A benchmark study of protein folding algorithms on nanobodies

Shibo Liang, Ziquan Liang, Zecheng Wu, Bingding Huang, Xin Wang

College of Big Data and Internet, Shenzhen Technology University, Shenzhen 518118, China

Shibo Liang, Ziquan Liang and Zecheng Wu contributed equally to this work.

Abstract

Nanobodies, also known as single domain antibodies or VHHs, are the artificial recombinant variable domains of heavy-chain-only antibodies. Nanobodies possess many unique properties, such as small size, good solubility, superior stability, rapid clearance from blood and deep tissue penetration. Therefore, nanobodies have emerged as promising tools for the diagnosis and treatment of diseases.

In recent years, many deep-learning based protein structure prediction methods have emerged that require only protein sequences input to obtain predictions of 3D protein structures with high credibility. Among them, AlphaFold2, RoseTTAFold, DeepAb, NanoNet and tFold performed excellently in the field of protein prediction or antibody/nanobody prediction. In this paper, we selected popular algorithms such as AlphaFold2, RoseTTAFold, DeepAb, tFold and NanoNet, and compared their performance on the prediction of 3D folded structure conformation of nanobody proteins.

In order to compare the performance of several algorithms in the prediction of nanobody protein structure, we selected 60 samples with known experimental 3D structures in the PDB database, and used these five prediction methods to predict their 3D structure from 2D amino acid sequences. After the prediction data is obtained, these dry lab prediction results are compared with the wet lab experiment results in the PDB database one by one, and finally the results are analyzed and discussed.

Key words

Nanobody, benchmark, protein folding, algorithms

Introduction

Proteins are amino acids that are dehydrated and condensed to form polypeptide chains, which are then coiled and folded to form large biomolecules with a spatial structure. The 3D structure of proteins is essential for understanding their biological functions, and understanding how amino acid sequences determine the 3D structure of proteins is a major challenge in the science community, also known as the "protein folding problem" [1].

Among proteins, antibodies are a class of proteins that people most want to study because of their important functions. An antibody (Ab), also known as an immunoglobulin (Ig), is a large Y-shaped protein secreted mainly by plasma cells. In most mammals (which included human beings), an antibody consists of four polypeptide chains, which are linked by interchain disulphide bonds. Among them, the two chains with larger molecular weights are called heavy chains (H), and the two chains with smaller molecular weights are called light chains (L). The heavy chain and the light chain are all consisted of relatively stable region (called the constant region (C)) and relatively variable region (called the variable region (V)). An antibody can also be parted into two antigen-binding fragments (Fabs, containing one VL, VH, CL, and CH1 domain each) as well as the crystallizable fragment (Fc) [2].

The V regions of the heavy and light chains are called VH and VL, respectively, and each has three highly variable regions of amino acid called complementarity determining regions (CDRs), namely CDR1, CDR2 and CDR3, of which CDR3 is the most variable. The three CDRs of VH and VL together form the antigen-binding site of Ig, which determines the specificity of the antibody and is the site where the antibody recognizes and binds to the antigen [3,4]. Besides the three CDR regions, the remaining amino acid composition and order are relatively conserved in the V region, which are called the framework regions (FRs). four framework regions are represented by FR1, FR2, FR3 and FR4 for each of VH or VL, respectively.

In 1993, Belgian scientists discovered a new antibody in the serum of camelids (later also dromedaries and alpacas) [5]. The structure of the new antibody missed the light chain constant region, heavy chain CH1 constant regions as well as two light chain variable regions. It was only composed of two heavy chain variable regions as well as heavy chain CH2 and CH3 constant regions. This antibody has a different structure from conventional antibodies and is known as heavy-chain antibody (HCAb) [5,6].

The variable domain of heavy chain of heavy-chain antibodies is shorted for VHH [7]. The molecular mass of VHH, which prepared by in vitro recombinant protein expression, is only around

15 kDa (kilodaltons), which is about one-tenth of that of conventional antibodies. VHH retains full antigen-binding capacity of the antibody and thus is also known as nanobody (Nb) [8].

Compared to traditional antibody fragments such as fragment of antigen binding (Fab) and single chain antibody fragment (scFv), nano-antibodies have several distinct advantages, such as weak immunogenicity, low production cost, great water solubility, superior tissue permeability, stability and this allows nano-antibodies to be used in many other applications [8,9]. This allows nanobody to access many sites that are inaccessible to other antibodies, such as inside tumor cells [10] and through the blood-brain barrier [11].

Currently, protein structures are determined experimentally by techniques such as X-ray crystallography [12], cryo-electron microscopy (cryo-EM) [13] and nuclear magnetic resonance (NMR) [14], which are all expensive and time-consuming. Over the last 60 years, these labor-intensive efforts have determined the structures of about 170,000 proteins, out of over 200 million known proteins in all life forms [15]. If a method could be developed to deduce the structure of a protein from its amino acid sequence alone precisely, it would be a great help to many biomedical research fields such as antibody drug discovery.

In recent years, deep-learning based protein structure prediction algorithms have emerged, which can obtain 3D protein structure prediction results by simply inputting protein 2D sequences. Among them, AlphaFold2 (AF2) firstly reached a relatively high accuracy [16], and RoseTTAFold [30], which was open-source released almost at the same time as AlphaFold2, has also proved its strong ability. In addition, DeepAb, NanoNet and Tencent AI lab's tFold platform also claimed a good performance in the field of protein prediction or antibody/nanobody prediction.

AlphaFold2 [16], developed by the Google DeepMind team in London, is a deep-learning based program that won the protein structures competition in CASP 14 in November 2020 [17]. Differing from their experiment gained structures by about only one atom's width, AF2 reaching an equivalent level of prediction observed by humans using sophisticated instruments such as cryo-EM. Compared to the AlphaFold1(AF1) developed by the same team in 2018, AlphaFold2 overcomes the limitation that AlphaFold1 was only used to predict protein structures on the CASP 13 dataset [18]. Also, the predicted protein structures obtained by AF2 are much more accurate than AF1 [19].

The structure of AlphaFold2 can be divided into three parts: feature extraction, encoder and decoder. After inputting a one-dimensional protein sequence, AlphaFold2 looks up the database to

search for its homologous sequence and obtains a feature representation of the sequence and amino acid pair through a series of methods; afterwards, the MSA [20] and pair matrices are constructed by the encoder and the information in the two matrices is updated; finally, the encoder encodes the three-dimensional coordinates of the protein structure using relative positions.

There are four main features of AlphaFold2. First, the database used for its neural network training is very large, including up to 1.7T size BFD database built by the authors of the AF2 paper, along with database Uniref 90, MGnify and Uniclust 30, etc.[24,26-29], which greatly improves the accuracy of the prediction model; second, the whole model of Evoformer and structure module part both use recycling, that is, the output is rejoined to the input for the information refinement; third, Alphafold2 extensively uses the transformer structure [21], it is MSA or residue-residue pair information update all uses the attention mechanism; Finally, a BERT-like masking operation [22] is used to add noise to the various input information to require a stable output. This helps to improve robustness and generalization [23].

Behind AlphaFold2 is the support of huge databases. AlphaFold2 uses multiple sequence databases, among which，Big Fantastic Database (BFD) [24] is the biggest. BFD is one of the largest publicly available collections of protein families，which consists of 65,983,866 families represented as MSAs and hidden Markov models (HMMs) covering 2,204,359,010 protein sequences from reference databases, metagenomes and metatranscriptomes.

In different steps of the algorithm，AlphaFold2 uses distinct databases. In the step of finding template structures at prediction time, AlphaFold2 used PDB [25], and PDB70 [26] clustering database. The PDB70 contains profile HMMs for a representative set of sequences from the PDB, filtered with a maximum pairwise sequence identity of 70%. It contained 35 000 profile HMMs with an average length of 234 match states.

