Prediction of single-mutation effects for fluorescent immunosensor engineering with an end-to-end trained protein language model

Akihito Inoue,[a#] Bo Zhu,[b#] Keisuke Mizutani,[c] Ken Kobayashi,[c] Takanobu Yasuda,[b] Alon Wellner,[d] Chang C. Liu[d] and Tetsuya Kitaguchi[b]*

[a] Graduate School of Life Science and Technology, Institute of Science Tokyo, 4259 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa 226-8501, Japan

[b] Laboratory for Chemistry and Life Science, Institute of Integrated Research, Institute of Science Tokyo, 4259 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa 226-8501, Japan

[c] School of Engineering, Institute of Science Tokyo, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

[d] Department of Biomedical Engineering, University of California, Irvine, CA, 92697, USA

[#] Equal contributions

*** Corresponding Author**

Tetsuya Kitaguchi, Ph. D.
Laboratory for Chemistry and Life Science, Institute of Integrated Research, Institute of Science Tokyo, 4259 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa 226-8501, Japan
E-mail: kitaguc.t.aa@m.titech.ac.jp

Keywords: immunosensor, quenchbody, nanobody, deep learning, protein language model

Abstract

Quenchbody (Q-body) is a fluorophore-labeled homogeneous immunosensor, in which the fluorophore is quenched by tryptophan (Trp) residues in the vicinity of the antigen-binding paratope and de-quenched in response to antigen binding. The development of Q-bodies against targets on demand remains challenging due to the large sequence space of the complementarity-determining regions (CDRs) related to antigen binding and fluorophore quenching. In this study, we pioneered a strategy using high-throughput screening and a protein language model (pLM) to predict the effects of mutations on fluorophore quenching with single amino acid resolution, thereby enhancing the performance of Q-bodies. We collected yeasts displaying nanobodies with high and low quenching properties for TAMRA fluorophore from a modified large synthetic nanobody library, followed by next-generation sequencing. The pre-trained pLM, connected with a single-layer perceptron, was trained end-to-end on the enriched CDR sequences. The achieved quenching prediction model focused on CDR1+3 performed best in evaluation with precision-recall curves. Using this model, we predicted and validated effective mutations in two anti-SARS-CoV-2 nanobodies, RBD1i13 and RBD10i14, that convert them into Q-bodies. For RBD1i13, three Trp mutants were predicted with high probability scores for quenching through in silico Trp scanning. These mutants were verified via yeast surface display to all show enhanced quenching. For RBD10i14, mutations at four positions close to an existing Trp gave high scores through in silico saturation mutagenesis scanning. Six out of eight high-score mutants, derived from two mutants at each of four positions, exhibited deeper quenching on yeast surface. Next, combined with the investigation of antigen binding of the mutants, we successfully achieved Q-bodies with enhanced responses. Overall, our strategy allows for the prediction of fluorescence responses based solely on the sequence of the antibody and will be essential for the rational selection and design of antibodies to achieve immunosensors with larger responses.

Introduction

Homogenous immunoassays using an antibody labeled with environment-sensitive dyes[1-3] or rhodamine-related fluorophores[4] have the advantage of simple experimental operation, but the engineering of the antibody is often required to make them function as an immunosensor. Quenchbody (Q-body) is a quench-based fluorescent homogeneous immunosensor, in which a fluorophore is labeled near the antigen binding site of the antibody from several animal species, including camelid-derived nanobodies, using an N-terminal tagged cysteine residue,[5-7] fluorophore-conjugated non-natural amino acid incorperation,[4] or coiled-coil forming E4/K4 peptides.[8] Q-body functions are based on antigen-dependent fluorescence recovery from quenching by tryptophan (Trp) residues of an antibody through photo-induced electron transfer, and deep quenching of the fluorophore in the absence of antigen is essential to achieve a large fluorescence response upon binding. Since some Trp residues are often conserved among antibodies,[9] such antibodies were converted into Q-bodies without extensive antibody engineering for the detection of various molecules ranging from proteins to small molecules.[10-11] To demonstrate its usability in various situations, the Q-body has been recently applied in unique applications such as droplet-based screening and monitoring[12-13] or intracellular imaging[14] by utilizing the characteristics of homogenous immunoassays. However, in many cases, because of the different position and number of native Trp residues, quenching of the fluorophore is negligible and small, resulting in little antigen-dependent fluorescence recovery. Therefore, the unpredictability of the fluorescence response from the amino acid sequence of each antibody causes difficulty in developing a Q-body against new targets on demand.

So far, we have attempted two approaches to obtain large fluorescence responses for Q-bodies. The first approach is to adapt antibody screening methods for finding antibodies suitable for Q-body responses.[15] When fluorophore-labeled antibodies were fabricated on the yeast cell surface by N-terminal labeling using coiled-coil forming peptide, antibodies suitable for Q-body conversion were successfully selected from a nanobody library against the target antigen based on quenching and de-quenching evaluations. The second approach is to increase the number of Trp residues in complementarity-determining regions (CDRs) of antibodies to deepen the quenching, which in some cases enhances fluorescence responses. Introduction of Trp mutations in the CDR[16-17] or selection from a Trp-rich antibody library whose CDRs were designed to increase the propensity of Trp residue was effective in fabricating a practical Q-body.[18] However, introduction of many additional Trp residues and presence of the large number of Trp residues in CDRs often result in reduced antigen-binding activity and unnecessary quenching, respectively, leading to small fluorescence recovery, necessitating the development of a new approach for efficient fabrication of Q-bodies.

In silico design and function prediction are useful in accelerating the protein design process and providing new engineering opportunities. For artificially or semi-artificially designed proteins, such as Q-bodies, deep neural network models may capture the relationship between the amino acid sequences and desired functions without the input of working mechanisms, which is ideal for reducing bias from limited knowledge and broadening the design space explored. Recently, after the successful invention of self-attention mechanism-based neural networks for natural language processing,[19] leveraging transformer-based or transformer-inspired protein language models (pLM) became a promising approach for the tasks of protein function predictions[20] beyond

structure prediction.[21-22] In this approach, sequence annotation, amino acid feature extraction, and multi-sequence alignment or clustering steps of the protein sequence are eliminated, which may be suitable for building an in silico prediction model for Q-body development and engineering. Through the combination of pre-trained pLMs and the output layer of a simple neural network, sequence-to-function predictors have been built after learning function-labeled amino acid sequence data.[20, 23-24]

In this study, we built an original model (NanoQ-model 1.0) based on a pLM to predict the effects that mutations in CDRs have on fluorophore quenching by end-to-end learning of the relationship between quenching and amino acid sequences of antibody CDRs. In silico screening of single mutations in an antibody using this original model allows for the efficient introduction of mutations that deepen quenching. Furthermore, detailed requirements for quenching, such as the position of the Trp residue and the preferred amino acids around the Trp residue, were also revealed. We believe that the screening method, quenching prediction model, and model training strategy using NGS data from high-throughput screening (HTS) presented in this work will facilitate the development of not only immunosensors but also other unique proteins with designated functions beyond those found in nature.

