Doctoral Thesis (Abridged)

博士論文（要約）

Structure-based Computational Design and Pose Prediction of Nanobodies

（構造ベースインシリコ VHH 抗体設計と結合ポーズの予測）

譚 俊禮

Tam Chun Lai

Doctoral Thesis (Abridged)

博士論文（要約）

Structure-based Computational Design and Pose Prediction of Nanobodies

（構造ベースインシリコVHH抗体設計と結合ポーズの予測）

譚 俊禮

Tam Chun Lai

Graduate School of Frontier Sciences

The University of Tokyo

Content

DEDICATION

To my mother, Tracy Wong.
this dissertation would not have been possible without your support.

To my wife, Jenny Liu.
for your love and patience during my studies.

And I dedicate this thesis to God for your wisdom in designing proteins.
Your designs were hard to study, but somehow interesting.

ACKNOWLEDGEMENTS

I would like to say thank you to my lab members in the Laboratory of Structural Bioinformatics in RIKEN (Yokohama), Japan. I appreciate Dr. Aditya K. Padhi and Dr. Ashutosh Kumar, who always generously supported me technically during my studies. I would like to say thank you to Dr. Mutsuko Niino and Dr. Takehisa Matsumoto at the Laboratory for Protein Functional and Structural Biology of Dr. Mikako Shirouzu in RIKEN, for their help in the experimental verification of the nanobody designs. Last but not least, I am very grateful to my supervisor, Dr. Kam Zhang, for accepting me as his student, his teaching and his understanding of my difficulty studying abroad, especially during the COVID19 pandemic.

ABSTRACT

Substantial growth in antibody drug development in the pharmaceutical industry is foreseeable due to the advantage over small molecule drugs in terms of, for example, specificity, biodegradability, and non-toxic metabolites from degradation. Out of the known antibody fragments in different molecular sizes, recently the nanobody, which is a heavy-chain only antibody from camelid species, has drawn considerable attention in antibody drug research due to its superior properties over the larger antibody fragments, such as high thermostability and versatility in choice of expression systems for production.

Until now, several experimental methods in antibody drug screening, for example, the classical animal immunization techniques and the more recent directed evolution methods (e.g. molecular displays), have been some of the standard practices in developing antibody drugs with good affinity and specificity to the target of interest. However, there are drawbacks to these experimental methods. For instance, one major disadvantage is the inability to rationally design antibodies to target a specific epitope of interest, where such control is often desirable because the functional alteration to the pharmaceutical targets is usually epitope-dependent. Moreover, the precise control of epitopes by rational antibody design minimizes off-target toxicity by avoiding binding to epitopes that inhibit other normal functions of the target.

In this study, we have explored the applications of computational nanobody design, which is an emerging technique in rational antibody design, on two pharmaceutically important targets, one targeting ELMO1-RhoG interaction, which is a key protein-protein interaction in signaling cancer cell migration and another targeting S2 of SARS-CoV-2,

in an attempt to develop a broad-spectrum antibody drug effective to SARS-CoV-2 mutant strains and SARS-related CoVs.

In the computational nanobody design targeting ELMO1-RhoG interaction, we have applied the dock-and-design approach by repurposing known nanobody structures from the PDB to bind ELMO1 on its interface with RhoG, which could theoretically quench the downstream signaling of cancer cell migration through the ELMO1/DOCK180 pathway, which normally induces actin polymerization and cell membrane protrusion for cell movements. We improved the selection of initial nanobody poses by applying an *in cerebro* guided optimization of antibody mode in PatchDock that used two positions on the CDR loops as distance constraints in nanobody-antigen docking, which lead us to initial poses with improved visual resemblance to known nanobody-antigen poses. We have adopted a new approach in pose selection we termed "pose-selection-by-design," which selected poses that generated binding energy funnels with good resemblance to the deep, funnel-shaped binding energy landscape commonly observed in protein-protein interactions. From our first batch of 20 designs tested for binding to ELMO1 by SPR binding assay, we have obtained one potential hit, nano-79, which showed weak binding to ELMO1. Based on nano-79 as an initial hit, we performed a second-round design to explore additional sequence variations that potentially improve binding affinity to our target. We have successfully obtained a set of designs which showed improved binding overall, with the best binder exhibiting a dissociation constant of 2uM to ELMO1.

During the current COVID-19 pandemic, due to the frequent emergence of SARS-CoV-2 mutant strains worldwide, there is a need to develop therapeutics that are tolerant to potential mutation escape of the SARS-CoV-2 variants. Currently, the majority of the spike-targeting antibodies developed bind at the RBD or its surrounding residues on S1. However, in general, RBD on S1 represents a relatively variable epitope compared with the S2 ectodomain. An antibody drug that targets a conserved epitope on S2 that is functionally important to the cell fusion and entry mechanism of SARS-CoV-2 could deliver a promising antibody drug that possesses a broad-spectrum neutralizing effect to the circulating and the to-be emerged mutant strains of SARS-CoV-2. We focused on one conserved structural epitope on S2 of SARS-CoV-2, which contains the proteolytic cleavage site S2' and is proximal to HR1 in its pre-fusion state, implying the functional importance of this epitope to the dissociation of S1 from S2, which is essential to the S-mediated host membrane fusion of SARS-CoV-2. We designed 21 nanobody structures that potentially bind to the epitope through an overall similar design approach as in the ELMO1-RhoG nanobody design. Preliminary result from SPR binding assay showed our designs did not bind SARS-CoV-2 S with dissociation constant less than 5uM, which needs further examination to improve their binding affinity.

Computational antibody design is still a relatively new technique in antibody drug development. There is a need for further methodological optimization to increase the hit rate of generating a binder with a detectable affinity for further affinity maturation. In structure-based computational antibody design, one of the difficulties lies in the pose selection from a large number of alternative poses generated by antibody-antigen docking, which directly affects the success of subsequent design simulations. Conceptually, designing native-like poses should have a better chance of developing a binder than designing poses far from the native. Followed by the two studies of computation nanobody design, we have explored the application of machine learning to improve the pose selection of nanobodies. With the calculation of features that consisted of a contact profile (e.g. CDR loop contacts) and an energy profile calculated by InterfaceAnalyzer from Rosetta and AnalyseComplex from FoldX, we have trained a binary classifier with the implementation of a gradient-boosted decision tree model, XGBoost, which can distinguish native-like from non-native-like poses with a given nanobody-antigen complex structure. To benchmark the performance of our binary classifier, we are currently comparing the performance of our model to ClusPro, the current state-of-the-art protein-protein docking algorithm, and DOVE, a competing method that distinguishes native and non-native protein-protein complex structures. Our model successfully ranked native-like nanobody poses with a significantly higher ranking than ClusPro, demonstrating the potential application of our nanobody pose prediction model to improve accuracy in native pose prediction of nanobody from protein-protein docking algorithms.

INTRODUCTION

Nanobody is a single-domain antibody truncated from the **v**ariable domain of the **h**eavy chain of the **h**eavy-chain-only antibody found in camelids, or VHH as its alternative name. With the first nanobody drug approved by the Food and Drug Administration of the United States in 2019, more nanobody drugs are now under clinical trials to target diverse therapeutic targets from viral infections to autoimmune disorders and carcinomas (Jovčevska and Muyldermans 2020; Morrison 2019; Muyldermans 2020).

The growing interest in nanobody could be explained by its several advantageous properties compared to conventional antibodies, for example, good solubility, unusually high thermal stability, readiness of recombinant production by bacterial expression and therefore the ease of design (Jovčevska and Muyldermans 2020; Hassanzadeh-Ghassabeh et al. 2013; Muyldermans 2020; Chanier and Chames 2019; Olson et al. 2019). Despite a small size of approximately 15 kDa, nanobody has non-compromised specificity and affinity compared with the full-length antibody, which are mediated by the three complementarity determining region (CDR) loops (H1, H2 and H3) that are anchored on its single-domain framework folded as beta-sandwich (Mitchell and Colwell 2018b, [a] 2018; Zavrtanik et al. 2018). Correlated to the sequence and structural features of the nanobody, nanobody possesses four framework mutations that improves its solubility (Mitchell and Colwell 2018b) and two disulfide bonds, one stabilizing intersheet and another intraloop of CDR3, which contributed to its thermal stability (Kunz et al. 2018).

To untap these desirable properties of the nanobody for the development of new antibody therapeutics, the ability to design new nanobody sequences is key. In a broad definition, antibody design is the derivation of an antibody sequence that binds to the target antigen. Until now, we have established several experimental methods for antibody design, with two methods of notable importance in the field. Firstly, animal immunization with purified antigens is a common routine for antibody discovery. When combined with next-generation sequencing, antibody sequences can be retrieved from the memory B-cell. Secondly, the application of directed evolution in the format of molecular display, which utilizes random mutagenesis and iterative panning to enrich a set of binding antibody sequences, is another popular routine in experimental antibody discovery (Prabakaran, Rao, and Wendt 2021; Laustsen et al. 2021). With the robustness of generating binding antibodies using molecular display methods in short turnover time and the availability of stable protocols of animal immunization, the two experimental methods are widely adopted in antibody discovery currently (Leenaars and Hendriksen 2005; Lee et al. 2007).

Nevertheless, there are disadvantages to these experimental methods. As a result of a common property of the two experimental methods which utilizes random mutations for affinity maturation, both methods share a major drawback: the inability of rationally designing antibody interaction targeted to a specific epitope of interest. Indeed, the ability to design interaction to a specific epitope surface is of high importance in antibody drug development because functional modulation of target proteins is usually epitope-dependent. For example, in the development of an antibody as a protein-protein interaction drug, it is necessary to control the binding orientation of the antibody because the binding epitope or the spatial overlap with either of the binding partners determines the efficacy of inhibition to the protein-protein interaction. In more demanding situations, it is necessary to minimize clashes of the antibody with other domains of the antigen and prevent cross-reactivity to isoforms of the target. To achieve such specific, epitope-oriented targeting with the control of binding orientation of antibody using the experimental methods in antibody discovery, one can expect the necessity of additional assays to multiple design candidates and structural modeling of their complex structures with the antigens to understand their structural-basis for further selections. Such massive modeling and screening is time-consuming but there is no guarantee in obtaining the design that binds to the desired epitope and binding orientation.

Representing a rationale-based approach, computational antibody design is an emerging antibody discovery method that is complementary to the existing experimental approaches in terms of its ability of designing novel antibody-antigen interaction bottom-up. As a relatively new method but offers precise control to epitope-targeting which is lacking in both animal immunization and molecular display techniques, we are now accumulating early examples and algorithms that succeed in this *in silico* antibody design approach.

Previously, there were several methodological concepts in computational antibody design that were proven successful, which mainly included hotspot design, CDR loop grafting, VDJ recombination mimic and the “dock-and-design” approach, which was the design method used in the two nanobody design examples of current study. A notable example of antibody design by hotspot design is the design of scaffold proteins to target the stem region of influenza of hemagglutinin (Fleishman, Whitehead, et al. 2011). Hotspot design, in principle, designs energetically dense interactions *de novo* or by learning from known protein-protein interaction to achieve effective increment to the binding affinity, followed by the redesign of residues surrounding the hotspot to accommodate the hotspot residues (Fleishman, Corn, et al. 2011). In the design method of influenza hemagglutinin antibody, *de novo* hotspot design was used. At the beginning, disembodied residue side chains were docked to the target to generate a *de novo* hotspot library. For a particular docked pose of a scaffold which showed shape complementarity to the target epitope, trials of hotspot grafting by looking up the hotspot library were iterated to obtain grafted side chains that were then filtered by hotspot energy and specificity to the target. The scaffold was further optimized by RosettaDesign and the final designs were selected by binding energy and shape complementarity. Two designs of binding affinity in nanomolar range were obtained after affinity maturation with directed evolution by yeast display. The crystal structure of one of their designs was almost identical according to their initial design.

In another example of antibody design by hotspot design (X. Liu et al. 2017), the hotspot interactions from a known complex structure of Keap1-Nrf2 were borrowed, to design Keap1-targeting variable fragment (Fv). With hotspot grafting, an initial hit with micromolar affinity was obtained. The initial hit was subjected to CDR loop grafting to diversitize the H3 loop, followed by a final sequence design of the CDR loops and scored by the Rosetta binding energy score, SASA and a shape complementarity score. Four binders were optimized down to nanomolar binding affinity to Keap1 while one of the crystal structures of the designs showed good agreement to the designed pose and interacting side chains on the CDR loops.

As a strategy used in design optimization of this example of Keap1 antibody, CDR loop grafting is another antibody design strategy that was proven successful. Attributed to the modular nature of an antibody that consisted of interchangeable CDR loops and the framework, as its name suggests, CDR loop grafting is the grating of CDR loops to alternative antibody frameworks to yield diversity of designs. Apart from the application in antibody repurposing, CDR loop crafting was also a common technique in the humanization of antibodies such as the grafting of CDR loops from murine antibodies to human antibody frameworks (Lo 2004). Compared with hotspot grafting that concerns residue side chains as the unit of design, CDR loop can be regarded as the design at the level of secondary structure involved in interaction, and therefore both techniques can be used complementarily to diversify CDR loop sequence because sequence design usually retain a proportion of residues unmutated. With a diverse number of structurally characterized CDR loops (Adolf-Bryfogle et al. 2015), CDR loop grafting enables the sampling of diverse CDR sequences and therefore the search space is more likely to bracket a potential binder, as shown in the successful application in the Keap1 antibody design.

In the RosettaAntibodyDesign (RAbD) utility of Rosetta Suite, CDR loop grafting was incorporated in their design algorithm that allowed the control of grafting of CDR loops of the light chains or the heavy chains from a library of CDR loops (Adolf-Bryfogle et al. 2018). In the main design workflow of the RAbD algorithm, with a given antibody-antigen complex structure as the input, CDRs are grafted from the populated structural clusters recorded in PyIgClassify to diversify the CDRs and framework combinations. After sequence design of the CDRs, designs are docked against the epitope. After a screening of the total energy score or the interface energy score according to the Metropolis Monte Carlo criterion, which either accept or reject the design, the algorithm returns to the initial CDR grafting step and iteratively runs for N cycles. Finally, top designs in terms of the total energy score or the interface energy score can be selected. In the retrospective validation of the RAbD method, the author demonstrated the successful recovery of 72% and 73% of contacting and non-contacting CDR residues to the interface from 60 antibody-antigen complexes. Further experimental validation showed RAbD successfully enhanced the affinity of a known HIV-neutralizing antibody by a maximum of 50 folds and the affinity of a hyaluronidase-targeting antibody by 12 folds.

Alternative to RAbD, OptMAVEn (Chowdhury, Allan, and Maranas 2018) and AbDesign (Lapidoth et al. 2015) were two directly competing antibody design algorithms where their methodological differences were compared in detail (Adolf-Bryfogle et al. 2018). For both OptMAVEn and AbDesign, they utilized a design strategy that mimicked natural V(D)J recombination to search for new constructs of antibody. In contrast to RAbD which starts with an antibody-antigen complex structure as the input, both versions of OptMAVEn (T. Li, Pantazes, and Maranas 2014; Chowdhury, Allan, and Maranas 2018) starts by constructing a new antibody through combining CDR3 with the variable (V) and joining (J) regions of heavy and light chains (lambda and kappa chains) from MAPs, a database of antibody parts chains (Pantazes and Maranas 2013), to generate new antibody constructs. This approach formed a major difference in terms of the basic unit of combination compared with RAbD, which takes individual CDR loops and the frameworks for combination. After the assembly, OptMAVEn then generates antibody poses by translating and rotating the assembled construct on the epitope designated for design. Antibody parts in poses with minimal clashes and a minimized interface energy were clustered according to the generated antibody poses, to look for design sequences that show the lowest interface energy in each cluster. Finally, sequence optimization and binding affinity estimation by MD simulation can be performed to validate the final designs. Computational validation by sequence recovery on Zika envelope protein- and lysozyme-targeted variable antibody fragments showed OptMAVEn-2.0 was able to design good affinity binders predicted by MD simulation and recovered a general of 45-65% native residue identity.