For MSA lookup during training and prediction, AlphaFold2 used Uniref90 [27] v.2020_01, BFD, Uniclust30 [28] v.2018_08 and MGnify [29] v.2018_12. The Uniref database (UniProt reference clusters) provides a set of sequence clusters from UniProt knowledge base and selected UniParc records to obtain complete coverage of sequence space at multiple resolutions (100%, 90% and 50% identity), while hiding redundant sequences. Uniref90 is constructed by clustering uniref100 sequences at 90% sequence identity level. The Uniclust30 database provide functionally homogeneous clusters of sequences at 30% sequence identity clustering depths, sets of representative sequences, MSAs of clusters, and annotations of all sequences with Pfam, SCOP,

and PDB matches.

MGnify provides a free to use platform for the assembly, analysis and archiving of microbiome data derived from sequencing microbial populations that are present in particular environments. For sequence distillation, AlphaFold2 used Uniclust30 v.2018_08 to construct a distillation structure dataset.

RoseTTAFold [30], which firstly published online to users in July 2021, developed by David Baker's team at the University of Washington. The network architecture of RoseTTAFold is similar to that of AlphaFold2, but claims better performance through a 3-track network in which information at the 1D sequence level, 2D distance map level and 3D coordinate level is sequentially transformed and integrated. In terms of database selection, AlphaFold2 differs from RoseTTAFold as Alphafold2 uses the BFD (1.7TB), pdb70 (56GB), uniref90 (58GB), and mgnif (64GB), params (3.5GB), small_bfd (17GB), pdb_mmcif (206GB) and uniclust30 (24.9GB) for a total database size of 2.2TB; RoseTTAFold, on the other hand, uses BFD (1.7TB), pdb100 (666GB) and Uniref30 (194GB) for a total database size of 2.6TB. [16,30] It can be seen that both are comparable in terms of the size of the databases used.

RoseTTAFold extends the AF2 two-track architecture with a third track, operating in three-dimensional coordinate to provide a tighter link between sequence, residue-residue distances/ orientations, and atomic coordinates. In this architecture, information flows back and forth between 1D amino acid sequence information, 2D distance metrics and 3D coordinates, enabling the network to reason together about relationships, distances and coordinates within and between sequences; in contrast, AlphaFold2 architecture only infers 3D structural coordinate after the 1D and 2D information processed in the neuron network. The network also enables rapid generation of accurate protein-protein complex models from sequence information alone, optimizing traditional modelling approaches (i.e., requiring modelling of individual subunits followed by docking). complex models from sequence information alone, short circuiting traditional approaches which require modeling of individual subunits followed by docking.) [31].
Although RoseTTAFold's innovative three-track model still does not perform as well as AlphaFold2 on the CASP14 due to training database and hardware impacts, RoseTTAFold still exhibits a fairly high level of predictive accuracy. Also, RoseTTAFold takes far fewer time and computer resources to produce predictions compared with AlphaFold2, which is a significant feature of RoseTTAFold.

DeepAb, developed by Jeffrey A. Ruffolo et al. at Johns Hopkins University, was published online

in June 2021. Its algorithm can be divided into two phases: the first phase is a deep residual convolutional network (ResNet) and Long Short-Term Memory Network (LSTM), which specifically predicts the relative distance and direction (2D structure) between pairs of residuals in the Fv region. The second stage is a distance and angle constrained approach based on the Rosetta minimization protocol for predicting 3D structures. The network requires only heavy and light chain sequences as input and is designed with interpretable attention components to provide insight into the prediction regions that the model focused on [32].

DeepAb was evaluated better than its original framework Deep H3 [33] on both the test set of therapeutically relevant antibodies and the RosettaAntibody benchmark set. [34].The DeepAb framework is derived from the CDR H3 ring structure prediction method, but builds on it with two innovations: One, to enrich the immunoglobulin sequences for supervised learning in the Fv region, DeepAb used a recurrent neural network (RNN) model to embed representations of general feature patterns in the immunoglobulin sequences, with a fixed implicit layer size of 64, allowing training on features that are not evident in a small subset of known structures, with the encoder in the RNN model using a bidirectional long and short-term memory network (Bi-LSTM) and the decoder uses a long and short-term memory network (LSTM) [35]. Secondly, DeepAb represents the variable region (Fv) structure as a set of inter-residue distances and orientations, specifically predicting inter-residue distances between three pairs of atoms and the set of inter-residue dihedrals and planar angles [32].

NanoNet, developed by Tomer Cohen et al. at the Hebrew University of Jerusalem, is the first modelling approach that optimized for VHH. The model is constructed using a large amount of data to train the neural network by using the VH domain of the antibody, the Vβ domain of the TCR and VHH [36].

NanoNet uses deep learning to make accurate end-to-end predictions of nanobody structures. While its primary goal is accurate Nb modelling, NanoNet can also accurately model the VH domains of antibodies and the Vβ domains of the TCR as they are included in the training data set. Its deep learning model accepts sequences (Nb, mAb VH domains or TCR Vβ domains) as input and generates coordinates for Cα atoms. Like AF2 and RoseTTAFold, NanoNet also employed direct end-to-end learning, allowing the network to learn complete 3D structures without the need to split the modelling problem into frames and CDR modelling [36]. This unique model that focusing only on Cα atomic coordinates allows it to predict antibody results with much better accuracy and significantly reduce computation time.

tFold is an AI-driven open platform for preclinical drug discovery and development, based on Tencent AI Lab's self-developed deep learning algorithms, with database and cloud computing support. On 17 November 2020, Tencent AI Lab used the "de novo folding" protein structure prediction method to help resolve the crystal structure of SRD5A2 [31] and effectively improved the accuracy of protein structure prediction through its self-developed algorithm "tFold" [37,38].

In order to accurately predict protein structures, tFold employs three innovative techniques to improve modelling accuracy. Firstly, the AI Lab team has developed 'multi-source fusion' technology to mine the co-evolutionary information in multiple sequence alignment (MSA); then, with the help of deep cross-attention residual network, tFold is then able to greatly improve the prediction of important 2D structural information, such as the residue-residue distance and orientation matrix. Finally, a novel 'Template-based Free Modeling' (TBFM) method is used to integrate the structural information in the 3D models generated by Free Modeling (FM) and Template-based Modeling (TBM), thus greatly improving the accuracy of the final 3D modeling [39].

To accurately and visually benchmark the above five protein prediction methods for nanoantibody prediction, we selected two popular methods to calculate and analyze the strengths and weaknesses of each method: RMSD values (Fig 1A) and TM-scores between experimental data (as the gold standard) and the prediction model, as well as drawing Rasch plots of the prediction model.

RMSD, root mean squared difference, is a parameter used to represent the difference in atomic positions between protein structures [40]. It is worth noting that the RMSD calculation needs to ensure that the atomic order and number in the two protein structure files are exactly the same, while the output of NanoNet is a protein structure file with only Cα atomic coordinates. This raises the problem of unequal number of atoms when comparing with the experimental file that containing all the atomic coordinates. Therefore, when calculating each model's RMSD of NanoNet with experimental value, we also deleted the non-Cα atomic coordinates in the experimental value file so that their RMSD value can be calculated.

In order to be able to add NanoNet to compare the differences in RMSD values between the five methods and the experimental values, we created copies of the protein structure files containing only Cα atomic coordinates of the full atomic protein structure prediction files obtained by the other four methods. We then recalculated their Cα-only RMSD values with the Cα-only experiment values

and obtained a new set of RMSD data with only Cα atomic coordinates. (Fig 1B)

The TM-score (template modeling score) is a measure of similarity between two protein structures [41]. The TM-score is designed to be a more accurate measure of the overall similarity of a full-length protein structure than the commonly used RMSD measure. The Rasch diagram is a method for visualizing the dihedral angles ψ and φ of the amino acid residues in the main chain of a protein structure. The Rasch diagram allows analysis of whether the conformation of the protein is plausible or not [42].

To compare the performance of the five algorithms on nanobody protein structure prediction, we selected 60 samples with experimentally known 3D structures in the PDB database, and used the five prediction methods mentioned above to input their 2D protein sequences. After obtaining a large number of three-dimensional atomic coordinate prediction data, we compare the prediction results of the five algorithms with the wet experimental values to obtain the RMSD values and the TM-score values. We will analyze and discuss our results below.

