ASPred: Identification of Antigen Specific B-cell receptors from single V(D)J sequences using Large Language Models

Karen Paco [1*] ,* **Mariana Paco Mendivil**[2*], **Zihao Zhang**[1], **Isabel Condori Roman**[1]

Sanaz Zebardast[1], **Peace Olatayimbo**[1], **Rashid M Alam**[1], **Christian Davila Ojeda**[1]

Dhruv Patel[1], **Jonathan Felix**[1], **Tristan Yang**[1], **Faisal bin Ashraf**[3], **Jordan Lay**[4]

Ilya Tolstorukov[1], **Karine Le Roch**[3], **Matthew H. Sazinsky**[5], **Jeniffer Hernandez**[1]

Stefano Lonardi[3], **Fernando L. Barroso da Silva**[6 7], **Animesh Ray**[1]

[1]Keck Graduate Institute - 535 Watson Dr. Claremont CA
[2]University of California, San Diego, San Diego, CA
[3]University of California, Riverside
[4]University of Cambridge, Cambridge CB2 1TN, UK
[5]Pomona College 645 N. College Avenue, Claremont, CA
[6]Universidade de São Paulo, Departamento de Ciências Biomoleculares,
Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Ribeirão Preto SP, Brazil
[7]Department of Chemical and Biomolecular Engineering, NC State University, Raleigh NC

Abstract

The rapid sequencing of antibody genes has accelerated vaccine development. However, predicting synthetic antibodies capable of binding and neutralizing novel antigens remains challenging due to a limited understanding of the rules of protein-protein interaction at the surface of an antigen to which its cognate antibody protein binds. While recent advances in single-cell sequencing of antibody-producing B-cells sequences have improved precision in mapping B-cell receptors (or BCRs, which are the membrane-bound forms of the antibodies) to their cognate antigens, there remain additional challenges. We have developed a computational strategy, the Antibody Specificity Predictor (ASPred), with which we have trained two Large Language Models (LLMs) with known sequences of antigen-BCR pairs to predict antigen-specific BCRs from the total BCR repertoire of immunized mice. By leveraging pattern recognition capabilities of LLMs we successfully classify novel B-cell receptors with a challenge antigen not represented in the training set, without the need for preselecting the B cells by antigen binding. The properties of the top 10 predicted candidates were validated by coarse-grained molecular dynamics simulations. These results suggest that sufficient information exists in BCR-antigen sequence pairs for LLMs to reliably predict antigen-antibody interaction specificity, potentially opening new avenues for the computational design of synthetic antibodies for vaccine and therapeutic development.

*email:karenpaco33@gmail.com, https://mpacomendivil.github.io/paco-bio

Preprint. Under review.

1 Introduction

Adaptive immunity depends on mutation and recombination at antibody heavy (H) and light (L) chain genes in immune cells [1] which generate the diversity [2, 3] followed by the selection and maturation of immune B-lymphocytes (B-cells) making the appropriate antibodies through signaling by cell-surface receptors (B-cell receptors or BCRs) that are composed of their respective antibody proteins in a membrane-localized form [4, 5]. B-cells are subsequently selected for maturation through a complex process of multicellular signaling to expand clonally, followed by immunoglobulin class switching by recombination to produce secreted antibody molecules with different constant regions as scaffolds (IgG, IgA, or IgE) but with the same variable H and L chain regions, respectively [6, 4, 3].

Monoclonal antibodies, which have become a cornerstone of medicine and biotechnology, are generally selected from pools of many antibodies by various *in vitro* methods including the hybridoma technology, phage or yeast-surface displays, and more recently by B-cell cloning [7, 8]. Rapid sequencing of genes encoding circulating antibodies following a virus infection has led to an unprecedented speed in producing monoclonal antibody vaccines against an emerging pandemic [9]. One approach to accelerate this capability even further is to computationally predict the sequence of synthetic antibodies that can potentially bind and neutralize a novel target antigen, a goal yet to be realized [10].

The space of antibody diversity is immense, of the theoretical order of at least 10^{16}, although in practice a far lower diversity is observed [11, 12, 3]. Diversity is generated during the maturation of B and T cell precursors in response to the presence of foreign antigens that are presented to these cells through cell surface displays. The germline antibody gene sequences at their V-D-J and V-J segments in H and L chain genes, respectively, undergo targeted cleavage, DNA nucleotide insertion/deletion mutations, and recombinational repair, producing numerous sequence variants in three segments of each H and L chain genes – the complementarity determining region or CDR 1 through 3, with maximum variation occurring in the CDR3 segment [6, 13]. These variant H and L chain CDRs, buried within the constant immunoglobulin M (IgM) scaffolds, assemble as the antibody precursors and are displayed as membrane-bound forms on B-cell surfaces, the BCRs [14]. Those cells exhibiting (and encoding) relatively strong binders to the antigen selectively proliferate in a process mimicking Darwinian evolution, and under selective pressure produce clones of highly specific antibody-producing mature B-cells [14].

Factors contributing to the selection and maturation of specific B-cell clones are complex and poorly understood [15, 16]. Only a small fraction (0.01–0.1%) of circulating B-cells in an antigen-inoculated individual display BCR sequences specific to the challenge antigen, and the structural relationships between the antigenic surface (epitope) and its corresponding antibody region (the paratope) that specifically binds to the epitope are obscure. These and other factors make identifying the origins of antibodies from the corresponding B-cell clones a challenging problem. Recent advances, such as sorting B-cells that bind to labeled antigens and sequencing their BCRs, have enabled the identification of antigen-specific BCRs [17, 18]. However, these techniques remain time-consuming, expensive, and laboratory-intensive. They are also technically complex, sensitive to experimental conditions, and can introduce biases in cell selection. These labor-intensive processes generate vast and complex datasets that require extensive computational analysis, yet often provide limited information about the functional relevance of the antibodies thus identified [19, 5].

Recent advances in supervised machine learning, particularly deep learning techniques, have enabled the prediction of novel protein 3D structures and protein complexes by training on vast protein sequence datasets [20–23]. Models like AlphaFold2[22] and RoseTTAFold[20] employ specialized architectures to capture evolutionary and structural features within protein sequences, effectively predicting protein folding and interactions. While protein language models (a form of LLMs) help in modeling sequence patterns to folded protein structures by capturing short- and long-range dependencies[24, 23, 21, 25] and correlating these dependencies with known protein structures, the reliable prediction of specific antigen-antibody pairs remains a largely unsolved problem. Challenges include the scarcity of antigen-antibody paired sequence datasets, the flexibility of the protein backbone and amino acid residues at binding interfaces, and potential biases arising from the clonal nature of antibody evolution[26–28].

