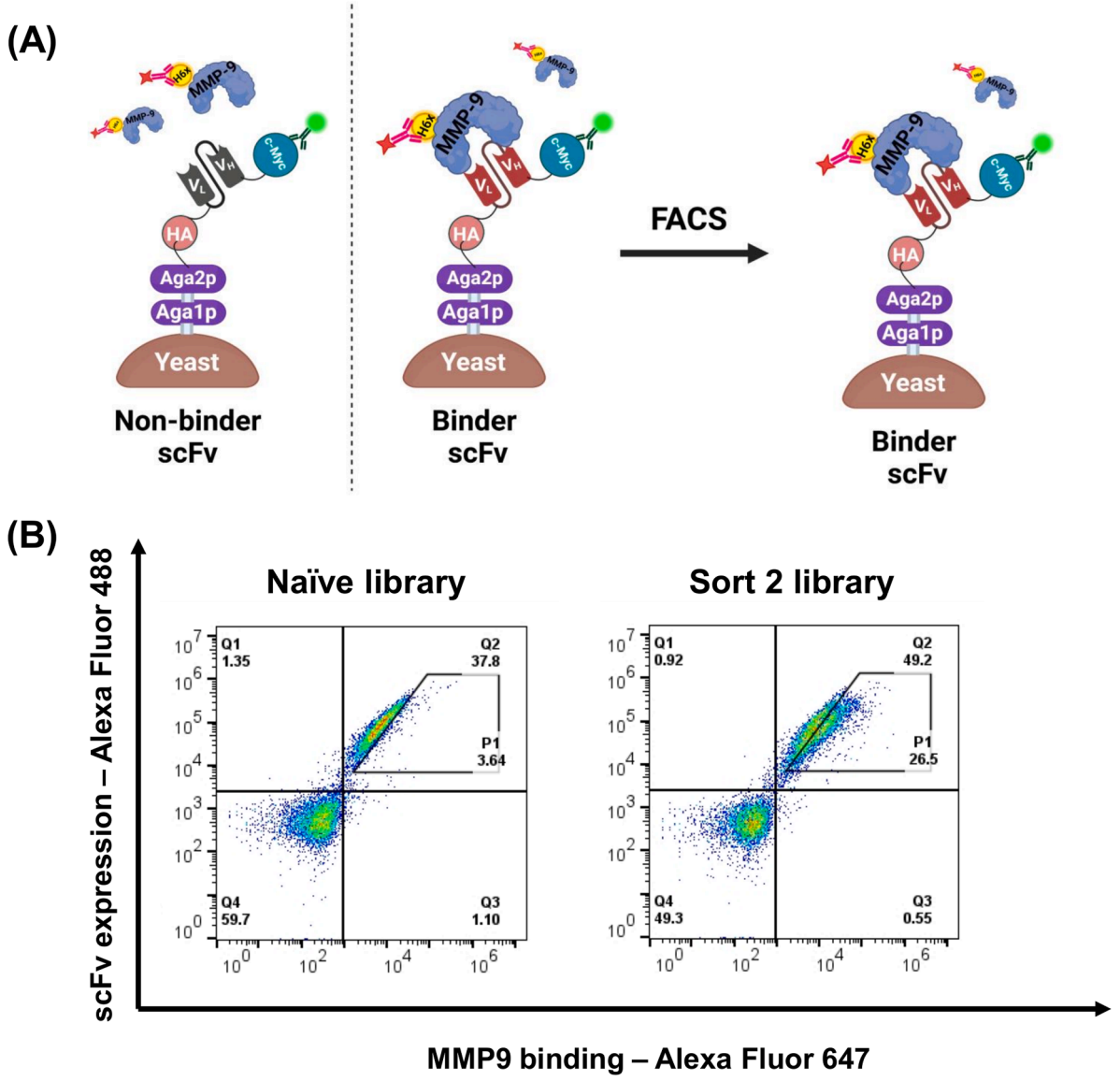


Fig. 2. FACS sorting of yeast cells displaying scFvs. A) Yeast cells displaying scFvs were directly incubated with fluorescent-conjugated c-myc antibody and the catalytic domain of MMP-9 (MMP-9cd) with 6xHis-tag, and then labeled with fluorescent-conjugated anti c-myc and anti-6xHis antibodies for quantitative analysis of expression and binding, respectively. B) Flow cytometry dual scatter plots of the naive scFv library and the screened scFv library after two rounds of FACS sorting toward MMP-9cd. The diagonal gate (P1) defines the enriched population of yeast cells displaying scFvs with high MMP-9cd binding affinity.

and 153 unique CDR-H3 sequences from the negative gate, indicated a higher frequency of certain amino acids in the CDR-H3 sequences of the final positive library compared to the naive library. Notably, the frequency of arginine (Arg) increased by approximately six-fold (Fig. 3**A,** Fig. S4). This observation is particularly intriguing given that the naive library was designed with a low frequency of Arg. The substantial enrichment of Arg in the binder-selected pool implies that its presence in the CDR-H3 loop likely contributes to the folding or structural stability of the heavy chain [53], enhancing the overall binding affinity since all collected scFv variants in the positive gate exhibit high expression levels. However, the enrichment of positively charged Arg residues could also be due to interactions with the negatively charged surface of MMP-9cd, particularly at the enzyme's active site and adjacent regions. This observation aligns with previous studies on CDR-H3 regions of high-affinity antibody binders, which have demonstrated that an increase in Arg residues occurs after affinity maturation. Additionally, a high Arg content in CDR-H3 has been correlated with increased nonspecific binding to the target antigen. Additionally, Arg side chains are expected to contribute favorably to binding energy in many protein-protein interactions [54]. In addition to Arg, other residues were enriched in the positive pool are mainly polar residues or those with polar side chains, such as Ser, Tyr, Gly, and Gln (Fig. 3**B**).

A thorough DNA sequencing analysis of all the unique sequences obtained for positive binders revealed that regardless of the length of CDR-H3, the combinations of Gly/Tyr/Ser residues are dominant in the screened scFv with improved MMP-9 binding (Fig. 3**C**). Overall, these results suggest that enrichment of Tyr residue, which is larger in size, likely facilitates favorable contacts with the MMP-9cd, while presence of smaller residues such as Ser and Gly may contribute to suitable conformations beneficial to high-affinity binding. This finding also aligns with previous evidence in the structures of Fabs screened from analogous minimalist libraries [55,56]. Taken together, these results suggest that high-affinity binding is best mediated by Tyr in combination with small, flexible residues like Gly and Ser.

The NGS analysis was also used to determine the frequency of CDR-H3 lengths among scFv variants that were screened as positive binders