Results

The scheme of how we perform the all-atom protein model comparison with the four protein folding prediction methods could be displayed as Figure 1A. Firstly, we searched for qualified nanobodies in PDB database, and downloaded copies of their protein sequence files (FASTA format) and protein structure files (PDB format); Secondly, we import protein sequence files into AlphaFold2, RoseTTAFold, DeepAb and tFold, and use these structure prediction methods to get the protein structure prediction results, respectively (PDB format file); Finally, the protein structure prediction results are imported into the RamachanDraw script to obtain the Ramachandran plot, and the protein structure file and protein structure prediction file downloaded from the PDB are imported into the PyMOL script to obtain the RMSD and TM scores of the four methods.

Figure1. The schematic diagram of the two kinds of comparison.

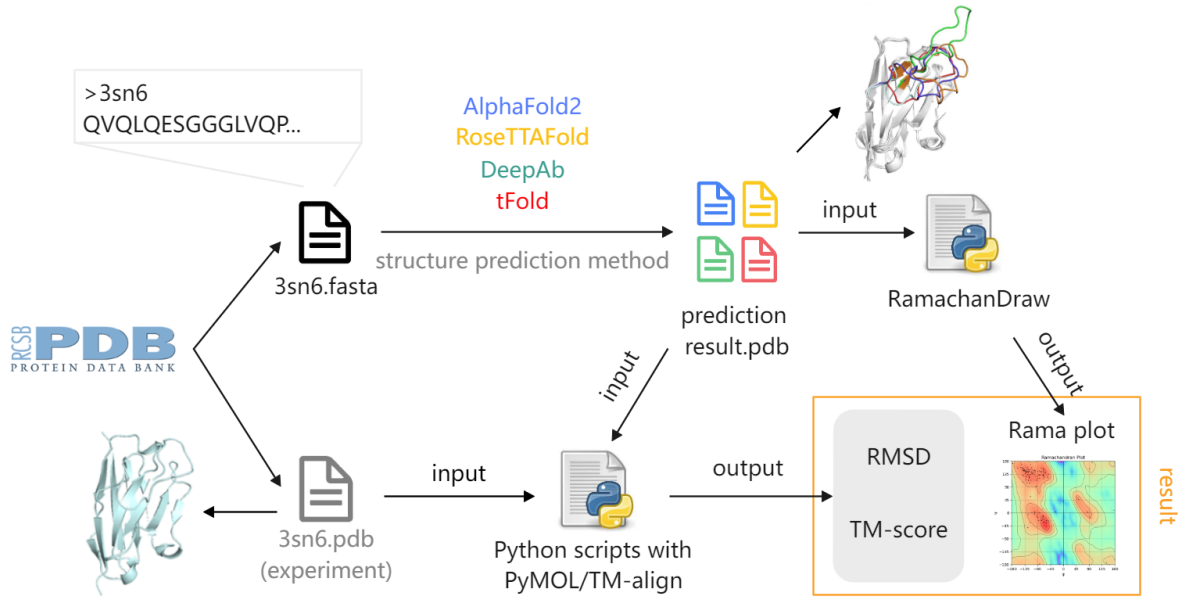

Figure1.A: Schematic diagram of calculating the scores of all-atom protein model with the four protein folding prediction methods.

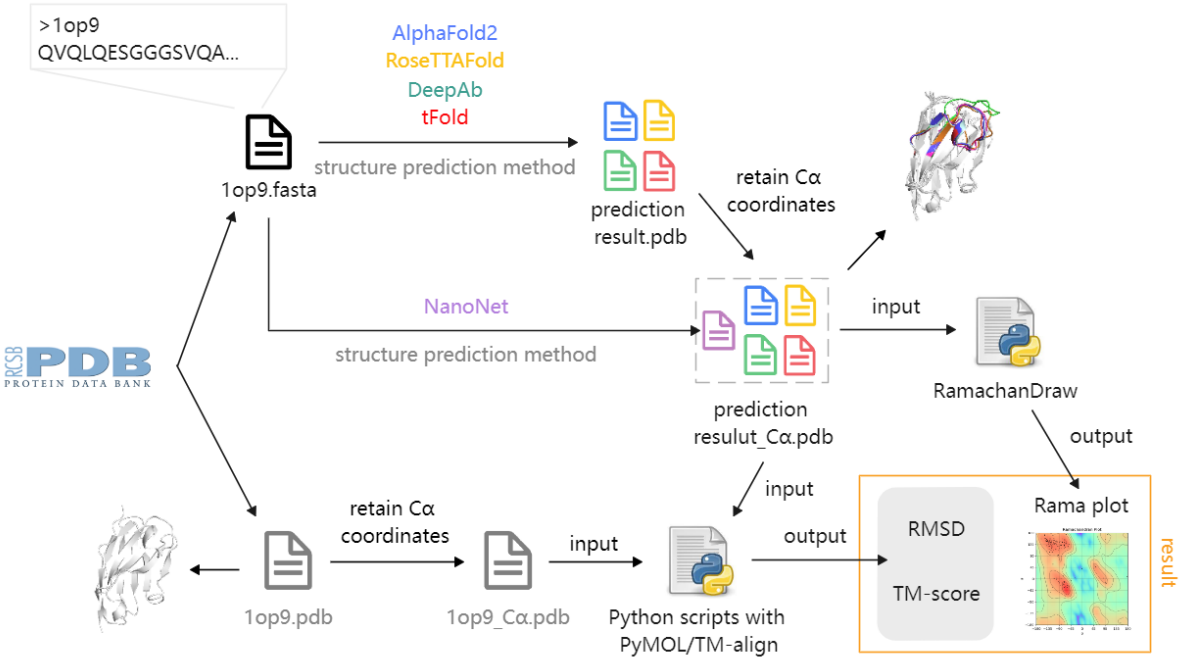

Figure1.B: Schematic diagram of calculating the scores of Cα-only protein model with the five protein prediction methods.