Similarly, AbDesign (Lapidoth et al. 2015) applied a similar approach through mimicking V(D)J recombination to assemble new antibody designs. The main workflow of the AbDesign algorithm starts by precomputation from natural antibodies for position-specific site matrices (PSSMs) and a torsion database of the V region, which consists of CDR1 and CDR2, and the CDR3. With the sequence identity and backbone angle respectively constrained by PSSMs and the torsion database, the V region and CDR3 were grafted to a chosen scaffold. Conformational representative of every cluster of assembled construct is docked to the target and sequence design is performed to optimize binding energy and stability of the construct. Finally, designs are further filtered by shape complementarity, packing quality and buried surface area. AbDesign was computationally validated by retrospectively comparing nine antibody designs to their native complex structures with the targets. AbDesign was able to reproduce side chains and backbone conformations observed in the native structures. Although in the examples, CDR conformations AbDesign sampled largely followed the canonical conformations, in the lysozyme-targeted design, it was able to sample a more diverse backbone conformation than the native structure. In another study, AbDesign was experimentally validated through the design of three scFVs that showed weak binding to insulin and the acyl-carrier protein 2 of *Mycobacterium tuberculosis (Baran et al. 2017)*. Affinity maturation of the designs by directed evolution through yeast display enhanced the binding affinity to the lowest K_d = 30nM.

With the above examples, computational antibody design demonstrated its unique ability in precisely designing antibody-antigen interactions. Yet, compared to the experimental methods in antibody discovery, it has relatively unpredictable hit rate potentially due to several weak points in the current design methodologies. For most of the previous examples of computational antibody design, they were structure-based methods where the 3D coordinates of the antibody-antigen complex structure were used for the design calculations to derive the design sequence.

Structure-based simulation has several aspects that contributed to the inaccuracy of structure-based computational antibody design. Firstly, there is the inaccuracy in modeling the structure, the flexibility and structural changes upon binding to the epitope of the CDR loops, which forms a majority of the paratope where the design is performed upon. Among the CDR loops, it is well-known that CDR3 is especially difficult to model due to more diverse sequences and structural conformations compared with CDR1 and CDR2 (Weitzner and Gray 2017; Adolf-Bryfogle et al. 2015; Marks and Deane 2017; Weitzner, Dunbrack, and Gray 2015; Shirai et al. 1998; Nishigami, Kamiya, and Nakamura 2016; Fernández-Quintero et al. 2019; Kumagai and Tsumoto 2002).

Secondly, as side chains were designed on the CDR loops, there is the need to predict the change in interface energy and the changes propagated to backbone angles (Davis et al. 2006), where inaccuracy exist due to imperfect energy functions (Xiong et al. 2014; Pokala and Handel 2001; Pillardy et al. 2001; Pokala and Handel 2005; Z. Li et al. 2013; Huang, Boyken, and Baker 2016; Pan and Kortemme 2021) and CDR loop modeling respectively. Moreover, the prediction of the energy and structural changes becomes more difficult when the number of design mutations increases (Andersson et al. 2016; Dehghanpoor et al. 2018; N. Zhang et al. 2020).

Thirdly, the decision making of picking initial poses for design is difficult. Previously, there were multiple prediction models developed for antibody pose prediction. In a broader sense, it is a sub-problem under the bigger problem of protein-protein interaction prediction that was addressed by protein-protein docking algorithms, with some of the docking algorithms dedicated an antibody-antigen mode to improve antibody pose prediction (van Zundert et al. 2016; Kozakov et al. 2017; Schneidman-Duhovny et al. 2005; Garzon et al. 2009; Sircar and Gray 2010). Representing a closely related class, a majority of earlier prediction methods solely developed for antibody-antigen interaction focused on separate predictions of the epitope and the paratope, instead of determining the 3D coordinates of the antibody-antigen complex (Norman et al. 2020).

More recently, a number of machine-learning prediction models for protein-protein interaction emerged, in which a majority of them utilized sequence, structure and evolutionary information as features to aid the prediction of nativeness of protein-protein interactions (Esmaielbeiki et al. 2016). Amongst these prediction methods, previous examples of computational antibody design by the “dock-and-design” approach mainly utilized protein-protein docking algorithms to suggest initial poses for design (Procko et al. 2013; Fleishman, Whitehead, et al. 2011; Fleishman, Corn, et al. 2011; Baran et al. 2017; Karanicolas et al. 2011; Strauch, Fleishman, and Baker 2014; Choi et al. 2014).

Due to the difference in assumption between ordinary protein-protein docking and "dock-and-design" for *de novo* antibody-antigen interface design, it is especially difficult to select initial poses in the latter situation because we usually assume a correct solution exists in usual protein-protein docking. However, it is not in the case for antibody design because the interaction is to-be designed. Despite the presumably non-binding nature between the antigen and the antibody before design, we ask protein-protein docking algorithms to suggest initial poses that harbors native-like qualities but in most situations the poses are by and large "non-native". Therefore, as a conceptual intermediates between native and non-native, the decision making in selecting these initial poses becomes elusive because the initial poses should be not as distinguishable between typical native and non-native poses. Altogether, these technical weak points have contributed to the relatively unpredictable success rate of computational antibody design.

Here, as the grand objective of this study, we aim at enhancing our experience in computational antibody design by exercising two examples of nanobody design. We designed nanobody to target ELMO1-RhoG, a key protein-protein interaction in the signaling of cell migration in cancer cells, and we designed nanobody to target the S2 ectodomain of SARS-CoV-2 spike protein. In both designs, we have revisited the "dock-and-design" approach with this particular subject of nanobody and suggested two technical optimizations in the design workflow: one on the generation of initial nanobody pose with the application of *in cerebro* learning from known nanobody-antigen complexes, and another on nanobody pose selection by energy landscape through design.

In contrast to previous studies which aid their computational design with experimental approach, solely with our computational design workflow, we have successfully obtained a set of ELMO1-binding nanobody with binding affinity in micromolar range, with the best binder at a dissociation constant of 2uM. Additionally, to further improve our ability in selecting initial poses that show native-like qualities for design, we explored the use of machine learning to guide the selection of native-like nanobody poses. Benchmarking of our nanobody pose prediction model showed the significantly better performance of our model compared with the current state-of-the-art protein-protein docking algorithm and a classifier of protein-protein complex nativeness in a similar class.

An improved methodology on computational antibody design can be generally applied to the development of antibody drugs targeting various protein antigens that are involved in different diseases. By pushing the success rate and streamlining such a computational approach in antibody discovery, time could be shortened to obtain a potent antibody drug which has a major significance during outbreaks of unknown infectious diseases. Moreover, unlike experimental methods in antibody discovery that required a real sample of antigen for experiments, computational antibody design can still be used to develop antibody drugs when the sample is difficult-to-obtain. Along with the increased application of next-generation sequencing on pathogen identification and the improved accuracies in protein structure prediction, the wider application of the computational antibody design is anticipated in future antibody discovery.

Here, with the two design examples and the development of the pose prediction model, this study has added to the early examples and suggested technical improvements mainly on initial pose selection for *de novo* antibody-antigen interface design. This study accelerates our progress of technical improvement in rationale-based antibody discovery by computational design.

Chapter 1 and 2 (pp. 22 to 59 and 103 to 133) of my doctoral thesis cannot be made public on the Internet for 5 years from the date of doctoral degree conferral because that parts are scheduled to be published within 5 years.

CHAPTER 3   NANOBODY POSE PREDICTION

INTRODUCTION

In spite of the rising number of successful examples of computational antibody design, compared with the experimental approaches such as animal immunization and directed evolution, the hit rate of computational antibody design remained relatively unpredictable. The relatively low hit rate of computational antibody design can be attributed to reasons such as the inaccuracies in CDR loops modeling (North, Lehmann, and Dunbrack 2011; Nishigami, Kamiya, and Nakamura 2016; Weitzner et al. 2014) and in the prediction of change in binding energy upon mutations at the interface (Gromiha, Yugandhar, and Jemimah 2017; Seeliger 2013; Z. Li et al. 2013). To accurately determine the binding energy change and the conformation of mutated CDR loops, there is a pre-existing assumption: the accuracy of the pose of the antibody, because it is the structural context for the downstream modeling of the interface and therefore directly influences the accuracy of binding affinity prediction.

In contrast to antibody pose generation for computational design, a native solution of pose exists for known antibody-antigen complexes. Compared with the non-native poses, the native poses of known antibody-antigen complexes have distinguishing characteristics such as the funnel-shaped binding energy landscape (Shen et al. 2008; Schueler-Furman et al. 2005), preference in residue propensity of the epitope (Ramaraj et al. 2012) and sometimes the paratope-epitope shape complementarity (Kuroda and Gray 2016; Yan and Huang 2019), which were useful features for antibody pose prediction or epitope prediction (London and Schueler-Furman 2008b; Soga et al. 2010; Dunbar et al. 2016). In *de novo* interface design of antibodies by “dock-and-design”, we can assume, with a given docked pose before design, there is weak if not merely no binding between the antibody and the target because the antibody is docked to an unrelated antigen. This non-binding presumption is also supported by the fact the CDR generally forms specific interaction to the epitope, which should be absent between an antibody with an unrelated antigen. Due to the weak if not no binding in most of the docked poses, the typical interface characteristics of antibody-antigen complexes is therefore less obvious or absent in docking used for design. This highlights the fundamental difference between ordinary pose prediction of experimentally verified antibody-antigen complex and the pose generation for computational antibody design. Due to the less obvious or absence of distinguishing features among the many docked poses, docking pose selection for design is arguably more difficult than ordinary pose prediction.

Nevertheless, although it is difficult to pick a pose for design, it is not impossible to generate a binder that agrees to the initial pose after design (Procko et al. 2013; Fleishman, Whitehead, et al. 2011; Baran et al. 2017; Karanicolas et al. 2011), meaning existing antibody-antigen docking methods have the ability to suggest “native-like” poses that are verifiable after design. Moreover, in these successful “dock-and-design” examples, the same design workflow that generated a majority of non-binders and a minority of binders implied the contribution of the initial docking to the success of design. In such a sense, we describe an initial pose that leads to a binding antibody through design as “native-like” because to a certain degree it harbors the quality that resembles a native pose but is still optimizable by design that strengthens the nativeness of the pose.

Here, to improve our ability of selecting native-like antibody poses for design, using nanobody as the subject, we have developed a gradient-boosted decision tree model that can distinguish native-like from non-native-like nanobody poses. Previously, there were many antibody pose prediction methods developed (Norman et al. 2020). In a broader scope, antibody pose prediction falls under the prediction of protein-protein interaction, which is the main question addressed by protein-protein docking algorithms. A number of the docking algorithms integrated functionalities dedicated to improving antibody-antigen docking, such as the usage of CDR loops as distant constraints (van Zundert et al. 2016; Kozakov et al. 2017; Schneidman-Duhovny et al. 2005; Garzon et al. 2009; Sircar and Gray 2010). From the latest 7th edition of Critical Assessment of Predicted Interactions (CAPRI), ClusPro is a top performing protein-protein docking algorithm which showed notable performance in the predictor group and the server group (Lensink et al. 2020).

Apart from docking, a majority of alternative approaches to predict antibody-antigen interaction belong to paratope and epitope prediction methods (Norman et al. 2020). Most of these paratope and epitope prediction methods do not suggest or evaluate the 3D coordinates of the predicted antibody-antigen complex, which is a major drawback compared to protein-protein docking methods because structural details of the interface are crucial to antibody design. Recently, with the increased usage of machine learning to study protein-protein interaction, for example, deep 3D convolution neural networks (Schneider et al. 2021; Wang et al. 2020), graph convolutional neural networks (Yue Cao and Shen 2020), tensor field neural network (Eismann et al. 2021), graph kernel (Geng et al. 2020) and logistic regression classifier (Tanemura, Pei, and Merz 2020), were developed to evaluate nativeness of binding pose using 3D coordinates of protein-protein complexes. These methods, which score a protein-protein complex with its 3D coordinates, were previously classified as partner specific interface predictor (Esmaielbeiki et al. 2016).

We benchmarked our decision tree model with the state-of-the-art protein-protein docking algorithm ClusPro and a 3D-CNN model that belongs to the specific class of predictor for protein-protein complex. To our best knowledge, our decision tree model is the first model in class that is dedicated to nanobody-antigen pose prediction, which is complementary to existing methods for antibody-antigen complex pose prediction due to the observable differences in binding modes between nanobody and conventional antibodies (Mitchell and Colwell 2018a, [b] 2018; Zavrtanik et al. 2018). The application of our decision tree model for initial pose selection aids the selection of “native-like” poses for nanobody design in the future.

MATERIALS AND METHODS

Data Collection and Preprocessing

Nanobody-antigen complex structures in a total of 371 unique PDB IDs were retrieved by searching VHH antibody that were with protein antigens, without constant region and a resolution cutoff of 3.5 Å from the SAbDab antibody structure database (Dunbar et al. 2014) in September, 2020 (Figure 32). For oligomeric structures, a biological assembly was picked so that each nanobody-antigen complex structure contains the interaction between one nanobody chain and one antigen chain. The nanobody-antigen complex structures were renumbered with PyIgClassify database (Adolf-Bryfogle et al. 2015) to standardize the numbering of CDR loops. The complex structures with any of the three CDR loops not recorded in the PyIgClassify database were removed to ensure the presence and the structural quality of the CDR loops. Pairwise structural alignment of all the collected antigen structures was performed by the superpose utility in CCP4 (Krissinel and Henrick 2004) to assess their structural redundancy. PDB IDs with similar antigen structures that have structural alignment quality score (Q) higher than 0.95 were removed to minimize information leakage during testing. After these preprocessing steps, a final total of 180 unique PDB IDs were retained.

Randomization of Orientation, Backbone and Side Chains

To prepare for docking, nanobody chains and antigen chains of the 180 complex structures were separated into independent PDB files. Because the nanobody and antigen coordinates separated in this way contained highly matching structural features in terms of orientation, backbone angles and side chain rotamers at the interface, to simulate the real-life situation of docking nanobody and antigen models that were modelled independently, these structural features were randomized. The orientation of all the nanobody and antigen chains were randomized by applying a random translation and rotation in the six axis of freedom. The lowest total energy structure was picked by sampling 1000 structures using a RosettaScript that performed backrub (Davis et al. 2006) and side chain packing to introduce changes to the backbone angle and side chain orientation of all the nanobody and antigen chains.

Initial Pose Generation and Refinement

Initial poses were generated by self-docking each nanobody to its native antigen by submitting to the ClusPro webserver (Kozakov et al. 2017) in the default mode and the antibody mode. Each initial pose generated from ClusPro was further refined by RosettaDock with an initial low-resolution centroid mode to sample orientation around the initial pose and a subsequent high-resolution, full-atom mode to optimize the orientation and side chain packing. We have instructed the generation of 100 refined poses per each initial pose and obtained a grand total of 3,338,574 refined poses.

Target Label Preparation

We aimed at developing a binary classifier that can distinguish native-like from non-native-like nanobody pose. We used DockQ (Basu and Wallner 2016), a benchmarked metric that estimates the quality of protein-protein complex models with reference to the native complex coordinates, to evaluate the quality of our docked nanobody poses. The DockQ score was benchmarked in the latest version of CAPRI (Lensink et al. 2020) that showed good resolving power to the CAPRI quality classes of protein-protein complex model, which includes from the worst to the best: “incorrect” (0.00-0.23), “acceptable” (0.23-0.49), “medium” (0.49-0.80) and “high” (0.80-1.00) where inside brackets are the ranges of DockQ scores corresponding to each class. We labeled our poses as “non-native-like” if DockQ score < 0.23 and otherwise “native-like”. This division was equivalent to a binary decision boundary that poses that were “non-native-like” poses in this study corresponded to “incorrect” solution in CAPRI while “native-like” poses in this study corresponded to “acceptable”, “medium” and “high” solutions in CAPRI (Table 6). The same decision boundary was set by our competing method DOVE, thus the same labeling method allowed the direct comparison of prediction performance with DOVE. After labeling, we have obtained a total of 106,391 native-like poses and 3,232,183 non-native-like poses.