Results

Collection of yeasts displaying nanobodies with high or low quenching for fluorophore

To build an original model for predicting quenching, yeasts displaying nanobodies with low quenching or high quenching properties were collected from an E4-tagged nanobody library on yeasts labeled with the FITC-K4-TAMRA peptide (Figure 1A). We first enriched yeasts displaying

nanobodies from the synthetic library (NbLib)[25] by using magnetic beads immobilized with antibodies against an affinity tag fused to the nanobody (Figure S1A), resulting in an increase in nanobody-displaying yeasts from 12.2 % to 65.9 % (Figure S1B). Then, the E4-tagged library was constructed by fusing the E4-tag to the N-terminal of the above enriched nanobodies (Figure S2A). When evaluating the quenching efficiency of the E4-tagged library on yeasts labeled with FITC-K4-TAMRA as previously described,[15] the mean of TAMRA/FITC ratio was 0.76 (calculated from Figure S2B left), which is lower than that of E4 only (≃1.24) (from Figure S2B right), suggesting

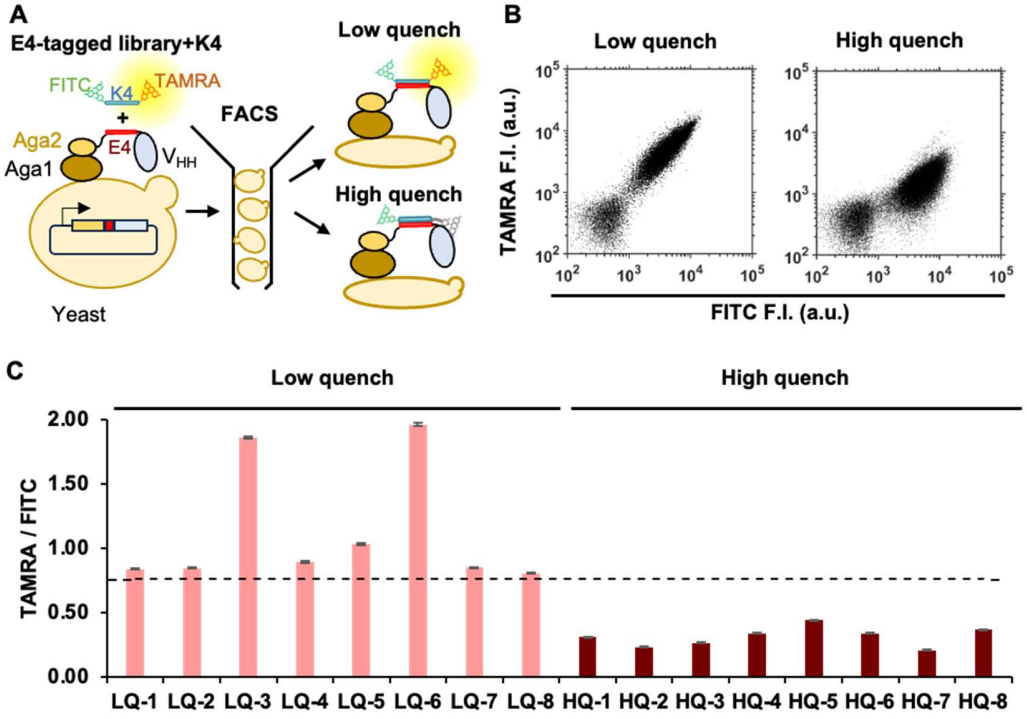

Figure 1. High-throughput screening of the nanobodies with low and high quenching properties against TAMRA. (A) Schematic image of sorting yeasts displaying TAMRA and FITC-labeled nanobodies by using E4/K4 assembly. TAMRA is used to evaluate quench and FITC is to correct for nanobody display. Aga, a-agglutinin. (B) Flow cytometric analysis of high-quenching and low-quenching nanobodies sorted by TAMRA/FITC ratio. (C) TAMRA/FITC ratio of eight nanobodies that randomly picked from high-quenching and low-quenching nanobodies. The bar graphs represent the means of TAMRA/FITC ratio ± standard error of mean. The dot line represents the mean of the TAMRA/FITC ratio of the E4-tagged library displayed on the yeast cell surface before sorting.

that the TAMRA is quenched by intrinsic amino acid residues such as Trp of the nanobodies in this library. After two rounds of selection based on the evaluation of quenching efficiency (Figure S2C, D), the mean of TAMRA/FITC ratio of the low-quenching nanobodies became 1.17 (from Figure 1B left), and that of the high-quenching nanobodies was 0.30 (from Figure 1B right), which were similar with that of E4 only ($\simeq$1.24) and a previously developed nano Q-body[15] ($\simeq$0.30), respectively. To evaluate the quenching efficiency of nanobodies on collected yeasts, we examined the TAMRA/FITC ratio of eight nanobodies on yeasts, which were randomly picked up from enriched high or low-quenching pools. The TAMRA/FITC ratios of all nanobodies from the low-quenching pool were over 0.76, and those from the high-quenching pool were below 0.76, with variations, suggesting that the nanobodies with low and high quenching were successfully sorted based on TAMRA/FITC ratio.

Building prediction models through end-to-end training

The training dataset was obtained by the next-generation sequencing (NGS) analysis of nanobodies with high and low quenching properties (Figure 2A left). Comparing the frequency of amino acids in the CDR, nanobodies with high and low quenching exhibited the different frequency against the initial E4-tagged library (Figure 2B, Figure S3). Especially in high-quenching nanobodies, Trp residues were enriched on the C-terminal of CDR1 and both ends of CDR3 (Figure 2C), and the same aromatic amino acids such as Phe and Tyr tended to be enriched at similar positions in CDR3. Interestingly, negatively charged amino acids such as Asp and Glu were less likely to be enriched in these positions. Next, to build a prediction model, a pre-trained pLM called ProtBert-BFD[20] was used (Figure 2A right). To build a binary classification model,

the pLM connected to a single-layer perceptron for probability score prediction was trained end-to-end with the amino acid sequence of high and low-quenching nanobodies. To compare the contribution to quenching efficiency among CDRs, three prediction models were trained with the