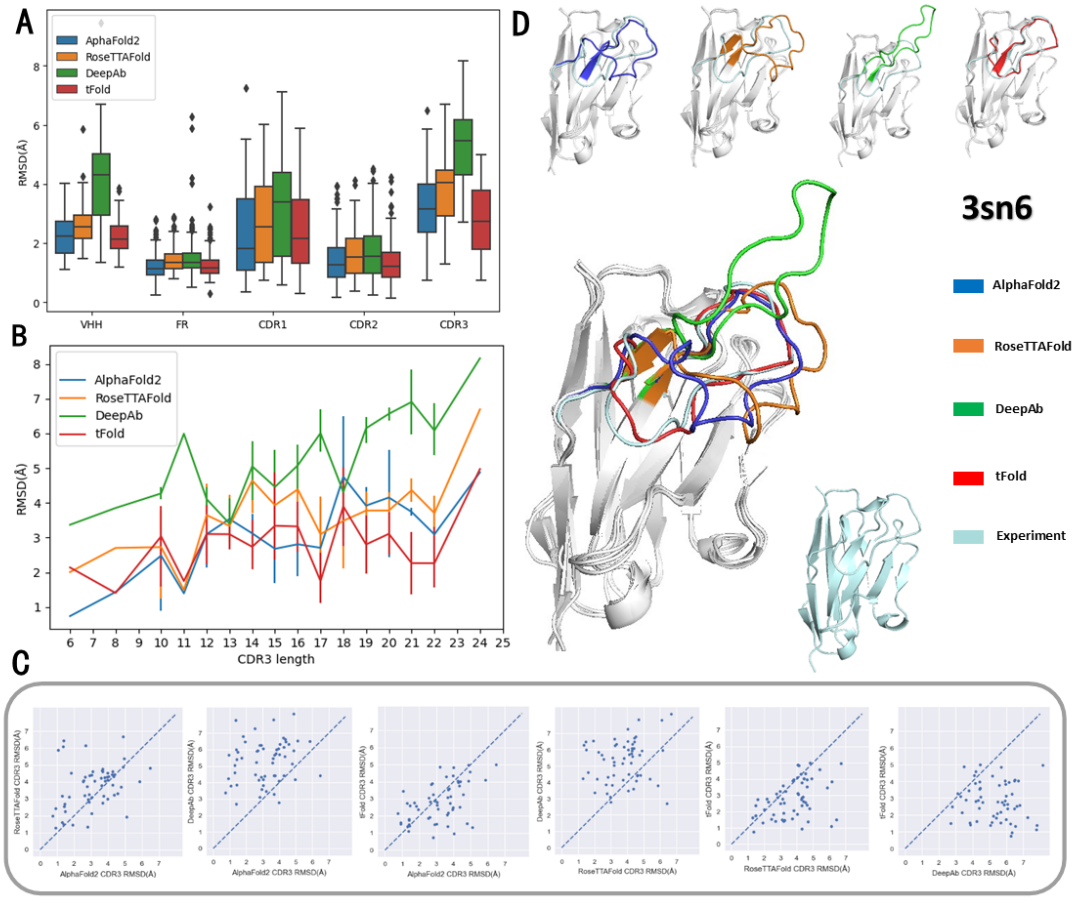

Figure 2: RMSD value and structure comparison of the four methods with the experimental structure (all-atoms comparison). (a) Boxplot of RMSD values on VHH full structure, FR frame regions, CDR1, CDR2 and CDR3 regions by the four algorithms (all atoms). (b) Comparison of CDR3 loops RMSD values with respect to the length of the CDR3 loops in the line graph by the four algorithms (all atoms). (c) Pair comparison of RMSD values of CDR3 loops by the four different algorithms in the scatter plot respectively (all atoms). (d) Visualization comparison of nanobody 3D structures in PyMOL (all atoms).

In the RMSD comparison, the worst RMSD results were obtained from DeepAb in the RMSD results (Fig. 2A, Supplementary Table 1) with a mean VHH RMSD of 4.19(±1.39) Å (see Table 1). In the frame region, the mean RMSD values of the four algorithms did not differ significantly. It can be seen that all the spatial structure of the frame region can be measured more accurately by all four methods compared with the CDR regions due to the relatively stable nature of the frame region (see Table 1).

In the CDR1 loops region, AlphaFold2 showed the best results with a mean value of 2.35(±1.54)

Å, followed by tFold2.44(±1.35) Å (Table 1), but also a relatively high outlier (3k3q and its CDR1 loop length is 19) of 7.25 Å can be seen for AlphaFold2 (see supplementary figure 1). The worst is DeepAb with a mean value of 3.14(±1.60) Å followed by RoseTTAFold with mean value of 2.80(±1.52) Å (Table 1).

Table 1: Mean RMSD value comparison of the four methods on whole VHH, Frame region, CDR1 region, CDR2 region and CDR3 region. (Full atom)

Method	VHH	FR	CDR1	CDR2	CDR3
AlphaFold2	2.25(±0.69)	1.19(±0.42)	2.35(±1.54)	1.52(±0.99)	3.17(±1.32)
RoseTTAFold	2.59(±0.77)	1.42(±0.40)	2.80(±1.52)	1.69(±0.95)	3.74(±1.27)
DeepAb	4.19(±1.39)	1.48(±0.64)	3.14(±1.60)	1.78(±1.05)	5.28(±1.28)
tFold	2.20(±0.59)	1.25(±0.39)	2.44(±1.35)	1.45(±0.96)	2.79(±1.14)

In the CDR2 regions, all of algorithm remain at a relatively low RMSD value and there is not significant difference between the four algorithms (Fig. 2A and Table 1). Also, from the statistics of the lengths of CDR1, CDR2 and CDR3, we found CDR2 regions got the smallest average length in our test data set (see supplementary figure 2 and supplementary table 4).

In the CDR3 regions, a large RMSD value difference begins to appear between the four methods. As the high variability nature of the CDR3 loops and the longer residue lengths compared to CDR1 and CDR2, CDR3 region is hard to predict. The best was tFold at 2.79(±1.14) Å, followed by AlphaFold2 at 3.17(±1.32) Å. The worst was DeepAb at 5.28(±1.28) Å, followed by RoseTTAFold with mean value3.74(±1.27) Å (Table 1).

In order to further study the relationship between CDR length and RMSD values, we also performed a comparison of the CDR3 loop's length with the resulting RMSD values (Fig. B). We can see that generally, there is a positive correlation between the length of CDR3 and RMSD values. The overall trend for the RMSD values increases with the length of the CDR3 loops and shows a sawtooth-like increase with tFold platform performed the best (smallest RMSD values) results and DeepAb performed the worst (largest RMSD values).

We also performed a one-to-one comparison of the RMSD values of the CDR3 loops for the four different methods, respectively (Fig2. C). In the comparison of AlphaFold2 with RoseTTAFold, it can be seen that the points are mostly concentrated around the diagonal, but overall AlphaFold2

has slightly better results than RoseTTAFold. In the comparison of AlphaFold2 with DeepAb, AF2 performed significantly better than DeepAb. When compare AlphaFold2 with tFold, tFold shows slight better than AlphaFold2. When RoseTTAFold versus DeepAb, RoseTTAFold overwhelmingly won the competition. In the comparison between RoseTTAFold and tFold, tFold had much better results than RoseTTAFold. When compare DeepAb with tFold, tFold won this comparison without doubt.

We also visualize the structure of the prediction results of each algorithm in PyMOL, high-lighted the CDR3 loop region, where DeepAb's prediction differed much from the structure of the experiment (Fig 2D). In contrast, the structure of tFold in the CDR3 loops was broadly consistent with that obtained from the experiments.

The scheme of how we perform the Cα-only protein model comparison with the five protein folding prediction methods could be displayed as Figure 1B. Firstly, we searched for qualified nanobody in the PDB database, and download copies of their protein sequence files (FASTA format) and protein structure files (PDB format); Secondly, the protein sequence files were imported into AlphaFold2, RoseTTAFold, DeepAb and tFold, and these structure prediction method was used to get the protein structure prediction results. Then, the obtained files were only retained Cα's coordinates after processing PDB files, thus, we get five kinds of software only including Cα Prediction results. Then, the protein sequence file is imported into Nanonet, and we use the Nanonet's structure prediction method to get the PDB file of Cα's coordinates. Finally, the protein structure prediction results are imported into the RamachanDraw script to obtain the Ramachandran plot, and the protein structure file and protein structure prediction file downloaded from the PDB are imported into the PyMOL script to obtain the RMSD and TM scores of the five methods.

As can be seen in the comparison of the boxplots of the Cα-only RMSD results (Figure 3A, Supplementary Table 2), DeepAb has the worst results. AlphaFold2, tFold and NanoNet all have better results. In the comparison of VHH RMSD, NanoNet performed best, but also with more outliers. In the frame region, the performance of the five methods was roughly the same, but there are also outliers, especially for DeepAb and NanoNet.

In the CDR1 ring region (Fig 3A), the best performer was NanoNet with a mean value of 1.32(±0.92) Å, followed by AlphaFold2 with a mean value of 1.37(±1.16) Å (Table 2). The worst

performer is still DeepAb with a mean value of 1.96(±1.16) Å, followed by RoseTTAFold with a mean value of 1.72(±1.03) Å, and in the middle is tFold with a mean value of 1.58(±0.97) Å.