Feature Engineering

We used InterfaceAnalyzer from Rosetta (Stranges and Kuhlman 2013) and AnalyseComplex from FoldX (Delgado et al. 2019) to calculate energy and contact features of each refined pose. Apart from the energy features that were automatically calculated by the two interface analyzing programs, we have calculated the CDR contact profiles of each refined pose in terms of the proportion of all CDR residues in the paratope and the proportion of interacting residue from each CDR loop compared with their own full lengths. In addition, we have used the aaDescriptors from the Peptides package (Osorio, Rondón-Villarreal, and Torres 2015), which is a collection of 66 descriptors that describe the physicochemical, electrostatic and topological properties of residues (Cruciani et al. 2004; Kidera et al. 1985; Sandberg et al. 1998; G. Liang and Li 2007; Feifei Tian, Zhou, and Li 2007; Mei et al. 2005; van Westen et al. 2013; Yang et al. 2010; Georgiev 2009; Zaliani and Gancia 1999), to describe the paratope and the epitope of our nanobody poses. For each individual property value, we calculated the summed value by adding up from all residues of the paratope, and independently repeated this summation for the epitope.

Model Training, Testing and Hyperparameter Optimization

We used XGBoost (T. Chen and Guestrin 2016), which is an efficient implementation of the gradient-boosted decision tree, to map our feature set to the binary label of pose nativeness (Table 7). We selected XGBoost due to its robustness in providing good prediction performance in various machine learning problems in the data scientist community Kaggle. We used k-fold cross-validation (k=5) to account for the randomness in the partitioning of the training set and the test set. In each round of testing, poses from 80% (144) of the PDBs were randomly selected as the training set and the remaining poses belonged to the test set. This random partition step was applied to the PDB IDs but not the refined poses to minimize information leakage from the training set to the test set because refined poses derived from the same initial pose might contain considerably similar feature values, which causes overestimation of the prediction performance of our model. Before benchmarking, we have optimized the hyperparameters of our XGBoost model by searching from hundreds of preliminary models with different combinations of the hyperparameters (Table 8). Due to a high imbalance of class labels, which is in the approximate ratio of 30:1 for non-native-like to native-like poses, area under the precision-recall curve (PR-AUC), instead of the commonly used ROC-AUC, of the test set prediction was used as the evaluation metric to pick the best hyperparameter combination (Boyd, Eng, and Page 2013; J. Davis and Goadrich 2006).

Benchmarking

To benchmark our model, we compared the prediction performance of the XGBoost model built with the optimized hyperparameter combination with the pose ranking from ClusPro (Kozakov et al. 2017) and DOVE (Wang et al. 2020). DOVE was chosen to benchmark our prediction model because it is a deep learning model that performs binary classification of native protein-protein interaction, which is of high similarity in terms of modeling method and objective to our method. Because both our model and DOVE used a predicted binary probability of nativeness as the output, we also compared the prediction performance with ClusPro, which uses ranking of pose as the output. We ranked the binary probability averaged from all refined pose derived from each initial pose in descending order as the ranking of each initial pose. Population of ranking of all native-like poses of all test set partitions from the 5-fold cross-validation was statistically compared with paired t-test.

Feature Importance Calculation

Feature importance was assigned as the SHAP value by the SHAP package (Lundberg and Lee 2017). SHAP package was commonly used to aid interpretation of machine learning models by assigning the SHAP value, which measures the directionality and the degree of contribution by each feature value to the predicted value in that particular sample. A positive SHAP value corresponds to a positive contribution of the feature value to the predicted value while a greater magnitude of the SHAP value corresponds to a higher contribution of the feature value to the predicted value. Importance of features were compared by ranking in descending order by their mean absolute SHAP value from the test set prediction.

RESULTS AND DISCUSSION

Benchmarking Results with ClusPro

To validate the performance of our nanobody pose prediction model, we have benchmarked the performance of our model with ClusPro (Kozakov et al. 2017), which is the current state-of-the-art protein-protein docking algorithm validated by the latest version of CAPRI (Lensink et al. 2020). For the 5-fold cross-validated prediction of the test set, our nanobody pose prediction model showed significantly higher ranking of native-like poses (p<1e-04) than ClusPro (Figure 33). For a majority of the native pose from the test set, they were ranked within top 10 by our nanobody pose prediction model while ClusPro ranked a majority of them within top 50, which showed the better performance in ranking native-like nanobody pose of our model.

Benchmarking Results with DOVE

To further compare the prediction performance of our nanobody pose prediction model, we have compared the prediction performance of our model with DOVE, a structure-based classifier of native protein-protein interaction by 3D convolution neural network. Due to a relatively large number of refined poses, the prediction of the poses by DOVE is still under progress. Preliminarily, a majority of native-like pose was predicted as a low probability of nativeness by DOVE.

Prediction Performance of the Best Single Model

We have analyzed the performance of an individual model with the best single model in terms of the highest PR-AUC$_{test}$ from the 5-fold cross-validation. Compared with the training set prediction, the prediction performance on the test set was considerably worse

(Figure 34). Together with the fact that our nanobody pose prediction model was able to prioritize native-like poses with significantly higher rankings than ClusPro, it implied our model performed better at distinguishing native-like poses from non-native-like poses within a native nanobody-antigen pair but performed less well on native-like pose prediction by using a single decision threshold on the binary probability.

Important Features Contributed to Prediction Performance

To understand the usefulness of individual features in distinguishing native-like from non-native-like poses in our nanobody pose prediction model, we have calculated the mean(|SHAP|) value of every feature in the test set prediction by the best single model, which was a measure of overall importance of each feature to the predicted values of the test set samples by the model (Figure 35). The feature that contributed most to the test set prediction by the model was the proportion of CDR residues in the paratope residues. Indeed, for a human to judge whether a nanobody-pose is native or non-naive, the degree of involvement of CDR residues at the interface is an important criteria to consider. Moreover, in the sampling of initial poses by ClusPro, a majority of the non-native-like pose was sampled from the default mode, which did not constrain distance between CDR loops and the interface. Therefore, it was consistent with our expectation that the proportion of CDR residues in the paratope was the top contributing feature to the predicted nativeness of nanobody pose.

The feature with second importance to the predicted nativeness of pose by our model was the interface energy density expressed in dG score (cross-interface) divided by buried surface area factored by a multiplication of 100. This interface energy density feature had a higher importance than the total binding energy score expressed in dG (separated), which was the difference in Rosetta energy score between the separated and the complexed form of a nanobody-antigen pair. We found no previous reports that compared the statistical distribution of energy score and buried surface area (SASA) between native pose and non-native-pose from self-docking nanobody.

To help explain the importance of energy density in distinguishing native-like and non-naive-like nanobody poses, compared with conventional antibody, native nanobody binds its antigen with a higher contact density to the epitope residues and a smaller paratope surface area, which potential gives rise to a higher shape complementarity to the epitope surface compared with conventional antibody-antigen complexes (Mitchell and Colwell 2018b). However, the average SASA of general protein-protein interaction was reported to be 800 ± 200 $Å^2$ (Chakrabarti and Janin 2002), which was not different by far from the estimated average SASA of nanobody paratope of 769 ± 201 $Å^2$ (Mitchell and Colwell 2018b). The relatively high degree of shape complementarity observed in native nanobody poses suggests there is a potentially more stringent requirement of steric clash at the interface, which could be associated with the torsional clash of epitope residues being the third important feature contributed to the predicted nativeness of nanobody poses.

In our attempt to use summed residue descriptors of paratope residues or epitope residues as additional features, we reason that they are helpful to the prediction of nativeness of nanobody pose because, expectedly, there should be detectable differences in the distribution of certain descriptor properties of the paratope or the epitope, such as hydrophobicity, between native-like and non-native-like poses. Indeed, some of these summed properties were regarded by our model as important features that contributed to the prediction of nanobody pose nativeness, with a few of them being comparable to features ordinarily used to assess antibody-antigen poses.

For example, the total hydrophobicity of the epitope measured in terms of kideraFactors (Kidera et al. 1985) was regarded as the 4th important feature, an important feature comparable to the ratio of interacting CDR3 residue to its length, which was ranked as 5th important. Compared with the ratio of interacting CDR2 residue to its length (16th important), summed properties, such as double-bend preference of the paratope (8th important), H-bonding capability of the epitope (9th important) and the 4th principal component of the topological descriptor ST-scales (12th important) (Yang et al. 2010), were regarded as relatively important features that contributed to the prediction of nanobody pose nativeness. Additionally, it appeared that the summed residue descriptors, which were derived from the count of individual residue species on the paratope and epitope, were more important than the count of individual residue species themselves because, within top 20 important features, only one feature of residue species count (glutamate count on the epitope, 18th important) was present versus the presence of nine summed residue descriptors. It implied the usefulness of the description of epitope and paratope as a whole by summation of residue descriptors to the prediction of nativeness of nanobody pose.

DISCUSSION

Computational protein design is one of the emerging methods to design new antibodies. However, the failure rate of the current methods of computational antibody design remained high. To improve the current practice of computational antibody design, there is a need to explore other variations in the methods to improve the success rate of generating antibodies with desired properties such as binding affinity and specificity. In this study, we have added another methodological variation to the field of computational antibody design, mainly by innovating the selection method of nanobody poses and designs.

Niche of Nanobody Design Method of Current Study

In this study, we have applied the “dock-and-design” approach to design nanobodies targeting ELMO1-RhoG interaction and S2 of SARS-CoV-2. The “dock-and-design” approach was previously applied in multiple examples of antibody design (Procko et al. 2013; Fleishman, Whitehead, et al. 2011; Fleishman, Corn, et al. 2011; Baran et al. 2017; Karanicolas et al. 2011; Strauch, Fleishman, and Baker 2014; Choi et al. 2014; T. Liang et al. 2021). As its name suggests, “dock-and-design” involves an initial step of antibody-antigen docking and a subsequent sequence design step to optimize the interface. In a broader sense, “dock-and-design” is the progenitor approach in *de novo* antibody design or antibody repurposing because nearly all the antibody design examples and algorithms discussed have incorporated “dock-and-design” as a certain part of their design workflows. Either preceding or following “dock-and-design”, additional design steps, such as CDR grafting and hotspot grafting, were added to increase the chance of obtaining a binding antibody by, in the respective examples here, diversifying the design construct and constraining the type of interactions at the interface.

Here, we refer “dock-and-design” to this primitive application of docking and sequence design without the addition of more design steps. There were several reasons we took this primitive approach of “dock-and-design” to design our nanobody under the existence of multiple antibody design methods and algorithms that were previously validated. Firstly, due to the observable differences between nanobody and the VH domain of conventional antibody in terms of binding mode and interface residue propensity (Mitchell and Colwell 2018b), we reasoned that the existing algorithms may bias towards the design of conventional antibody but not nanobody. For example, when we tested the robustness of PatchDock to generate high-quality poses of nanobody for design, we observed the deviation of poses from known nanobody-antigen complexes, which was exactly our motivation to optimize nanobody pose generation through *in cerebro* learning.

Secondly, we wanted to maintain the framework sequence of nanobody which is more conserved and harbors four mutations that enhance solubility compared to VH domain of conventional antibody (Mitchell and Colwell 2018b), therefore it was not necessary to design the framework from scratch by combinatorial design. Indeed, our collaborator did not report severe expression or solubility problems of the several dozens of designs, implying the tolerance of the nanobody framework to contain the design mutations on CDRs.

Thirdly, because the “dock-and-design” steps form the core of a majority of antibody design methods, a similar approach of optimization should be translatable in improving existing and future design methods. With the simplicity of the “dock-and-design” approach, it is also easier for people who want to perform computational antibody design but with less experience in intensive scripting to learn computational antibody design.

Indeed, the two main design concepts introduced in this study can be easily applied with minimal knowledge in scripting. Firstly, we proposed the use of the two-points constraint, one on CDR1 and one on CDR2, to automate the generation of high-quality nanobody poses that resemble known nanobody-antigen complexes with PatchDock. Without this optimization, PatchDock generally gave docked pose where the framework of the nanobody tends to lean on the epitope surface because the whole CDR3 was specified as the distance constraint by default. The considerable contact between the framework and the epitope is undesirable because framework residues, which are conserved, should in principle not be designed. Alternatively, if framework residues at the interface are not designed, the contacts are expectedly non-specific. With our optimized distance constraints in PatchDock, we have minimized the proportion of the undesirable poses and successfully generated poses which involve all CDR1, CDR2 and CDR3 in the interface as seen from the proportion of CDR contacts from the ELMO1-targeting nanobody designs.

As another design concept introduced by this study, we proposed the “pose-selection-by-design” approach for initial pose selection. In previous design examples which used PatchDock for the selection of initial pose, selecting a number of initial pose from top complementarity scores, typically in the range of 10^2-10^3, was a common practice (Procko et al. 2013; Fleishman, Whitehead, et al. 2011; Fleishman, Corn, et al. 2011; Baran et al. 2017; Karanicolas et al. 2011; Strauch, Fleishman, and Baker 2014; Choi et al. 2014). In previous design examples with docking placed at the upstream of workflow, additional design and selection steps, such as hotspot grafting, were applied to trim down the number of initial poses before a final sequence design step. Compared to selecting only a few poses from PatchDock, these additional pose selection steps are more favorable because firstly, they increase the diversity of initial poses and secondly, complementarity score calculated by surface shape matching from PatchDock is low resolution and therefore the score alone does not imply designability of the pose.

Here, we regard our “pose-selection-by-design” as an alternative strategy to trim down the number of initial poses which took advantage of incorporating an increased pose diversity while simultaneously assessing their potential to develop into a binder. “Pose-selection-by-design” was inspired by the characteristic binding energy landscape from protein-protein interaction, which can be viewed as the change of binding energy as two interacting protein partners approach to the native binding conformation. A typical binding energy landscape of protein-protein interaction has a deep funnel shape towards the native conformation (Ravikumar, Huang, and Yang 2012; Tovchigrechko and Vakser 2001; Alsallaq and Zhou 2007; Schug and Onuchic 2010; Schueler-Furman et al. 2005; London and Schueler-Furman 2008a, 2007; Ruvinsky and Vakser 2008). The typical size of the binding energy funnel was previously estimated to be 6-8 Å, starting from where an obvious decrease in binding energy becomes apparent as the conformation approaches a lower RMSD to the native conformation (Hunjan et al. 2008). This estimation of the funnel size is consistent with the previous description of the non-overlapping nature of alternative binding modes in protein-protein interactions (Kundrotas and Vakser 2013). When the binding partners continue to approach the native conformation and surpasses approximately 2 Å, a steep drop to a local energy minimum was observed, which could be explained by the release of energy from side chain packing which is possible only with the close distance of binding partners (Schueler-Furman et al. 2005).

FunHunt, a support vector machine model for native protein-protein interaction, included features from the characteristic energy landscape of protein-protein interaction to classify the nativeness of protein-protein complex (London and Schueler-Furman 2008b). Their benchmarking showed the native conformation of 50 out of 52 known protein-protein complexes and 12 CAPRI targets were correctly classified, demonstrating the usefulness of features from binding energy funnels in the prediction of native protein-protein interaction.

In this study, however, the energy landscape used for pose selection, which was expressed in interface energy of designs against RMSD of the nanobody backbone before and after the design, was different by nature compared with the energy funnel of docked pose from known protein-protein interactions in two major aspects. Firstly, the initial pose before design was a predicted pose and therefore was not experimentally proven, implying the RMSD before and after design was not a measure of correctness of the conformation. Secondly, in the interface energy-RMSD plot, every point represents a different design which harbors a different CDR sequence at the interface, which is different from ordinary protein-protein docking where each docked pose has a relatively constant residue content at the interface. Due to these major differences in nature of the energy landscape from “dock-and-design” and sole docking, it was projected that an energy landscape from “dock-and-design” can be fuzzier because it resembles an overlay of binding energy-RMSD plots of multiple designs harboring different CDR sequences and therefore when different designs were docked into exactly the same pose, should give a range of binding energy.