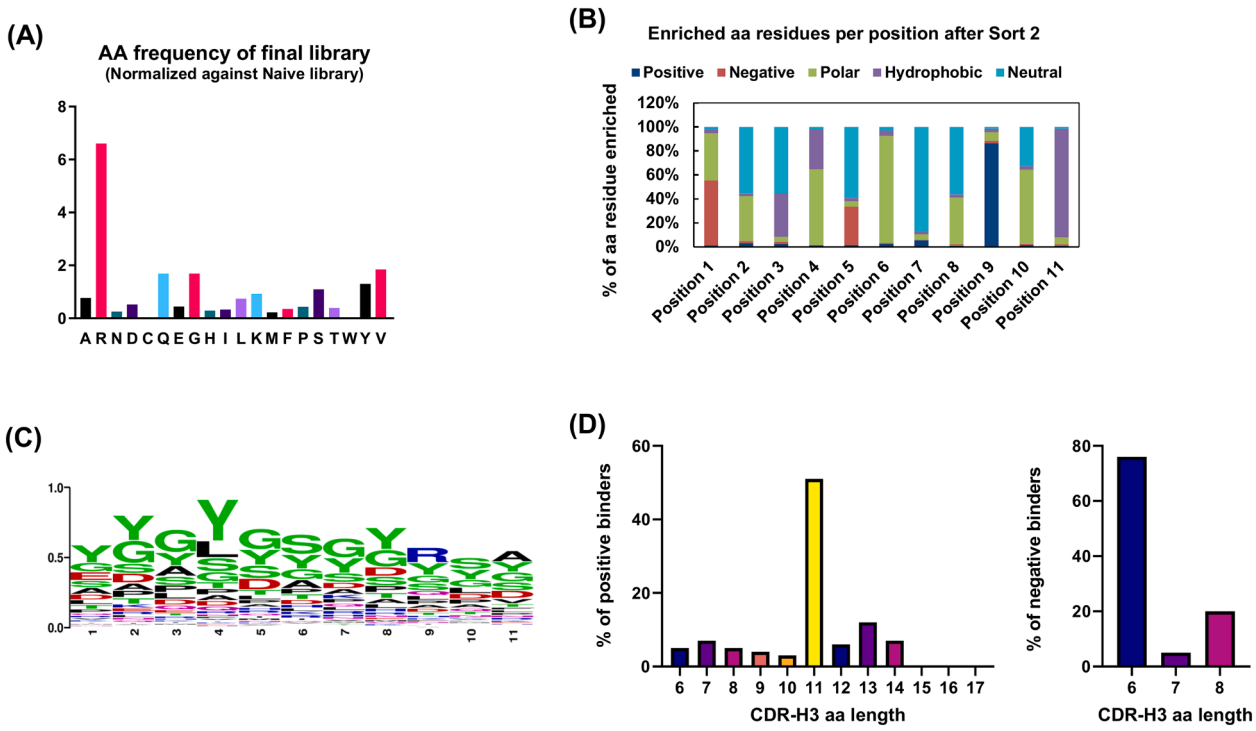


Fig. 3. Next-generation sequencing of binders and non-binders to MMP-9. A) The bar graph displays the frequency of amino acids in the CDR-H3 loop of the final scFv library, normalized to the naive library. Each bar's height indicates the relative enrichment or depletion of a specific amino acid in the final library. B) This bar graph represents the proportion of various amino acid types (Positive: R, K, H, Negative: D, E, Hydrophobic: A, I, L, M, F, W, V, Neutral: G, C, P, and Polar: N, Q, S, T, Y) at each position within the CDR-H3 region of the scFv variants with 11 aa lengths. Each bar corresponds to a specific position in the CDR-H3 region, showing the relative frequency of each amino acid type. The height of each colored segment within a bar indicates the prevalence of that particular amino acid type at the given position. C) The sequence logo illustrates the amino acid frequencies at each position within the CDR-H3 loop of the final scFv library with 11 aa residues. The height of each letter is proportional to its frequency at a given position, with taller letters indicating higher frequency. D) The distribution of CDR-H3 lengths among positive and negative MMP-9cd binders.

and non-binders to MMP-9cd. Interestingly, CDR-H3 sequences with a length of 11 amino acids appeared the most frequently in the pool of positive binders. This was followed by CDR-H3 lengths of 13, 14, and 12 residues, respectively. In contrast, non-binders exhibited an enrichment of CDR-H3 lengths at 6 and 8 amino acids (Fig. 3**D**). This is consistent with the length of CDR-H3 region (11 amino acids) in REGA-3G12 antibody, an inhibitor of MMP-9cd [57,58].

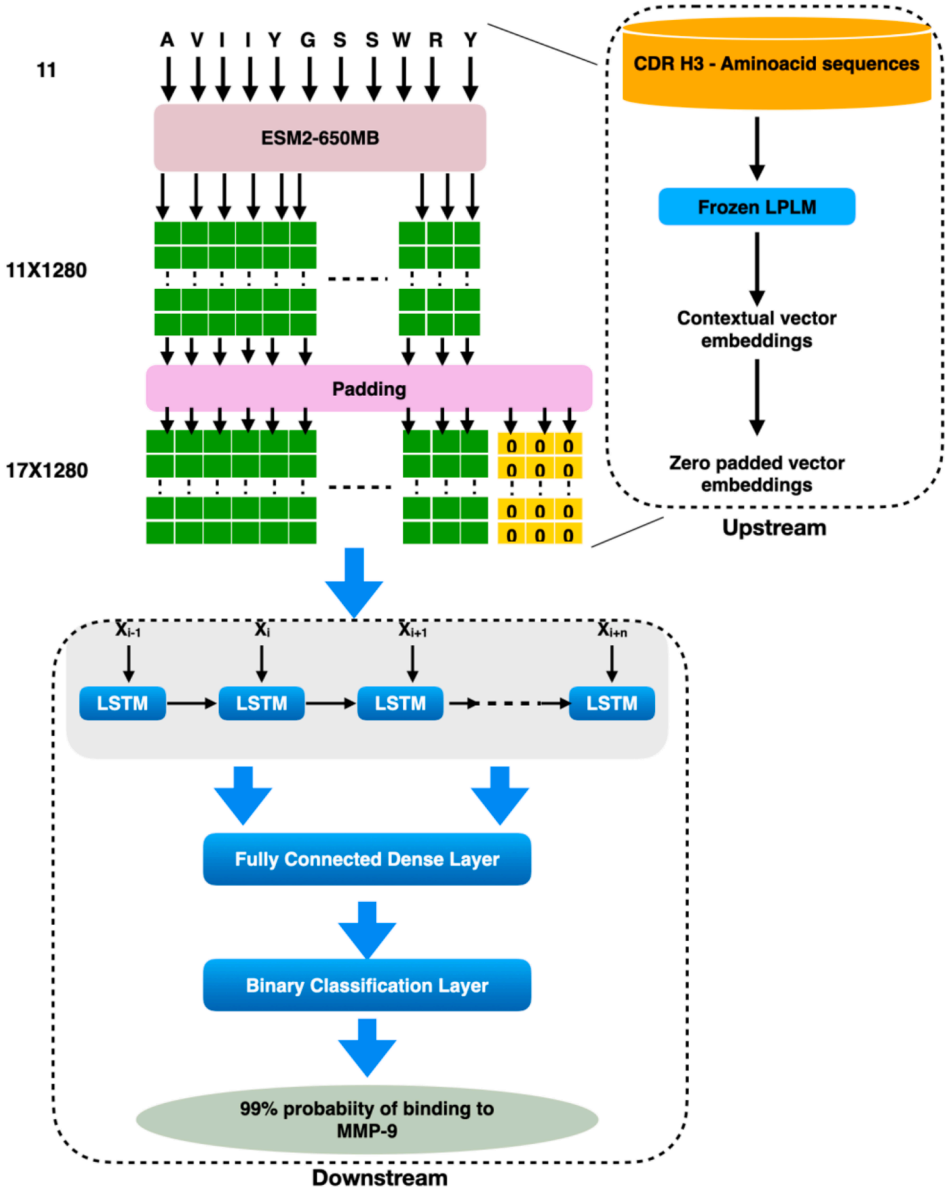
The CDR-H3 sequences from the negative binder population of scFv variants showed a high abundance of leucine (Leu), Phenylalanine (Phe), and methionine (Met) (Fig. S5). These amino acids are classified as non-polar and uncharged side chains, typically considered less favorable for making strong molecular interactions with MMP-9cd. Thus, there is a potential correlation between CDR-H3 length and the binding affinity of isolated scFv variants with improved binding to MMP-9. The enrichment of shorter CDR-H3 lengths in the non-binder population and the dominance of non-polar amino acid residues indicate a reduced capacity for forming key contacts with the MMP-9's binding pockets.