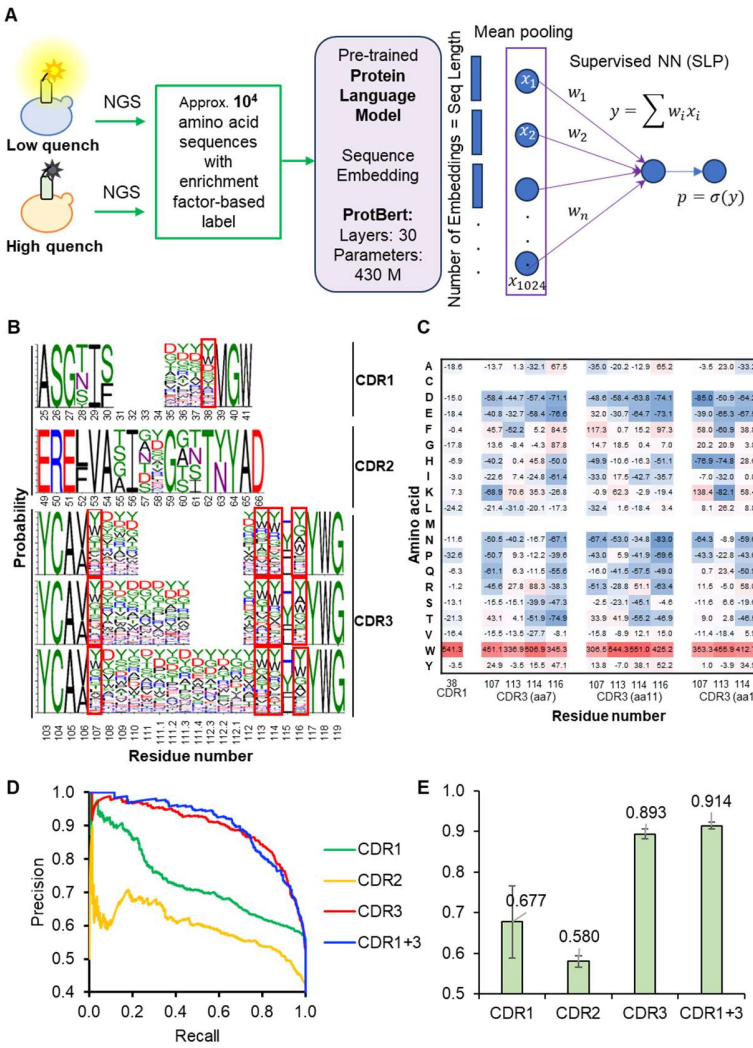

Figure 2. Building original models and evaluating the performance. (A) Schematic image of building an original model for predicting quenching. (B) Web logo of high quenching nanobodies after 2 rounds selections (HQ2). The red boxes represent the position where Trp residues were enriched. (C) Heatmap shows the frequency of amino acid changes in the enriched position. Red: increased; Blue: decreased. (D) Representative PR curve of CDR1, CDR2, CDR3, and CDR1+3 models. (E) The bar graphs represent the means of the AUC value in Figure 2C and the other two PR curves ± standard deviation (n=3).

sequences of CDR1, CDR2, or CDR3 separately. When the prediction performance was evaluated by plotting the PR curve (Figure 2D, 2E), the CDR3-based model showed the largest AUC value, followed by the CDR1- and CDR2-based models, suggesting that CDR3, CDR1, and CDR2 contributes to quenching of TAMRA in that order. To expand the sequence space that can be predicted with the model, we built the additional model using the combination of CDR1 and CDR3 (CDR1+3 model) which showed a relatively higher AUC value with lower SD than that of the CDR3-based model in the PR curve (Figure 2E). Therefore, we decided to use the CDR1+3 model named NanoQ-model 1.0 for predicting effective mutation on quenching, leading to improving the performance of immunosensor.

Prediction of effective mutation on quenching

To test whether the NanoQ-model 1.0 can be used to nominate mutations that achieve quenching, we carried out an in silico screening based on probability scores for high-quenching mutations in nanobodies outside the training dataset. Two nanobodies against the receptor-binding domain (RBD) of the SARS-CoV-2 spike (S) protein, RBD1i13 and RBD10i14, were selected. These two nanobodies were obtained using OrthoRep-driven antibody evolution from parental clones RBD1 and RBD10, which are derived from NbLib.[26] Taking advantage of the fact that RBD1i13 does not have a Trp residue in CDR1 and 3, we performed in silico Trp scanning on RBD1i13 using the NanoQ-model 1.0. We found that several Trp mutations, such as I107W, L113W, and H115W, showed increased scores (Figure 3A). For RBD10i14 carrying one Trp residue in CDR3, the initial score of the wildtype was already as high as 0.67, and there was no obvious score increase during in silico Trp scanning (Figure S4A). Therefore, we further

performed in silico single-site saturation mutagenesis on RBD10i14 to consider mutations beyond Trp that would deepen quenching by the existing Trp residue at position 113. We found several mutations that increase the score, especially mutations introduced at Y112.1, E112, T114, and P116, which showed a relatively greater increase in score than mutations at other positions, suggesting that amino acids around the Trp residue are important for quenching (Figure 3C). Therefore, the