Despite the differences in physical meaning of the energy landscapes in two cases, we explored the utility of the energy landscape from “dock-and-design” to select initial pose with the following arguments. Firstly, we did observe binding energy landscapes from designs that closely resembled the deep funnel-shaped energy landscape observed in typical protein-protein interactions, which could be explained by either the design converged to a small number of sequence variants or the binding was tolerant to multiple sequence variants from design. In the case of sequence convergence, it was a good sign for selection because RosettaDesign has found the design solution. In the case of sequence tolerance by the epitope, it potentially implies a certain degree of non-specific binding but because getting a binder is the first goal to achieve, specificity can be further designed based on a non-specific binder, which is also a practice in antibody design by directed evolution using a minimalist residue library (Kelly et al. 2018; Xu et al. 2013; Birtalan et al. 2008) and in antibody affinity maturation by nature (Shehata et al. 2019).

Secondly, we also observed the sampling of binding energy minima at low RMSD range less than 2 Å for some designs. Indeed, we were aware that a potentially binding design may not necessarily have binding energy minimum at low RMSD range because the initial pose used for RMSD calculation may not necessarily be the native conformation. Therefore, by the approach of selecting funnel-shaped energy landscapes from design, we might have lost a portion of designs that did not show funnel-shaped energy landscapes. However, for the alternative poses we observed that showed binding energy minimum at a low RMSD, it indicated the agreement between initial pose generation by PatchDock and the docking during design by RosettaDock. This phenomenon was a positive signal for us to select the pose because there were more energy landscapes from other designs that were more elusive, which sampled a narrow RMSD range or a narrow binding energy range (Figure 11).

Thirdly, we reason that our “pose-selection-by-design” method for initial pose selection is a heuristic approach because we used the design results from quick design to guide initial pose selection. “Pose-selection-by-design” minimized our assumptions in the selection of initial pose that will turn into a binding design compared with the practice of confinement of the number of initial poses in other design studies. By successfully obtaining nanobody designs targeted to ELMO1, we would like to confirm the sensitivity of the design theory above to the success rate of computational antibody design in the future.

To further improve the success rate of our nanobody design workflow, we will incorporate the modeling of flexibility of CDR loops to improve the prediction of interface interaction, which will in turn improve the accuracy of binding affinity prediction and the segregation of a binding design with alternative designs.

Niche of Nanobody Pose Prediction Method of Current Study

In the two examples of nanobody design of this study, we have improved the quality of initial pose by *in cerebro* learning and mimicking nanobody-antigen complexes through optimization of distance constraints of CDRs in docking by PatchDock. Initial pose selection that applied knowledge from *in cerebro* judgement has an obvious advantage. In contrast to structure-based small molecule design which usually put visual judgment as the last stage of design, due to a relatively small number of initial poses in antibody design, it is relatively feasible to perform visual judgment at the initial stage. An initial trimming of the number of poses by visual judgment enriches a number of high-quality poses. Such initial trimming ensures the efficient use of computation power for designing poses that delivers good design results and therefore less computation power is wasted on designing poses that will not be picked at the final stage of visual judgement.

However, any kind of visual judgment has several major drawbacks. Firstly, there is subjectivity in visual judgement. Therefore, visual judgement by different individuals will lead to inconsistent evaluation outcomes. Secondly, there is difficulty in communicating the desirable and undesirable visual features learned *in cerebro.* Due to this communication difficulty, thirdly, it is difficult to popularize the knowledge from visual learning and thus limited the scale of the application on antibody design by other people. Indeed, rather than plainly describing our observation on how nanobodies usually bind their antigens, which was also described in previous reviews of nanobody binding modes (Mitchell and Colwell 2018b, [a] 2018; Zavrtanik et al. 2018), we have translated our visual learning to executable instructions to the PatchDock program. Therefore, firstly, the regeneration of the high-quality poses was automated and secondly, people can directly apply the same instruction without the need of understanding verbally from our description.

Despite our attempt to bridge the gap between communication and application, because the PatchDock optimization was ultimately based on our visual learning, there are biases in the predicted binding mode of poses generated from the optimization. This was the motivation of developing our nanobody pose prediction model, which was trained and vigorously tested upon a dataset derived from self-docking a majority currently known nanobody to their native antigens with ClusPro, the current state-of-the-art docking algorithm verified by the latest version of CAPRI (Lensink et al. 2020), and benchmarked against the performance of a classifier of similar class. Recently, there were several classifiers developed to distinguish native from non-native protein-protein complex structures (Schneider et al. 2021; Wang et al. 2020; Yue Cao and Shen 2020; Eismann et al. 2021; Geng et al. 2020; Tanemura, Pei, and Merz 2020). To our best knowledge, our model is the first model in class that is dedicated to the prediction of native nanobody pose, which is of special niche because there are distinctive features nanobody-antigen interaction in terms of binding mode, interface residue propensity, shape of epitope surface, residue conservation of the framework and structural diversity of CDR loops compared with conventional antibody-antigen interaction (Mitchell and Colwell 2018b, [a] 2018; Zavrtanik et al. 2018).

Although it was pointed out that nanobody-antigen interaction resembles more to general protein-protein interaction than to antibody-antigen interaction (Zavrtanik et al. 2018), a specialist model that is solely trained with examples of a specific type of protein-protein interaction should give better prediction performance compared with a generalist model that learned from examples with diverse patterns of feature distribution. Compared with the classifiers that were trained by automatic feature learning, such as by 3D convolution in DOVE, our model took the conventional feature engineering approach that utilized energy terms from previously validated energy functions of Rosetta and FoldX and calculated the contact profile which consisted of proportions of CDR loops contacts and residue count.

Indeed, CDR contact features, which were generally not incorporated to aid prediction in other classifiers of general protein-protein interaction, were important features that contributed substantially to the prediction of our model, which could be a potential source of the difference in prediction performance. However, although ClusPro used CDR information in its antibody mode, our model performed significantly better than ClusPro in ranking the native-like poses. It could be explained by either our model has captured more accurately the distinctive pattern of usage of CDR loops in nanobodies compared with conventional antibodies, or the contribution from other features that were not used pose ranking in ClusPro antibody docking.

Additionally, we have explored the use of sum of aaDescriptors of all residues in paratope or epitope as features to our model, to see if the properties described by the aaDescriptors by treating the paratope or the epitope as a whole could be the distinguishing features to native-like and non-native-like poses. Comparison of SHAP value showed several summation of aaDescriptors of epitope or paratope, such as the summed hydrophobicity of epitope residues and the summed double-bend preference of the paratope residues, had comparable importance compared with interface energy and CDR contact features. Features important to the prediction of native-like nanobody pose learned by model interpretation in this study enriched our understanding of nanobody-antigen interaction and could guide future improvement in nanobody pose prediction. The application of our nanobody pose prediction model will aid the selection of initial poses that show native-like properties for further design.

CONCLUSION

In this study, we have exercised the structure-based computational nanobody designs on two therapeutic targets. In the design workflow, we have introduced two optimizations to initial pose selection of nanobody. We have automated the generation of nanobody poses which resemble native nanobody poses through *in cerebro* visual learning. We have attempted a new pose selection strategy called “pose-selection-by-design” which uses design results to guide initial pose selection. In the design of ELMO1-targeting nanobody, we have successfully designed a set of binding nanobodies with the best binding affinity verified in the micromolar range. In the design of S2-targeting nanobody for SARS-CoV-2, we attempted the design of a broad-spectrum antibody drug by targeting the conserved ectodomain of SARS-CoV-2 S. We continue to examine the potential cause of undetectable binding between our S-targeting nanobody designs to enable the future development of a broad-spectrum therapeutics to SARS-related CoVs. To improve our ability to select nanobody poses for design, we have developed a nanobody pose prediction method which outperformed the current state-of-the-art method of protein-protein docking algorithm and a deep learning method of a similar class. This study represented one of the early examples in the field of computational antibody design and contributed to increasing our knowledge in this emerging methodology for antibody discovery.

REFERENCES

Adolf-Bryfogle, Jared, Oleks Kalyuzhniy, Michael Kubitz, Brian D. Weitzner, Xiaozhen Hu, Yumiko Adachi, William R. Schief, and Roland L. Dunbrack Jr. 2018. "RosettaAntibodyDesign (RAbD): A General Framework for Computational Antibody Design." *PLoS Computational Biology* 14 (4): e1006112.

Adolf-Bryfogle, Jared, Qifang Xu, Benjamin North, Andreas Lehmann, and Roland L. Dunbrack Jr. 2015. "PyIgClassify: A Database of Antibody CDR Structural Classifications." *Nucleic Acids Research* 43 (Database issue): D432–38.

Ali, Fedaa, Amal Kasry, and Muhamed Amin. 2021. "The New SARS-CoV-2 Strain Shows a Stronger Binding Affinity to ACE2 due to N501Y Mutant." *Medicine in Drug Discovery* 10 (June): 100086.

Alouane, Tarek, Meriem Laamarti, Abdelomunim Essabbar, Mohammed Hakmi, El Mehdi Bouricha, M. W. Chemao-Elfihri, Souad Kartti, et al. 2020. "Genomic Diversity and Hotspot Mutations in 30,983 SARS-CoV-2 Genomes: Moving Toward a Universal Vaccine for the 'Confined Virus'?" *Pathogens* 9 (10). https://doi.org/10.3390/pathogens9100829.

Alsallaq, Ramzi, and Huan-Xiang Zhou. 2007. "Energy Landscape and Transition State of Protein-Protein Association." *Biophysical Journal* 92 (5): 1486–1502.

Andersson, Erik, Rebecca Hsieh, Howard Szeto, Roshanak Farhoodi, Nurit Haspel, and Filip Jagodzinski. 2016. "Assessing How Multiple Mutations Affect Protein Stability Using Rigid Cluster Size Distributions." In *2016 IEEE 6th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, 1–6.

Ashkenazy, Haim, Shiran Abadi, Eric Martz, Ofer Chay, Itay Mayrose, Tal Pupko, and Nir Ben-Tal. 2016. "ConSurf 2016: An Improved Methodology to Estimate and Visualize Evolutionary Conservation in Macromolecules." *Nucleic Acids Research* 44 (W1): W344–50.

Baran, Dror, M. Gabriele Pszolla, Gideon D. Lapidoth, Christoffer Norn, Orly Dym, Tamar Unger, Shira Albeck, Michael D. Tyka, and Sarel J. Fleishman. 2017. "Principles for Computational Design of Binding Antibodies." *Proceedings of the National Academy of Sciences of the United States of America* 114 (41): 10900–905.

Barlow, Kyle A., Shane Ó Conchúir, Samuel Thompson, Pooja Suresh, James E. Lucas, Markus Heinonen, and Tanja Kortemme. 2018. "Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein–Protein Binding Affinity upon Mutation." *The Journal of Physical Chemistry. B* 122 (21): 5389–99.

Barnes, Christopher O., Claudia A. Jette, Morgan E. Abernathy, Kim-Marie A. Dam, Shannon R. Esswein, Harry B. Gristick, Andrey G. Malyutin, et al. 2020. "SARS-CoV-2

Neutralizing Antibody Structures Inform Therapeutic Strategies." *Nature* 588 (7839): 682–87.

Basu, Sankar, and Björn Wallner. 2016. "DockQ: A Quality Measure for Protein-Protein Docking Models." *PloS One* 11 (8): e0161879.

Birtalan, Sara, Yingnan Zhang, Frederic A. Fellouse, Lihua Shao, Gabriele Schaefer, and Sachdev S. Sidhu. 2008. "The Intrinsic Contributions of Tyrosine, Serine, Glycine and Arginine to the Affinity and Specificity of Antibodies." *Journal of Molecular Biology* 377 (5): 1518–28.

Boyd, Kendrick, Kevin H. Eng, and C. David Page. 2013. "Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals." In *Machine Learning and Knowledge Discovery in Databases*, 451–66. Springer Berlin Heidelberg.

Cai, Yongfei, Jun Zhang, Tianshu Xiao, Hanqin Peng, Sarah M. Sterling, Richard M. Walsh Jr, Shaun Rawson, Sophia Rits-Volloch, and Bing Chen. 2020. "Distinct Conformational States of SARS-CoV-2 Spike Protein." *Science* 369 (6511): 1586–92.

Cao, Yanan, Lin Li, Zhimin Feng, Shengqing Wan, Peide Huang, Xiaohui Sun, Fang Wen, Xuanlin Huang, Guang Ning, and Weiqing Wang. 2020. "Comparative Genetic Analysis of the Novel Coronavirus (2019-nCoV/SARS-CoV-2) Receptor ACE2 in Different Populations." *Cell Discovery* 6 (1): 11.

Cao, Yue, and Yang Shen. 2020. "Energy-Based Graph Convolutional Networks for Scoring Protein Docking Models." *Proteins* 88 (8): 1091–99.

Chakrabarti, Pinak, and Jol Janin. 2002. "Dissecting Protein-Protein Recognition Sites." *Proteins* 47 (3): 334–43.

Chanier, Timothée, and Patrick Chames. 2019. "Nanobody Engineering: Toward Next Generation Immunotherapies and Immunoimaging of Cancer." *Antibodies (Basel, Switzerland)* 8 (1). https://doi.org/10.3390/antib8010013.

Chen, Jiahui, Rui Wang, Menglun Wang, and Guo-Wei Wei. 2020. "Mutations Strengthened SARS-CoV-2 Infectivity." *Journal of Molecular Biology* 432 (19): 5212–26.

Chen, Rita E., Xianwen Zhang, James Brett Case, Emma S. Winkler, Yang Liu, Laura A. VanBlargan, Jianying Liu, et al. 2021. "Resistance of SARS-CoV-2 Variants to Neutralization by Monoclonal and Serum-Derived Polyclonal Antibodies." *Nature Medicine* 27 (4): 717–26.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. KDD '16. New York, NY, USA: Association for Computing Machinery.

Choi, Yoon Sup, Soomin Yoon, Kyung-Lock Kim, Jiho Yoo, Parkyong Song, Minsoo Kim, Young-Eun Shin, et al. 2014. “Computational Design of Binding Proteins to EGFR Domain II.” *PloS One* 9 (4): e92513.

Chowdhury, Ratul, Matthew F. Allan, and Costas D. Maranas. 2018. “OptMAVEn-2.0: De Novo Design of Variable Antibody Regions against Targeted Antigen Epitopes.” *Antibodies* 7 (3): 23.

Cohen, Jon. 2021. “The Dream Vaccine.” *Science* 372 (6539): 227–31.

Collier, Dami A., Anna De Marco, Isabella Atm Ferreira, Bo Meng, Rawlings Datir, Alexandra C. Walls, Jessica Bassi, et al. 2021. “SARS-CoV-2 B. 1.1. 7 Escape from mRNA Vaccine-Elicited Neutralizing Antibodies.” *MedRxiv*. https://www.medrxiv.org/content/10.1101/2021.01.19.21249840v3.full-text.

Conway, Patrick, Michael D. Tyka, Frank DiMaio, David E. Konerding, and David Baker. 2014. “Relaxation of Backbone Bond Geometry Improves Protein Energy Landscape Modeling.” *Protein Science: A Publication of the Protein Society* 23 (1): 47–55.

Cruciani, Gabriele, Massimo Baroni, Emanuele Carosati, Monica Clementi, Roberta Valigi, and Sergio Clementi. 2004. “Peptide Studies by Means of Principal Properties of Amino Acids Derived from MIF Descriptors.” *Journal of Chemometrics* 18 (34): 146–55.