3.3. Performance of the DL model for predicting MMP-9cd binding of scFv variants

A stratified 10-fold cross validation was performed on the training sequences. Metrics for the LSTM classifiers trained on features extracted from ESM-2 (650MB, 3B, and 15B) and AntiBERTy models (Fig. 4). The precision indicates the percentage of CDRH-3 sequences predicted by the model to be binders that were experimentally verified as binders. The recall indicates the percentage of experimentally verified binders the models detected. The F1 score is the harmonic mean of precision and recall. The LSTM model trained on features extracted from LPLMs is very effective at predicting the binding affinity of CDR-H3 to MMP-9 with precision close to or above 99 %. The LSTM model trained on features extracted from the largest ESM model (ESM-2–15B) has the highest out-of-sample F1 score (99.11 %). However, its F1 score exceeds that of the smallest model, AntiBERTy, with 26 M parameters, by only 0.67 %. Considering the memory usage and the speed of inference for larger models, one might prefer the smaller AntiBERTy model over the larger ESM models for this application.

Force plots were created using the Shapley values generated by the DeepSHAP algorithm. The plot consists of arrows that represent each feature's contribution. The length of the arrow is proportional to the binding contribution's magnitude. Positive contributions were shown in red while negative contributions were shown in blue. Indices were used to distinguish between the same amino acids in different positions of CDR-H3 (Fig. 5**A**). For instance, REGA-3G12 (CDR-H3: AVIIYGSSWRY) was predicted to be a positive binder, with the Gly6 and Val2 residues primarily contributing to this binding prediction by the model. On the other hand, M0072 (CDR-H3: GAWYL), a non-binder scFv antibody to the active MMP-9cd [59], was correctly predicted as a negative binder. The Leu5 residue mainly drove this non-binding prediction, as this residue frequently appeared in non-binder sequences discussed in the NGS analysis section. The residue-position mapping also uncovers specific interactions between residues and positions critical for understanding the scFv variants' binding mechanisms to MMP-9cd (Fig. 5**B**).

Analyzing Shapley values extracted from the ML model can be useful in identifying specific residues and positions in the CDR-H3 that enhance binding to MMP-9cd. The average (global) Shapley value for a specific position represents the mean contribution of that position to the scFv binding affinity to MMP-9cd, across different possible amino acids (Fig. S6A). The positions with high positive average Shapley values (1



Embedding Model	Classifier	Mean Cross Validation			Mean Test Performance		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score
650MB ESM2	LSTM	98.88%	100.00%	99.43%	98.02%	98.82%	98.41%
3B ESM2	LSTM	99.84%	100.00%	99.92%	98.44%	99.33%	98.88%
15B ESM2	LSTM	99.90%	99.95%	99.92%	98.65%	99.58%	99.11%
AntiBERTy	LSTM	99.95%	99.89%	99.92%	98.24%	98.61%	98.43%

Fig. 4. A computational model for predicting binding affinity of CDR-H3 to MMP-9. The pre-trained LPLMs were used to extract the vector representations (embeddings) for the CDR-H3s, which were padded to achieve a uniform length. These embeddings were used to train a downstream LSTM model with LSTM, dense, and binary classification layers. The REGA-3G12 CDR-H3 (AVIIYGSSWRY) with a length of 11 amino acids was passed to the pre-trained LPLM ESM-2 650MB. This model produced 11 × 1280 vector representations, with 1280 vectors corresponding to each amino acid. To achieve a uniform length of 17 × 1280 (where 17 is the maximum length of CDR-H3 in the training dataset), the embeddings were zero-padded. These zero-padded embeddings were input into the downstream LSTM model, which had been previously trained to predict CDR-H3 binding affinity to MMP-9. The model predicted a 99 % probability of binding to MMP-9cd for the REGA-3G12 CDR-H3 (AVIIYGSSWRY).

through 6) are critical and could be prime targets for mutations while designing targeted libraries to improve binding affinity. Conversely, positions with low or negative Shapley values are associated with low or no binding to MMP-9cd. The average global Shapley value per amino acid residue represents the mean contribution of that residue to the scFv variants' binding to the MMP-9cd (Fig. S6B). The bar chart reveals that the amino acid residues Cys, Phe, and Met, in general, have negative effects on binding irrespective of their locations. Positional variability captures the fluctuations in Shapley values for each position within the CDR-H3 region (Fig. S6C). The contributions of positions 1 through 7 can vary greatly depending on the specific amino acid at these positions. The amino acid variability at these positions highlights the flexibility of

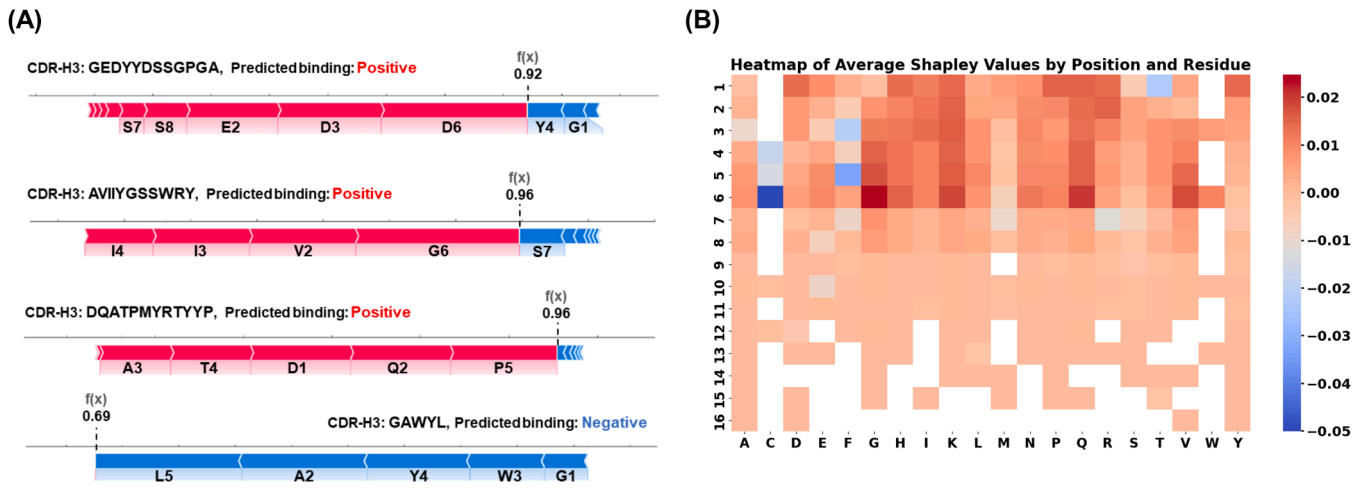


Fig. 5. Residue-Position Mapping. A) The Shapley plot illustrates the final prediction of the machine learning model. Red arrows indicate specific amino acids that positively contribute to the binding prediction, while blue arrows represent amino acid residues that negatively participate in the binding prediction. B) The heatmap represents the impact of various amino acid residues at specific positions on binding affinity. Higher Shapley values (warmer colors) show positive contributions to MMP-9 binding, highlighting crucial interactions between residues and positions. Conversely, lower Shapley values (cooler colors) indicate negative contributions.

these residues for binding interactions to the MMP-9cd target and could be targeted for further optimization.