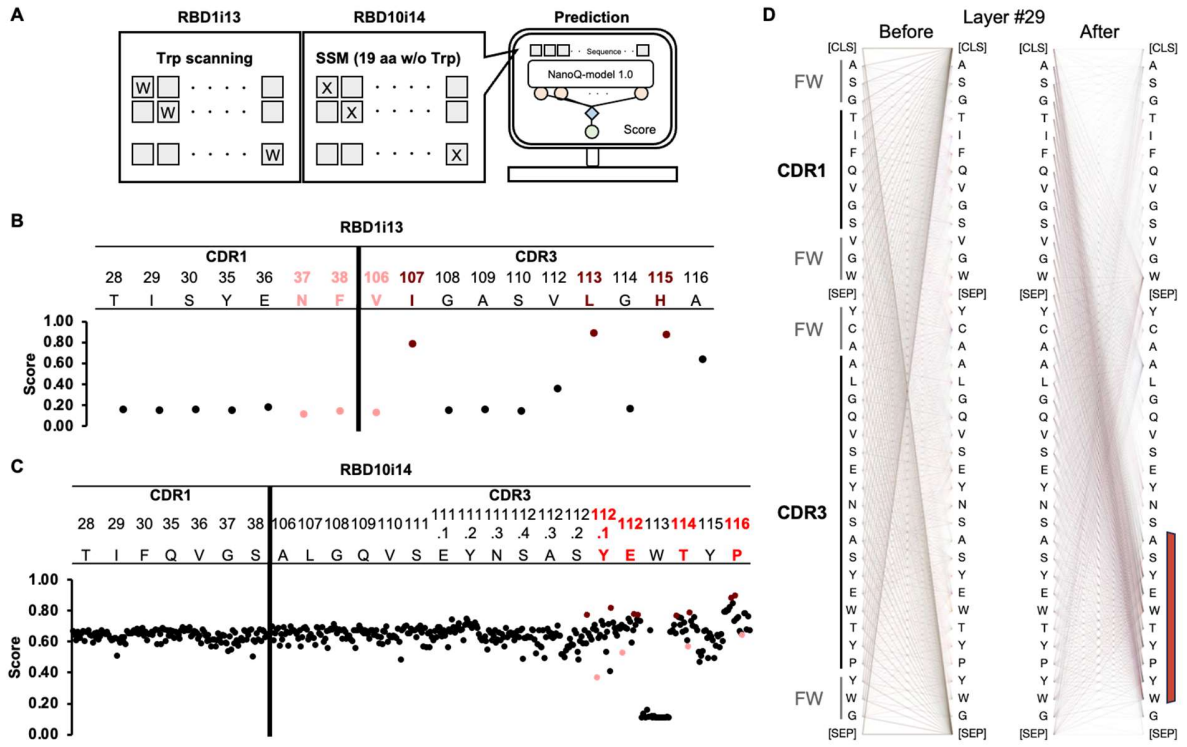

Figure 3. Prediction of effective mutants on quenching. (A) Schematic image of predicting single mutation effect on quenching in silico. (B) The probability score during in silico Trp scanning on RBD1i13. The 3 highest scores were dark red, and 3 lowest scores were pink (C) The probability score during in silico single saturation mutagenesis on RBD10i14. The four positions selected for the validation were red. The 2 highest scores at the same position were dark red, and the 1 lowest score was pink. (D) Representative attention visualization on RBD10i14-E112R nanobody before and after end-to-end training. Comparison of the whole layer #29 from Figure S5 and S6 between the pre-trained ProtBert-BFD (Before) and NanoQ-model 1.0 (After). Red bar show the region with significantly increased attention. The intensity of the line indicates attention weights. FW, framework region. [CLS], classification token representing the N-terminal in pLM; [SEP], separation token.

NanoQ-model 1.0 is able to give prediction scores with position and sequence dependency at single-amino acid resolution. The self-attention visualization[27] was performed to understand which regions of amino acids were focused more after the end-to-end training using our sequence-quenching dataset (Figure 3D, Figure S5 and S6). We found the attention changed significantly in the downstream layers of the model and observed clear attention focusing on the residues surrounding the Trp in our NanoQ-model 1.0 compared with the template model ProtBert-BFD.

Validation of predicted mutations

To validate the predictions above, the quenching efficiency of mutants exhibiting high and low probability scores were evaluated on yeast cell surface (Figure 4A). In the Trp scanning mutants of RBD1i13, all three mutants (I107W, L113W, H115W) exhibiting the highest score showed a lower TAMRA/FITC ratio than wildtype, reflecting deeper quenching of TAMRA. Likewise, two (N37W, V106W) out of three mutants exhibiting the lowest score showed comparable ratio with wildtype, although one (F38W), contrary to prediction, showed lower ratio (Figure 4B). On the other hand, in the Trp scanning mutants on RBD10i14 carrying one native Trp residue, all six mutants regardless of high or low score showed lower TAMRA/FITC ratios than that of wildtype (Figure S4B). These results suggest that the prediction in Trp scanning is easier for the nanobody without Trp residue, and the prediction of higher quenching mutations with higher scores is more accurate than that for lower quenching mutations with lower scores in general.

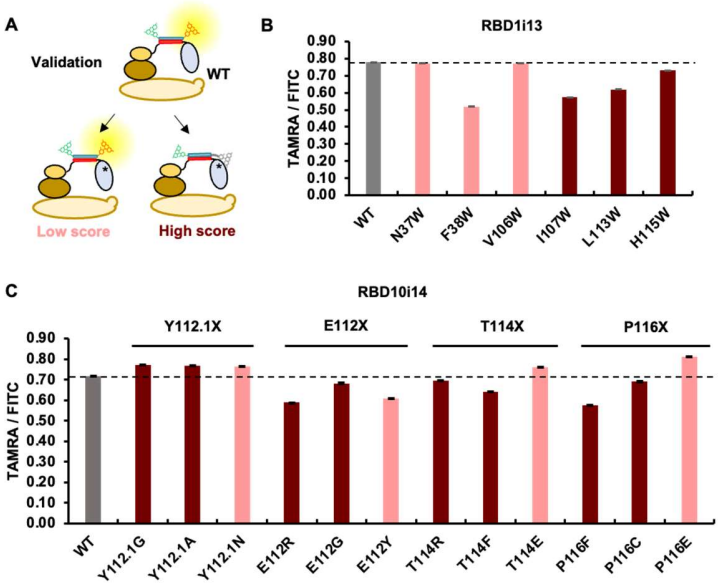

Figure 4. Validation of effective mutation on yeast cell surface. (A) Schematic image of validating the prediction on yeast cell surface. (B, C) TAMRA/FITC ratio of mutants selected during in silico Trp scanning (B) and in silico single saturation mutagenesis (C). The bar graphs represents the mean of TAMRA/FITC ratio ± standard error of mean (SEM). The highest scores were dark red and the lowest scores were colored pink, from the prediction in Figure 3.

To validate the predictions of in silico single-site saturation mutagenesis on RBD10i14, we focused on four positions, Y112.1, E112, T114, and P116, where the introduction of mutations led to higher scores. When two mutants with the highest score and one mutant with the lowest score at each position were evaluated on yeast cell surface, six (E112R, E112G, T114R, T114F, P116F, P116C) out of eight mutants with the highest score showed lower TAMRA/FITC ratio than wildtype, and three (Y112.1N, T114E, P116E) out of four mutants with the lowest score showed higher TAMRA/FITC ratio (Figure 4C). This prediction and its validation show that the prediction accuracy is 75 % for both high and low scores, suggesting that the quenching efficiency of the Trp-excluded single saturation mutagenesis for nanobodies was predicted with reliable accuracy using in silico saturation mutagenesis.

Validation with fabricated Q-body

To confirm whether the mutants showing deeper quenching improve the fluorescence response, we evaluated two independent functions involved in fluorescence response: antigen binding and fluorescence recovery. When we evaluated the antigen binding activity on the yeast cell surface with biotinylated RBD and streptavidin-phycoerythrin (PE) (Figure S7A), L1113W and H115W from RBD1i13, and E112R from RBD10i14 showed comparable PE/FITC ratio with wildtype (Figure S7B, C), suggesting that antigen binding activity was maintained in these mutants. Since

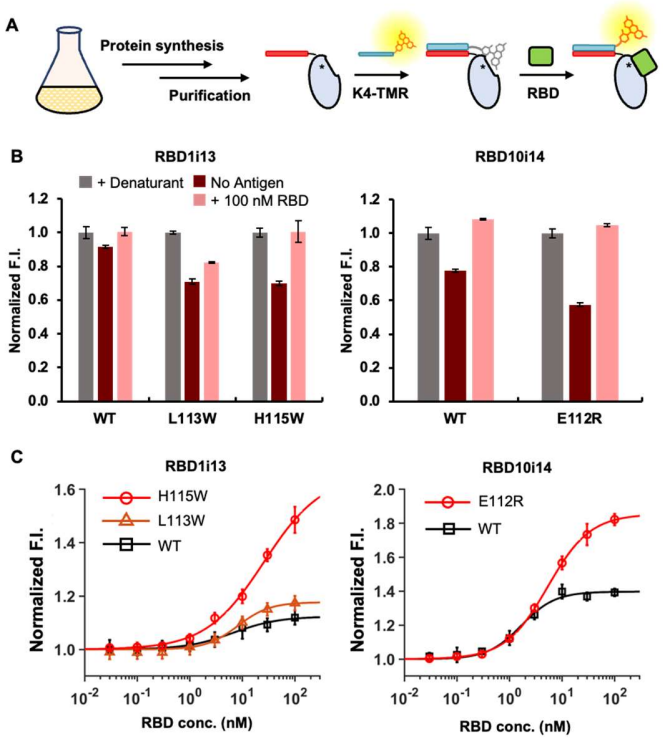

Figure 5. Validation with fabricated Q-bodies. (A) Schematic image of fabricating Q-body using E4-nanobody recognizing RBD expressed in *E. coli*. (B) Fluorescence response of mutants retaining the antigen binding. (C) Dose-response curve of mutants. All data are shown as mean ± standard deviation (n=3).