Dai, Wentao, Aiping Wu, Liangxiao Ma, Yi-Xue Li, Taijiao Jiang, and Yuan-Yuan Li. 2016. “A Novel Index of Protein-Protein Interface Propensity Improves Interface Residue Recognition.” *BMC Systems Biology* 10 (Suppl 4): 112.

Davies, Nicholas G., Sam Abbott, Rosanna C. Barnard, Christopher I. Jarvis, Adam J. Kucharski, James D. Munday, Carl A. B. Pearson, et al. 2021. “Estimated Transmissibility and Impact of SARS-CoV-2 Lineage B.1.1.7 in England.” *Science* 372 (6538). https://doi.org/10.1126/science.abg3055.

Davis, Ian W., W. Bryan Arendall 3rd, David C. Richardson, and Jane S. Richardson. 2006. “The Backrub Motion: How Protein Backbone Shrugs When a Sidechain Dances.” *Structure* 14 (2): 265–74.

Davis, Jesse, and Mark Goadrich. 2006. “The Relationship between Precision-Recall and ROC Curves.” In *Proceedings of the 23rd International Conference on Machine Learning*, 233–40. ICML ’06. New York, NY, USA: Association for Computing Machinery.

Day, Troy, Sylvain Gandon, Sébastien Lion, and Sarah P. Otto. 2020. “On the Evolutionary Epidemiology of SARS-CoV-2.” *Current Biology: CB* 30 (15): R849–57.

Dehghanpoor, Ramin, Evan Ricks, Katie Hursh, Sarah Gunderson, Roshanak Farhoodi, Nurit Haspel, Brian Hutchinson, and Filip Jagodzinski. 2018. “Predicting the Effect of

Single and Multiple Mutations on Protein Structural Stability." *Molecules* 23 (2). https://doi.org/10.3390/molecules23020251.

Delgado, Javier, Leandro G. Radusky, Damiano Cianferoni, and Luis Serrano. 2019. "FoldX 5.0: Working with RNA, Small Molecules and a New Graphical Interface." *Bioinformatics* 35 (20): 4168–69.

Di Caro, A., F. Cunha, N. Petrosillo, and N. J. Beeching. 2021. "SARS-CoV-2 Escape Mutants and Protective Immunity from Natural Infections or Immunizations." *Clinical Microbiology*. https://www.sciencedirect.com/science/article/pii/S1198743X21001464.

Dunbar, James, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M. Deane. 2014. "SAbDab: The Structural Antibody Database." *Nucleic Acids Research* 42 (Database issue): D1140–46.

Dunbar, James, Konrad Krawczyk, Jinwoo Leem, Claire Marks, Jaroslaw Nowak, Cristian Regep, Guy Georges, Sebastian Kelm, Bojana Popovic, and Charlotte M. Deane. 2016. "SAbPred: A Structure-Based Antibody Prediction Server." *Nucleic Acids Research* 44 (W1): W474–78.

Eismann, Stephan, Raphael J. L. Townshend, Nathaniel Thomas, Milind Jagota, Bowen Jing, and Ron O. Dror. 2021. "Hierarchical, Rotation-Equivariant Neural Networks to Select Structural Models of Protein Complexes." *Proteins* 89 (5): 493–501.

Esmaielbeiki, Reyhaneh, Konrad Krawczyk, Bernhard Knapp, Jean-Christophe Nebel, and Charlotte M. Deane. 2016. "Progress and Challenges in Predicting Protein Interfaces." *Briefings in Bioinformatics* 17 (1): 117–31.

Fernández-Quintero, Monica L., Johannes Kraml, Guy Georges, and Klaus R. Liedl. 2019. "CDR-H3 Loop Ensemble in Solution - Conformational Selection upon Antibody Binding." *mAbs* 11 (6): 1077–88.

Fleishman, Sarel J., Jacob E. Corn, Eva-Maria Strauch, Timothy A. Whitehead, John Karanicolas, and David Baker. 2011. "Hotspot-Centric de Novo Design of Protein Binders." *Journal of Molecular Biology* 413 (5): 1047–62.

Fleishman, Sarel J., Timothy A. Whitehead, Damian C. Ekiert, Cyrille Dreyfus, Jacob E. Corn, Eva-Maria Strauch, Ian A. Wilson, and David Baker. 2011. "Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin." *Science* 332 (6031): 816–21.

Focosi, Daniele, and Fabrizio Maggi. 2021. "Neutralising Antibody Escape of SARS-CoV-2 Spike Protein: Risk Assessment for Antibody-Based Covid-19 Therapeutics and Vaccines." *Reviews in Medical Virology*, no. rmv.2231 (March). https://doi.org/10.1002/rmv.2231.

Garcia-Beltran, Wilfredo F., Evan C. Lam, Kerri St Denis, Adam D. Nitido, Zeidy H. Garcia, Blake M. Hauser, Jared Feldman, et al. 2021. “Multiple SARS-CoV-2 Variants Escape Neutralization by Vaccine-Induced Humoral Immunity.” *Cell*, March. https://doi.org/10.1016/j.cell.2021.03.013.

Garzon, José Ignacio, José Ramón Lopéz-Blanco, Carles Pons, Julio Kovacs, Ruben Abagyan, Juan Fernandez-Recio, and Pablo Chacon. 2009. “FRODOCK: A New Approach for Fast Rotational Protein–protein Docking.” *Bioinformatics* 25 (19): 2544–51.

Geng, Cunliang, Yong Jung, Nicolas Renaud, Vasant Honavar, Alexandre M. J. J. Bonvin, and Li C. Xue. 2020. “iScore: A Novel Graph Kernel-Based Function for Scoring Protein-Protein Docking Models.” *Bioinformatics* 36 (1): 112–21.

Georgiev, Alexander G. 2009. “Interpretable Numerical Descriptors of Amino Acid Space.” *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 16 (5): 703–23.

Goicoechea, Silvia M., Sahezeel Awadia, and Rafael Garcia-Mata. 2014. “I’m Coming to GEF You: Regulation of RhoGEFs during Cell Migration.” *Cell Adhesion & Migration* 8 (6): 535–49.

Greaney, Allison J., Andrea N. Loes, Katharine H. D. Crawford, Tyler N. Starr, Keara D. Malone, Helen Y. Chu, and Jesse D. Bloom. 2021. “Comprehensive Mapping of Mutations to the SARS-CoV-2 Receptor-Binding Domain That Affect Recognition by Polyclonal Human Serum Antibodies.” *bioRxiv*. https://doi.org/10.1101/2020.12.31.425021.

Greaney, Allison J., Tyler N. Starr, Pavlo Gilchuk, Seth J. Zost, Elad Binshtein, Andrea N. Loes, Sarah K. Hilton, et al. 2021. “Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain That Escape Antibody Recognition.” *Cell Host & Microbe* 29 (1): 44–57.e9.

Gromiha, M. Michael, K. Yugandhar, and Sherlyn Jemimah. 2017. “Protein-Protein Interactions: Scoring Schemes and Binding Affinity.” *Current Opinion in Structural Biology* 44 (June): 31–38.

Grubaugh, Nathan D., Emma B. Hodcroft, Joseph R. Fauver, Alexandra L. Phelan, and Muge Cevik. 2021. “Public Health Actions to Control New SARS-CoV-2 Variants.” *Cell* 184 (5): 1127–32.

Gupta, Gaorav P., and Joan Massagué. 2006. “Cancer Metastasis: Building a Framework.” *Cell* 127 (4): 679–95.

Guruprasad, Lalitha. 2021. “Human SARS CoV-2 Spike Protein Mutations.” *Proteins* 89 (5): 569–76.

Hassanzadeh-Ghassabeh, Gholamreza, Nick Devoogdt, Pieter De Pauw, Cécile Vincke, and Serge Muyldermans. 2013. “Nanobodies and Their Potential Applications.” *Nanomedicine* 8 (6): 1013–26.

Hoffmann, Markus, Hannah Kleine-Weber, and Stefan Pöhlmann. 2020. “A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells.” *Molecular Cell* 78 (4): 779–84.e5.

Honegger, A., and A. Plückthun. 2001. “Yet Another Numbering Scheme for Immunoglobulin Variable Domains: An Automatic Modeling and Analysis Tool.” *Journal of Molecular Biology* 309 (3): 657–70.

Hornak, Viktor, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. 2006. “Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters.” *Proteins* 65 (3): 712–25.

Huang, Po-Ssu, Scott E. Boyken, and David Baker. 2016. “The Coming of Age of de Novo Protein Design.” *Nature* 537 (7620): 320–27.

Hu, Jie, Pai Peng, Kai Wang, Liang Fang, Fei-Yang Luo, Ai-Shun Jin, Bei-Zhong Liu, Ni Tang, and Ai-Long Huang. 2021. “Emerging SARS-CoV-2 Variants Reduce Neutralization Sensitivity to Convalescent Sera and Monoclonal Antibodies.” *Cellular & Molecular Immunology* 18 (4): 1061–63.

Hunjan, Jagtar, Andrey Tovchigrechko, Ying Gao, and Ilya A. Vakser. 2008. “The Size of the Intermolecular Energy Funnel in Protein-Protein Interactions.” *Proteins* 72 (1): 344–52.

Hussain, Mushtaq, Nusrat Jabeen, Anusha Amanullah, Ayesha Ashraf Baig, Basma Aziz, Sanya Shabbir, and Fozia Raza. 2020. “Structural Basis of SARS-CoV-2 Spike Protein Priming by TMPRSS2.” *bioRxiv*. https://doi.org/10.1101/2020.04.21.052639.

Jeliazkov, Jeliazko R., Rahel Frick, Jing Zhou, and Jeffrey J. Gray. 2021. “Robustification of RosettaAntibody and Rosetta SnugDock.” PloS One 16 (3): e0234282.

Jovčevska, Ivana, and Serge Muyldermans. 2020. “The Therapeutic Potential of Nanobodies.” *BioDrugs: Clinical Immunotherapeutics, Biopharmaceuticals and Gene Therapy* 34 (1): 11–26.

Karanicolas, John, Jacob E. Corn, Irwin Chen, Lukasz A. Joachimiak, Orly Dym, Sun H. Peck, Shira Albeck, et al. 2011. “A de Novo Protein Binding Pair by Computational Design and Directed Evolution.” *Molecular Cell* 42 (2): 250–60.

Kashyap, Vivek K., Anupam Dhasmana, Andrew Massey, Sudhir Kotnala, Nadeem Zafar, Meena Jaggi, Murali M. Yallapu, and Subhash C. Chauhan. 2020. “Smoking and COVID-19: Adding Fuel to the Flame.” *International Journal of Molecular Sciences* 21 (18). https://doi.org/10.3390/ijms21186581.

Katoh, Hironori, Kiyo Hiramoto, and Manabu Negishi. 2006. “Activation of Rac1 by RhoG Regulates Cell Migration.” *Journal of Cell Science* 119 (Pt 1): 56–65.

Katoh, Hironori, and Manabu Negishi. 2003. “RhoG Activates Rac1 by Direct Interaction with the Dock180-Binding Protein Elmo.” *Nature* 424 (6947): 461–64.

Kelly, Ryan L., Doris Le, Jessie Zhao, and K. Dane Wittrup. 2018. “Reduction of Nonspecificity Motifs in Synthetic Antibody Libraries.” *Journal of Molecular Biology* 430 (1): 119–30.

Kidera, Akinori, Yasuo Konishi, Masahito Oka, Tatsuo Ooi, and Harold A. Scheraga. 1985. “Statistical Analysis of the Physical Properties of the 20 Naturally Occurring Amino Acids.” *Journal of Protein Chemistry* 4 (1): 23–55.

Kistler, Kathryn E., and Trevor Bedford. 2021. “Evidence for Adaptive Evolution in the Receptor-Binding Domain of Seasonal Coronaviruses OC43 and 229e.” *eLife* 10 (January). https://doi.org/10.7554/eLife.64509.

Kozakov, Dima, David R. Hall, Bing Xia, Kathryn A. Porter, Dzmitry Padhorny, Christine Yueh, Dmitri Beglov, and Sandor Vajda. 2017. “The ClusPro Web Server for Protein-Protein Docking.” *Nature Protocols* 12 (2): 255–78.

Krissinel, E., and K. Henrick. 2004. “Secondary-Structure Matching (SSM), a New Tool for Fast Protein Structure Alignment in Three Dimensions.” *Acta Crystallographica. Section D, Biological Crystallography* 60 (Pt 12 Pt 1): 2256–68.

Kumagai, Izumi, and Kouhei Tsumoto. 2002. “Antigen-Antibody Binding.” In *Encyclopedia of Life Sciences*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1038/npg.els.0001117.

Kumari, Rashmi, Rajendra Kumar, Open Source Drug Discovery Consortium, and Andrew Lynn. 2014. “G_mmpbsa--a GROMACS Tool for High-Throughput MM-PBSA Calculations.” *Journal of Chemical Information and Modeling* 54 (7): 1951–62.

Kundrotas, Petras J., and Ilya A. Vakser. 2013. “Protein-Protein Alternative Binding Modes Do Not Overlap.” *Protein Science: A Publication of the Protein Society* 22 (8): 1141–45.

Kunz, Patrick, Katinka Zinner, Norbert Mücke, Tanja Bartoschik, Serge Muyldermans, and Jörg D. Hoheisel. 2018. “The Structural Basis of Nanobody Unfolding Reversibility and Thermoresistance.” *Scientific Reports* 8 (1): 7934.

Kuroda, Daisuke, and Jeffrey J. Gray. 2016. “Shape Complementarity and Hydrogen Bond Preferences in Protein-Protein Interfaces: Implications for Antibody Modeling and Protein-Protein Docking.” *Bioinformatics* 32 (16): 2451–56.

Lapidoth, Gideon D., Dror Baran, Gabriele M. Pszolla, Christoffer Norn, Assaf Alon, Michael D. Tyka, and Sarel J. Fleishman. 2015. “AbDesign: An Algorithm for Combinatorial

Backbone Design Guided by Natural Conformations and Sequences." *Proteins* 83 (8): 1385–1406.

Lauring, Adam S., and Emma B. Hodcroft. 2021. "Genetic Variants of SARS-CoV-2—What Do They Mean?" *JAMA: The Journal of the American Medical Association* 325 (6): 529–31.

Laustsen, Andreas H., Victor Greiff, Aneesh Karatt-Vellatt, Serge Muyldermans, and Timothy P. Jenkins. 2021. "Animal Immunization, in Vitro Display Technologies, and Machine Learning for Antibody Discovery." *Trends in Biotechnology*, March. https://doi.org/10.1016/j.tibtech.2021.03.003.

Lee, Carol M. Y., Niccolo Iorno, Frederic Sierro, and Daniel Christ. 2007. "Selection of Human Antibody Fragments by Phage Display." *Nature Protocols* 2 (11): 3001–8.

Leenaars, Marlies, and Coenraad F. M. Hendriksen. 2005. "Critical Steps in the Production of Polyclonal and Monoclonal Antibodies: Evaluation and Recommendations." *ILAR Journal / National Research Council, Institute of Laboratory Animal Resources* 46 (3): 269–79.

Lensink, Marc F., Nurul Nadzirin, Sameer Velankar, and Shoshana J. Wodak. 2020. "Modeling Protein-protein, Protein-peptide, and Protein-oligosaccharide Complexes: CAPRI 7th Edition." *Proteins: Structure, Function, and Bioinformatics* 88 (8): 916–38.

Liang, Guizhao, and Zhiliang Li. 2007. "Factor Analysis Scale of Generalized Amino Acid Information as the Source of a New Set of Descriptors for Elucidating the Structure and Activity Relationships of Cationic Antimicrobial Peptides." *QSAR & Combinatorial Science*. https://doi.org/10.1002/qsar.200630145.

Liang, Tianjian, Hui Chen, Jiayi Yuan, Chen Jiang, Yixuan Hao, Yuanqiang Wang, Zhiwei Feng, and Xiang-Qun Xie. 2021. "IsAb: A Computational Protocol for Antibody Design." *Briefings in Bioinformatics*, April. https://doi.org/10.1093/bib/bbab143.