3.4. The engineered scFv variants showed improvement in MMP-9cd binding compared to N-TIMP-1

The individual scFv variants isolated after FACS screening with an improved binding affinity towards MMP-9cd were isolated, grown, and induced for scFv expression on the yeast and MMP-9 binding. Yeast cell surface display serves as a valuable high throughput screening platform for quantitative analysis of the expression level of stable scFv mutants displayed on the yeast [53,60] and binding to soluble MMP proteins using immunolabeling and flow cytometry. A significant increase in expression levels (up to three-fold) and MMP-9cd binding up to six-fold was achieved compared to N-TIMP-1, a known tight binder of MMP-9cd [41](Fig. 6**A,** Fig. S7). This finding represents a significant achievement in developing high-affinity binders for MMP-9cd after two rounds of FACS. The flow cytometry dual scatter plots also represent the scFv expression and MMP-9cd binding level for these scFv variants compared to N-TIMP-1 and REGA-3G12 (**Fig. 6B,** Fig. S8), demonstrating higher expression and MMP-9cd binding values compared to N-TIMP1. Additionally, these scFv variants were tested for binding toward MMP-3cd, another MMP family member that shares similarities in the active site with MMP-9cd but is different in sequence, structure, and function. The results showed that all isolated scFv clones isolated after screening toward MMP-9cd exhibited a higher binding to MMP-9cd compared to MMP-3cd up to three-fold (Fig. S9). This visual evidence further supports the significant increase in binding affinity and expression levels of the scFvs isolated from the synthetic scFv antibody library.

The Sanger sequencing of isolated clones (Table 1), shows that clones containing negatively charged residues such as Asp, and positively charged residues like Lys, His, and Arg exhibit higher expression compared to those that do not contain these residues. Overall, an increase in charged and hydrophilic residues, along with a decrease in hydrophobic residues, improved solubility of scFv variants consistent with previous observations in this area [61,62].

3.5. The length and charge of the scFv variants' CDR-H3 have a positive effect on MMP-9 binding

AlphaFold2 was used for structural modeling of protein complexes of engineered scFv variants with binding improvement toward MMP-9cd.

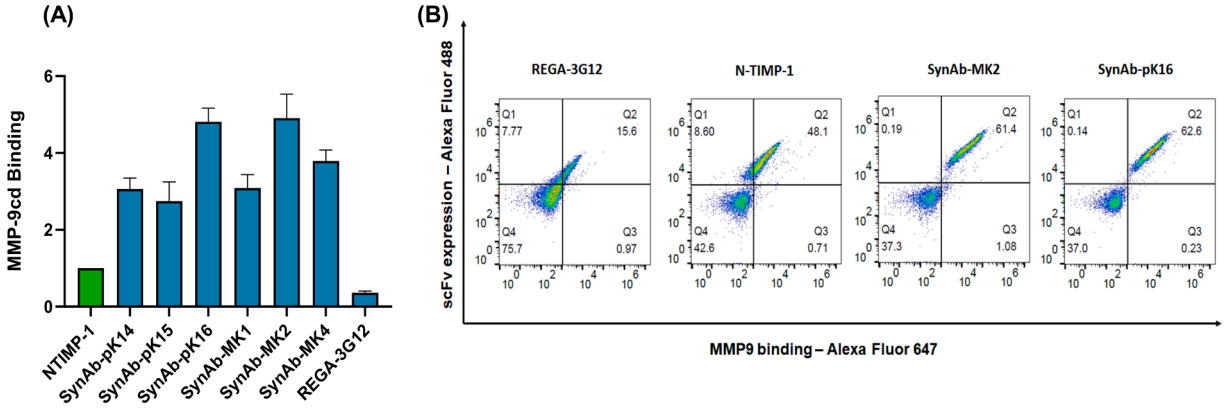


Fig. 6. scFv variants isolated after two sequential FACS screening for MMP-9cd binding. A) The bar graph shows the mean fluorescence intensity for 6xHis-MMP-9cd binding to scFv variants, adjusted for background and normalized to N-TIMP-1 which was used as a positive control for 6xHis-MMP-9cd binding. Yeast-displayed scFv variants were incubated with 300 nM soluble 6xHis-MMP-9cd protein in all experiments. Each data point represents the mean of triplicate samples, with error bars and the standard error of the mean (SEM) displayed for each data point. B) Flow cytometry scatter plots illustrate several isolated yeast-displayed scFv variants with enhanced MMP-9cd binding activity, using N-TIMP-1 as a reference. The x-axis (APC channel) represents binding to 6xHis-MMP-9cd (300 nM), while the y-axis (FITC channel) shows scFv expression levels.

Table 1
Sanger sequencing of isolated scFv antibodies. Sequences of the different CDR regions in the light and heavy chains are shown in the table. Kabat numbering is used for numbering the residues in each antibody.

CDR sequences						
Clone names	VL			VH		
	L1	L2	L3	H1	H2	H3
SynAb-pK14	RASQSVSSSYLA	GASSRAT	QQYGSSPSTF	GFTFSNAWMS	RIKSKTDGGTTDY	PSFSAQPPYYS
SynAb-pK15	RASQSVSSSYLA	GASSRAT	QQYGSSPSTF	GSGYSFTSYW	GIIYPGDSDTRY	GDALDPPMYS
SynAb-pK16	RASQSISSYLN	ASSSLQS	QQSYSTPLTF	GFTFSNAWMS	RIKSKTDGGTTDY	IQVAKDGQSKTA
SynAb-MK1	RASQSISSYLN	ASSSLQS	QQSYSTPLTF	GSGYSFTSYW	GIIYPGDSDTRY	DYSASSSGE
SynAb-MK2	RASQSISSYLN	AASSLQS	QQSYSTPLTF	GGTFSSYAIS	GGIIPIFGTANY	HPIYSGHKGSGG
SynAb-MK4	RASQSVSSSYLA	GASSRAT	QQYGSSPSTF	GFTFSNAWMS	RIKSKTDGGTTDY	DQATPMYRTYYP

AlphaFold2 Multimer uses a complex deep-learning architecture, specifically designed to process the complex interactions between amino acids [63], by integrating all interacting chains within a single prediction run, effectively simulating the entire complex [64]. This feature makes AlphaFold2 Multimer an exceptional tool for predicting complex protein-protein interactions like those between antibodies and antigens [65]. However, the limited availibity of sufficient co-evolution data restricts the accuracy of predicting antibody-antigen interactions as antibodies bind strongly to antigens due to process like somatic hypermutation and affinity maturation [66]. The hypermutation and affinity maturation processes involve the rapid evolution and selection of antibody mutants within an organism to improve binding affinity, distinct from the long-term evolutionary processes that co-evolution data typically capture. Thus, analysis of relative metrics of AlphaFold2 for the modeled scFv variants revealed high confidence in the accuracy of the model protein complex structures. For instance, all scFv/MMP-9cd structures have a predicted Local Distance Difference Test (pLDDT) score exceeding 85, indicating AlphaFold's high confidence in the positions of each atom within the predicted structures. Furthermore, the predicted template modeling (pTM) score, which assesses both the accuracy of the entire predicted protein structure and its similarity to a known reference structure, is above 0.5 for all structures, showing a high degree of accuracy (Table S1).

The active site and exosite regions of MMP-9cd exhibit a distinct negative charge and a concave geometry, though they present a flatter conformation compared to other MMPs (Fig. 7). The surface charge and geometry of the CDR-H3 loop of scFv antibody variants complement the active site of MMP-9. Most of the engineered scFV variants including top isolated mutants with upto five-fold improvement of MMP-9 binding compared to N-TIMP-1, SynAb-pK16, and SynAb-MK2, are predicted to have positively charged CDR-H3 regions that align well with the negatively charged area of the MMP-9 active site and its convex shape. This suggests that these variants have specialized adaptations for binding to MMP-9cd. This finding is consistent with the positive charge and concave geometry of the CDR-H3 loop in REGA-3G12, an MMP-9 inhibitory antibody, which allows it to effectively fit into the active site (Fig. 7).