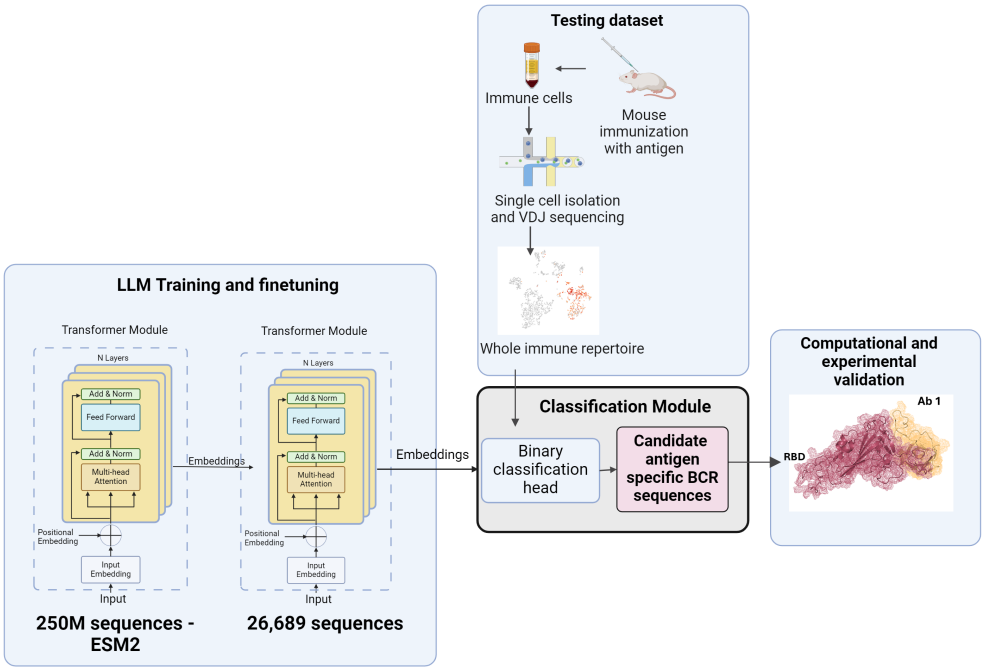

Figure 1: Diagram of method for identification of antigen-specific B-cell receptors

Here we address the challenge of identifying antigen-specific BCR sequences without physically pre-selecting B cells for antigen binding. We demonstrate that a fine-tuned Large Language Model, ASPred, can classify a single B-cell V(D)J sequence repertoire of mice immunized with a challenge antigen, successfully identifying antigen-specific BCR sequences. The binding properties of the 10 top candidates that were obtained from ASPred were tested to form complexes with the RBD from the wildtype SARS-CoV-2 by different molecular simulation approaches. These findings suggest that sufficient information exists within BCR-antigen sequence pairs for an LLM to uncover sequence dependencies and reliably predict antigen-antibody interaction specificity.

2 Results

2.1 Prediction of antigen-specific B-cell receptors

To identify antigen-specific B cell receptors (BCRs) from the total repertoire of peripheral blood mononuclear cells (PBMC), we employed three distinct approaches. The first approach utilized the clustering tool *InterClone* [29], which introduces a novel method to cluster antibodies sharing antigenic targets based on their complementarity-determining region (CDR) sequences, with *MMSeq2* facilitating effective sequence clustering through homology alignment and gap management. We constructed a dataset comprising 11,917 known SARS-CoV-2-specific antibody heavy chain sequences sourced from *CovAbDab* [30] and an additional 310 heavy chain sequences obtained from our own experimental single-cell testing dataset. Using a clustering threshold of 70% for the CDR similarity index (SID) and requiring 90% coverage, we identified 96 candidate sequences from our single-cell testing dataset, which clustered with known SARS-CoV-2 antibody sequences.

The remaining two approaches were based on advanced transformer models: the Evolutionary Scale Modeling (*ESM-2*)[32] and the Protein structure-sequence T5 (*Prot-T5*)[31]. For both models, we utilized pre-trained weights from established protein transformer architectures, specifically `esm2_t33_650M_UR50D` and `prot_t5_xl_uniref50`. We incorporated a binary classification head in each model to differentiate between SARS-CoV-2-specific and non-specific antibodies.

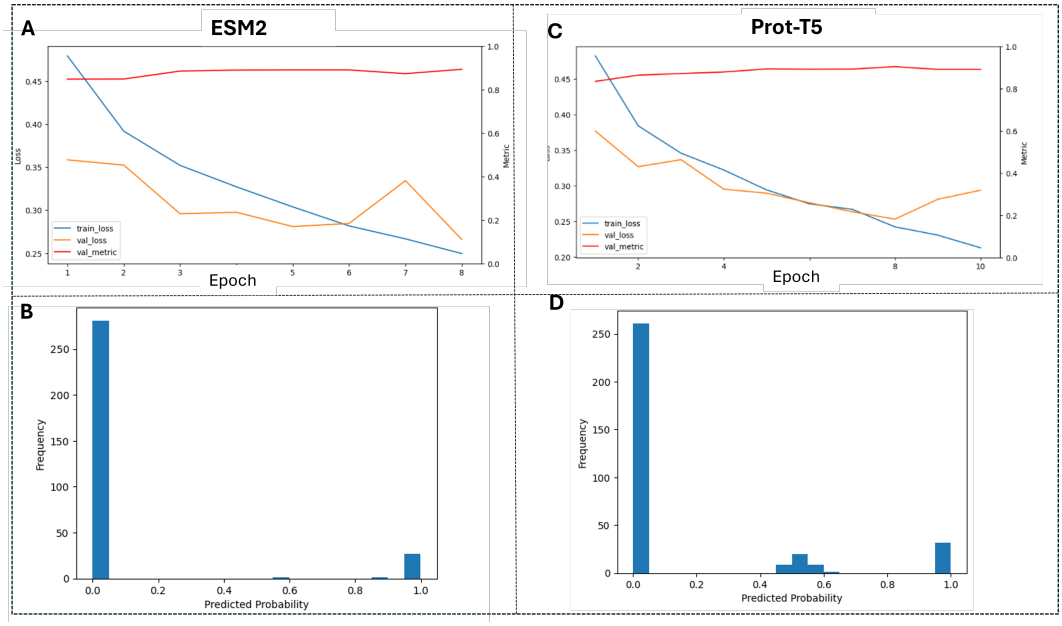

Figure 2: Predicted probability for the ESM-2 model (A and B) [21] and ProT-T5 (C and D)] [31].

Our antibody language models were fine-tuned on a comprehensive dataset of 26,689 sequences, which included a balanced representation of both positive (antigen-specific) and negative (non-specific) samples. To enhance performance and reduce computational burden during the fine-tuning process, we utilized a technique known as Low-Rank Adaptation[33], or LoRA. This method involves fixing the weights of the pre-trained model while incorporating trainable low-rank decomposition matrices into each layer of the Transformer architecture. By doing so, LoRA significantly decreases the number of parameters that require adjustment for downstream tasks. This efficient adaptation process enables our models to effectively learn task-specific nuances while maintaining low computational overhead.