Li, Cheng, Xiaolong Tian, Xiaodong Jia, Jinkai Wan, Lu Lu, Shibo Jiang, Fei Lan, Yinying Lu, Yanling Wu, and Tianlei Ying. 2021. "The Impact of Receptor-Binding Domain Natural Mutations on Antibody Recognition of SARS-CoV-2." *Signal Transduction and Targeted Therapy* 6 (1): 132.

Liebschner, Dorothee, Pavel V. Afonine, Matthew L. Baker, Gábor Bunkóczi, Vincent B. Chen, Tristan I. Croll, Bradley Hintze, et al. 2019. "Macromolecular Structure Determination Using X-Rays, Neutrons and Electrons: Recent Developments in Phenix." *Acta Crystallographica. Section D, Structural Biology* 75 (Pt 10): 861–77.

Li, Qianqian, Jianhui Nie, Jiajing Wu, Li Zhang, Ruxia Ding, Haixin Wang, Yue Zhang, et al. 2021. “SARS-CoV-2 501Y.V2 Variants Lack Higher Infectivity but Do Have Immune Escape.” *Cell*, February. https://doi.org/10.1016/j.cell.2021.02.042.

Li, Tong, Robert J. Pantazes, and Costas D. Maranas. 2014. “OptMAVEn--a New Framework for the de Novo Design of Antibody Variable Region Models Targeting Specific Antigen Epitopes.” *PloS One* 9 (8): e105954.

Liu, Xiaofeng, Richard D. Taylor, Laura Griffin, Shu-Fen Coker, Ralph Adams, Tom Ceska, Jiye Shi, Alastair D. G. Lawson, and Terry Baker. 2017. “Computational Design of an Epitope-Specific Keap1 Binding Antibody Using Hotspot Residues Grafting and CDR Loop Swapping.” *Scientific Reports* 7 (January): 41306.

Liu, Zhuoming, Laura A. VanBlargan, Louis-Marie Bloyet, Paul W. Rothlauf, Rita E. Chen, Spencer Stumpf, Haiyan Zhao, et al. 2021. “Identification of SARS-CoV-2 Spike Mutations That Attenuate Monoclonal and Serum Antibody Neutralization.” *Cell Host & Microbe* 29 (3): 477–88.e4.

Li, Yang, Mingliang Ma, Qing Lei, Feng Wang, Ziyong Sun, Xionglin Fan, and Sheng-Ce Tao. 2020. “Linear Epitope Landscape of SARS-CoV-2 Spike Protein Constructed from 1,051 COVID-19 Patients.” *Infectious Diseases (except HIV/AIDS)*. medRxiv. https://doi.org/10.1101/2020.07.13.20152587.

Li, Zhixiu, Yuedong Yang, Jian Zhan, Liang Dai, and Yaoqi Zhou. 2013. “Energy Functions in de Novo Protein Design: Current Challenges and Future Prospects.” *Annual Review of Biophysics* 42 (February): 315–35.

Lo, Benny K. C. 2004. “Antibody Humanization by CDR Grafting.” *Methods in Molecular Biology* 248: 135–59.

London, Nir, and Ora Schueler-Furman. 2007. “Assessing the Energy Landscape of CAPRI Targets by FunHunt.” *Proteins* 69 (4): 809–15.

———. 2008a. “Funnel Hunting in a Rough Terrain: Learning and Discriminating Native Energy Funnels.” *Structure* 16 (2): 269–79.

———. 2008b. “FunHunt: Model Selection Based on Energy Landscape Characteristics.” *Biochemical Society Transactions* 36 (Pt 6): 1418–21.

Lundberg, Scott, and Su-In Lee. 2017. “A Unified Approach to Interpreting Model Predictions.” *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1705.07874.

Marks, C., and C. M. Deane. 2017. “Antibody H3 Structure Prediction.” *Computational and Structural Biotechnology Journal* 15 (January): 222–31.

McInnes, Leland, John Healy, and Steve Astels. 2017. “Hdbscan: Hierarchical Density Based Clustering.” *The Journal of Open Source Software* 2 (11): 205.

Mei, Hu, Zhi H. Liao, Yuan Zhou, and Shengshi Z. Li. 2005. “A New Set of Amino Acid Descriptors and Its Application in Peptide QSARs.” *Biopolymers* 80 (6): 775–86.

Mitchell, Laura S., and Lucy J. Colwell. 2018a. “Comparative Analysis of Nanobody Sequence and Structure Data.” *Proteins* 86 (7): 697–706.

———. 2018b. “Analysis of Nanobody Paratopes Reveals Greater Diversity than Classical Antibodies.” *Protein Engineering, Design & Selection: PEDS* 31 (7-8): 267–75.

Moore, John P., and Paul A. Offit. 2021. “SARS-CoV-2 Vaccines and the Growing Threat of Viral Variants.” *JAMA: The Journal of the American Medical Association* 325 (9): 821–22.

Morrison, Chris. 2019. “Nanobody Approval Gives Domain Antibodies a Boost.” *Nature Reviews. Drug Discovery* 18 (7): 485–87.

Muyldermans, Serge. 2020. “Applications of Nanobodies.” *Annual Review of Animal Biosciences*, November. https://doi.org/10.1146/annurev-animal-021419-083831.

Nishigami, Hiroshi, Narutoshi Kamiya, and Haruki Nakamura. 2016. “Revisiting Antibody Modeling Assessment for CDR-H3 Loop.” *Protein Engineering, Design & Selection: PEDS* 29 (11): 477–84.

Norman, Richard A., Francesco Ambrosetti, Alexandre M. J. J. Bonvin, Lucy J. Colwell, Sebastian Kelm, Sandeep Kumar, and Konrad Krawczyk. 2020. “Computational Approaches to Therapeutic Antibody Design: Established Methods and Emerging Trends.” *Briefings in Bioinformatics* 21 (5): 1549–67.

North, Benjamin, Andreas Lehmann, and Roland L. Dunbrack Jr. 2011. “A New Clustering of Antibody CDR Loop Conformations.” *Journal of Molecular Biology* 406 (2): 228–56.

Nour, Islam, Ibrahim O. Alanazi, Atif Hanif, and Saleh Eifan. 2021. “Molecular Adaptive Evolution of SARS-COV-2 Spike Protein in Saudi Arabia.” *Saudi Journal of Biological Sciences*, March. https://doi.org/10.1016/j.sjbs.2021.02.077.

Olson, Mark A., Patricia M. Legler, Daniel Zabetakis, Kendrick B. Turner, George P. Anderson, and Ellen R. Goldman. 2019. “Sequence Tolerance of a Single-Domain Antibody with a High Thermal Stability: Comparison of Computational and Experimental Fitness Profiles.” *ACS Omega* 4 (6): 10444–54.

Osorio, Daniel, Paola Rondón-Villarreal, and Rodrigo Torres. 2015. “Peptides: A Package for Data Mining of Antimicrobial Peptides.” *Small* 12: 44–444.

Ou, Junxian, Zhonghua Zhou, Ruixue Dai, Jing Zhang, Wendong Lan, Shan Zhao, Jianguo Wu, et al. 2020. “Emergence of RBD Mutations in Circulating SARS-CoV-2 Strains Enhancing the Structural Stability and Human ACE2 Receptor Affinity of the Spike Protein.” *bioRxiv*. https://doi.org/10.1101/2020.03.15.991844.

Pantazes, Robert J., and Costas D. Maranas. 2013. “MAPs: A Database of Modular Antibody Parts for Predicting Tertiary Structures and Designing Affinity Matured Antibodies.” *BMC Bioinformatics* 14 (May): 168.

Pan, Xingjie, and Tanja Kortemme. 2021. “Recent Advances in de Novo Protein Design: Principles, Methods, and Applications.” *The Journal of Biological Chemistry*, March, 100558.

Patel, Manishha, and Jean-François Côté. 2013. “Ras GTPases’ Interaction with Effector Domains: Breaking the Families' Barrier.” *Communicative & Integrative Biology* 6 (4): e24298.

Pillardy, J., C. Czaplewski, A. Liwo, J. Lee, D. R. Ripoll, R. Kaźmierkiewicz, S. Oldziej, et al. 2001. “Recent Improvements in Prediction of Protein Structure by Global Optimization of a Potential Energy Function.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (5): 2329–33.

Planas, Delphine, Timothée Bruel, Ludivine Grzelak, Florence Guivel-Benhassine, Isabelle Staropoli, Françoise Porrot, Cyril Planchais, et al. 2021. “Sensitivity of Infectious SARS-CoV-2 B.1.1.7 and B.1.351 Variants to Neutralizing Antibodies.” *Nature Medicine*, March, 1–8.

Poh, Chek Meng, Guillaume Carissimo, Bei Wang, Siti Naqiah Amrun, Cheryl Yi-Pin Lee, Rhonda Sin-Ling Chee, Siew-Wai Fong, et al. 2020. “Two Linear Epitopes on the SARS-CoV-2 Spike Protein That Elicit Neutralising Antibodies in COVID-19 Patients.” *Nature Communications* 11 (1): 2806.

Pokala, Navin, and Tracy M. Handel. 2001. “Review: Protein Design—Where We Were, Where We Are, Where We’re Going.” *Journal of Structural Biology* 134 (2): 269–81.

———. 2005. “Energy Functions for Protein Design: Adjustment with Protein–Protein Complex Affinities, Models for the Unfolded State, and Negative Design of Solubility and Specificity.” *Journal of Molecular Biology* 347 (1): 203–27.

Pomplun, Sebastian. 2021. “Targeting the SARS-CoV-2-Spike Protein: From Antibodies to Miniproteins and Peptides.” *RSC Medicinal Chemistry* 12 (2): 197–202.

Prabakaran, Ponraj, Sambasiva P. Rao, and Maria Wendt. 2021. “Animal Immunization Merges with Innovative Technologies: A New Paradigm Shift in Antibody Discovery.” *mAbs* 13 (1): 1924347.

Procko, Erik, Rickard Hedman, Keith Hamilton, Jayaraman Seetharaman, Sarel J. Fleishman, Min Su, James Aramini, et al. 2013. “Computational Design of a Protein-Based Enzyme Inhibitor.” *Journal of Molecular Biology* 425 (18): 3563–75.

Ramaraj, Thiruvarangan, Thomas Angel, Edward A. Dratz, Algirdas J. Jesaitis, and Brendan Mumey. 2012. “Antigen-Antibody Interface Properties: Composition, Residue

Interactions, and Features of 53 Non-Redundant Structures." *Biochimica et Biophysica Acta* 1824 (3): 520–32.

Ravikumar, Krishnakumar M., Wei Huang, and Sichun Yang. 2012. "Coarse-Grained Simulations of Protein-Protein Association: An Energy Landscape Perspective." *Biophysical Journal* 103 (4): 837–45.

Raybould, Matthew I. J., Aleksandr Kovaltsuk, Claire Marks, and Charlotte M. Deane. 2020. "CoV-AbDab: The Coronavirus Antibody Database." *Bioinformatics* , August. https://doi.org/10.1093/bioinformatics/btaa739.

Ruvinsky, Anatoly M., and Ilya A. Vakser. 2008. "Chasing Funnels on Protein-Protein Energy Landscapes at Different Resolutions." *Biophysical Journal* 95 (5): 2150–59.

Sandberg, M., L. Eriksson, J. Jonsson, M. Sjöström, and S. Wold. 1998. "New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids." *Journal of Medicinal Chemistry* 41 (14): 2481–91.

Sasaki, Michihito, Kentaro Uemura, Akihiko Sato, Shinsuke Toba, Takao Sanaki, Katsumi Maenaka, William W. Hall, Yasuko Orba, and Hirofumi Sawa. 2021. "SARS-CoV-2 Variants with Mutations at the S1/S2 Cleavage Site Are Generated in Vitro during Propagation in TMPRSS2-Deficient Cells." *PLoS Pathogens* 17 (1): e1009233.

Schneider, Constantin, Andrew Buchanan, Bruck Taddese, and Charlotte M. Deane. 2021. "DLAB - Deep Learning Methods for Structure-Based Virtual Screening of Antibodies." *bioRxiv*. https://doi.org/10.1101/2021.02.12.430941.

Schneidman-Duhovny, Dina, Yuval Inbar, Ruth Nussinov, and Haim J. Wolfson. 2005. "PatchDock and SymmDock: Servers for Rigid and Symmetric Docking." *Nucleic Acids Research* 33 (Web Server issue): W363–67.

Schoof, Michael, Bryan Faust, Reuben A. Saunders, Smriti Sangwan, Veronica Rezelj, Nick Hoppe, Morgane Boone, et al. 2020. "An Ultrapotent Synthetic Nanobody Neutralizes SARS-CoV-2 by Stabilizing Inactive Spike." *Science* 370 (6523): 1473–79.

Schueler-Furman, Ora, Chu Wang, Phil Bradley, Kira Misura, and David Baker. 2005. "Progress in Modeling of Protein Structures and Interactions." *Science* 310 (5748): 638–42.

Schug, Alexander, and José N. Onuchic. 2010. "From Protein Folding to Protein Function and Biomolecular Binding by Energy Landscape Theory." *Current Opinion in Pharmacology* 10 (6): 709–14.

Seeliger, Daniel. 2013. "Development of Scoring Functions for Antibody Sequence Assessment and Optimization." *PloS One* 8 (10): e76909.

Shehata, Laila, Daniel P. Maurer, Anna Z. Wec, Asparouh Lilov, Elizabeth Champney, Tingwan Sun, Kimberly Archambault, et al. 2019. "Affinity Maturation Enhances

Antibody Specificity but Compromises Conformational Stability." *Cell Reports* 28 (13): 3300–3308.e4.

Shen, Yang, Ioannis Ch Paschalidis, Pirooz Vakili, and Sandor Vajda. 2008. "Protein Docking by the Underestimation of Free Energy Funnels in the Space of Encounter Complexes." *PLoS Computational Biology* 4 (10): e1000191.

Shirai, H., N. Nakajima, J. Higo, A. Kidera, and H. Nakamura. 1998. "Conformational Sampling of CDR-H3 in Antibodies by Multicanonical Molecular Dynamics Simulation." *Journal of Molecular Biology* 278 (2): 481–96.

Shi, Rui, Chao Shan, Xiaomin Duan, Zhihai Chen, Peipei Liu, Jinwen Song, Tao Song, et al. 2020. "A Human Neutralizing Antibody Targets the Receptor-Binding Site of SARS-CoV-2." *Nature* 584 (7819): 120–24.

Sikora, Mateusz, Sören von Bülow, Florian E. C. Blanc, Michael Gecht, Roberto Covino, and Gerhard Hummer. 2020. "Map of SARS-CoV-2 Spike Epitopes Not Shielded by Glycans." https://doi.org/10.1101/2020.07.03.186825.

Sircar, Aroop, and Jeffrey J. Gray. 2010. "SnugDock: Paratope Structural Optimization during Antibody-Antigen Docking Compensates for Errors in Antibody Homology Models." *PLoS Computational Biology* 6 (1): e1000644.

Soga, Shinji, Daisuke Kuroda, Hiroki Shirai, Masato Kobori, and Noriaki Hirayama. 2010. "Use of Amino Acid Composition to Predict Epitope Residues of Individual Antibodies." *Protein Engineering, Design & Selection: PEDS* 23 (6): 441–48.

Song, Wenfei, Miao Gui, Xinquan Wang, and Ye Xiang. 2018. "Cryo-EM Structure of the SARS Coronavirus Spike Glycoprotein in Complex with Its Host Cell Receptor ACE2." *PLoS Pathogens* 14 (8): e1007236.

Starr, Tyler N., Allison J. Greaney, Amin Addetia, William W. Hannon, Manish C. Choudhary, Adam S. Dingens, Jonathan Z. Li, and Jesse D. Bloom. 2021. "Prospective Mapping of Viral Mutations That Escape Antibodies Used to Treat COVID-19." *Science* 371 (6531): 850–54.