These isolated scFvs possess 12 amino acid residues in the CDR-H3 loop. The length and amino acid composition of this region confer a convex shape to these isolated scFvs. These distinct features present a favorable opportunity for the rational design of scFv variants, particularly by focusing on the CDR-H3 length and amino acid composition of murine or human antibodies. These antibodies have shorter CDR-H3 regions compared to camelid antibodies, making them suitable for enhanced binding to the relatively flat active site of MMP-9cd. This explains the frequent appearance of charge and polar amino acid residues in strong binders are predominantly charged and polar. Additionally, the structural studies provide evidence for the longer CDR-H3

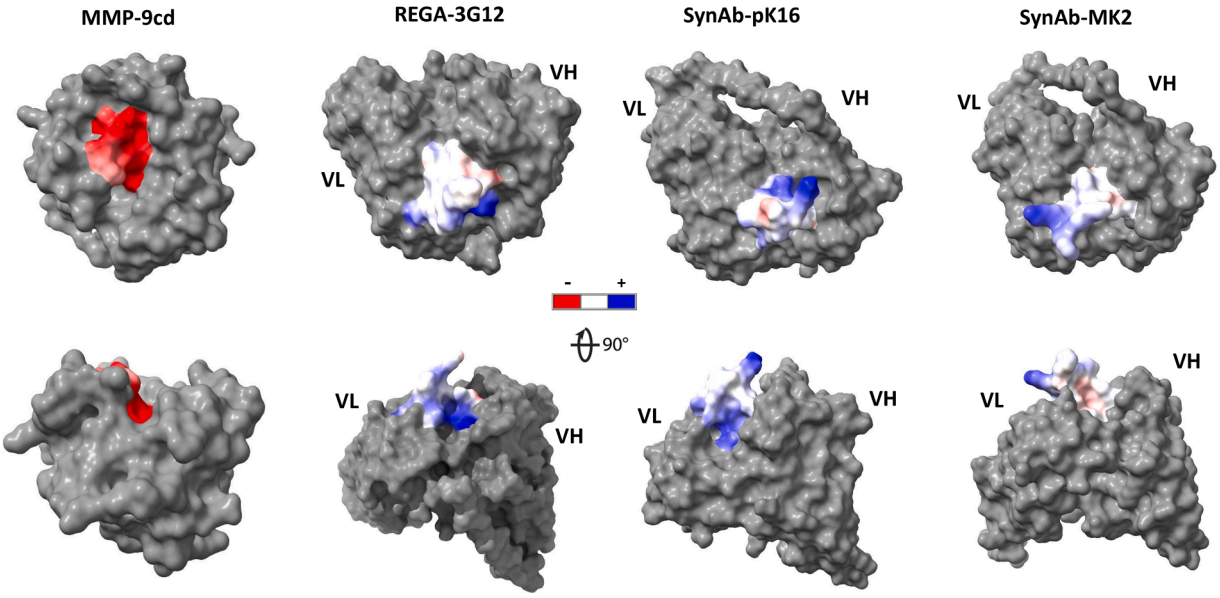


Fig. 7. Surface charge distribution of CDR-H3 on different scFvs and MMP-9cd. The surface charge is visually represented by color, with red indicating negatively charged regions, blue showing positively charged areas, and white denoting neutral regions. The 90° rotation offers a detailed view of the charge distribution, revealing the three-dimensional electrostatic landscape and geometry of each protein's CDR-H3. The active site of MMP-9cd displays a prominent negatively charged area, suggesting a strong potential for electrostatic interactions with positively charged CDR-H3 sequences. As a positive control, REGA-3G12's CDR-H3 domain shows a neutral to positive charge distribution, which is well-suited for electrostatic interactions with MMP-9cd.

regions in high-affinity binders compared to non-binders or weak binders, which often feature shorter CDR-H3 loops with a higher proportion of non-polar residues. This pattern indicates that the charge distribution in the variable regions of antibodies, particularly within the CDR regions, plays a crucial role in determining their antigen-binding capability as well as their overall stability and folding [67].

3.6. The scFv variants with improved MMP-9 binding show close contacts between CDR-H3 with MMP-9cd active site

Studying the protein complex structure of scFv variants with MMP-9 revealed important contacts between CDR-H3 and the active site of MMP-9cd or neighboring loops (exosites). The interactions of the scFv CDR-H3 region with MMP-9cd determines residues and their specific position responsible for improving MMP-9cd binding, as well as potentially determining the binding location in proximity to the enzyme active site. Protein complexes predicted by AlphaFold multimer for REGA-3G12, a known positive binder, and SynAb-pK16 and SynAb-MK2, the two highest binders to MMP-9, reveal interesting interactions between the scFvs and MMP-9cd (Fig. 8). Consistent with the global Shapley values, the first six positions of CDR-H3 appeared to contribute significantly to MMP-9cd binding, with position 6 having the closest contacts. Interestingly, for all of these scFvs, position 6 is found to be important in binding to the active site residues or neighboring residues of MMP-9cd, specifically involving Gly104, Asp233, and Gly231 in REGA-3G12, SynAb-pK16, and SynAb-MK2, respectively. This highlights the significance of specific amino acids at position 6 in driving effective binding interactions with MMP-9cd.

The SynAb-MK2 variant makes several hydrogen bond (H-bonds) with MMP-9cd residues close to the active site. The scFv residue Gly231 at position 6 of CDR-H3 makes an H-bond with residues Ile92 and Pro90 located at the exosite of MMP9-cd (Fig. 8**. B**) similar to Gly104, an amino acid at position 6 in the REGA-3G12 variant (Fig. 8**. A**), emphasizing the importance of Gly residue, one of the most frequent residues appeared in deep sequencing and position 6 of CDR-H3. The SynAb-pK16 variant makes several critical interactions with MMP-9cd, particularly at position 6 of its CDR-H3, which is Asp233 (Fig. 8**. C**). This residue plays a key role in binding to the exosite of MMP-9cd. Additionally, SynAb-pK16 makes further contacts through positions 4 and 5 of its CDR-H3, which are Ala231 and Lys232, respectively.

Regardless of position 6, as suggested by global Shapley values, positions 1, 4, and 5 also play significant roles in interacting with MMP-9cd in these scFv variants. This highlights the importance of these positions in directing the binding towards MMP-9cd. In summary, these findings emphasize the role of CDR-H3, particularly the first six positions, in enhancing binding between scFvs and MMP-9cd. Further, it highlights a potential hot spot residue at position 6 for interaction with MMP-9cd and underlines the importance of other residues (positions 1, 4, and 5) for proper binding orientation.

4. Discussion

A synthetic antibody library previously engineered for improved nonspecific binding [25]was used to screen, isolate, and analyze the scFv antibody variants with improved binding to MMP-9. The scFv antibody variants showed improved binding to MMP-9 compared to endogenous MMP-9 inhibitor, N-TIMP-1, and other MMP-9 inhibitory antibodies such as REGA-3G12 [57,58]. The role of both length and amino acid composition in the CDR-H3 loop in binding to MMP-9 was investigated using experimental and computational analysis, considering the