The *ESM-2* model was fine-tuned with a dropout probability of 0.2, a learning rate of $2e^{-5}$, a weight decay of 0.01, a batch size of 4, and was trained for 10 epochs, achieving an accuracy of 90% on the test dataset. Similarly, the *Prot-T5* model was fine-tuned with a dropout probability of 0.1, a learning rate of $2e^{-5}$, a batch size of 8, and trained for 8 epochs, yielding an accuracy of 88% on its test dataset. By applying a threshold of 0.4 on the predicted probabilities, the *ESM-2* model identified 60 candidates, while the *Prot-T5* model identified over 100 candidates. The selection of sequences with the highest predicted probabilities indicates that our binary classification system effectively distinguishes between binder and non-binder antibodies, with most probabilities clustering around the extremes of 0 (non-antigen-specific) and 1 (antigen-specific).

We merged the three sets of candidate sequences, highlighting those selected by multiple models. From this, we identified the top ten SARS-CoV-2-specific antibodies for further analysis and validation of their binding to the receptor-binding domain (RBD) of the SARS-CoV-2 spike protein.

2.2 Antibody-antigen docking validates predictions

A multi-computational approach at the molecular level, integrating docking, a constant-pH coarse-grained (CG) Monte Carlo (MC), and atomistic Molecular Dynamics (MD) simulations, was employed for the *in-silico* validation of the ASPred top candidates.

For the docking validation procedure, we initially pre-processed the candidate antibodies and the RBD, addressing any structural issues before docking the structures using ClusPro 2.0.[34] Multiple conformations were obtained for each antibody-RBD pair. The best conformation from each docking was then selected for further analysis of binding energy using Prodigy[35]. We discovered that antibodies 1 and 157 predicted to be antigen-specific exhibit the strongest binding affinity to the RBD antigen compared to the other candidates, with a binding energy of -14.5 kcal/mol. The typical

Table 1: Top 10 SARS-CoV-2 antibody candidates from ASPred, including their predicted binding probabilities and variable heavy region sequences (CDR1, CDR2, and CDR3). These selected antibodies were further evaluated using the molecular simulation protocols.

ID	CDR1	CDR2	CDR3	Predicted Probability
Ab1	GYTFTSYY	INPSNGGT	TRNEGHYFDY	0.990054846
Ab53	GYTFTSYW	INPSTGYT	ASSYYYGSSYYAMDY	0.999145985
Ab77	GFTFSSYA	ISSGGSYT	ARPFYYGSSYFDY	0.853913486
Ab117	GFTFSNYW	IRLKSNNYAT	TRDDYYAMDY	0.594344974
Ab131	GYTFTDYA	ISTYNGNT	AYGNYWYFDV	0.977651179
Ab157	GYTFTSYV	INPYNDGT	ARDGNYWYFDV	0.989455283
Ab172	SYTFTDYA	ISTYYGNT	ARGDGNYDFAY	0.989846647
Ab192	GYSFTGYT	INPYNGGT	AREGYRYDVEGLDY	0.992514193
Ab200	GYTFTSYW	INPSNGRT	ASRGYDEGYAMDY	0.991556287
Ab242	GYSFTDYN	IDPYNGGT	ARDHYGNYGDFDY	0.998622537

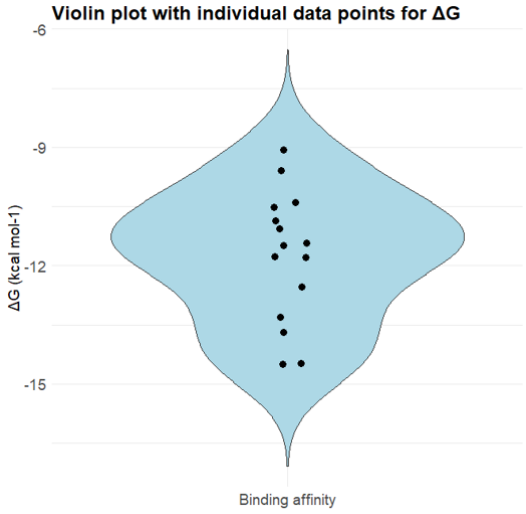

Figure 3: Binding affinities of the top 10 antibody candidates specific to RBD. All K_d values were obtained from Prodigy. [36]

binding energy range for a specific antigen-antibody pair is -8 to -15 kcals/mole, nearer the latter for more specific binding. These results suggest that the predicted antibodies should bind relatively well to the antigens, as determined by docking and binding energy estimates.

2.3 Relative binding affinities at physiological pH condition

The computed free energies of interaction from the Fast cOarse-grained pRotein-proTein modEl (FORTE)[37] were used to assess the relative binding affinities between the top 10 candidates and the wild-type RBD at pH 7.4 under physiological 1:1 salt concentration. Figure 4 presents a comparative analysis of these top candidates, all of which exhibit binding affinities comparable with values typical of high-efficiency monoclonal antibodies designed by other simulation protocols [37, 38]. Notably, ASPred predicts antibodies with even higher affinities than those achieved by other multiscale *in silico* approaches[37]. Among the top 10, the strongest candidate found by FORTE is ID 117, closely followed by IDs 53 and 157, all demonstrating binding affinities surpassing those in previous studies[37, 38]. This finding further validates the success of ASPred in identifying ideal antibodies.

An additional important attribute of a designed antibody is its structural stability. FPTS[39] was employed to determine the averaged net charge of each titratable group of the antibodies under these physicochemical conditions, and these values were then used to calculate the electrostatic stability of the top 10 candidates at the same structural configurations used in the FORTE simulations. This approach provides insights into the stability profiles of each antibody, complementing the binding