Starr, Tyler N., Allison J. Greaney, Adam S. Dingens, and Jesse D. Bloom. 2021. "Complete Map of SARS-CoV-2 RBD Mutations That Escape the Monoclonal Antibody LY-CoV555 and Its Cocktail with LY-CoV016." *Cell Reports. Medicine* 2 (4): 100255.

Stranges, P. Benjamin, and Brian Kuhlman. 2013. "A Comparison of Successful and Failed Protein Interface Designs Highlights the Challenges of Designing Buried Hydrogen Bonds." *Protein Science: A Publication of the Protein Society* 22 (1): 74–82.

Strauch, Eva-Maria, Sarel J. Fleishman, and David Baker. 2014. "Computational Design of a pH-Sensitive IgG Binding Protein." *Proceedings of the National Academy of Sciences of the United States of America* 111 (2): 675–80.

Sun, Chunyun, Long Chen, Ji Yang, Chunxia Luo, Yanjing Zhang, Jing Li, Jiahui Yang, Jie Zhang, and Liangzhi Xie. 2020. "SARS-CoV-2 and SARS-CoV Spike-RBD Structure and Receptor Binding Comparison and Potential Implications on Neutralizing Antibody and Vaccine Development." *bioRxiv*. https://doi.org/10.1101/2020.02.16.951723.

Tanemura, Kiyoto A., Jun Pei, and Kenneth M. Merz Jr. 2020. "Refinement of Pairwise Potentials via Logistic Regression to Score Protein-Protein Interactions." *Proteins* 88 (12): 1559–68.

Tang, Tiffany, Miya Bidon, Javier A. Jaimes, Gary R. Whittaker, and Susan Daniel. 2020. "Coronavirus Membrane Fusion Mechanism Offers a Potential Target for Antiviral Development." *Antiviral Research* 178 (April): 104792.

Tian, Fang, Bei Tong, Liang Sun, Shengchao Shi, Bin Zheng, Zibin Wang, Xianchi Dong, and Peng Zheng. 2021. "Mutation N501Y in RBD of Spike Protein Strengthens the Interaction between COVID-19 and Its Receptor ACE2." *bioRxiv*. https://doi.org/10.1101/2021.02.14.431117.

Tian, Feifei, Peng Zhou, and Zhiliang Li. 2007. "T-Scale as a Novel Vector of Topological Descriptors for Amino Acids and Its Application in QSARs of Peptides." *Journal of Molecular Structure* 830 (1): 106–15.

Tortorici, M. Alejandra, Martina Beltramello, Florian A. Lempp, Dora Pinto, Ha V. Dang, Laura E. Rosen, Matthew McCallum, et al. 2020. "Ultrapotent Human Antibodies Protect against SARS-CoV-2 Challenge via Multiple Mechanisms." *Science* 370 (6519): 950–57.

Tovchigrechko, A., and I. A. Vakser. 2001. "How Common Is the Funnel-like Energy Landscape in Protein-Protein Interactions?" *Protein Science: A Publication of the Protein Society* 10 (8): 1572–83.

Vangone, Anna, Raffaele Spinelli, Vittorio Scarano, Luigi Cavallo, and Romina Oliva. 2011. "COCOMAPS: A Web Application to Analyze and Visualize Contacts at the Interface of Biomolecular Complexes." *Bioinformatics* 27 (20): 2915–16.

Volz, Erik, Verity Hill, John T. McCrone, Anna Price, David Jorgensen, Áine O'Toole, Joel Southgate, et al. 2021. "Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity." *Cell* 184 (1): 64–75.e11.

Walls, Alexandra C., Young-Jun Park, M. Alejandra Tortorici, Abigail Wall, Andrew T. McGuire, and David Veesler. 2020. "Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein." *Cell* 181 (2): 281–92.e6.

Walls, Alexandra C., M. Alejandra Tortorici, Joost Snijder, Xiaoli Xiong, Berend-Jan Bosch, Felix A. Rey, and David Veesler. 2017. "Tectonic Conformational Changes of a

Coronavirus Spike Glycoprotein Promote Membrane Fusion." *Proceedings of the National Academy of Sciences of the United States of America* 114 (42): 11157–62.

Wang, Xiao, Genki Terashi, Charles W. Christoffer, Mengmeng Zhu, and Daisuke Kihara. 2020. "Protein Docking Model Evaluation by 3D Deep Convolutional Neural Networks." *Bioinformatics* 36 (7): 2113–18.

Weisblum, Yiska, Fabian Schmidt, Fengwen Zhang, Justin DaSilva, Daniel Poston, Julio Cc Lorenzi, Frauke Muecksch, et al. 2020. "Escape from Neutralizing Antibodies by SARS-CoV-2 Spike Protein Variants." *eLife* 9 (October). https://doi.org/10.7554/eLife.61312.

Weitzner, Brian D., Roland L. Dunbrack Jr, and Jeffrey J. Gray. 2015. "The Origin of CDR H3 Structural Diversity." *Structure* 23 (2): 302–11.

Weitzner, Brian D., and Jeffrey J. Gray. 2017. "Accurate Structure Prediction of CDR H3 Loops Enabled by a Novel Structure-Based C-Terminal Constraint." *Journal of Immunology* 198 (1): 505–15.

Weitzner, Brian D., Daisuke Kuroda, Nicholas Marze, Jianqing Xu, and Jeffrey J. Gray. 2014. "Blind Prediction Performance of RosettaAntibody 3.0: Grafting, Relaxation, Kinematic Loop Modeling, and Full CDR Optimization." *Proteins* 82 (8): 1611–23.

Westen, Gerard Jp van, Remco F. Swier, Jörg K. Wegner, Adriaan P. Ijzerman, Herman Wt van Vlijmen, and Andreas Bender. 2013. "Benchmarking of Protein Descriptor Sets in Proteochemometric Modeling (part 1): Comparative Study of 13 Amino Acid Descriptor Sets." *Journal of Cheminformatics* 5 (1): 41.

Wibmer, Constantinos Kurt, Frances Ayres, Tandile Hermanus, Mashudu Madzivhandila, Prudence Kgagudi, Brent Oosthuysen, Bronwen E. Lambson, et al. 2021. "SARS-CoV-2 501Y.V2 Escapes Neutralization by South African COVID-19 Donor Plasma." *Nature Medicine* 27 (4): 622–25.

Wu, Aiping, Lulan Wang, Hang-Yu Zhou, Cheng-Yang Ji, Shang Zhou Xia, Yang Cao, Jing Meng, et al. 2021. "One Year of SARS-CoV-2 Evolution." *Cell Host & Microbe* 29 (4): 503–7.

Xia, Shuai, Meiqin Liu, Chao Wang, Wei Xu, Qiaoshuai Lan, Siliang Feng, Feifei Qi, et al. 2020. "Inhibition of SARS-CoV-2 (previously 2019-nCoV) Infection by a Highly Potent Pan-Coronavirus Fusion Inhibitor Targeting Its Spike Protein That Harbors a High Capacity to Mediate Membrane Fusion." *Cell Research* 30 (4): 343–55.

Xiong, Peng, Meng Wang, Xiaoqun Zhou, Tongchuan Zhang, Jiahai Zhang, Quan Chen, and Haiyan Liu. 2014. "Protein Design with a Comprehensive Statistical Energy Function and Boosted by Experimental Selection for Foldability." *Nature Communications* 5 (October): 5330.

Xu, Yingda, William Roach, Tingwan Sun, Tushar Jain, Bianka Prinz, Ta-Yi Yu, Joshua Torrey, et al. 2013. "Addressing Polyspecificity of Antibodies Selected from an in Vitro Yeast Presentation System: A FACS-Based, High-Throughput Selection and Analytical Tool." *Protein Engineering, Design & Selection: PEDS* 26 (10): 663–70.

Yang, Li, Mao Shu, Kaiwang Ma, Hu Mei, Yongjun Jiang, and Zhiliang Li. 2010. "ST-Scale as a Novel Amino Acid Descriptor and Its Application in QSAM of Peptides and Analogues." *Amino Acids* 38 (3): 805–16.

Yan, Yumeng, and Sheng-You Huang. 2019. "Pushing the Accuracy Limit of Shape Complementarity for Protein-Protein Docking." *BMC Bioinformatics* 20 (Suppl 25): 696.

Yuan, Meng, Hejun Liu, Nicholas C. Wu, Chang-Chun D. Lee, Xueyong Zhu, Fangzhu Zhao, Deli Huang, et al. 2020. "Structural Basis of a Shared Antibody Response to SARS-CoV-2." *Science* 369 (6507): 1119–23.

Yu, Fei, Rong Xiang, Xiaoqian Deng, Lili Wang, Zhengsen Yu, Shijun Tian, Ruiying Liang, Yanbai Li, Tianlei Ying, and Shibo Jiang. 2020. "Receptor-Binding Domain-Specific Human Neutralizing Monoclonal Antibodies against SARS-CoV and SARS-CoV-2." *Signal Transduction and Targeted Therapy* 5 (1): 212.

Zahradník, Jiří, Shir Marciano, Maya Shemesh, Eyal Zoler, Jeanne Chiaravalli, Björn Meyer, Orly Dym, Nadav Elad, and Gideon Schreiber. 2021. "SARS-CoV-2 RBD in Vitro Evolution Follows Contagious Mutation Spread, yet Generates an Able Infection Inhibitor." *bioRxiv*. https://doi.org/10.1101/2021.01.06.425392.

Zaliani, A., and E. Gancia. 1999. "MS-WHIM Scores for Amino Acids: A New 3D-Description for Peptide QSAR and QSPR Studies." *Journal of Chemical Information and Computer Sciences* 39 (3): 525–33.

Zavrtanik, Uroš, Junoš Lukan, Remy Loris, Jurij Lah, and San Hadži. 2018. "Structural Basis of Epitope Recognition by Heavy-Chain Camelid Antibodies." *Journal of Molecular Biology* 430 (21): 4369–86.

Zhang, Ning, Yuting Chen, Haoyu Lu, Feiyang Zhao, Roberto Vera Alvarez, Alexander Goncearenco, Anna R. Panchenko, and Minghui Li. 2020. "MutaBind2: Predicting the Impacts of Single and Multiple Mutations on Protein-Protein Interactions." *iScience* 23 (3): 100939.

Zhang, Yang. 2008. "I-TASSER Server for Protein 3D Structure Prediction." *BMC Bioinformatics* 9 (January): 40.

Zhou, Daming, Wanwisa Dejnirattisai, Piyada Supasa, Chang Liu, Alexander J. Mentzer, Helen M. Ginn, Yuguang Zhao, et al. 2021. "Evidence of Escape of SARS-CoV-2 Variant B.1.351 from Natural and Vaccine-Induced Sera." *Cell*, February. https://doi.org/10.1016/j.cell.2021.02.037.

Zost, Seth J., Pavlo Gilchuk, James Brett Case, Elad Binshtein, Rita E. Chen, Joseph P. Nkolola, Alexandra Schäfer, et al. 2020. “Potently Neutralizing and Protective Human Antibodies against SARS-CoV-2.” *Nature* 584 (7821): 443–49.

Zundert, G. C. P. van, J. P. G. L. M. Rodrigues, M. Trellet, C. Schmitz, P. L. Kastritis, E. Karaca, A. S. J. Melquiond, M. van Dijk, S. J. de Vries, and A. M. J. J. Bonvin. 2016. “The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes.” *Journal of Molecular Biology* 428 (4): 720–25.

Chapter 1 and 2 (pp. 22 to 59 and 103 to 133) of my doctoral thesis cannot be made public on the Internet for 5 years from the date of doctoral degree conferral because that parts are scheduled to be published within 5 years.

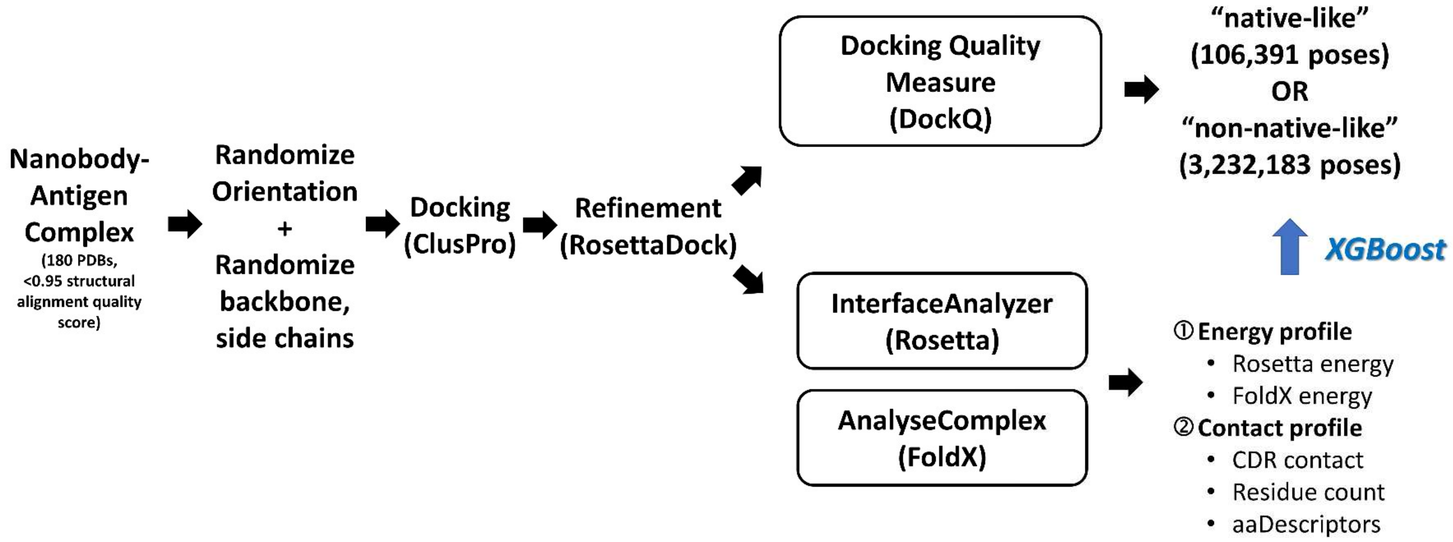
Nanobody-Antigen Complex
(180 PDBs, <0.95 structural alignment quality score)
Randomize Orientation + Randomize backbone, side chains
Docking (ClusPro)
Refinement (RosettaDock)
Docking Quality Measure (DockQ)
"native-like" (106,391 poses) OR "non-native-like" (3,232,183 poses)
XGBoost
InterfaceAnalyzer (Rosetta)
AnalyseComplex (FoldX)
① Energy profile
• Rosetta energy
• FoldX energy
② Contact profile
• CDR contact
• Residue count
• aaDescriptors

Figure 32: Workflow from data collection, data preprocessing to feature and label preparation for the nanobody pose prediction model.