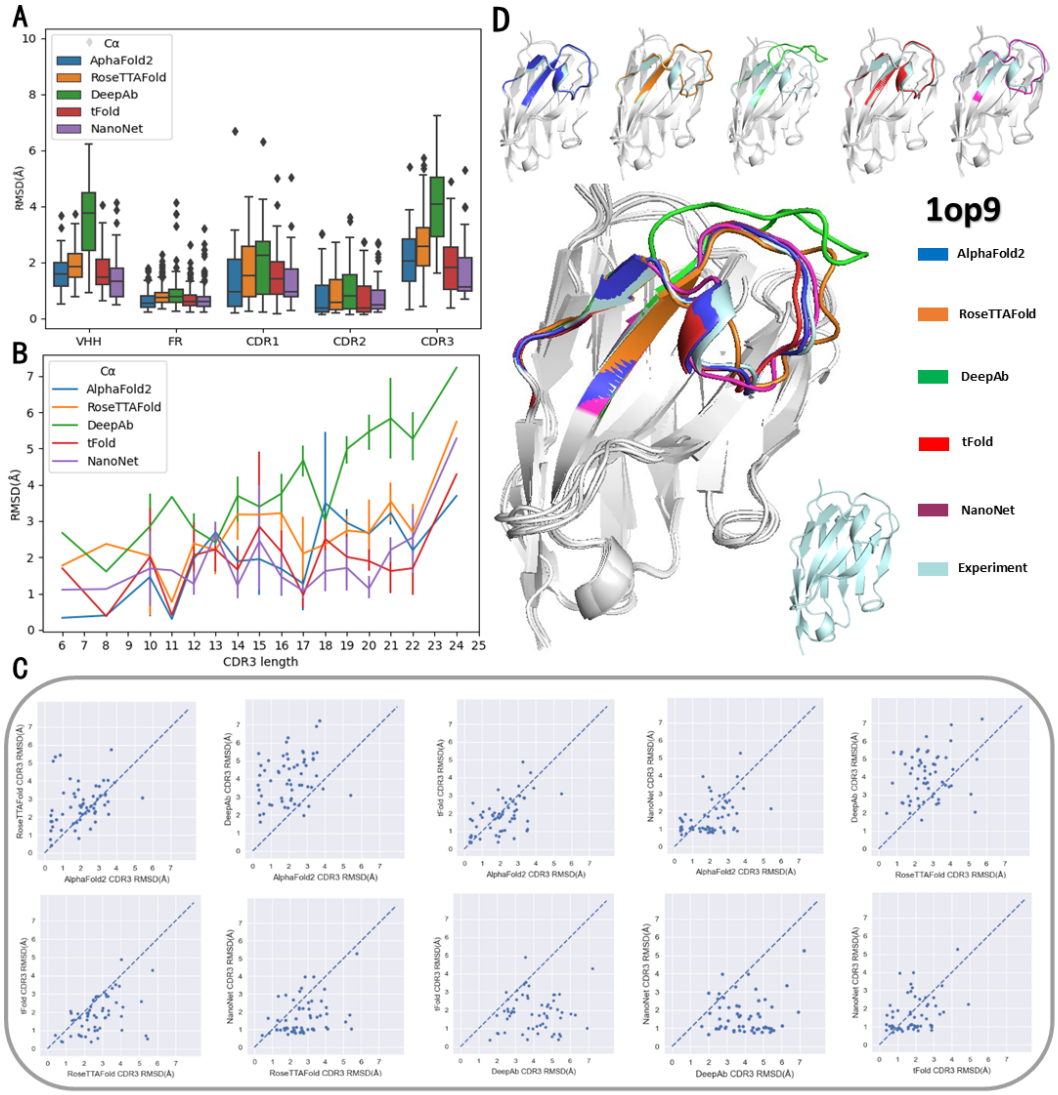

Figure 3: RMSD value and structure comparison of the five methods with the experimental structure (Cα-only comparison). (a) Boxplot of RMSD values on VHH full structure, FR frame regions, CDR1, CDR2 and CDR3 regions by the five algorithms (Cα-only). (b) Comparison of CDR3 loops mean RMSD values with respect to the length of the CDR3 loops in the line graph by the five algorithms (Cα-only). (c) Pair comparison of RMSD values of CDR3 loops by the five different algorithms in the scatter plot respectively (Cα-only). (d) Visualization comparison of nanobody 3D structures in PyMOL (Cα-only).

In the CDR2 ring region (Fig 3A), the best performer was tFold with a mean value of

0.72(±0.71) Å, followed by NanoNet with a mean value of 0.79(±0.67) Å. DeepAb had the worst performance with a mean value of 1.06(±0.86) Å, followed by RoseTTAFold with a mean value of 0.90(±0.71) Å, and ranked in the middle was AlphaFold2 with a mean value of 0.77(±0.77) Å.

Table 2: Mean RMSD value comparison of the five methods on whole VHH, Frame region, CDR1 region, CDR2 region and CDR3 region. (Cα-only atoms)

Method	VHH	FR	CDR1	CDR2	CDR3
AlphaFold2	1.60(±0.65)	0.65(±0.33)	1.37(±1.16)	0.77(±0.77)	2.07(±1.12)
RoseTTAFold	1.90(±0.66)	0.80(±0.34)	1.72(±1.03)	0.90(±0.71)	2.68(±1.13)
DeepAb	3.62(±1.52)	0.88(±0.51)	1.96(±1.16)	1.06(±0.86)	4.04(±1.28)
tFold	1.63(±0.66)	0.68(±0.34)	1.58(±0.97)	0.72(±0.71)	1.86(±0.99)
NanoNet	1.41(±0.82)	0.67(±0.42)	1.32(±0.92)	0.79(±0.67)	1.67(±0.97)

In the CDR3 ring region (Fig 3A), NanoNet has the most outstanding performance, and the mean value of NanoNet in the CDR3 ring is 1.67(±0.97) Å, but from Figure A, we can see that NanoNet also has an outlier with a large RMSD value in the CDR3 ring. Behind NanoNet is tFold with a mean value of 1.86(±0.99) Å. The worst performer is still DeepAb with a mean value of 4.04(±1.28) Å. In front of DeepAb is RoseTTAFold with a mean value of 2.68(±1.13) Å, and AlphaFold2 is in the middle with a mean value of 2.07(±1.12) Å. (Table 2)

In the graph of RMSD value versus CDR3 loop's length (Fig. 3B), NanoNet has better results. In the comparison of the CDR3 loops RMSD values of the results of different methods (Fig. 3C), NanoNet has better results, and the difference results between NanoNet and tFold and AlphaFold2 are not large, and these points are mostly concentrated in the diagonal.

In the comparison of DeepAb with the other methods, the results of DeepAb are always poorer. When compare between AlphaFold2 and RoseTTAFold, AlphaFold was slightly better than RoseTTAFold. In the comparison between AlphaFold2 and tFold, the difference between the two results is not too big. For the comparison of AlphaFold2 and NanoNet, NanoNet gave slightly better results than AlphaFold2. In the comparison between RoseTTAFold and tFold, the results of tFold were better. When RoseTTAFold versus NanoNet, NanoNet was better than RoseTTAFold. Finally, in the comparison between tFold and NanoNet, NanoNet got

slightly better results.

In the PyMOL structure visualization (Fig. 3D), we also high-lighted the CDR3 loop region. The predictions of AlphaFold2, RoseTTAFold, tFold and NanoNet were all relatively close to the experimental structure, with the structure in the CDR3 loop region of DeepAb being far from the experimentally obtained structure.

In the comparison of full atom methods, tFold performed best in the TM-score followed by AlphaFold2 tightly. DeepAb also performed worst with RoseTTAFold in the middle (Fig 4A). This result is consistent with the RMSD evaluation. In the comparison of ca-only atom methods (Fig 4B), It can be seen from the results that the overall results of NanoNet exceeded all other methods in the ca-only comparison, but there was outlier of NanoNet with low scores (Supplementary Table 3).

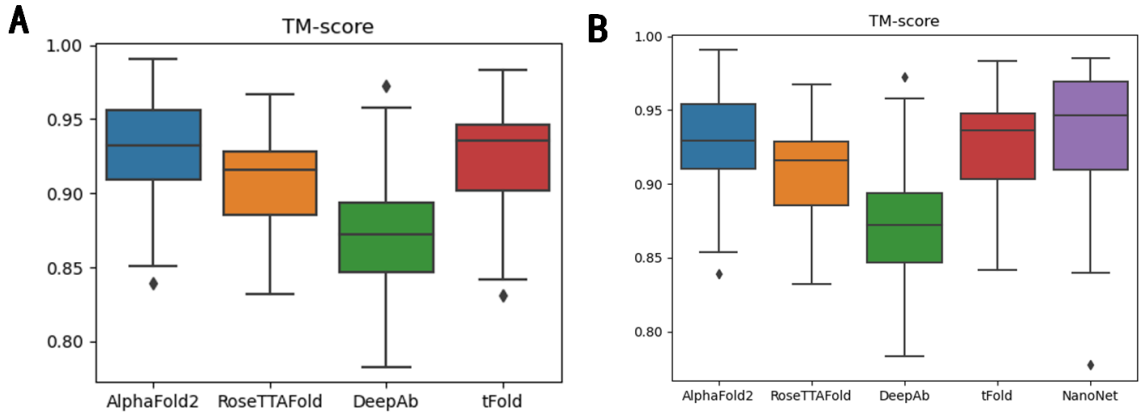

Figure 4. Comparison of boxplots of TM-score of each algorithm by the all-atom structure and the Cα-only structure. (a) Boxplots of TM-score of each algorithm by the all-atom structure. (b) Boxplots of TM-score of each algorithm by the Cα-only structure.