validation on the yeast cell surface has been completed, E4-tagged nanobody was produced by *E. coli* to fabricate dye/antibody complex (Figure 5A). We found that all mutants in solution showed deeper quenching than wildtype, which is consistent with that on the yeast cell surface (Figure 5B), and showed larger fluorescence response than wildtype in the dose-dependent manner (Figure 5B, 5C). The larger maximum fluorescence response and the lower limit of detection (LOD) values were confirmed in all 3 mutants, while an acceptable increase in EC_{50} was observed (Table S1). It was shown that the response on the yeast surface correlated with that in solution, and that the response and sensitivity of the fluorescent homogeneous immunosensor based on nanobody were successfully improved by introducing mutations exhibiting the high predicted probability score during in silico screening.

Discussion

In the present study, we successfully built an original model, NanoQ-model 1.0, based on a pLM to capture the relationship between quenching and antibody sequence, enabling the prediction of single mutation effects on quenching. We found Trp residues at position 38 of CDR1, and position 107, 113, 114 and 116 of CDR3 were enriched in high quenching mutants (Figure 2B), which is significant compared with other amino acids (Figure 2C). These positions are approximately 12-17 Å away from the N-terminus, and in the vicinity of the rigid structure of framework regions, which are less flexible than the center of CDR (Figure S8). The enrichment may be due to the optimal distance for the dye to access Trp residue and the firm interaction with dye caused by less movement of Trp residue, making it suitable for quenching. Indeed, this is also true for previously fabricated Q-body employing anti-methotrexate (MTX) nanobody, where Trp

residue at position 37 of CDR1 contributes to quenching[5]. Therefore, the position of Trp residues contributing to quenching is similar among nanobodies regardless of their amino acid sequence.

It is observed that introducing a Trp residue is likely to produce a mutant with deeper quenching (Figure 3B, S4A). Interestingly, our model also successfully predicted the effective mutations around existing Trp residue, including mutation to Arg or Phe residue (Figure 4C). Since these amino acids generally do not contribute to quenching,[28] they seem to work for deepening the quenching by the existing Trp residue. In the case of Phe mutation, since it is an aromatic amino acid, it probably interacts with TAMRA through pi-pi stacking to change its orientation between the existing Trp residue and TAMRA, and deepening quenching. On the other hand, since Arg residue is involved in non-specific interaction with other molecules in antibodies via many different types of interactions including pi-cation interaction,[29-31] it would also interact with TAMRA, changing the orientation and deepening the quenching as well.

We were able to predict the effect of mutations on quenching for nanobodies obtained from NbLib, including RBD1i13 and RBD10i14 that were affinity matured from NbLib clones. Further building the model employing large antibody libraries such as naïve single-chain variable region (scFv) libraries as well as nanobody libraries will allow the prediction of quenching for improving the performance of the Q-bodies based on antibodies from many species. Nevertheless, antibodies with quenching do not always show complete fluorescence recovery by antigen. For example, the L113W mutant of RBD1i13 showed less fluorescence recovery compared to H115W mutant upon the addition of antigen regardless of deep quenching. Since the fabricated Q-body employing this mutant exhibited the EC_{50} values comparable to wildtype (Table S1), the lower fluorescence recovery was not caused by the decrease in antigen binding activity derived from the introduction

of mutation but due to insufficient release of TAMRA from quenching by the addition of antigen. We consider that this is because the Trp residue in position 113 is slightly far from the antigen binding site in RBD1i13, leading to the insufficient release of TAMRA from L113W even after binding with the antigen.

In addition to the prediction of quenching, de-quenching is also required. However, it is difficult to perform because there is no universal antigen for all antibodies and fluorescence recovery respond to antigens cannot be evaluated and learned. Since the key amino acid residues for antigen binding were analyzed or predictable through affinity-based screening[32] and the BERT-based antibody paratope prediction model,[33] we expected de-quenching will also be predictable in our future study. Therefore, our original model for the prediction of the quantitative fluorescence response of Q-body will be the first step in selecting the optimal mutation from antibody sequence alone.

Conclusions

In conclusion, we have successfully improved the performance of Q-body by utilizing the prediction model based on the protein language model at the resolution of single amino acid mutation. The performance of Q-bodies was based on the trap of fluorophore to antibody and release triggered via antigen-binding, and our model has the potential to be extended to the prediction of the antigen-dependent switching behavior of small molecules such as peptides and chemicals from antibodies. Thus, this breakthrough will enable the precise prediction of switching behaviors in small molecules, including anti-cancer drugs, and holds promise for applications beyond diagnostics, extending into the field of antibody-drug development.

Materials and Methods

General procedure

Escherichia coli strains XL-10 Gold (Agilent, Santa Clara, CA, USA) and SHuffle T7 Express lysY (New England Biolabs Japan, Tokyo, Japan) were used for general cloning and protein expression, respectively. EBY100 (ATCC, Manassas, VA, USA) was used to develop an E4-tagged library. Yeast-Display Nanobody Library (NbLib) was purchased from Kerefast (Boston, MA, USA) and used as previously described.[25] pYD1-E4-MTXVHH was constructed as previously described,[15] and the oligonucleotides (**Table S2**) were synthesized by Eurofins Genomics (Tokyo, Japan).

Constructing E4-tagged synthetic nanobody library

The synthetic nanobody library for sorting based on quenching was constructed by fusing the E4-tag to the N-terminal of the pre-developed synthetic nanobody library (NbLib). Firstly, to enrich the yeasts displaying nanobody, two rounds of MACS were performed. Briefly, 1.0×10^{10} induced yeast was washed and resuspended in phosphate-buffered saline (PBS) containing 0.1 % BSA (PBS-A) and then incubated with 250 µL Dynabeads M-280 Streptavidin ($6\text{-}7\times10^{8}$ /mL) at 4 ℃ for 2 h. After removing the yeast-expressing nanobodies that bound nonspecifically to magnetic beads (Pre-selection 1), the left yeast was incubated with an additional 250 µL Dynabeads M-280 Streptavidin which was preincubated with biotinylated anti-hemagglutinin (HA) IgG (Fujifilm-Wako Pure Chemicals, Japan) at 4 ℃ for 2 h, followed by washing three times with 1 mL PBS-A (Pre-selection 2). Afterward, the magnetic beads were collected to 20 mL SD (-W) medium and cultured at 4 ℃, 225 rpm, overnight, and the plasmid was extracted from ~1 mL of the saturated yeast by general yeast miniprep. The nanobody sequences were amplified from

extracted plasmid using Inf_AgeI_Krese_back and Inf_BamHI_Kruse_for, and the overhang sequences including E4 or FLAG-tag were amplified from pYD1-E4-MTXVHH using pYD1_back_long and Adapter_E4_for, or Adapter_flag_back and pYD1_for_long, respectively (Table S2, Library construction). Then, the three fragments were fused by overlap PCR, and the DNA fragment was mixed with AgeI- and BamHI-digested pYD1-E4-MTXVHH at the ratio of 3:1 (w/w), followed by transforming them into EBY100 by electroporation (E4-tagged library) as previously described.[34]