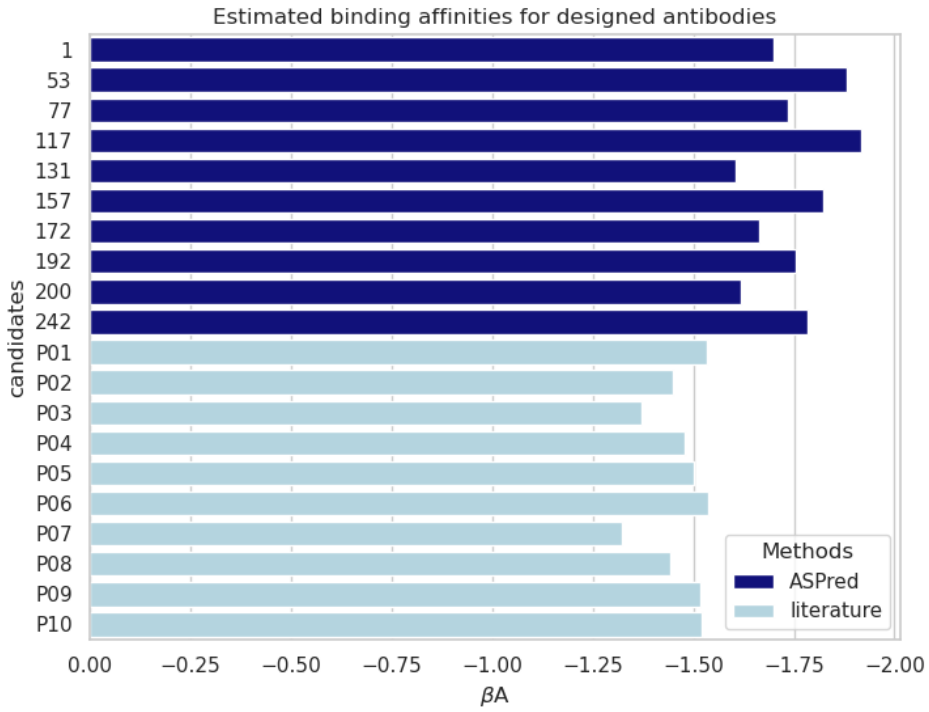

Figure 4: Computed binding affinities of the top 10 best antibody candidates specific to the wildtype RBD by FORTE. The minima free energy of interaction values (βA) were measured for the SARS-CoV-2 RBD-mAbs (heavy chain only) complexation at pH 7.4 and 150 mM of NaCl. β is the thermodynamic beta ($\beta = kT$ where k is the Boltzmann constant and T is the temperature). The estimated error is 0.01 kT. Data for the antibodies P01 to P10 was obtained running FORTE for the structures done by Neamtu *et al.*[37].

affinity analysis to ensure robust antibody design. The main results, shown in Figure 5, indicate that all antibody chains designed by ASPred display electrostatic stability values comparable to those of other anti-SARS-CoV-2 antibodies, with some cases showing even greater stability. Notably, candidate ID 117, one of the top candidates with the highest relative binding affinity (see Figure 4), also demonstrates a high level of electrostatic stability. This conformational stability may enhance its likelihood of successful binding in *in vitro* assays, supporting its potential as a strong candidate for further experimental validation.

3 Discussion

The identification of antigen-specific B cell receptors (BCRs) through a combination of clustering techniques and advanced transformer models represents a promising avenue for discovering antibodies within entire immune repertoires. While our approach serves as a proof of concept for targeting SARS-CoV-2, it also lays the foundation for applying the ASPred algorithm to other antigens. This versatility offers a significant advantage, as the methods and principles established in this study can be broadly applied across various infectious diseases if sufficient data on antibody specificity is available.

Currently, despite the abundant information on whole immune repertoires, there is a noticeable deficiency in data regarding antibody specificity. One potential direction for future exploration is leveraging existing immune repertoire datasets to uncover novel antibody sequences that may have been missed due to technical limitations and experimental constraints commonly found in laboratory settings.

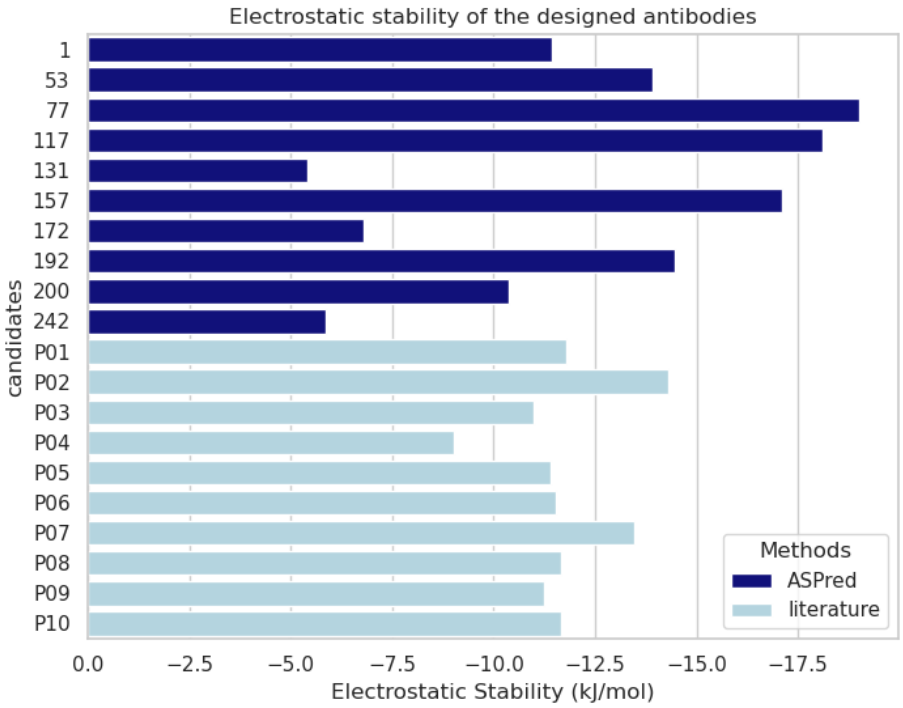

Figure 5: Simulated electrostatic stability for the designed ASPred antibodies (heavy chain) at pH 7.4. Salt concentration was fixed at 150 mM. Data for the antibodies P01 to P10 was obtained by performing the calculations for the structures done by Neamtu *et al.* [37].

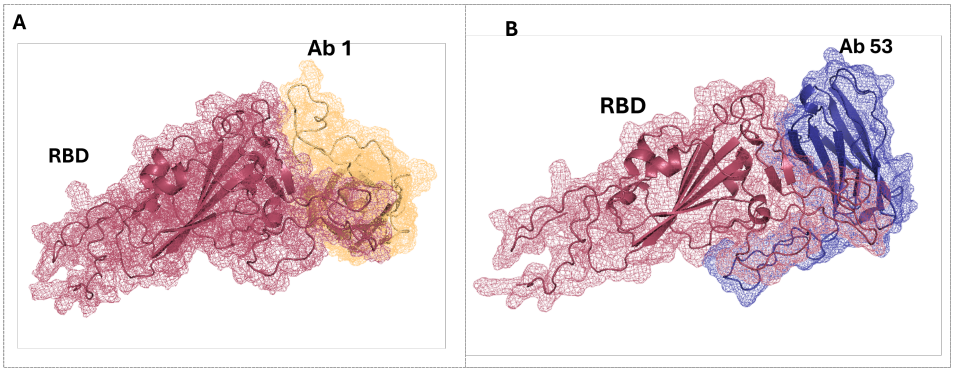

Figure 6: The three-dimensional structure of two antibody candidates bound to RBD

In the realm of transformer-based methods, fine-tuning the models *ESM-2* and *Prot-T5* has shown effectiveness in distinguishing between antigen-specific and non-specific antibodies. The high accuracy rates achieved by both models demonstrate their capacity to capture complex relationships within antibody sequences. This finding was further corroborated by *in silico* validations conducted with additional molecular simulation methods. These complementary approaches strengthen the reliability of ASPred's predictions and underscore the robustness of one of the candidates (ID 117) for potential *in vitro* assays. However, concerns about generalizability arise since these models were primarily trained on known antibodies, and their performance with entirely novel sequences warrants further investigation. Expanding the training datasets to include a more diverse range of antibodies, particularly those developed through synthetic biology, could enhance the robustness of these classification systems. This enhancement would facilitate the identification of critical antibody sequences across a wider array of pathogens. If these models successfully capture the biophysical principles underlying antibody-antigen interactions, we could soon identify a broader spectrum of antigen-specific antibodies relevant to multiple disease states—a development that would have significant implications for vaccine and drug discovery.