What's more, in the visualization of the predicted results we found that RoseTTAFold did not have relatively good results when the CDR3 ring of certain nanobodies had a β-turn or a similar structure to the β-turn (see Supplementary Figure 3A). This may be an important reason for the poorer results of RoseTTAFold compared to AlphaFold2.

Uncertainty of DeepAb's prediction results

We used DeepAb to make three predictions of the amino acid sequence of the same single heavy chain nanobody and compared the results of the three predictions of DeepAb. We found a certain randomness in the results of DeepAb's predictions compared with other methods that are

all with good reproducibility (see supplementary Figure 3B). This randomness was more obvious in the CDR3 ring and sometimes in other regions as well. We do not know whether it is a phenomenon that occurs in the selection of the optimal solution of the results by DeepAb or a prediction bug of the code of DeepAb.

Discussion

1. **Performance on predictive accuracy**

The accuracy of the five deep learning algorithms used in the benchmark study of nanobody folding algorithms for predicting the VHH were: NanoNet > (better than) tFold > AlphaFold2 > RoseTTAFold >> (much better than) DeepAb based on RMSD value (Fig 2A, 3A). The main prediction errors were all on the CDR loop regions, in particular CDR3 loops, which has the greatest impact over the overall prediction results, then followed by CDR1 loops.

In this highly variable region, AlphaFold2 surprisingly did not give the best results (Fig 2B, 3B). Meanwhile, the DeepAb algorithm, which is designed for traditional antibody Fv structure prediction showed the worst results. However, the NanoNet algorithm gave the best results for both the overall framework as well as the local variable regions (Fig 3A and B, Table 2). It can be suggested that to absorb nanobody folding database or/and algorithms such as NanoNet would be a possible way to improve protein folding algorithms such as AlphaFold2 in the performance of the prediction of highly variable regions. We also got the same ranking order in TM-score comparison (Fig 4A and B).

2. **Time spending**

It is also worth noting that the prediction speed of NanoNet folding algorithm is the fastest in this model prediction. We input 60 sequences and the prediction only takes 4.65 seconds in all as NanoNet can process the prediction paralleled.

In contrast, the average time to predict a single nanobody sequence using DeepAb on a local server with four NVIDIA Tesla V100 SXM2 GPU in single-chain mode was around 3 minutes. The average time taken to predict a single nanobody using AlphaFold2's Colab notebook was around 5 minutes (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb). The average time for a single nanobody sequence prediction using RoseTTAFold's online

platform Robetta (https://robetta.bakerlab.org, last accessed date: 14 July 2022) is around 1~55 minutes, and each account is limited to 20 predictions per day.

The time for a single nanobody sequence prediction using the tFold platform （https://drug.ai.tencent.com/console/en/tfold?type=predict, last accessed date: 28 July 2022） is usually 20~30 minutes and sometimes takes 9 or 16 hours. Inconveniently, there are only 10 tasks can be submitted per account per day and only 1 result can be downloaded per 24 hours for the tFold platform at the time we submit this paper, which is a great disadvantage of this online platform and we hope Tencent AI lab would allow more task to be submit and download each day in future (Supplementary Figure 4).

3. Robustness

At the functional level used by the algorithms, NanoNet only supports the prediction of the spatial position of the five atoms [N, Cα, C, O, Cβ] of each amino acid backbone structure of the nanobody (VHH) formed by the variable region VH of the heavy chain of the common Y-shaped lgG antibody. DeepAb supports the prediction of the VL and VH double chains of the variable region (Fv) of the antibody or only the single chain VH or VL. Other three folding algorithms AlphaFold2, RoseTTAFold and tFold support the prediction of the entire atomic structure of the protein antibody.

Meanwhile, we also tried a non-nanobody protein haemoglobin (PDB ID: 2wy4) as the input of the five algorithms to test the robustness of each algorithm, we found that NanoNet was almost unusable for this haemoglobin protein prediction (RMSD=14.08Å), while DeepAb's prediction result was even completely unreasonable (RMSD=35.14Å), but AlphaFold2, RoseTTAFold and tFold all could predict its 3D structure correctly (RMSD < 3Å) (Supplementary Figure 5).

In addition, in the folding prediction experiments for nanobody 1i3u, we found that DeepAb, RoseTTAFold and tFold all did not allow wildcard character 'X' in the amino acid sequence. This 'X' would raise ‘ValueError’ and made the program stopped. So, we delete this invalid character 'X' in the fasta sequence and solved this problem. Meanwhile, in the character study of input amino acid sequences, we found that some of the nanobody fasta files downloaded from the PDB database had the Polyhistidine-tag [43,44] at the end of their sequences. If the files with these Polyhistidine-tag were directly used for folding structure prediction, the prediction error from NanoNet would be RMSD=2.87(±1.97) Å, which is much higher than RMSD=1.38(±0.82) Å obtained by deleting the tag.