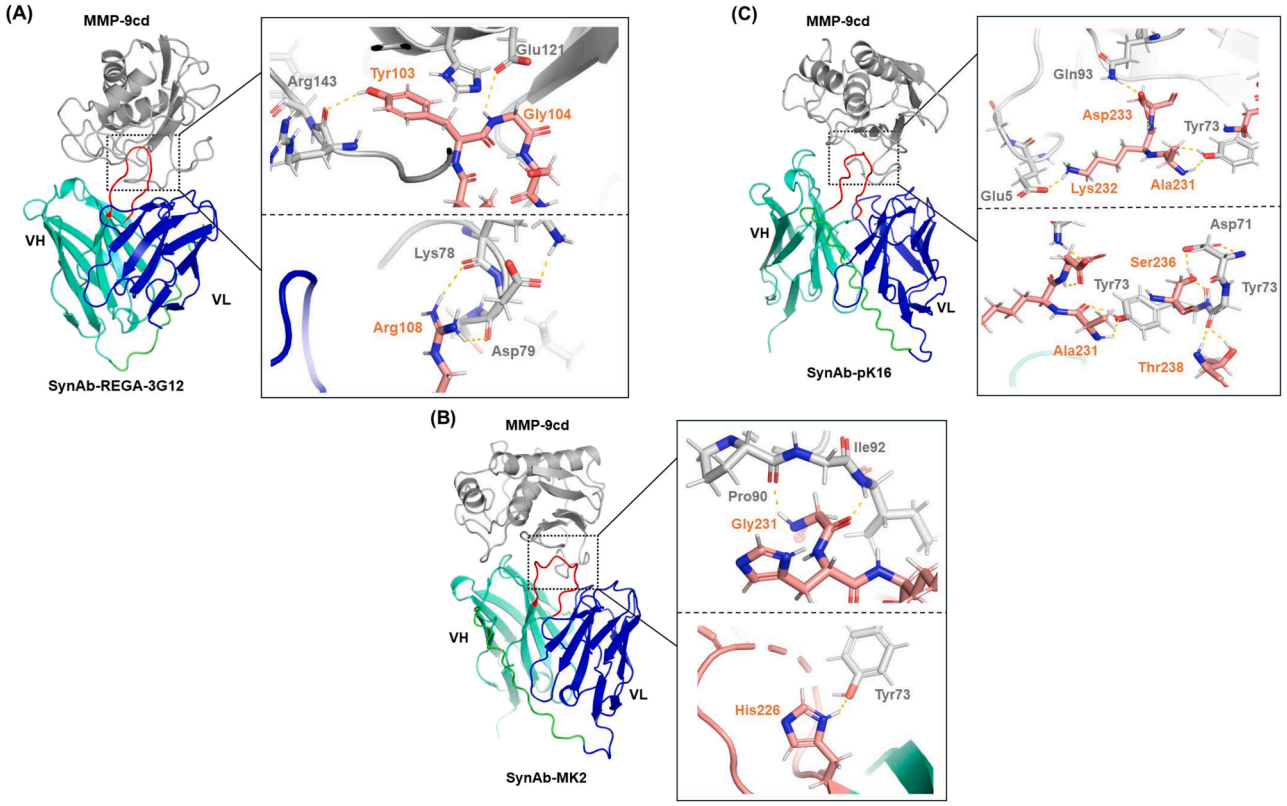


Fig. 8. Binding interactions between CDR-H3 variants and MMP-9cd. The structure of MMP-9cd (in grey) complexed with the scFv antibody variants is depicted, with the light chain (VL) shown in dark blue, the heavy chain (VH) in cyan, and CDR-H3 in dark pink. (A) The REGA-3G12 variant illustrates an interaction at position 6 of CDR-H3, where Gly104 forms a hydrogen bond with residues of MMP-9cd. (B) The SynAb-MK2 variant features Gly231 at position 6 of CDR-H3 forming hydrogen bonds with Ile92 and Pro90, located at the exosite of MMP-9cd. (C) The SynAb-pK16 variant with residue Asp233 at position 6 of CDR-H3 highlighted, which plays a crucial role in binding to the exosite of MMP-9cd, with additional interactions involving positions 4 (Ala231) and 5 (Lys232).

significance of CDR-H3 in antibody-antigen interactions. The results of next-generation DNA sequencing analysis revealed patterns in the CDR-H3 of scFvs binders with polar residues such as Ser, and Tyr being dominant in frequency. The most favorable length of CDR-H3 for MMP-9 binding was found to be 11 amino acids which was consistent with computational modeling and structural studies in binding to the MMP-9 catalytic site. Further, the importance of CDR-H3 charge was highlighted as the selected scFv variants with improved MMP-9 binding showed positive surface charges which match the negative charge in the MMP-9cd active site.

A machine learning strategy based on recent advances in developing protein language models to predict protein structure and function was used to generate a model that predicts scFv variants binding to MMP-9 based on CDR-H3 sequences as the key driver of binding with high accuracy. The fine-tuned MLM model can be used to highlight significant residues for binding to MMP-9 and similar targets. The developed models and knowledge could be translated to other protein binders.

The MMP activity could be blocked by targeting the neighboring regions outside of the catalytic domain, known as exosites [68]. Unlike the catalytic domains, which are similar across MMPs, different MMPs have distinct exosites. Targeting exosites using synthetic antibodies was previously used to provide selective inhibition of specific MMPs which led to the development of mAbs targeting MMP-9, such as AB0041 and AB0046, as well as their humanized version, GS-5745 [69]. Using the developed language model tools to optimize key regions of scFv antibodies focusing on CDR-H3 regions could facilitate overcoming the limitations in targeting specific MMPs.

CRediT authorship contribution statement

Elham Khorasani Buxton: Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis. **Maryam Raeeszadeh Sarmazdeh:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Iftichar Kalanther:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation. **Sachin Kumar:** Visualization. **Masoud Kalantar:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation.

Declaration of Competing Interest

The authors have no conflict of interest.

Acknowledgment

We would like to thank Prof. Dane Wittrup (Chemical Engineering-MIT) for the generous gift of the nonspecific synthetic antibody library. M. R.-S. has funding support from NIH R03AG070511 and NIH R21HD109743. Figs. 1 and 2A were created with BioRender.com.

Author contributions

M.R. -S., E. K.-B; conception, M.K.; performed the experiments, I.K.: performed machine learning models, S. K., M.K.: performed computational analysis, M.K., I.K., E. K. -B, M.R.-S., data analysis, writing, and editing the manuscript drafts. M.R.-S. funding and supervision.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.10.005.

Data Availability

The authors have shared a public GitHub repository: (https://github.com/Iftikhar2/Engineered_CDRH3_MMP9_Binding?tab=readme-ov-file) containing the data used for training, as well as the trained LSTM models. A sample code is provided, allowing users to input either a single CDRH3 sequence or multiple CDR-H3s (via a CSV file) to predict the CDR-H3-MMP9 binding affinity. Additionally, the repository contains experimental datasets and associated code for engineered CDR-H3 binding to MMP-9, including structural predictions, protein binding models, and relevant documentation. Further data supporting the findings of this study are available from the corresponding author upon reasonable request.