Insights derived from docking studies and molecular dynamics simulations have proven crucial in elucidating the binding interactions between selected antibody candidates and the receptor-binding domain (RBD) of the SARS-CoV-2 spike protein. One candidate exhibited a strong binding affinity, indicated by a binding energy of -14.5 kcal/mol, aligning well with predictions made by ASPred. However, it is important to interpret these results cautiously; binding energy alone does not fully capture the biological relevance of these interactions. Factors such as conformational stability, entropy changes during binding, and dynamic interactions within a physiological context are essential to fully understand the true efficacy.

In conclusion, while our findings contribute to advancing the field of computational immunology, they also underscore inherent challenges and limitations. Addressing these challenges—such as potential biases in training data and the complexities of real-world biological interactions—is crucial for effectively applying these methodologies in therapeutic development. Future research should focus on integrating more diverse datasets and exploring various biological contexts, while continuously validating computational predictions through experimental means. This comprehensive approach will be essential for establishing robust and effective antibody therapies against multiple pathogens, thereby extending the applicability of our methodologies well beyond SARS-CoV-2.

4 Methods

Ethics statement The animal work was conducted with the approval of the UC Riverside Institutional Animal Care and Use Committee (IACUC). All animal procedures were performed according to approved guidelines.

Immunization and sample collection Mice were housed at the University of California Riverside vivarium. BALB/c female mice at 6 weeks of age were given subcutaneous injections of 100 μL per day in the back of the neck on days 0 and 14 of antigens emulsified in aluminum hydroxide 2 % (full-length spike protein SARS-CoV-2 nCov-2019). Blood samples were collected post-immunization and on days 14 and 28. On day 28 mice were deeply anesthetized with isoflurane and blood was drawn by cardiac puncture. Mice were immediately euthanized by cervical dislocation according to IACUC guidelines. Peripheral Blood Mononuclear Cells (PBMCs) were isolated using the direct human PBMC isolation kit (StemCell Technologies) and cryopreserved at -80°C for further work. Analysis of antigen immunogenicity in mice by Enzyme Linked-Immunosorbent Assay (ELISA): ELISA was carried out in 96-half-area well plates from (Greiner Bio-One), plates were coated with the full-length spike protein from SARS-CoV-2 nCov-2019 (0.4 μ g/well) using sodium carbonate coating buffer (0.05M, pH 9.6) and allowed to incubate overnight at 4°C. Plates were washed (2x) with PBST (PBS with 0.05% Tween-20) and once with PBS. Plates were blocked with a 5% non-fat dry milk solution for 1 hour. Mouse plasma samples were added (dilutions: 1:5,), and each dilution was incubated with the immobilized antigens for 1 hour at room temperature with continuous shaking. The plates were washed 5 min each 2x with PBST and 1x with PBS. The bound mouse IgG antibodies were then detected with horseradish peroxidase (HRP) conjugated anti-mouse IgG secondary antibody (EMD Millipore Corp., Catalog No. AP112P, Lot: 3855607) diluted at 1:2,500 in PBS by incubating at RT for 1 hr, and the plates were washed 2x in PBST, rinsed 1x with PBS. The enzymatic reaction

was initiated by adding the OPD Peroxidase substrate (Sigma Alrich, P9187-50SET) to the wells and allowing the reaction to proceed for 5 minutes and the developed color was measured using an ELISA microplate reader (NanoQuant Infinite M200, Tecan) at 450 nm after 5 and 30 minutes.

GEM Generation and construction of gene expression next-generation sequencing libraries PBMCs were used without antigen labeling and sorting. Single-cell suspensions were mixed with nuclease-free water and $5'$ single-cell RNA master mixture, then loaded into a Chromium chip with barcoded gel beads and partitioning oil. The chip was placed in the Chromium controller to generate gel beads in emulsion (GEMs). cDNA was obtained from 100 μ l GEMs/sample by reverse-transcription reactions: 53 ° C for 45 min, 85 ° C for 5 min, then maintained at 4 °C. cDNA products were purified and cleaned using Dynabeads. cDNA was amplified by PCR: 98°C for 45s; 98 °C for 20 s, 63 °C for 30 s, 72 °C for 1 min and amplified for 16 cycles; then, 72 °C for 1 min. Amplified PCR products were purified using SPRIselect reagent kit (B23317, Beckman Coulter). The concentration of the cDNA library was determined by Qubit dsDNA HS Assay Kit (Invitrogen) and Bioanalyzer (Agilent, 2100). Single Cell RNA-Seq V(D)J and 5' gene expression library was performed using the Chromium Next GEM Single Cell 5' Reagent Kits v2 (Dual Index)(CG000331, Rev E, 10X Genomics) and Dual Index kit TT set A (PN- 1000215, 10X Genomics) according to the manufacturer's instructions. For unlabeled and unsorted samples, the target was estimated at 5000 cells.

Identification of antigen-specific BCRs using interclone and large language models Raw sequencing data were assembled and annotated using Cell Ranger 7.2.0, and multi-mode was used to process gene expression and VDJ data simultaneously. Mouse transcriptome (GRCm38) provided by Cell Ranger was used as a reference for the process.

Clustering was performed using the source code of InterClone, which employed MMSeqs2 for clustering. The dataset being clustered was constructed with 11,917 known SARS-CoV-2 specific antibody heavy chain sequences from CovAbDab [30] and 310 heavy chain sequences from the single-cell sequencing data.

Two transformer models were constructed from pre-trained weights of two protein transformer models (esm2_33_650M_UR50D and prot_t5_xl_uniref50)[32, 31] with a binary classification head. The models were fine-tuned using the known SARS-CoV-2 specific sequences mentioned above and 14,772 randomly selected known antibody sequences on other targets from PLAbDab.[40] The fine-tuning process was optimized using Low-rank Adaption to reduce the number of trainable parameters and resource consumption. ESM model[32] was fine-tuned with 0.2 dropout probability, 2 $e^{-}5$ learning rate, 0.01 weight decay, batch size of 4, and trained for 10 epochs.

Prediction of antibody structure using Igfold Variable heavy chain protein sequences obtained bt ASPred were folded using IgFold,[1] and the structures were refined with PyRosetta.[41] Antibody sequences were renumbered according to the Chothia scheme[1].

Docking of antibody candidates with RBD The RBD protein was folded using AlphaFold [22] and used as ligands, while antibody structures served as receptors in the ClusPro docking web server, [34] with the antibody mode and non-CDR masking options enabled.