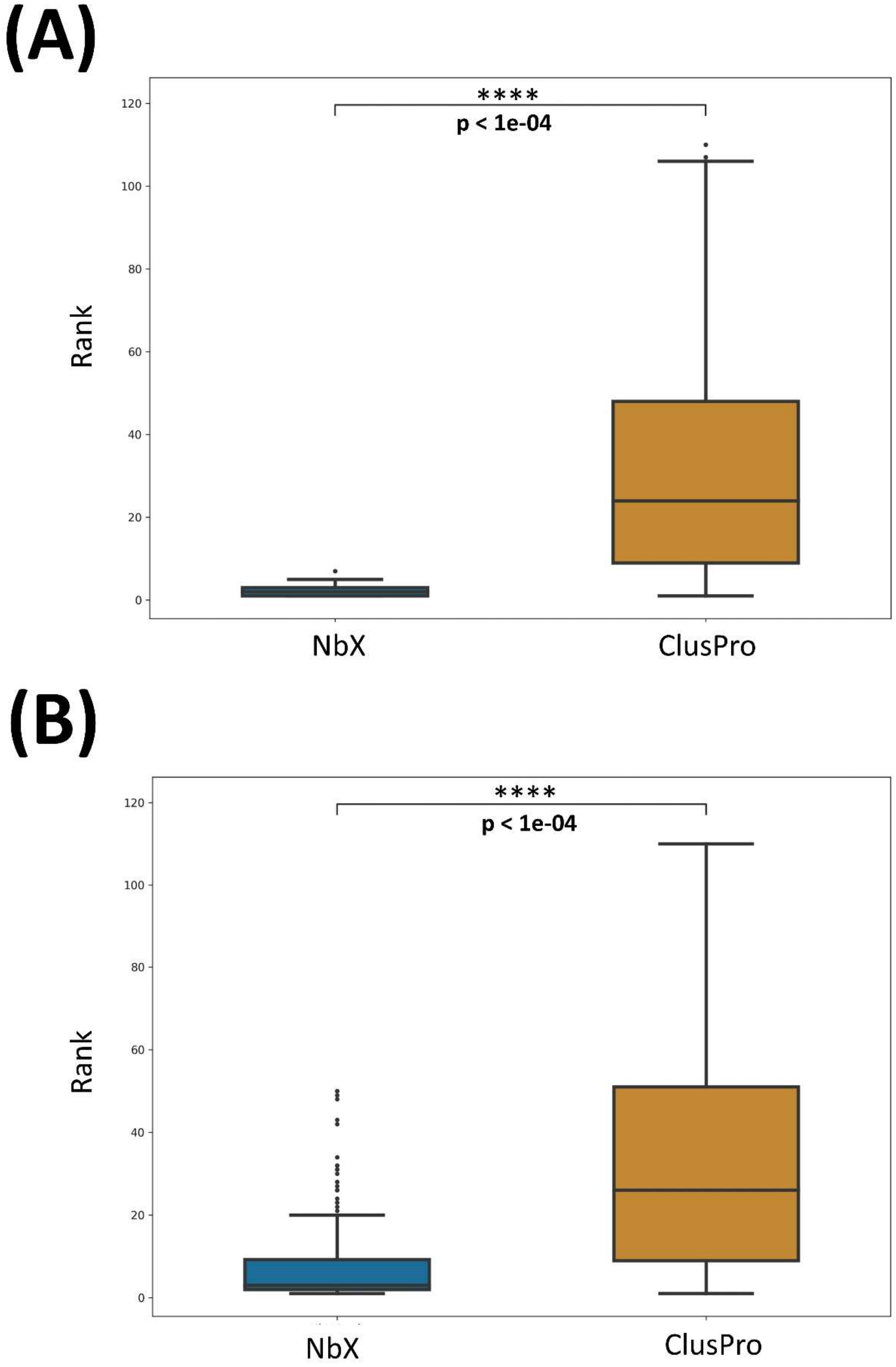

Figure 33: Prediction performance comparison between the nanobody pose prediction model (NbX) and ClusPro on the 5-fold validated prediction of (A) training set and (B) test set.

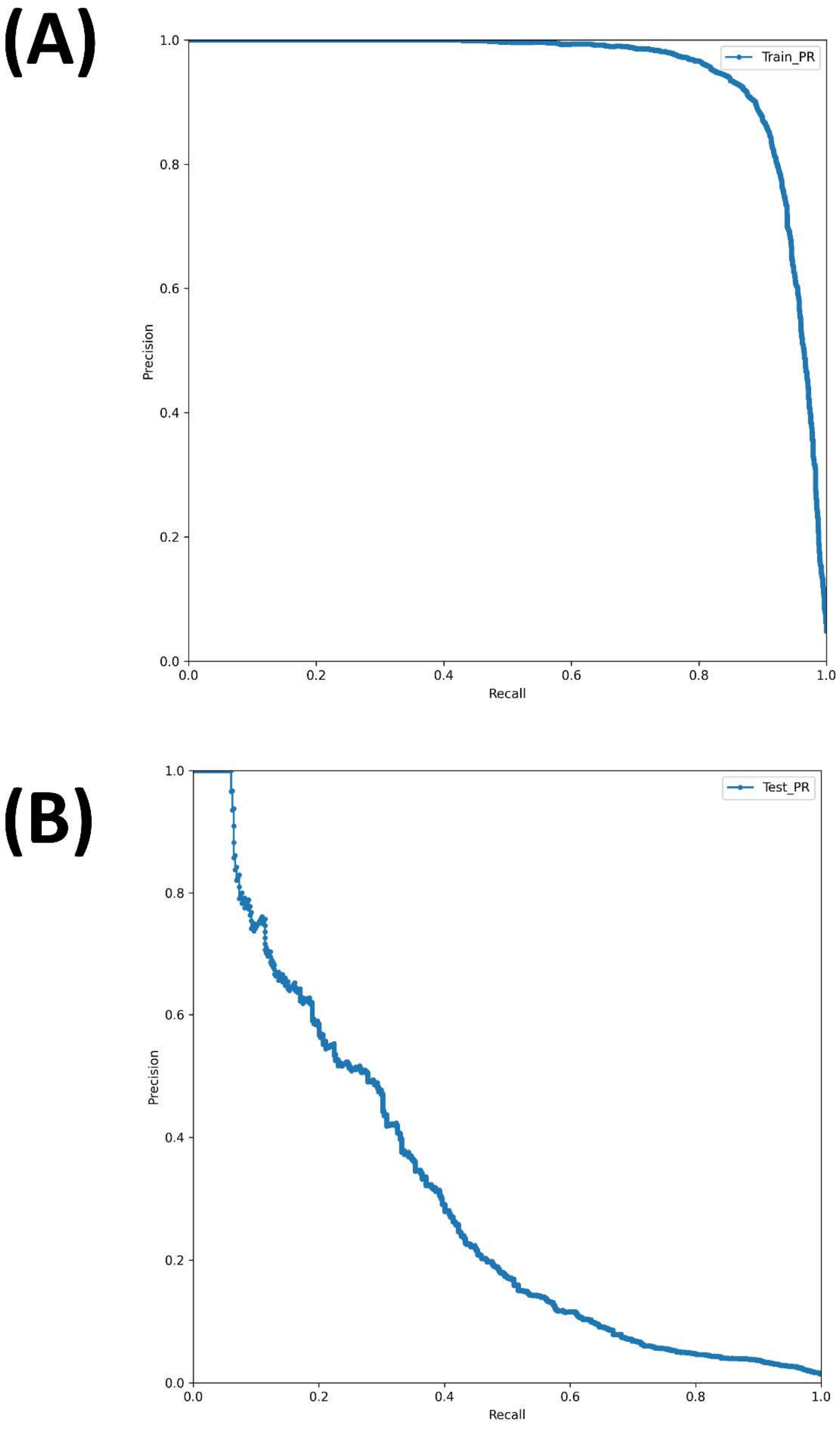

Figure 34: Precision-recall curve of the best single model on the prediction of (A) training set and (B) test set.

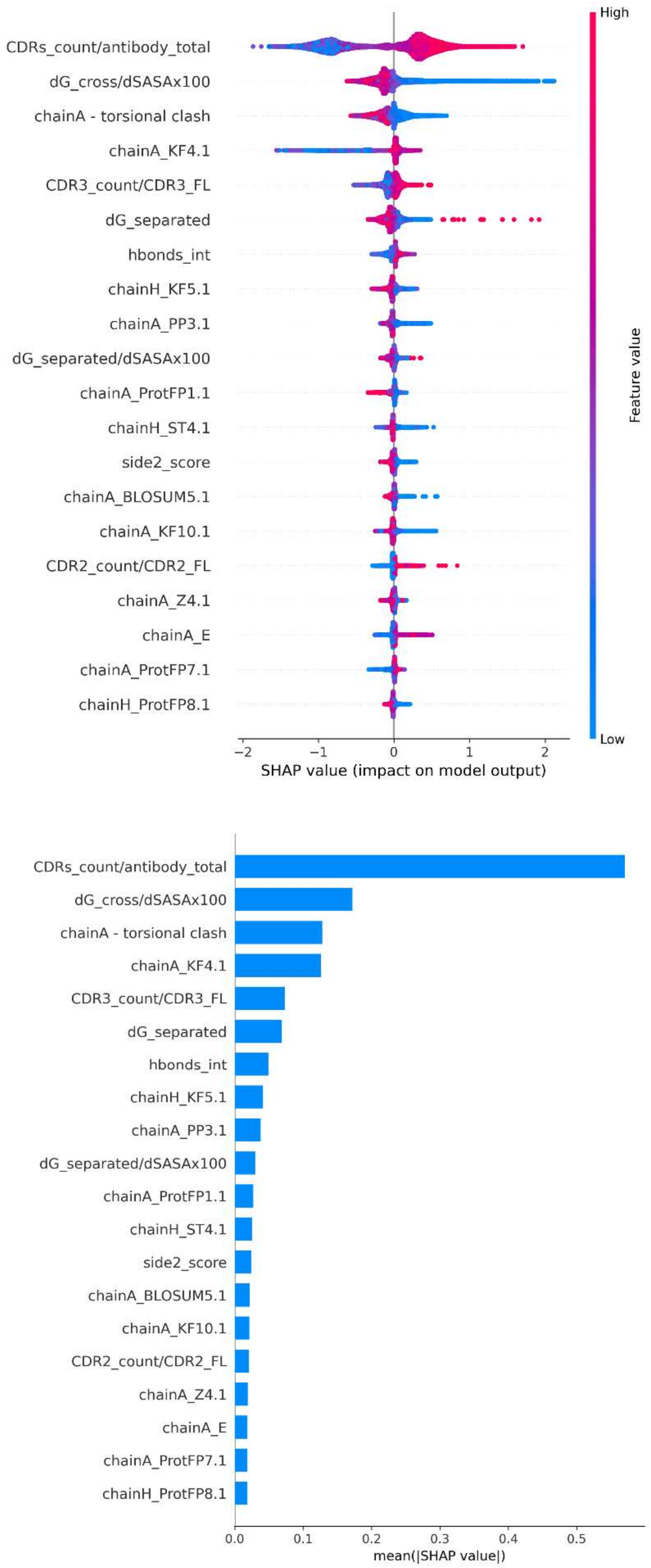

Figure 35: Features with top importance contributed to the prediction of test set of the best single model. (A) Summary plot showing SHAP values of individual predictions with the annotation of feature values. (B) Summary plot showing the mean(|SHAP|) of the features.

Table 1: Distribution of the three CDR loops of the 164 nanobody chains in PyIgClassify clusters. The number after hyphen of cluster names represents the length of CDR. The total number do not add up to 164 because there were unclassified CDR loops with unique structural conformations.

H1

	Count	% CDR
H1-10	6	4.1
H1-11	2	1.4
H1-13	132	91.0
H1-14	1	0.7
H1-16	2	1.4
H1-17	1	0.7
H1-18	1	0.7
total	145	100.0

H2

	Count	% CDR
H2-8	3	1.9
H2-9	38	24.7
H2-10	106	68.8
H2-11	2	1.3
H2-13	3	1.9
H2-15	2	1.3
total	154	100

H3

	Count	% CDR
H3-5	4	2.7
H3-6	3	2.0
H3-7	2	1.3
H3-8	3	2.0
H3-9	5	3.3
H3-10	5	3.3
H3-11	4	2.7
H3-12	6	4.0
H3-13	4	2.7
H3-14	18	12.0
H3-15	9	6.0
H3-16	11	7.3
H3-17	22	14.7
H3-18	9	6.0
H3-19	10	6.7
H3-20	13	8.7
H3-21	10	6.7
H3-22	5	3.3
H3-23	2	1.3
H3-24	1	0.7
H3-25	1	0.7
H3-26	3	2.0
total	150	100

Table 2: The combinatorial selection scheme used for selection of the 16 final ELMO1-targeting nanobody designs of the first-round design.

Criteria Sets	MM/PBSA	Funnel visual score	Min. ddg-binding score	Number of mutations	Number of redundant sequence	SASA
1	negative	min 6				
2	best 20					
3		min 7				
4			best 20			
5	negative	min 5		max 8		
6	best 20 (within max 8 mutation)			max 8		
7		min 7		max 8		
8			best 20 (within max 8 mutation)	max 8		
9	negative				min 8	
10		min 7			min 8	
11			best 20 (within min 8 redundant sequence)		min 8	
12						best 20

Table 3: The combinatorial selection scheme used for selection of the 20 designs from the second-round design. An additional of three designs were selected by PCA analysis of sequence space with the unselected designs, which sum up to a final number of 23 designs from second-round design.

Criteria set	MM/PBSA	FlexddG	Final sorting	Number of design selected
1	Best 50%	Best 50%	FlexddG	10
2	Best 5	-	MM/PBSA	5
3	-	Best 5	FlexddG	5

Table 4: Examples of SARS-CoV-2 S-targeting antibody currently developed.

	Target	**Blocks ACE2 interaction?**	**Neutralizing?**	**Type of antibody**	**Isolated or identified by**	**Best affinity achieved**
(Wrapp et al. 2020)	RBD	Yes	Yes	Nanobody	Llama immunization, IgG FC-fusion	Kd = 39 nM
(Nieto et al. 2020)	RBD	Yes	Yes	Nanobody	Alpaca immunization -> E. coli surface display	Kd = 0.63 nM
(Chi et al. 2020)	RBD	Yes	Yes	Nanobody (humanized)	Phage display	Kd = 0.7 nM
(Chi et al. 2020)	N-terminal domain (NTD)	No	Yes	Monoclonal antibodies (mAbs)	Isolated B cells from patients, sequenced VH and VL genes	EC50 = 0.607 ug/ml (in vitro neutralization of live SARS-CoV-2 in Vro-E6 cells)
(Zheng et al. 2020)	S2 (1048-1206)	(Not reported)	(Not reported)	Monoclonal antibodies (mAbs)	Previously identified panel of murine mAbs generated using SARS-CoV S (1029-1192)	(Not reported)
(Zost et al. 2020)	NTD, RBD, S2	(Not reported)	(Not reported)	Monoclonal antibodies (mAbs) (a large panel of antibodies)	B cell sorting, single-cell sequencing, cloning, recombinant expression, assays	(Not reported)
(Barnes et al. 2020)	RBD (most)	Yes	Yes	IgG (polyclonal) and Fab	Direct purification of IgG from patient plasma, protease digestion yielded Fab fragment.	(Not reported)

Table 5: Selection scheme of the S2-targeting nanobody.

Set	Criteria	Number of design selected
1	Lowest min. ddg-binding : Best 3	3
2	Lowest FlexddG (ddG) : Best 3	3
3	Lowest FlexddG (design dG) : Best 3	3
4	Good redock energy funnel shape : Best 3	3
5	Lowest total Score (REU) : Best 3	3
6	Highest average ΔSASA : Best 3	3
7	Lowest number of mutations : Best 3	3
Total		21

Table 6: Binary labeling of “native-like” and “non-native-like” and their corresponding CAPRI label and DockQ score ranges.

CAPRI	DockQ	Labeling
Incorrect	0.00 – 0.23	Non-native-like (0)
Acceptable	0.23 – 0.49	Native-like (1)
Medium	0.49 – 0.80	
High	0.80 – 1.00	

Table 7: Settings in modeling and benchmarking the nanobody pose prediction model.

Feature	Label	Set	Number of PDBs	Model	Classification Type	Metrics for picking best model	Validation Method	Ranking
Energy and Contact Profiles	“native-like” OR “non-native-like”	Training	80% (144 PDBs)	XGBoost (gradient boosted decision tree model)	Binary	PR-AUC$_{test}$	K-fold validation (k=5)	Mean classification probability (refined pose)
		Test	20% (36 PDBs)					

Table 8: Searching ranges of hyperparameters for optimization of model performance.

Hyperparameters	Search Range
“maxdepth”	4 – 12
“nestimators”	50 – 1000
“learningrate”	0.01 – 0.10
“subsample”	0.8 – 1.0
“colsamplebytree”	0.6 – 1.0
“gamma”	0 – 10