4. Relevance of the predicted outcomes and algorithm differences

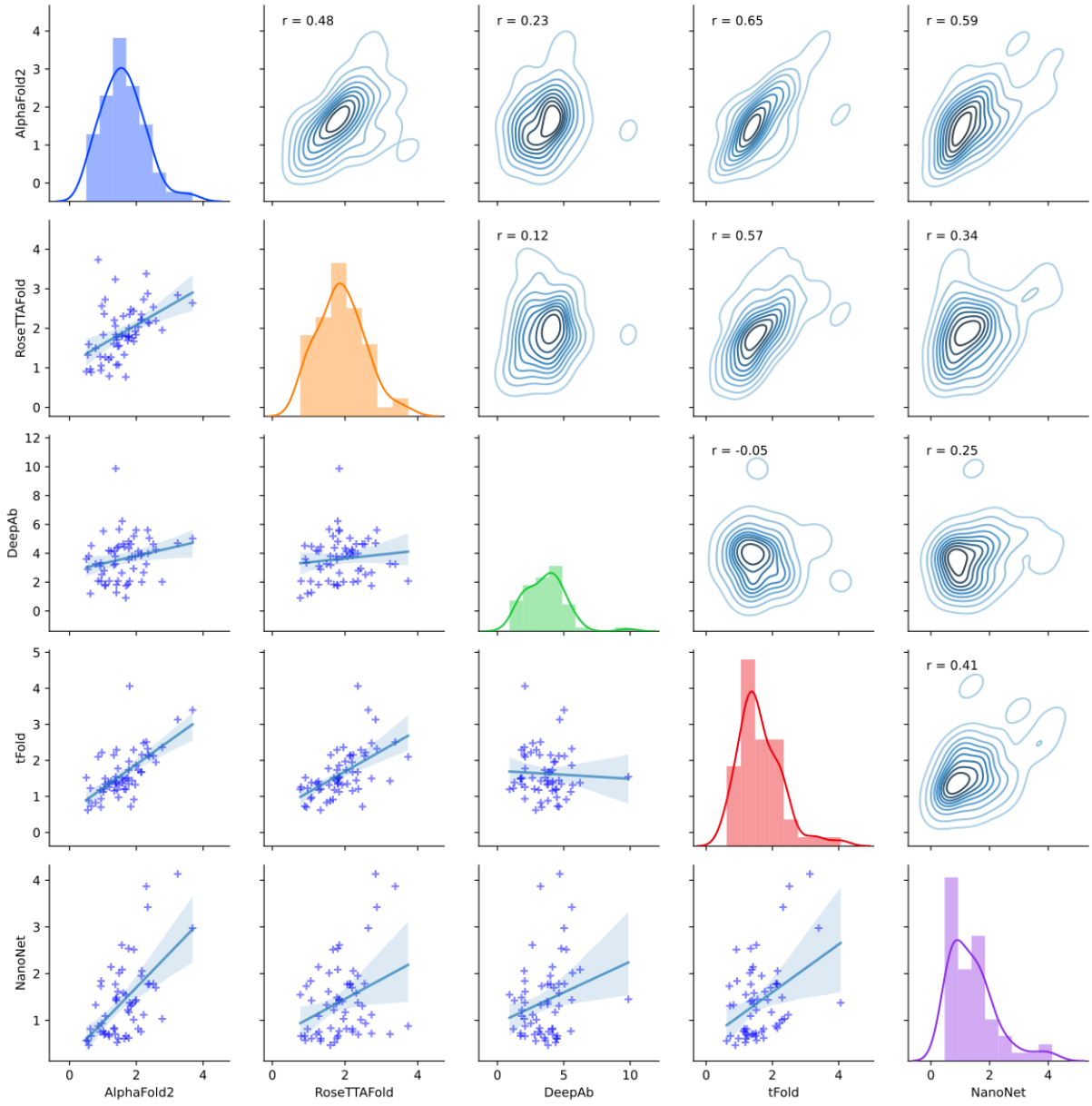

Figure 5. Pair plot of Pearson correlation of predicted nanobodies between 5 protein folding algorithms. Below the diagonal of the matrix, the scatter plot with the linear correlation is performed for the RMSD value of pairwise protein folding algorithms. Above the diagonal of the matrix, the kernel density of the data is plot for the scatters of the pairwise protein folding algorithms. Also, the correlation value r was calculated and marked above. It can be seen that the kernel density of two algorithms with higher correlation is more concentrated around diagonal in shape (as shown of the nuclear density map of AlphaFold2 and tFold in upper right of Figure 5, r=0.65). The RMSD density bar graph of each algorithm is plot on the diagonal.

Correlation analysis of RMSD results would be a simple but efficient way to help us to see if there is a close relationship between every two algorithms, as nowadays the machine learning methods, especially the protein folding related deep learning methods are all more or less black boxed for their core code. As a result, as Figure 5, AlphaFold2 had the highest correlation with tFold (r=0.65); followed by AlphaFold2 vs. NanoNet (r=0.59), RoseTTAFold vs. tFold (r=0.57); and then AlphaFold2 vs. RoseTTAFold (r=0.48).

AlphaFold2 performs best on frame region in both total-atom and Cα-only structure predictions. Meantime, AF2's predictions on both CDR1 and CDR2 are still in the top set, but on CDR3 prediction, AF2's performance was just in the moderate level. Overall, in terms of accuracy, AlphaFold2 is good for predicting the overall structure of all-atom proteins, especially on stable regions such as frame regions of antibodies.

RoseTTAFold and AlphaFold2 are both the improvements of the AlphaFold1 algorithm [30]. Their algorithm structures are partially similar according to the algorithmic framework described in the articles; however, the correlation of their results was not very high. We suppose that its added 3D-track in the 3-track block as well as the difference of the database contribute in this difference as we described in the introduction part of this paper.

In our experiment, the best result of full-atom nanobody folding algorithms was not obtained by these two algorithms. The tFold algorithm developed by Tencent AI lab, which was released earlier than the two, unexpectedly obtained better results in the full-atom structure comparison [37, 38]. Its substantive folding step is also divided into the three steps of Homology modelling like AlphaFold [16]. From the correlation values, the prediction results of the algorithm are very similar to AlphaFold2 and RoseTTAFold. Also, the accuracy of tFold's prediction for CDR3 regions are significantly better than AlphaFold2 and RoseTTAFold, which suggests tFold may especially optimized for highly variable regions such as CDR3. This is also an important point for antibody drug discovery and conform to Tencent's AI-aid drug development aim.

NanoNet is a lightweight folding algorithm that based on ResNet. However, its prediction results are very similar to those of AlphaFold2 which used more sophisticated structures with attention algorithm. The data in Table 2 show that it has the highest accuracy especially in CDR3

regions, which proved its ability in the nanobody folding structure prediction. At the same time, when it comes to the predicting time costs, the lightweight NanoNet achieved remarkable fast speed, owing to its simplified Cα-only structure.

The folding algorithm with the worst results in this study was DeepAb. Like its previous version framework DeepH3 [33], they are both folding algorithms designed for the traditional antibody Fv regions and are not well suited for predicting the folding structure of nanobodies. We are not so shocked by this result as its algorithm is mainly developed for double-stranded antibodies, and its training datasets is also mainly double-chained antibodies. Although there is a single-chain mode in DeepAb, it has performed poorly in the predictions of nanobodies. It also demonstrated that the algorithm and the training data set have a great influence on the accuracy of the prediction results.

Perspectives

In this new emerging field, new algorithms will continue to emerge, which could be either better in algorithm architecture, or/and faster in prediction speed, and better accuracy in terms of RMSD value. We look forward to more new algorithms, new architectures with higher accuracy as well as faster speed for the protein and nanobody 3D structure predictions, especially for the high-variability region predictions such as CDR3 regions.

Methods

Obtaining the pdb experimental file:

From PDB official website [25], we were able to obtain the amino acid sequences of the nanobodies and the experimental pdb files of the nanobodies with the duplicated entries removed and only kept the best resolution ones. Then, we input these amino acid sequences of nanobodies to the five structure prediction algorithms respectively to generate the perdition pdb files. Finally, the wet-lab experimental pdb files were used as gold standard to compare with the prediction results pdb files to calculate RMSD and TM-score.

Calculation of RMSD

We input the experimental pdb files and the prediction results pdb files into PyMOL [45,46] (https://github.com/schrodinger/pymol-open-source) to calculate the RMSD value between gold standard and each algorithm's prediction result. PyMOL can automatically align the atoms between the predicted model pdb results and the experimental pdb file. When calculating RMSD with the default outlier rejection cutoff as 2.0 Angstrom in PyMOL's align parameter, the atoms in the two input pdb files that have a large difference in distance are removed and then the RMSD calculation is performed [45,46]. This default filtering threshold would remove the outlier automatically, but this can lead to an undistinguishable RMSD results that deleted all abnormal values. In order to ensure fair and reasonable results and to keep the maximum number of atoms in the experimental pdb file for the RMSD calculation for the benchmark, we then set this cutoff value to 10 in the PyMOL's align method to ensure that we do not delete the atoms with large differences up to 10, so that we can calculate the RMSD value with all atoms included to ensure a more accurate and reliable benchmark result.

CDR rings segmentation by the IMGT scheme

The nanobody protein sequence is divided into 7 regions, FR1, CDR1, FR2, CDR2, FR3, CDR3, and FR4 by the annotation method provided by IMGT [47].

Pdb file partitioning based on CDR loops

The annotation method provided by IMGT scheme divided the amino acid sequences of nanobody proteins into a total of seven different regions of FRs [1-4] and CDRs [1-3]. Based on the information of the amino acid sequences of these seven regions, we also segmented the experimental pdb files obtained from the official website of PDB into seven different regions of FRs and CDRs. We also split the processed experimental structure pdb file with only Cα atoms

and the experimental structure pdb file obtained from the official website of the PDB into CDR regions. After segmentation, each frame or CDR loop region generates a separate pdb file. That is, each nanobody pdb file was divided into seven pdb files with different FR or CDR regions. Since the results of NanoNet only have the 3D coordinates of Cα atoms [36], in order to compare with NanoNet, we processed the pdb files to retain only Cα atoms, only the information of the five atoms [N, CA, C, O, CB] was kept in the pdb file.

Calculation of the RMSD of the CDR loops:

PyMOL allows us to automatically align full-atoms and residues together to calculate RMSD, which can help us to compare a part of the structure from experiment to the whole structure in the database automatically. We directly calculate the RMSD value by comparing the whole pdb files of the prediction results of various algorithms with the experimental structure pdb files of each region of the segmented CDR, which is consistent with the result of first performing the CDR partitioning of the pdb files of the prediction results of the various algorithms and then performing the RMSD calculation with the partitioned pdb files from experiments.

TM-score calculation

The TM-score is calculated using the TM-align [41]. The experimental pdb file and the predicted pdb file are passed into the TM-align method to calculate the TM-score.