References

[1] Raeeszadeh-Sarmazdeh M, Do LD, Hritz BG. Metalloproteinases and their inhibitors: potential for the development of new therapeutics. Cells 2020;9.
[2] Radisky ES. Extracellular proteolysis in cancer: proteases, substrates, and mechanisms in tumor progression and metastasis. J Biol Chem 2024;300(6):107347. https://doi.org/10.1016/J.JBC.2024.107347.
[3] Radisky ES, Raeeszadeh-Sarmazdeh M, Radisky DC. Therapeutic potential of matrix metalloproteinase inhibition in breast cancer. J Cell Biochem 2017;118:3531–48.
[4] Kalantar M, Hilpert GA, Mosca ER, Raeeszadeh-Sarmazdeh M. Engineering metalloproteinase inhibitors: tissue inhibitors of metalloproteinases or antibodies, that is the question. Curr Opin Biotechnol 2024;86:103094. https://doi.org/10.1016/J.COPBIO.2024.103094.
[5] Razai AS, Eckelman BP, Salvesen GS. Selective inhibition of matrix metalloproteinase 10 (MMP10) with a single-domain antibody. J Biol Chem 2020;295:2464–72.
[6] Appleby TC, Greenstein AE, Hung M, Liclican A, Velasquez M, Villasenor AG, et al. Biochemical characterization and structure determination of a potent, selective antibody inhibitor of human MMP9. J Biol Chem 2017;292(16):6810–20. https://doi.org/10.1074/jbc.M116.760579.
[7] Chen K-HE, Chen C, Lopez T, Radecki KC, Bustamante K, Lorenson MY, et al. Use of a novel camelid-inspired human antibody demonstrates the importance of MMP-14 to cancer stem cell function in the metastatic process. Oncotarget 2018;9(50):29431–44. https://doi.org/10.18632/ONCOTARGET.25654.
[8] Nam DH, Ge X. Generation of highly selective MMP antibody inhibitors. Methods Mol Biol 2018;1731:307–24.
[9] Kinder M, Greenplate AR, Grugan KD, Soring KL, Heeringa KA, McCarthy SG, et al. Engineered protease-resistant antibodies with selectable cell-killing functions. J Biol Chem 2013;288(43):30843–54. https://doi.org/10.1074/jbc.M113.486142.
[10] Lopez T, Mustafa Z, Chen C, Lee KB, Ramirez A, Benitez C, et al. Functional selection of protease inhibitory antibodies. Proc Natl Acad Sci USA 2019;116(33):16314–9. https://doi.org/10.1073/PNAS.1903330116/SUPPL_FILE/PNAS.1903330116.SAPP.PDF.
[11] Sargunas PR, Spangler JB. Joined at the hip: the role of light chain complementarity determining region 2 in antibody self-association. Proc Natl Acad Sci USA 2022;119(28):e2208330119. https://doi.org/10.1073/PNAS.2208330119/ASSET/2C6D0849-4F8F-4B72-8626-97F8BBF84BFD/ASSETS/IMAGES/LARGE/PNAS.2208330119FIG01.JPG.
[12] Boder ET, Raeeszadeh-Sarmazdeh M, Price JV. Engineering antibodies by yeast display. Arch Biochem Biophys 2012;526:99–106.
[13] Raeeszadeh-Sarmazdeh Maryam, Boder ET. In: Traxlmayr MW, editor. Yeast Surface Display: New Opportunities for a Time-Tested Protein Engineering System. *Yeast Surface Display*. New York, NY: Springer US; 2022. p. 3–25. https://doi.org/10.1007/978-1-0716-2285-8_1.
[14] Sargunas PR, Spangler JB. Full speed AHEAD in antibody discovery. *17:10* Nat Chem Biol 2021 2021;17(10):1011–2. https://doi.org/10.1038/s41589-021-00838-y.
[15] Richards DA. Exploring alternative antibody scaffolds: antibody fragments and antibody mimics for targeted drug delivery. Drug Discov Today Technol 2018;30:35–46.
[16] Lu RM, Hwang YC, Liu IJ, Lee CC, Tsai HZ, Li HJ, et al. Development of therapeutic antibodies for the treatment of diseases. J Biomed Sci 2020 27:1 2020;27(1):1–30. https://doi.org/10.1186/S12929-019-0592-Z.
[17] Peng WJ, Zhang JQ, Wang BX, Pan HF, Lu MM, Wang J. Prognostic value of matrix metalloproteinase 9 expression in patients with non-small cell lung cancer. Clin Chim Acta 2012;413(13–14):1121–6. https://doi.org/10.1016/J.CCA.2012.03.012.
[18] Xu Y, Li Z, Jiang P, Wu G, Chen K, Zhang X, et al. The Co-expression of mmp-9 and tenascin-C is significantly associated with the progression and prognosis of pancreatic cancer. Diagn Pathol 2015;10(1):1–8. https://doi.org/10.1186/S13000-015-0445-3/TABLES/5.
[19] Vadillo-Ortega F, Estrada-Gutiérrez G. Role of matrix metalloproteinases in preterm labour. BJOG 2005;112(SUPPL. 1):19–22. https://doi.org/10.1111/J.1471-0528.2005.00579.X.
[20] Liu H, Wang J, Wang H, Tang N, Li Y, Zhang Y, et al. Correlation between matrix metalloproteinase-9 and endometriosis. Int J Clin Exp Pathol 2015;8(10):13399.

[21] Weigel MT, Krämer J, Schem C, Wenners A, Alkatout I, Jonat W, et al. Differential expression of MMP-2, MMP-9 and PCNA in endometriosis and endometrial carcinoma. Eur J Obstet Gynecol Reprod Biol 2012;160(1):74–8. https://doi.org/10.1016/j.ejogrb.2011.09.040.

[22] Holliger P, Hudson PJ. Engineered antibody fragments and the rise of single domains. *2005 23:9* Nat Biotechnol 2005;23(9):1126–36. https://doi.org/10.1038/nbt1142.

[23] Paemen L, Martens E, Masure S, Opdenakker G. Monoclonal antibodies specific for natural human neutrophil gelatinase B used for affinity purification, quantitation by Two-Site ELISA and inhibition of enzymatic activity. Eur J Biochem 1995;234(3):759–65. https://doi.org/10.1111/J.1432-1033.1995.759_A.X.

[24] Appleby TC, Greenstein AE, Hung M, Liclican A, Velasquez M, Villasenor AG, et al. Biochemical characterization and structure determination of a potent, selective antibody inhibitor of human MMP9. J Biol Chem 2017;292(16):6810–20. https://doi.org/10.1074/JBC.M116.760579.

[25] Kelly RL, Le D, Zhao J, Wittrup KD. Reduction of nonspecificity motifs in synthetic antibody libraries. J Mol Biol 2018;430(1):119–30. https://doi.org/10.1016/J.JMB.2017.11.008.

[26] Weitzner BD, Dunbrack RL, Gray JJ. The origin of CDR H3 structural diversity. Structure 2015;23(2):302–11. https://doi.org/10.1016/J.STR.2014.11.010.

[27] Nam DH, Rodriguez C, Remacle AG, Strongin AY, Ge X. Active-site MMP-selective antibody inhibitors discovered from convex paratope synthetic libraries. Proc Natl Acad Sci USA 2016;113(52):14970–5. https://doi.org/10.1073/PNAS.1609375114/SUPPL_FILE/PNAS.201609375SI.PDF.

[28] Sela-Passwell N, Kikkeri R, Dym O, Rozenberg H, Margalit R, Arad-Yellin R, et al. Antibodies targeting the catalytic zinc complex of activated matrix metalloproteinases show therapeutic potential. *2011 18:1* Nat Med 2011;18(1):143–7. https://doi.org/10.1038/nm.2582.

[29] Fischer T, Riedl R. Inhibitory antibodies designed for matrix metalloproteinase modulation. *2019, Vol. 24, Page 2265* Molecules 2019;24(12):2265. https://doi.org/10.3390/MOLECULES24122265.

[30] Marshall DC, Lyman SK, McCauley S, Kovalenko M, Spangler R, Liu C, et al. Selective allosteric inhibition of MMP9 is efficacious in preclinical models of ulcerative colitis and colorectal cancer. PLoS One 2015;10(5):e0127063. https://doi.org/10.1371/JOURNAL.PONE.0127063.

[31] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model *(1979)* Science 2023;379(6637):1123–30.

[32] Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, et al. Large language models generate functional protein sequences across diverse families. *2023 41:8* Nat Biotechnol 2023;41(8):1099–106. https://doi.org/10.1038/s41587-022-01618-2.

[33] Valentini G, Malchiodi D, Gliozzo J, Mesiti M, Soto-Gomez M, Cabri A, et al. The promises of large language models for protein design and modeling. Front Bioinforma 2023;3:1304099. https://doi.org/10.3389/FBINF.2023.1304099/BIBTEX.

[34] Mardikoraem M, Woldring D. Protein fitness prediction is impacted by the interplay of language models, ensemble learning, and sampling methods. Pharmaceutics 2023;15(5):1337. https://doi.org/10.3390/PHARMACEUTICS15051337/S1.

[35] Chen J, Woldring DR, Huang F, Huang X, Wei GW. Topological deep learning based deep mutational scanning. Comput Biol Med 2023;164:107258. https://doi.org/10.1016/J.COMPBIOMED.2023.107258.

[36] Bepler T, Berger B. Learning the protein language: evolution, structure, and function. Cell Syst 2021;12(6):654–669.e3. https://doi.org/10.1016/J.CELS.2021.05.017.