Binding affinities for the complexation of the antibody candidates with RBD The "Fast cOarse-grained pRotein-proTein modEl" (FORTE)[37] is a coarse-grained biophysical model specifically designed to simulate protein-protein interactions, allowing for the dynamic adjustment of amino acid charges on titratable groups based on the surrounding environment at a specified pH (input as a parameter). The core of this model is the "Fast Proton Titration Scheme" (FPTS)[39] combined with the ability to translate and rotate macromolecules using the Metropolis MC method.[42] All calculations with FORTE were performed at pH 7.4, 150 mM NaCl, and 298 K. This model enables molecular simulations that are computationally faster than more complex models, providing relative binding affinities at a reduced computational cost. This allows for the comparison across various molecular systems and/or physicochemical conditions.

The heavy chains of the top candidates predicted by ASPred were folded using IgFold and then used as input for the FORTE simulations. To ensure comparability with previous studies,[37] the wildtype RBD structure was constructed via SWISS-MODEL workspace, using the NCBI reference

sequence NC_045512 (accession YP_009724390.1) as a template. This approach facilitated a direct comparison of the binding affinities between ASPred-predicted antibodies and previously characterized ones.[37] The free energies of interaction (or binding affinities) were calculated as a function of the macromolecules' separation distances by analyzing their center-to-center pair radial distribution functions. These values were sampled in histogram form during the MC production phase, providing detailed distributions of the probability of finding the two molecules at different separation distances. To compare binding affinities across systems, we adopted the free energy minima (βA) observed in these simulations, which represent the most stable interaction points. This approach is a simple and consistent basis for evaluating relative affinities between the top candidates.

Following the equilibration phase, each system underwent at least 3×10^9 MC steps for production-phase sampling. To account for variability, three independent replicates were conducted for each system, allowing for the estimation of statistical errors for a proper comparison between the simulated systems.

Electrostatic stability of the antibody candidates The electrostatic stabilities of the antibody heavy chains were calculated directly from the averaged net charges obtained in the FPTS simulations, using the IgFold-generated conformations. In line with previous studies[37], stability values were determined by evaluating the Coulombic contributions of individual titratable groups for each protein structure under specific conformation and defined physical–chemical conditions. This approach allows for an assessment of how electrostatic interactions may influence the stability of each antibody candidate.

Molecular dynamics simulations of one antibody candidate with RBD The system was first established and equilibrated according to conventional molecular dynamics procedures. Proteins were positioned at the center of a cubic box, with a clearance of 0.2 nm from the boundaries. This box was filled with the TIP3P water model and enriched with 0.15 M of Na^+ and Cl^- ions, applying the AMBER99SB-ILDN protein force field. An energy minimization step was executed to allow the ions to achieve stability. Following this, temperature and pressure equilibration phases were conducted at 310 K and 1 bar, each lasting for 100 ps. After these procedures were completed, comprehensive molecular dynamics simulations were carried out using the designated force field for 100 ns.Periodic boundary conditions were used, and electrostatic interactions were calculated by the particle mesh Ewald method.The resulting molecular dynamics trajectory files were analyzed after the removal of periodic boundary conditions. The stability of each simulated complex was assessed through root mean square deviation calculations for the backbone and visual examination.

Funding

The work described in this report was supported by the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH) under the award number 1R01AI169543.

References

[1] Jeffrey A. Ruffolo, Lee-Shin Chu, Sai Pooja Mahajan, and Jeffrey J. Gray. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature Communications*, 14(1):2389, April 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38063-x. URL https://www.nature.com/articles/s41467-023-38063-x.

[2] Kenneth B. Hoehn, Anna Fowler, Gerton Lunter, and Oliver G. Pybus. The Diversity and Molecular Evolution of B-Cell Receptors during Infection. *Molecular Biology and Evolution*, 33(5):1147–1157, May 2016. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/msw015. URL https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msw015.

[3] Eleonora Market and F. Nina Papavasiliou. V(D)J Recombination and the Evolution of the Adaptive Immune System. *PLoS Biology*, 1(1):e16, October 2003. ISSN 1545-7885. doi: 10.1371/journal.pbio.0000016. URL https://dx.plos.org/10.1371/journal.pbio.0000016.

[4] Juan Carlos Yam-Puc, Lingling Zhang, Yang Zhang, and Kai-Michael Toellner. Role of B-cell receptors for B-cell development and antigen-induced differentiation. *F1000Research*, 7:429, April 2018. ISSN 2046-1402. doi: 10.12688/f1000research.13567.1. URL https://f1000research.com/articles/7-429/v1.

[5] Alexey Ferapontov, Marjan Omer, Isabelle Baudrexel, Jesper Sejrup Nielsen, Daniel Miotto Dupont, Kristian Juul-Madsen, Philipp Steen, Alexandra S. Eklund, Steffen Thiel, Thomas Vorup-Jensen, Ralf Jungmann, Jørgen Kjems, and Søren Egedal Degn. Antigen footprint governs activation of the B cell receptor. *Nature Communications*, 14(1):976, February 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-36672-0. URL https://www.nature.com/articles/s41467-023-36672-0.

[6] Ganesh E. Phad, Dora Pinto, Mathilde Foglierini, Murodzhon Akhmedov, Riccardo L. Rossi, Emilia Malvicini, Antonino Cassotta, Chiara Silacci Fregni, Ludovica Bruno, Federica Sallusto, and Antonio Lanzavecchia. Clonal structure, stability and dynamics of human memory B cells and circulating plasmablasts. *Nature Immunology*, 23(7):1076–1085, July 2022. ISSN 1529-2908, 1529-2916. doi: 10.1038/s41590-022-01230-1. URL https://www.nature.com/articles/s41590-022-01230-1.

[7] Alessandro Pedrioli and Annette Oxenius. Single B cell technologies for monoclonal antibody discovery. *Trends in Immunology*, 42(12):1143–1158, December 2021. ISSN 14714906. doi: 10.1016/j.it.2021.10.008. URL https://linkinghub.elsevier.com/retrieve/pii/S1471490621002131.

[8] Jonathan D. Kaunitz. Development of Monoclonal Antibodies: The Dawn of mAb Rule. *Digestive Diseases and Sciences*, 62(4):831–832, April 2017. ISSN 0163-2116, 1573-2568. doi: 10.1007/s10620-017-4478-1. URL http://link.springer.com/10.1007/s10620-017-4478-1.

[9] Camila H. Coelho, Nathaniel Bloom, Sydney I. Ramirez, Urvi M. Parikh, Amy Heaps, Scott F. Sieg, Alex Greninger, Justin Ritz, Carlee Moser, Joseph J. Eron, Judith S. Currier, Paul Klekotka, David A. Wohl, Eric S. Daar, Jonathan Li, Michael D. Hughes, Kara W. Chew, Davey M. Smith, Shane Crotty, and the Accelerating COVID-19 Therapeutic Interventions and Vaccines–2/A5401 (ACTIV-2/A5401) Study Team. SARS-CoV-2 monoclonal antibody treatment followed by vaccination shifts human memory B cell epitope recognition suggesting antibody feedback, November 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.11.21.567575.