Analyzing and sorting yeast-display Q-body based on quenching

The expression ratio of the pre-developed library before and after pre-selection was confirmed by flow cytometric analysis. Briefly, 5.0×10^6 induced yeast was collected by centrifugation at 3,000 × g for 1 min at 4 °C, and the pellet was washed with 1 mL PBS-A. After the pellet was resuspended in 100 μL PBS-A with 1 μL of anti-HA IgG (wako), followed by incubation for 30 min at RT, the pellet was resuspended in 100 μL PBS-A with 1 μL of goat anti-mouse IgG (Fc)-FITC (American Qualex International, USA), followed by incubation for 30 min at 4 °C. After the yeast was resuspended in 500 μL of PBS containing 5% ImmunoBlock (KAC, Japan) (PBS-B), flow cytometric analysis was performed using an SH-800 cell sorter (Sony, Tokyo, Japan) with a blue laser (488 nm) and a detection filter (525/50).

High-quenching and low-quenching nanobodies were selected from the E4-tagged library by cell sorting. Firstly, after inducing the display, the yeast cells (~1×10^8 cells) were collected by centrifugation at 3,000 × g for 1 min at 4 °C, and the pellet was washed with 1 mL PBS-B. After the pellet was resuspended in 1 mL PBS-B, FITC-K4-TAMRA was added at a final concentration of 100 nM, followed by incubation for 1 h at 4 °C. After two washings with 2 mL PBS-B, the

yeast-displayed Q-body was resuspended in 8 mL of PBS-B to perform cell sorting using an SH-800 cell sorter (Sony). In the first selection, 7.9×10^7 events were measured to select 2.5×10^5 cells of high quenching nanobody and 2.0×10^5 cells of low quenching nanobody. In the second selection, 5.9×10^6 events of high quenching nanobody were measured to select 2.5×10^5 cells, and 6.5×10^6 events of low quenching nanobody were measured to select 1.4×10^5 cells using the same gate in the first selection. A blue laser (488 nm) and two detection filters (525/50 and 585/30) were used to measure the fluorescence intensities of FITC and TAMRA, respectively. The obtained data were analyzed using the control software, and the means of FITC and TAMRA whose fluorescence intensity is over 10^3 were calculated to check the quenching. The yeast cells selected by fluorescence-activated cell sorting (FACS) were cultured in 2 mL SD (-W, -U) medium for 2 days, and subsequently cultivated on SD (-W, -U) agar plates for 2 days. A single colony was picked, and a nanobody display was performed.

Next-generation sequence (NGS) and Bioinformatic analysis

NGS analysis was performed to confirm the nanobody sequences comprehensively. Initially, the plasmids encoding the nanobody sequence were extracted from ~1 mL of the saturated yeast by general yeast miniprep, and the sequences were amplified by PCR, in which a barcode sequence was added at the edge of the DNA fragment to discriminate each library (Table S2, NGS). After performing agarose gel purification, the concentration was quantified using the QuantiFluor dsDNA System. Next, the purified DNA and NEBNext Ultra II End Repair/dA-tailing Module (NEB) were mixed to perform the phosphorylation and dA-tailing at the ends of the DNA, followed by purification using AMPPure XP Beads (SQK-LSK114 kit). Then, an adapter sequence was added to the end of the DNA by reacting with Ligation Adapter (SQK-LSK114 kit) and NEBNext

Quick T4 DNA Ligase for 10 minutes at RT, followed by purification again using AMPPure XP Beads. Finally, MinION Flow Cell (R10.4.1) (Oxford Nanopore Technologies) was primed with a solution containing flush buffer, flush tether (SQK-LSK114 kit), and UltraPure BSA (ThermoFisher) to apply the ligation reaction mixture and Library beads (SQK-LSK114 kit). The sequence was read on a MinION sequencing device.

The basecalling was performed using Guppy with supper accurate (SUP) model. Then, the obtained fastq data was prepared for data processing using FASTQ Groomer (v1.1.5) on Galaxy.[35] The fastq data derived from the initial E4-tagged library, Selection 1H, Selection 1L, Selection 2H (High quench) and Selection 2L (Low quench) was sorted with barcode sequence (9/11 bases, Table S3) using Manipulate FASTAQ (v1.1.5) (Maximum error rate = 0.0), followed by trimming with the same barcode sequence using Cutadapt (v4.4) (Maximum error rate = 0.12).[36] Next, CDR1, CDR2, CDR3 and CDR1+3 sequence was extracted by character sequences (Table S3) of the FR before and after CDR sequence using Cutadapt. They were translated from frame 1, and the sequences were further sorted in Excel based on the number of sequences and amino acids in the original NbLib sequence[25] to eliminate sequence errors and the number of valid sequences was calculated (Table S4). Sequence logo (v3.5.0)[37] was then used to indicate the frequency of the amino acid sequence at each location.

Building quench predicting classifier (predictor) based on protein language model

Initially, the enrichment factor (EF) was calculated for the CDR1, CDR2, CDR3, and the combination of CDR1 and CDR3 according to the following formula (eqn 1).

$$EF = \frac{(Frequecy\ in\ Selection\ 2)\times(Total\ number\ of\ reads\ in\ Selection\ 1)}{(Frequecy\ in\ Selection\ 1)\times(Total\ number\ of\ reads\ in\ Selection\ 2)} \quad \text{(eqn 1)}$$

Then, the training dataset for end-to-end training of the pLM was created using the sequences in selection 2 with an EF greater than 1, followed by labeling them as 1.0 for high quenching and 0.0 for low quenching (Table S5). The training dataset was divided into training (60 %), validation (20 %), and test dataset (20 %), respectively. The seed for random data split was set to 42, 4259, and 88 for the repeated training of each dataset. A pre-trained pLM, ProtBert-BFD,[20] with 420 million parameters was used in our study, and a single-layer perceptron with one hidden layer was used as the output neural network. The maximum of the sequence length was set to 512 and the batch size was 32 or 64. The trainings were repeated until the validation loss didn't decrease after five consecutive rounds. The test dataset was used to depict the ROC/PR curves for model evaluation. The multiscale attention visualization was performed using the BertViz tool (v1.4.0).[27]

Q-body preparation and dose-dependency measurement of anti-RBD nanobody

The quenching activity was evaluated for the nanobodies exhibiting high score on yeast cell surface. Initially, pYD1-E4-RBD1i13 and pYD1-E4-RBD10i14 were constructed from pYD1-E4-MTXVHH by exchanging the MTXVHH with RBD1i13 or RBD10i14 (Table S6). Then, the nanobody sequence containing the mutation was prepared by overlapping PCR from pYD1-E4-RBD1i13 or pYD1-E4-RBD10i14 and mixed with NheI- and BamHI-restricted pYD1-vector at the ratio of 3:1 to transformed into EBY100 as previously described.[15]

The antigen-binding activity of the anti-RBD nanobody was evaluated by reacting with biotinylated RBD, which was made using NHS-PEG4 biotin (ThermoFisher) at a final concentration of 10 nM for 1 hour. Evaluation of quenching efficiency and their flow cytometric analysis were the same with the previous research.[15] TAMRA / FITC and PE / FITC were calculated from the mean values of TAMRA, FITC and PE, which are measured employing the

yeasts showing the fluorescence intensity as more than 10^3. Standard error (SEM) of TAMRA / FITC and PE / FITC was calculated from the mean, standard deviation, and number of analyses of FITC and TAMRA or PE calculated as the following formula (eqn 2).

$$SEM = \frac{Mean\ (TAMRA\ or\ PE)}{Mean\ (FITC)} \times \sqrt{\left(\frac{SD\ (FITC)}{Mean\ (FITC)}\right)^2 + \left(\frac{SD\ (TAMRA\ or\ PE)}{Mean\ (TAMRA\ or\ PE)}\right)^2} \times \frac{1}{\sqrt{No.of\ Analysis}}$$

(eqn2)

The mutants showing deeper quenching and antigen binding ability comparable to that of wildtype were transferred to *E. coli* expression vectors containing the E4-tag (Table S6) by general cloning using Inf_E4_AgeI_back and Inf_Kruse_XhoI_for (Table S4, Sub-cloning), and their function as Q-bodies was evaluated as previously described.[8]

Acknowledgments

We thank Dr. Toru Hisabori from the Tokyo Institute of Technology for critical discussion and advice on this work. This work was supported in part by JSPS KAKENHI Grant Numbers JP22H05176 (T.K.), JP24K01264 (T.K.), JP21K14468 (B.Z.), JP24H01123 (B.Z.), JP23K13607 (T.Y.), JP21J21386 (A.I.), and JP22KJ1271 (A.I.) from the Japan Society for the Promotion of Science, the Pre-Research Unit System of the Institute of Innovative Research, Tokyo Institute of Technology (B.Z.), Grant-in-Aid for Pioneering Research from Organization for Fundamental Research of Institute of Innovative Research, Tokyo Institute of Technology (B.Z.), the Cooperative Research Program of 'Network Joint Research Center for Materials and Devices' (K.K.), and the Leadership Off-Campus Project of Tokyo Tech Academy for Leadership (A.I.).

Additional information

Supporting Information accompanied this paper is available separately.

Competing interests: B.Z., T.Y., and T.K. received honoraria from HikariQ Health, Inc. for another unrelated project.

Author contributions

A.I., B.Z., and T.K. conceived the project, designed the experiments, and wrote the manuscript. A.I. performed the experiments. B.Z. designed the scripts, performed the in silico experiments, and developed the Google Colab version of the model. K.M. and K.K. designed the scripts. T.Y. supported performing experiments. K.M., K.K. and T.Y. edited the manuscript. A.W. and C.C.L. provided the experimental resources. T.K. supervised the project. All authors have approved the final version of the manuscript.

References

(1) Renard, M.; Belkadi, L.; Hugo, N.; England, P.; Altschuh, D.; Bedouelle, H., Knowledge-based design of reagentless fluorescent biosensors from recombinant antibodies. *J. Mol. Biol.* **2002,** *318*, 429−442.

(2) Brient-Litzler, E.; Pluckthun, A.; Bedouelle, H., Knowledge-based design of reagentless fluorescent biosensors from a designed ankyrin repeat protein. *Protein Eng. Des. Sel.* **2010,** *23*, 229−241.

(3) de Picciotto, S.; Dickson, P. M.; Traxlmayr, M. W.; Marques, B. S.; Socher, E.; Zhao, S.; Cheung, S.; Kiefer, J. D.; Wand, A. J.; Griffith, L. G.; Imperiali, B.; Wittrup, K. D., Design Principles for SuCESsFul Biosensors: Specific Fluorophore/Analyte Binding and Minimization of Fluorophore/Scaffold Interactions. *J. Mol. Biol.* **2016,** *428*, 4228−4241.

(4) Abe, R.; Ohashi, H.; Iijima, I.; Ihara, M.; Takagi, H.; Hohsaka, T.; Ueda, H., "Quenchbodies": quench-based antibody probes that show antigen-dependent fluorescence. *J Am. Chem. Soc.* **2011,** *133*, 17386−17394.

(5) Inoue, A.; Ohmuro-Matsuyama, Y.; Kitaguchi, T.; Ueda, H., Creation of a Nanobody-Based Fluorescent Immunosensor Mini Q-body for Rapid Signal-On Detection of Small Hapten Methotrexate. *ACS Sens.* **2020,** *5*, 3457−3464.

(6) Jeong, H. J.; Kawamura, T.; Dong, J. H.; Ueda, H., Q-Bodies from Recombinant Single-Chain Fv Fragment with Better Yield and Expanded Palette of Fluorophores. *ACS Sens.* **2016,** *1*, 88−94.

(7) Zhu, B.; Nosaka, N.; Kanamaru, S.; Dong, J.; Dai, Y.; Inoue, A.; Yang, Y.; Kobayashi, K.; Kitaguchi, T.; Iwasaki, H.; Koike, R.; Wakabayashi, K.; Ueda, H., Rapid and sensitive SARS-CoV-2 detection using a homogeneous fluorescent immunosensor Quenchbody with crowding agents. *Analyst* **2022,** *147*, 4971−4979.

(8) Yasuda, T.; Inoue, A.; Kitaguchi, T.; Ueda, H., Rapid construction of fluorescence quenching-based immunosensor Q-bodies using alpha-helical coiled-coil peptides. *Chem. Commun.* **2021,** *57*, 8206−8209.

(9) Kabat, E. A.; Wu, T. T.; Bilofsky, H., Attempts to Locate Residues in Complementarity-Determining Regions of Antibody Combining Sites That Make Contact with Antigen. *Proc. Natl. Acad. Sci. U. S. A.* **1976,** *73*, 617−619.

(10) Dong, J.; Ueda, H., Recent Advances in Quenchbody, a Fluorescent Immunosensor. *Sensors*

2021, *21*, 1223.

(11) Jeong, H. J., Quenchbodies That Enable One-Pot Detection of Antigens: A Structural Perspective. *Bioengineering* **2023,** *10*, 1262.