Comparison of Nanonet with other methods:

Since the output of NanoNet is only the coordinates of Cα, but the output of other methods does not only have Cα, in the comparison between NanoNet and other methods, we processed the file and only retained the coordinates of Cα for RMSD calculation; only the information of the five atoms [N, CA, C, O, CB] was retained in the pdb file. Although this method deleted other atoms on a residue, we could still obtain the basic spatial structure of the predicted results, and only retaining the Cα atom has little effect on the spatial structure of the Nanobody. This transformation did not change the 3D coordinates of atoms in the pdb file, only the visualization results in PyMOL is sometimes a little bit different from those of the full atoms (see Supplementary Figure 6).

Ramachandran plot

We used Drug RamachanDraw to draw the Ramachandran plot (https://github.com/alxdrcirilo/RamachanDraw) [48].The Ramachandran plot can be drawn by passing the pdb file of the prediction result into the Ramachandran plotting tool in Python.

References

[1] Dill, K. A. & MacCallum, J. L. The protein-folding problem, 50 years on. science 338, 1042–1046 (2012).

[2] Chiu, M. L., Goulet, D. R., Teplyakov, A. & Gilliland, G. L. Antibody Structure and Function: The Basis for Engineering Therapeutics. Antibodies 8, 55 (2019).

[3] Chothia, C. & Lesk, A. M. Canonical structures for the hypervariable regions of immunoglobulins. Journal of molecular biology 196, 901–917 (1987).

[4] Kabat, E. A. & Wu, T. T. Attempts to locate complementarity-determining residues in the variable positions of light and heavy chains. Annals of the New York Academy of Sciences 190, 382–393 (1971).

[5] Hamers-Casterman, C. et al. Naturally occurring antibodies devoid of light chains. Nature 363, 446–448 (1993).

[6] Hassanzadeh-Ghassabeh, G., Devoogdt, N., De Pauw, P., Vincke, C. & Muyldermans, S. Nanobodies and their potential applications. Nanomedicine 8, 1013–1026 (2013).

[7] Nguyen, V. K., Hamers, R., Wyns, L. & Muyldermans, S. Camel heavy-chain antibodies: diverse germline VHH and specific mechanisms enlarge the antigen-binding repertoire. The EMBO journal 19, 921–930 (2000).

[8] Muyldermans, S. Single domain camel antibodies: current status. Reviews in Molecular Biotechnology 74, 277–302 (2001).

[9] Nguyen, V. K., Desmyter, A. & Muyldermans, S. Functional heavy-chain antibodies in Camelidae. (2001).

[10] Jovčevska, I. & Muyldermans, S. The therapeutic potential of nanobodies. BioDrugs 34, 11–26 (2020).

[11] Li, T. et al. Cell-penetrating anti-GFAP VHH and corresponding fluorescent fusion protein VHH-GFP spontaneously cross the blood-brain barrier and specifically recognize astrocytes: application to brain imaging. The FASEB Journal 26, 3969–3979 (2012).

[12] Ilari, A. & Savino, C. Protein structure determination by x-ray crystallography. Bioinformatics 63–87 (2008).

[13] Milne, J. L. et al. Cryo-electron microscopy–a primer for the non-microscopist. The FEBS journal 280, 28–45 (2013).

[14] Clore, G. M. & Gronenborn, A. M. Determination of three-dimensional structures of proteins and nucleic acids in solution by nuclear magnetic resonance spectroscop. Critical reviews in

biochemistry and molecular biology 24, 479–564 (1989).

[15] McPherson, A. & Gavira, J. A. Introduction to protein crystallization. Acta Crystallographica Section F: Structural Biology Communications 70, 2–20 (2014).

[16] Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).

[17] Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. A large-scale experiment to assess protein structure prediction methods. Proteins: Structure, Function, and Bioinformatics 23, ii–iv (1995).

[18] Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. Proteins: Structure, Function, and Bioinformatics 87, 1011–1020 (2019).

[19] AlQuraishi, M. AlphaFold at CASP13. Bioinformatics 35, 4862–4865 (2019).

[20] Pei, J. Multiple protein sequence alignment. Current opinion in structural biology 18, 382–386 (2008).

[21] Vaswani, A. et al. Attention is all you need. Advances in neural information processing systems 30, (2017).

[22] Hsu, W.-N. et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing 29, 3451–3460 (2021).

[23] AlQuraishi, M. Machine learning in protein structure prediction. Current opinion in chemical biology 65, 1–8 (2021).

[24] BFD. https://bfd.mmseqs.com/.

[25] Sussman, J. L. et al. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. Acta Crystallographica Section D: Biological Crystallography 54, 1078–1084 (1998).

[26] Steinegger, M. et al. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 20, 473 (2019).

[27] Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31, 926–932 (2015).

[28] Mirdita, M. et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 45, D170–D176 (2017).

[29] Mitchell, A. L. et al. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res. 48, D570–D578 (2020).

[30] Baek, M. et al. Accurate prediction of protein structures and interactions using a 3-track neural network. Science 373, 871–876 (2021).

[31] Mirdita, M. et al. ColabFold: making protein folding accessible to all. Nature Methods 1–4 (2022).

[32] Ruffolo, J. A., Sulam, J. & Gray, J. J. Antibody structure prediction using interpretable deep learning. Patterns 3, 100406 (2022).

[33] Ruffolo, J. A., Guerra, C., Mahajan, S. P., Sulam, J. & Gray, J. J. Geometric potentials from deep learning improve prediction of CDR H3 loop structures. Bioinformatics 36, i268–i275 (2020).

[34] Sircar, A., Kim, E. T. & Gray, J. J. RosettaAntibody: antibody variable region homology modeling server. Nucleic acids research 37, W474–W479 (2009).

[35] Huang, Z., Xu, W. & Yu, K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015).

[36] Cohen, T., Halfon, M. & Schneidman-Duhovny, D. NanoNet: Rapid end-to-end nanobody modeling by deep learning at sub angstrom resolution. http://biorxiv.org/lookup/doi/10.1101/2021.08.03.454917 (2021) doi:10.1101/2021.08.03.454917.

[37] Shen T, Wu J, Lan H, et al. Whenhomologous sequences meet structural decoys: Accuratecontact prediction by tFold in CASP14— (tFold for CASP14contact prediction). Proteins. 2021;89(12):1901-1910. doi:10.1002/prot.26

[38] Zheng L, Lan H, Shen T, et al. tFold-TR: Combining Deep Learning Enhanced Hybrid Potential Energy for Template-Based Modeling Structure Refinement[J]. arXiv preprint arXiv:2105.04350, 2021.

[39] Zheng, L. et al. tFold-TR: Combining Deep Learning Enhanced Hybrid Potential Energy for Template-Based Modeling Structure Refinement. arXiv preprint arXiv:2105.04350 (2021).

[40] Carugo, O. How root-mean-square distance (rmsd) values depend on the resolution of protein structures that are compared. Journal of applied crystallography 36, 125–128 (2003).

[41] Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic acids research 33, 2302–2309 (2005).

[42] Hooft, R. W., Sander, C. & Vriend, G. Objectively judging the quality of a protein structure from a Ramachandran plot. Bioinformatics 13, 425–430 (1997).

[43] Hengen, P. “Purification of His-Tag fusion proteins from Escherichia coli.” Trends in biochemical sciences vol. 20,7 (1995): 285-6. doi:10.1016/s0968-0004(00)89045-3

[44] Gavin, Anne-Claude et al. “Functional organization of the yeast proteome by systematic analysis of protein complexes.” Nature vol. 415,6868 (2002): 141-7. doi:10.1038/415141a

[45] DeLano, W.L. Pymol, Version 2.5; Delano Scientific: San Carlos, CA, USA, 2002.

[46] Schrodinger, L.L.C. The Pymol Molecular Graphics System, Version 1.7.2.2; Delano

Scientific: San Carlos, CA, USA, 2015.

[47] Lefranc, M.-P. et al. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Developmental & Comparative Immunology 27, 55–77 (2003).

[48] Hooft, RWW, Sander, C. & Vriend, G. Objectively judging the quality of a protein structure from a Ramachandran plot. Bioinformatics 13, 425–430 (1997).