[10] Alissa M. Hummer, Brennan Abanades, and Charlotte M. Deane. Advances in computational structure-based antibody design. *Current Opinion in Structural Biology*, 74:102379, June 2022. ISSN 0959440X. doi: 10.1016/j.sbi.2022.102379. URL https://linkinghub.elsevier.com/retrieve/pii/S0959440X22000586.

[11] CA Jr Janeway, P Travers, and M Walport. The generation of diversity in immunoglobulins. In *Immunobiology: The Immune System in Health and Disease*. New York: Garland Science; 2001, 5th edition edition, 2001. URL https://www.ncbi.nlm.nih.gov/books/NBK27140/.

[12] Yuval Elhanati, Zachary Sethna, Quentin Marcou, Curtis G. Callan, Thierry Mora, and Aleksandra M. Walczak. Inferring processes underlying B-cell repertoire diversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1676):20140243, September 2015. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2014.0243. URL https://royalsocietypublishing.org/doi/10.1098/rstb.2014.0243.

[13] Aleksandr Kovaltsuk, Matthew I. J. Raybould, Wing Ki Wong, Claire Marks, Sebastian Kelm, James Snowden, Johannes Trück, and Charlotte M. Deane. Structural diversity of B-cell receptor repertoires along the B-cell differentiation axis in humans and mice. *PLOS Computational Biology*, 16(2):e1007636, February 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007636. URL https://dx.plos.org/10.1371/journal.pcbi.1007636.

[14] Adam Nathan McShane and Dessislava Malinova. The Ins and Outs of Antigen Uptake in B cells. *Frontiers in Immunology*, 13:892169, April 2022. ISSN 1664-3224. doi: 10.3389/fimmu.2022.892169. URL https://www.frontiersin.org/articles/10.3389/fimmu.2022.892169/full.

[15] Nilushi S. De Silva and Ulf Klein. Dynamics of B cells in germinal centres. *Nature Reviews Immunology*, 15(3):137–148, March 2015. ISSN 1474-1733, 1474-1741. doi: 10.1038/nri3804. URL https://www.nature.com/articles/nri3804.

[16] Alexander D. Gitlin, Ziv Shulman, and Michel C. Nussenzweig. Clonal selection in the germinal centre by regulated proliferation and hypermutation. *Nature*, 509(7502):637–640, May 2014. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature13300. URL https://www.nature.com/articles/nature13300.

[17] Jim Boonyaratanakornkit and Justin J. Taylor. Techniques to Study Antigen-Specific B Cell Responses. *Frontiers in Immunology*, 10:1694, July 2019. ISSN 1664-3224. doi: 10.3389/fimmu.2019.01694. URL https://www.frontiersin.org/article/10.3389/fimmu.2019.01694/full.

[18] Ian Setliff, Andrea R. Shiakolas, Kelsey A. Pilewski, Amyn A. Murji, Rutendo E. Mapengo, Katarzyna Janowska, Simone Richardson, Charissa Oosthuysen, Nagarajan Raju, Larance Ronsard, Masaru Kanekiyo, Juliana S. Qin, Kevin J. Kramer, Allison R. Greenplate, Wyatt J. McDonnell, Barney S. Graham, Mark Connors, Daniel Lingwood, Priyamvada Acharya, Lynn Morris, and Ivelin S. Georgiev. High-Throughput Mapping of B Cell Receptor Sequences to Antigen Specificity. *Cell*, 179(7):1636–1646.e15, December 2019. ISSN 00928674. doi: 10.1016/j.cell.2019.11.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867419312243.

[19] Henry A. Utset, Jenna J. Guthmiller, and Patrick C. Wilson. Bridging the B Cell Gap: Novel Technologies to Study Antigen-Specific Human B Cell Responses. *Vaccines*, 9(7):711, July 2021. ISSN 2076-393X. doi: 10.3390/vaccines9070711. URL https://www.mdpi.com/2076-393X/9/7/711.

[20] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. Van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, August 2021. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abj8754. URL https://www.science.org/doi/10.1126/science.abj8754.

[21] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, April 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2016239118. URL https://pnas.org/doi/full/10.1073/pnas.2016239118.

[22] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2.

[23] Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577 (7792):706–710, January 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-019-1923-7. URL https://www.nature.com/articles/s41586-019-1923-7.

[24] Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners, December 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.12.15.422761.

[25] Ethan C. Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M. Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, 16(12):1315–1322, December 2019. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-019-0598-1. URL https://www.nature.com/articles/s41592-019-0598-1.

[26] Jeliazko R. Jeliazkov, Adnan Sljoka, Daisuke Kuroda, Nobuyuki Tsuchimura, Naoki Katoh, Kouhei Tsumoto, and Jeffrey J. Gray. Repertoire Analysis of Antibody CDR-H3 Loops Suggests Affinity Maturation Does Not Typically Result in Rigidification. *Frontiers in Immunology*, 9:413, March 2018. ISSN 1664-3224. doi: 10.3389/fimmu.2018.00413. URL http://journal.frontiersin.org/article/10.3389/fimmu.2018.00413/full.

[27] Victor Ovchinnikov, Joy E Louveau, John P Barton, Martin Karplus, and Arup K Chakraborty. Role of framework mutations and antibody flexibility in the evolution of broadly neutralizing antibodies. *eLife*, 7: e33038, February 2018. ISSN 2050-084X. doi: 10.7554/eLife.33038. URL https://elifesciences.org/articles/33038.

[28] Chu'nan Liu, Lilian M. Denzler, Oliver E.C. Hood, and Andrew C.R. Martin. Do antibody CDR loops change conformation upon binding? *mAbs*, 16(1):2322533, December 2024. ISSN 1942-0862, 1942-0870. doi: 10.1080/19420862.2024.2322533. URL https://www.tandfonline.com/doi/full/10.1080/19420862.2024.2322533.

[29] Jan Wilamowski, Zichang Xu, Hendra S Ismanto, Songling Li, Shunsuke Teraguchi, Mara Anais Llamas-Covarrubias, Xiuyuan Lu, Sho Yamasaki, and Daron M Standley. InterClone: Store, Search and Cluster Adaptive Immune Receptor Repertoires, August 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.07.31.501809.

[30] Matthew I J Raybould, Aleksandr Kovaltsuk, Claire Marks, and Charlotte M Deane. CoV-AbDab: the coronavirus antibody database. *Bioinformatics*, 37(5):734–735, May 2021. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btaa739. URL https://academic.oup.com/bioinformatics/article/37/5/734/5893556.

[31] Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual Language Model for Protein Sequence and Structure, July 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.07.23.550085.

[32] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan Dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/10.1126/science.ade2574.

[33] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, 2021. URL https://arxiv.org/abs/2106.09685.

[34] Dima Kozakov, David R Hall, Bing Xia, Kathryn A Porter, Dzmitry Padhorny, Christine Yueh, Dmitri Beglov, and Sandor Vajda. The ClusPro web server for protein–protein docking. *Nature Protocols*, 12(2):255–278, February 2017. ISSN 1754-2189, 1750-2799. doi: 10.1038/nprot.2016.169. URL https://www.nature.com/articles/nprot.2016.169.