(12) Ito, Y.; Sasaki, R.; Asari, S.; Yasuda, T.; Ueda, H.; Kitaguchi, T., Efficient Microfluidic Screening Method Using a Fluorescent Immunosensor for Recombinant Protein Secretions. *Small* **2023,** *19*, e2207943.

(13) Zhu, B.; Du, Z.; Dai, Y. C.; Kitaguchi, T.; Behrens, S.; Seelig, B., Nanodroplet-Based Reagent Delivery into Water-in-Fluorinated-Oil Droplets. *Biosensors* **2023,** *13*, 768.

(14) Dai, Y. C.; Sato, Y.; Zhu, B.; Kitaguchi, T.; Kimura, H.; Ghadessy, F. J.; Ueda, H., Intra Q-body: an antibody-based fluorogenic probe for intracellular proteins that allows live cell imaging and sorting. *Chem Sci* **2022,** *13*, 9739−9748.

(15) Inoue, A.; Yasuda, T.; Zhu, B.; Kitaguchi, T.; Murakami, A.; Ueda, H., Evaluation and selection of potent fluorescent immunosensors by combining fluorescent peptide and nanobodies displayed on yeast surface. *Sci. Rep.* **2021,** *11*, 22590.

(16) Ning, X.; Yasuda, T.; Kitaguchi, T.; Ueda, H., Construction of Fluorescent Immunosensor Quenchbody to Detect His-Tagged Recombinant Proteins Produced in Bioprocess. *Sensors* **2021,** *21*, 4993.

(17) Liang, Y. F.; Li, J. D.; Fang, R. Y.; Xu, Z. L.; Luo, L.; Chen, Z. J.; Yang, J. Y.; Shen, Y. D.; Ueda, H.; Hammock, B.; Wang, H., Design of an Antigen-Triggered Nanobody-Based Fluorescence Probe for PET Immunoassay to Detect Quinalphos in Food Samples. *Anal. Chem.* **2023,** *95*, 12321−12328.

(18) Dong, J.; Oka, Y.; Jeong, H. J.; Ohmuro-Matsuyama, Y.; Ueda, H., Detection and destruction of HER2-positive cancer cells by Ultra Quenchbody-siRNA complex. *Biotechnol. Bioeng.* **2020,**

117, 1259−1269.

(19) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I., Attention Is All You Need. In *Advances in Neural Information Processing Systems (Curran Associates, 2017)*, p 5998−6008.

(20) Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; Bhowmik, D.; Rost, B., ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *Ieee T Pattern Anal* **2022,** *44*, 7112−7127.

(21) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millan, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D., Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021,** *373*, 871−876.

(22) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D., Highly accurate protein structure prediction with AlphaFold. *Nature* **2021,** *596*, 583−589.

(23) Schmirler, R.; Heinzinger, M.; Rost, B., Fine-tuning protein language models boosts predictions across diverse tasks. *Nat. Commun.* **2024,** *15*, 7407.

(24) Porebski, B. T.; Balmforth, M.; Browne, G.; Riley, A.; Jamali, K.; Furst, M.; Velic, M.;

Buchanan, A.; Minter, R.; Vaughan, T.; Holliger, P., Rapid discovery of high-affinity antibodies via massively parallel sequencing, ribosome display and affinity screening. *Nat. Biomed. Eng.* **2024,** *8*, 214−232.

(25) McMahon, C.; Baier, A. S.; Pascolutti, R.; Wegrecki, M.; Zheng, S.; Ong, J. X.; Erlandson, S. C.; Hilger, D.; Rasmussen, S. G. F.; Ring, A. M.; Manglik, A.; Kruse, A. C., Yeast surface display platform for rapid discovery of conformationally selective nanobodies. *Nat. Struct. Mol. Biol.* **2018,** *25*, 289−296.

(26) Wellner, A.; McMahon, C.; Gilman, M. S. A.; Clements, J. R.; Clark, S.; Nguyen, K. M.; Ho, M. H.; Hu, V. J.; Shin, J. E.; Feldman, J.; Hauser, B. M.; Caradonna, T. M.; Wingler, L. M.; Schmidt, A. G.; Marks, D. S.; Abraham, J.; Kruse, A. C.; Liu, C. C., Rapid generation of potent antibodies by autonomous hypermutation in yeast. *Nat. Chem. Biol.* **2021,** *17*, 1057−1064.

(27) Vig, J., A Multiscale Visualization of Attention in the Transformer Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (Association for Computational Linguistics, 2019)*, p 37−42.

(28) Marme, N.; Knemeyer, J. P.; Sauer, M.; Wolfrum, J., Inter- and intramolecular fluorescence quenching of organic dyes by tryptophan. *Bioconjug. Chem.* **2003,** *14*, 1133−1139.

(29) Birtalan, S.; Zhang, Y.; Fellouse, F. A.; Shao, L.; Schaefer, G.; Sidhu, S. S., The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. *J. Mol. Biol.* **2008,** *377*, 1518−1528.

(30) Tiller, K. E.; Li, L.; Kumar, S.; Julian, M. C.; Garde, S.; Tessier, P. M., Arginine mutations in antibody complementarity-determining regions display context-dependent affinity/specificity trade-offs. *J. Biol. Chem.* **2017,** *292*, 16638−16652.

(31) Kumar, K.; Woo, S. M.; Siu, T.; Cortopassi, W. A.; Duarte, F.; Paton, R. S., Cation-pi

interactions in protein-ligand binding: theory and data-mining reveal different roles for lysine and arginine. *Chem Sci* **2018,** *9*, 2655−2665.

(32) Saka, K.; Kakuzaki, T.; Metsugi, S.; Kashiwagi, D.; Yoshida, K.; Wada, M.; Tsunoda, H.; Teramoto, R., Antibody design using LSTM based deep generative model from phage display library for affinity maturation. *Sci. Rep.* **2021,** *11*, 5852.

(33) Li, S.; Meng, X.; Li, R.; Huang, B.; Wang, X., NanoBERTa-ASP: predicting nanobody paratope based on a pretrained RoBERTa model. *BMC Bioinformatics* **2024,** *25*, 122.

(34) Benatuil, L.; Perez, J. M.; Belk, J.; Hsieh, C. M., An improved yeast transformation method for the generation of very large human antibody libraries. *Protein Eng. Des. Sel.* **2010,** *23*, 155−159.

(35) Blankenberg, D.; Gordon, A.; Von Kuster, G.; Coraor, N.; Taylor, J.; Nekrutenko, A.; Galaxy, T., Manipulation of FASTQ data with Galaxy. *Bioinformatics* **2010,** *26*, 1783−1785.

(36) Martin, M., Cutadapt removes adapter sequences from high-throughput sequence reads. *EMBnet J* **2011,** *17*, 10−12.

(37) Crooks, G. E.; Hon, G.; Chandonia, J. M.; Brenner, S. E., WebLogo: a sequence logo generator. *Genome Res.* **2004,** *14*, 1188−1190.