[37] Toumaian MR, Raeeszadeh-Sarmazdeh M. Engineering tissue inhibitors of metalloproteinases using yeast surface display. Methods Mol Biol 2022;2491:361–85. https://doi.org/10.1007/978-1-0716-2285-8_19.

[38] Bonadio A, Oguche S, Lavy T, Kleifeld O, Shifman J. Computational design of matrix metalloproteinase-9 (MMP-9) resistant to auto-cleavage. Biochem J 2023;480(14):1097–107. https://doi.org/10.1042/BCJ20230139.

[39] Bolt AJ, Do LD, Raeeszadeh-Sarmazdeh M. Bacterial expression and purification of human matrix metalloproteinase-3 using affinity chromatography. J Vis Exp 2022;2022(181). https://doi.org/10.3791/63263.

[40] Ahmadighadykolaei H, Lambert JA, Raeeszadeh-Sarmazdeh M. TIMP-1 protects tight junctions of brain endothelial cells from MMP-mediated degradation. Pharm Res 2023;40(9):2121–31. https://doi.org/10.1007/S11095-023-03593-Y/FIGURES/7.

[41] Hosseini A, Kumar S, Hedin K, Raeeszadeh-Sarmazdeh M. Engineering minimal tissue inhibitors of metalloproteinase targeting MMPs via gene shuffling and yeast surface display. Protein Sci 2023;32(12):e4795. https://doi.org/10.1002/PRO.4795.

[42] Ruffolo, J.A.; Gray, J.J.; Sulam, J. Deciphering Antibody Affinity Maturation with Language Models and Weakly Supervised Learning. 2021.

[43] Maaten L, van der; Hinton G. Visualizing data using T-SNE. J Mach Learn Res 2008;9(86):2579–605.

[44] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80. https://doi.org/10.1162/NECO.1997.9.8.1735.

[45] Elabd H, Bromberg Y, Hoarfrost A, Lenz T, Franke A, Wendorff M. Amino acid encoding for deep learning applications. BMC Bioinforma 2020;21(1):1–14. https://doi.org/10.1186/S12859-020-03546-X/FIGURES/4.

[46] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017:4766–75. *2017-December*.

[47] H.T.T. Nguyen; H. Cao, K.V.T.N.; N. D.K.Pham Evaluation of Explainable Artificial Intelligence: SHAP, LIME, and CAM, 2021.

[48] Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *2022 19:6* Nat Methods 2022;19(6):679–82. https://doi.org/10.1038/s41592-022-01488-1.

[49] Liu C, Denzler LM, Hood OEC, Martin ACR. Do antibody CDR loops change conformation upon binding? MAbs 2024;16(1). https://doi.org/10.1080/19420862.2024.2322533.

[50] Padlan EA. Anatomy of the antibody molecule. Mol Immunol 1994;31(3):169–217. https://doi.org/10.1016/0161-5890(94)90001-9.

[51] Xu JL, Davis MM. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. Immunity 2000;13(1):37–45. https://doi.org/10.1016/S1074-7613(00)00006-6.

[52] Zemlin M, Klinger M, Link J, Zemlin C, Bauer K, Engler JA, et al. Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. J Mol Biol 2003;334(4):733–49. https://doi.org/10.1016/J.JMB.2003.10.007.

[53] Shusta EV, Kieke MC, Parke E, Kranz DM, Wittrup KD. Yeast polypeptide fusion surface display levels predict thermal stability and soluble secretion efficiency. J Mol Biol 1999;292(5):949–56. https://doi.org/10.1006/JMBI.1999.3130.

[54] Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280(1):1–9. https://doi.org/10.1006/JMBI.1998.1843.

[55] Fellouse FA, Barthelemy PA, Kelley RF, Sidhu SS. Tyrosine plays a dominant functional role in the paratope of a synthetic antibody derived from a four amino acid code. J Mol Biol 2006;357(1):100–14. https://doi.org/10.1016/J.JMB.2005.11.092.

[56] Birtalan S, Zhang Y, Fellouse FA, Shao L, Schaefer G, Sidhu SS. The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. J Mol Biol 2008;377(5):1518–28. https://doi.org/10.1016/J.JMB.2008.01.093.

[57] Martens E, Leyssen A, Van Aelst I, Fiten P, Piccard H, Hu J, et al. A monoclonal antibody inhibits gelatinase B/MMP-9 by selective binding to part of the catalytic domain and not to the fibronectin or zinc binding domains. Biochim Et Biophys Acta (BBA) - Gen Subj 2007;1770(2):178–86. https://doi.org/10.1016/J.BBAGEN.2006.10.012.

[58] Love EA, Sattikar A, Cook H, Gillen K, Large JM, Patel S, et al. Developing an antibody–drug conjugate approach to selective inhibition of an extracellular protein. ChemBioChem 2019;20(6):754–8. https://doi.org/10.1002/CBIC.201800623.

[59] Rezhdo A, Lessard CT, Islam M, Van Deventer JA. Strategies for enriching and characterizing proteins with inhibitory properties on the yeast surface. Protein Eng, Des Sel 2023;36:1–14. https://doi.org/10.1093/PROTEIN/GZAC017.

[60] Raeeszadeh-Sarmazdeh M, Greene KA, Sankaran B, Downey GP, Radisky DC, Radisky ES. Directed evolution of the metalloproteinase inhibitor TIMP-1 reveals That Its N- and C-terminal domains cooperate in matrix metalloproteinase recognition. J Biol Chem 2019;294(24):9476–88. https://doi.org/10.1074/JBC.RA119.008321.

[61] Lawrence MS, Phillips KJ, Liu DR. Supercharging proteins can impart unusual resilience. J Am Chem Soc 2007;129(33):10110–2. https://doi.org/10.1021/JA071641Y/SUPPL_FILE/JA071641YSI20070628_045815.PDF.

[62] Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM. Rationalization of the effects of mutations on peptide andprotein aggregation rates. *2003 424:6950* Nature 2003;424(6950):805–8. https://doi.org/10.1038/nature01891.

[63] Zhu W, Shenoy A, Kundrotas P, Elofsson A. Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. Bioinformatics 2023;39(7). https://doi.org/10.1093/BIOINFORMATICS/BTAD424.

[64] Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2. *2022 13:1* Nat Commun 2022;13(1):1–11. https://doi.org/10.1038/s41467-022-28865-w.

[65] Harmalkar A, Lyskov S, Gray JJ. Reliable protein-protein docking with AlphaFold, rosetta, and replica-exchange. Elife 2024;13. https://doi.org/10.7554/ELIFE.94029.1.

[66] Mishra AK, Mariuzza RA. Insights into the structural basis of antibody affinity maturation from next-generation sequencing. Front Immunol 2018;9(FEB):330276. https://doi.org/10.3389/FIMMU.2018.00117/BIBTEX.

[67] Makowski EK, Chen H, Lambert M, Bennett EM, Eschmann NS, Zhang Y, et al. Reduction of therapeutic antibody self-association using yeast-display selections and machine learning. MAbs 2022;14(1). https://doi.org/10.1080/19420862.2022.2146629.

[68] Levin M, Udi Y, Solomonov I, Sagi I. Next generation matrix metalloproteinase inhibitors — novel strategies bring new prospects. Biochim Et Biophys Acta (BBA) - Mol Cell Res 2017;1864(11):1927–39. https://doi.org/10.1016/J.BBAMCR.2017.06.009.

[69] Marshall DC, Lyman SK, McCauley S, Kovalenko M, Spangler R, Liu C, et al. Selective allosteric inhibition of MMP9 is efficacious in preclinical models of ulcerative colitis and colorectal cancer. PLoS One 2015;10(5):e0127063. https://doi.org/10.1371/JOURNAL.PONE.0127063.