[35] Anna Vangone and Alexandre Bonvin. PRODIGY: A Contact-based Predictor of Binding Affinity in Protein-protein Complexes. *BIO-PROTOCOL*, 7(3), 2017. ISSN 2331-8325. doi: 10.21769/BioProtoc.2124. URL https://bio-protocol.org/e2124.

[36] Li C. Xue, João Pglm Rodrigues, Panagiotis L. Kastritis, Alexandre Mjj Bonvin, and Anna Vangone. PRODIGY: a web server for predicting the binding affinity of protein–protein complexes. *Bioinformatics*, 32(23):3676–3678, December 2016. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btw514. URL https://academic.oup.com/bioinformatics/article/32/23/3676/2525629.

[37] Andrei Neamtu, Francesca Mocci, Aatto Laaksonen, and Fernando L. Barroso Da Silva. Towards an optimal monoclonal antibody with higher binding affinity to the receptor-binding domain of SARS-CoV-2 spike proteins from different variants. *Colloids and Surfaces B: Biointerfaces*, 221:112986, January 2023. ISSN 09277765. doi: 10.1016/j.colsurfb.2022.112986. URL https://linkinghub.elsevier.com/retrieve/pii/S0927776522006701.

[38] Nicholas C. Wu, Meng Yuan, Sandhya Bangaru, Deli Huang, Xueyong Zhu, Chang-Chun D. Lee, Hannah L. Turner, Linghang Peng, Linlin Yang, David Nemazee, Andrew B. Ward, and Ian A. Wilson. A natural mutation between SARS-CoV-2 and SARS-CoV determines neutralization by a cross-reactive antibody, September 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.09.21.305441.

[39] Andre Azevedo Reis Teixeira, Mikael Lund, and Fernando Luís Barroso Da Silva. Fast Proton Titration Scheme for Multiscale Modeling of Protein Solutions. *Journal of Chemical Theory and Computation*, 6(10):3259–3266, October 2010. ISSN 1549-9618, 1549-9626. doi: 10.1021/ct1003093. URL https://pubs.acs.org/doi/10.1021/ct1003093.

[40] Brennan Abanades, Tobias H Olsen, Matthew I J Raybould, Broncio Aguilar-Sanjuan, Wing Ki Wong, Guy Georges, Alexander Bujotzek, and Charlotte M Deane. The Patent and Literature Antibody Database (PLAbDab): an evolving reference set of functionally diverse, literature-annotated antibody sequences and structures. *Nucleic Acids Research*, 52(D1):D545–D551, January 2024. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkad1056. URL https://academic.oup.com/nar/article/52/D1/D545/7424429.

[41] Sidhartha Chaudhury, Sergey Lyskov, and Jeffrey J. Gray. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, 26(5):689–691, March 2010. ISSN 1367-4811, 1367-4803. doi: 10.1093/bioinformatics/btq007. URL https://academic.oup.com/bioinformatics/article/26/5/689/212442.

[42] Sergios Theodoridis. Monte Carlo Methods. In *Machine Learning*, pages 731–769. Elsevier, 2020. ISBN 978-0-12-818803-3. doi: 10.1016/B978-0-12-818803-3.00026-X. URL https://linkinghub.elsevier.com/retrieve/pii/B978012818803300026X.

Appendix / supplemental material

Table 2: SARS-CoV-2 antibody candidates: Thermodynamic Properties

ID	ΔG (kcal mol^{-1})	Kd (M) at °C	NIS charged
Ab1	-14.5	2.30E-11	17.97
Ab53	-12.5	6.30E-10	17.23
Ab77	-9.6	9.00E-08	20.41
Ab117	-10.9	1.10E-08	20.27
Ab131	-13.3	1.80E-10	18.02
Ab157	-14.5	2.20E-11	20.57
Ab172	-11.8	2.30E-09	18.60
Ab192	-11.4	4.10E-09	20.14
Ab200	-13.7	9.30E-11	18.75
Ab242	-11.1	6.70E-09	18.71

Table 3: SARS-CoV-2 antibody candidates: Interaction Counts

ID	ICs charged-charged	ICs charged-polar	ICs polar-polar
Ab1	6	15	1
Ab53	2	13	9
Ab77	2	4	8
Ab117	7	11	9
Ab131	5	10	2
Ab157	5	16	5
Ab172	4	4	3
Ab192	4	5	6
Ab200	11	10	5
Ab242	6	10	9

Table 4: SARS-CoV-2 antibody candidates: Additional Interaction Features

ID	ICs charged-apolar	ICs polar-apolar	NIS apolar
Ab1	8	31	38.31
Ab53	18	24	35.81
Ab77	16	16	39.46
Ab117	19	17	37.16
Ab131	16	23	37.81
Ab157	27	27	36.88
Ab172	16	18	37.80
Ab192	28	15	38.19
Ab200	26	20	36.81
Ab242	16	19	37.07

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes] .

   Justification: The abstract contains information of the scope of the project and future directions.

   Guidelines:

   - The abstract clearly identifies the challenge in antibody and vaccine development—specifically, the complexity in predicting synthetic antibodies that can bind and neutralize novel antigens.
   - The abstract emphasizes that their method does not require preselecting B cells based on antigen binding, which is a significant technological leap. This indicates the potential to streamline the discovery process, making it more efficient.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are address in the discussion.

   Guidelines:

   - Discussion mentions the limitations of our approach, limitations such as: biases in training data and the complexities of real-world biological interactions—is crucial for effectively applying these methodologies in therapeutic development.
   - We mentioned the need to use other testing datasets for validation.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We have included information on the LLM implementation and training. The details on the LLM architecture are described in methods.

   Guidelines:

   - We have included details of LLM implementation, architecture and training. Parameters for LLM training such as learning rate, epoch, batch size and final accuracy obtained by the models were described in results section of the paper.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes] .

   Justification: We have included the information to reproduce our aproach in the methods section.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Github provided in supplemental information.

   Guidelines:

   - We have included codes for training of LLMs and the identification of BCRs using InterClone.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setting is presented in the methods description and results section

   Guidelines:

   - The full details are provided with the code and supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Appropriate information is included in results section.

   Guidelines:

   - We have included plots and appropriate information on supplemental material. Our paper is still in progress and we are working on including other testing datasets.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have included this information in supplemental data

   Guidelines:

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted conform with the NeurIPS Code of Ethics.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: Research is primarily scientific and technical in nature, focusing on methodological advancements on therapeutic discovery than immediate societal applications.

    Guidelines:

    -

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: We have not included this.

    Guidelines:

    -

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have included citations and references corresponding to models and datasets used in the paper.

Guidelines:

- We cited the original papers that produced the code packages and datasets.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification:

Guidelines:

-

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: does not involved human subjects.

Guidelines:

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: paper does not involve crowdsourcing nor research with human subjects.

Guidelines: