A Data-Driven Approach To Antigen-Antibody Complex Structure Modeling Using Labeled VHH Antibodies

Takashi Nagata[1,2], **Hiroyuki Yamazaki**[1,3], **Ryota Maeda**[1,3], **Hirofumi Tsuruta**[1], **Ryotaro Tamura**[1], **Akihiro Imura**[1,3]

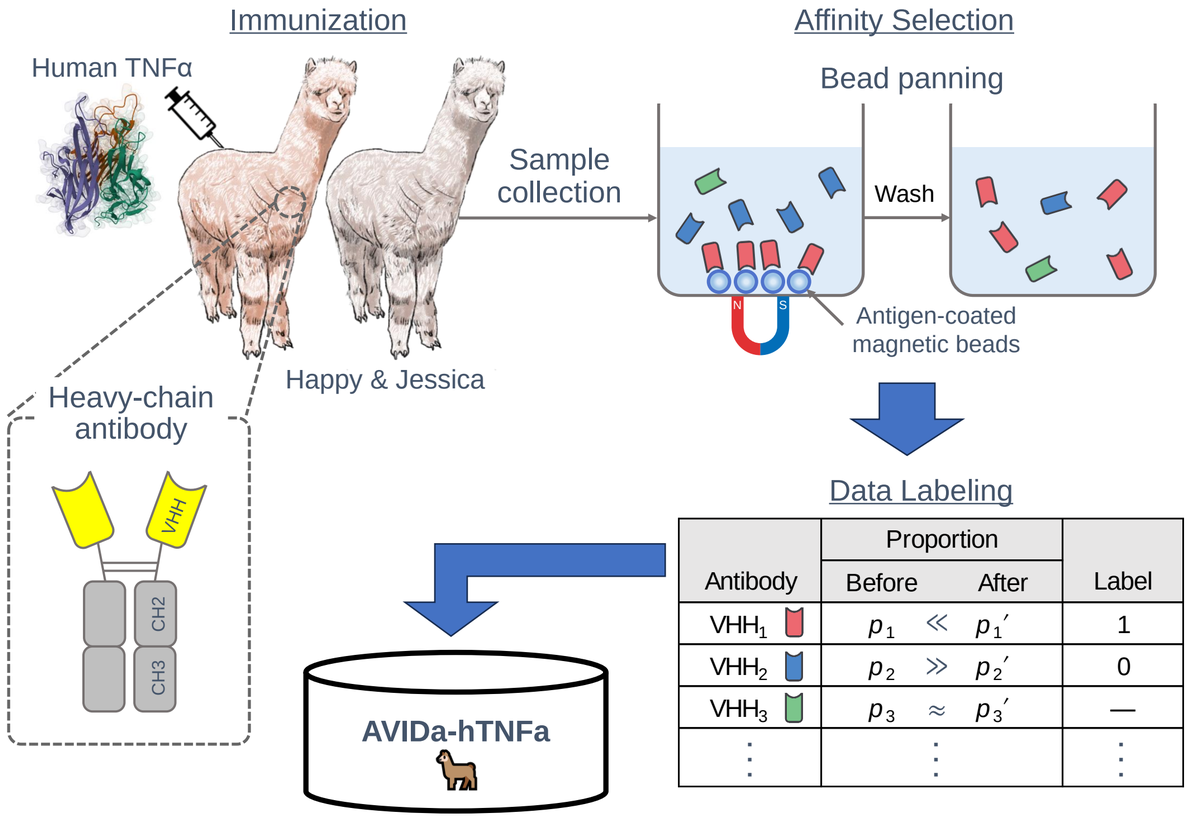
[1]COGNANO, Inc., [2]Institute of Advanced Energy, Kyoto University, [3]Biorhodes, Inc.

Motivation & Dataset Overview

- Recent advances in AI-based protein structure prediction have significantly improved accuracy for individual proteins; however, modeling antigen–antibody complexes in three dimensions remains challenging, particularly in capturing the precise antigen-binding interface.
- To expand available data on antigen–antibody interactions, we immunized alpacas with several antigens and constructed VHH (single-domain antibody) libraries. We previously released large-scale VHH datasets targeting IL-6 [1] and the SARS-CoV-2 spike protein [2].
- In this study, we present a new dataset, AVIDa-hTNFa, comprising anti-human TNFα VHH clones, further extending our antibody sequence and binding data resources.

AVIDa-hTNFa: Antigen-VHH Interaction Dataset

Dataset generation [1,2]



Generation of Custom Multiple Sequence Alignments (MSAs)

We selected 26 TNFα-binding VHH clones with less than 93% sequence identity from our in-house antibody library. For each clone, we performed 3D structure prediction using **three different input conditions** in LocalColabFold [3]:

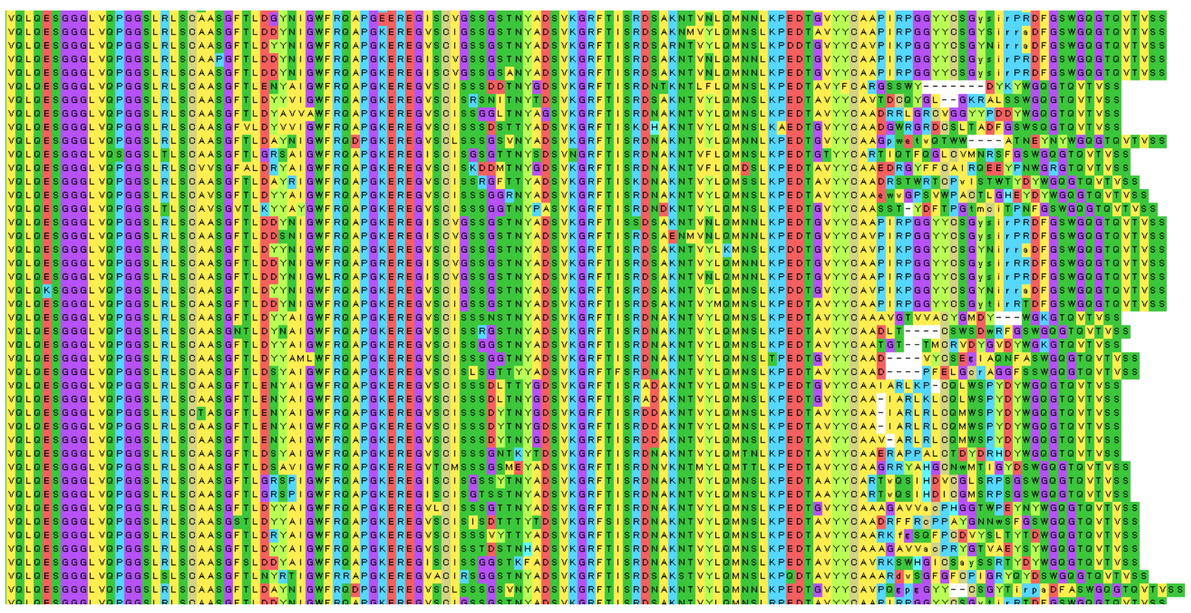
1. **Default setting**: MSAs were automatically generated by searching public sequence databases, and template-based modeling was applied.
2. **Custom MSA only**: MSAs were generated from the AVIDa-hTNFa dataset using MMseqs2, and no template structures were used.
3. **Concatenated MSA**: The custom MSA and the public database-derived MSA were merged, and predictions were again performed without template structures.

This design allowed direct comparison between standard and dataset-informed inputs, isolating the impact of custom sequence data on predicted structural accuracy and variability.

List of the Selected Clones

Happy	ONU1	ONU162	ONU262	ONU272	ONU273	ONU287	ONU377	ONU487	ONU639	ONU681	ONU1115	ONU1277	ONU1287	ONU1581
Jessica	ONU2	ONU11	ONU54	ONU57	ONU95	ONU105	ONU219	ONU302	ONU354	ONU424	ONU557	ONU1334		

Custom MSAs Derived From the AVIDa-hTNFa Library



Amino acid sequences were visualized by MEGA X software [4]

Comparative Analysis

We compared the predicted structures under the three input conditions based on sequence coverage, per-residue pLDDT (predicted Local Distance Difference Test) scores, predicted aligned error (PAE), and overall 3D conformation.

Statistical analyses of pLDDT values were performed using the Mann–Whitney U test and the Kruskal–Wallis test, with Bonferroni correction applied for multiple comparisons. To assess region-specific differences in prediction confidence, each residue of the VHH sequences was annotated using the IMGT numbering scheme via the ANARCI tool [5].

COGNANO

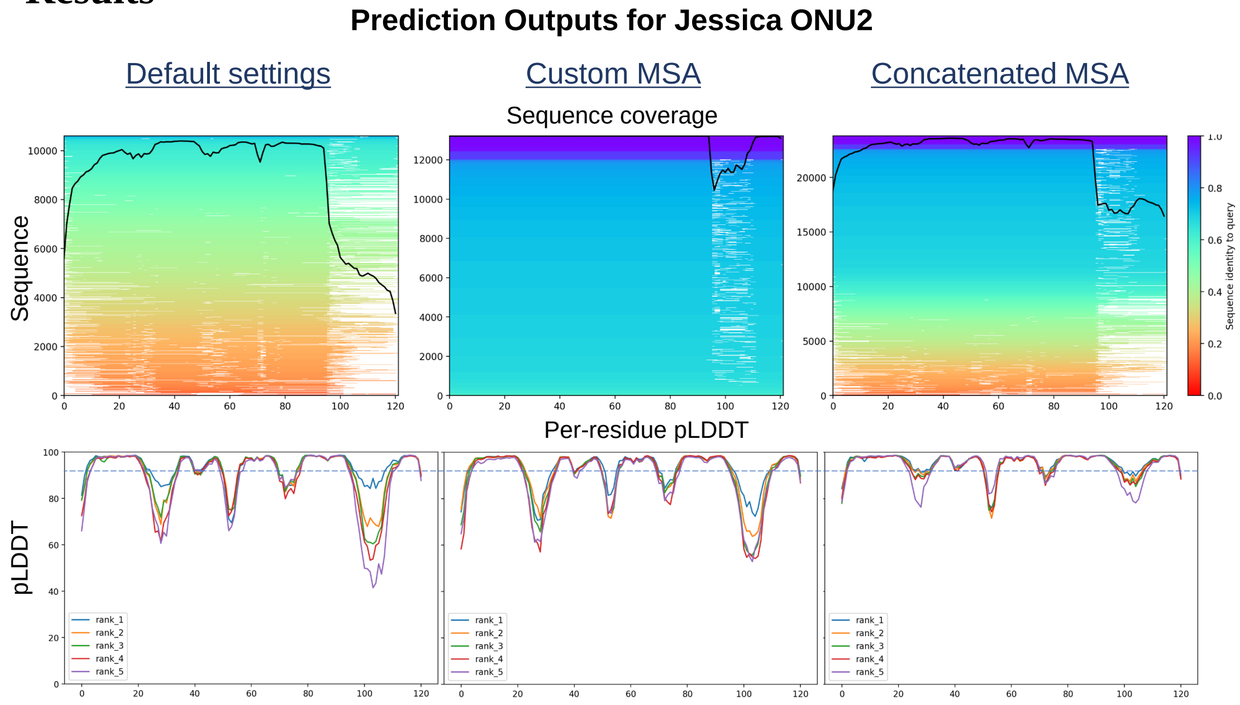
ICLR

Project page

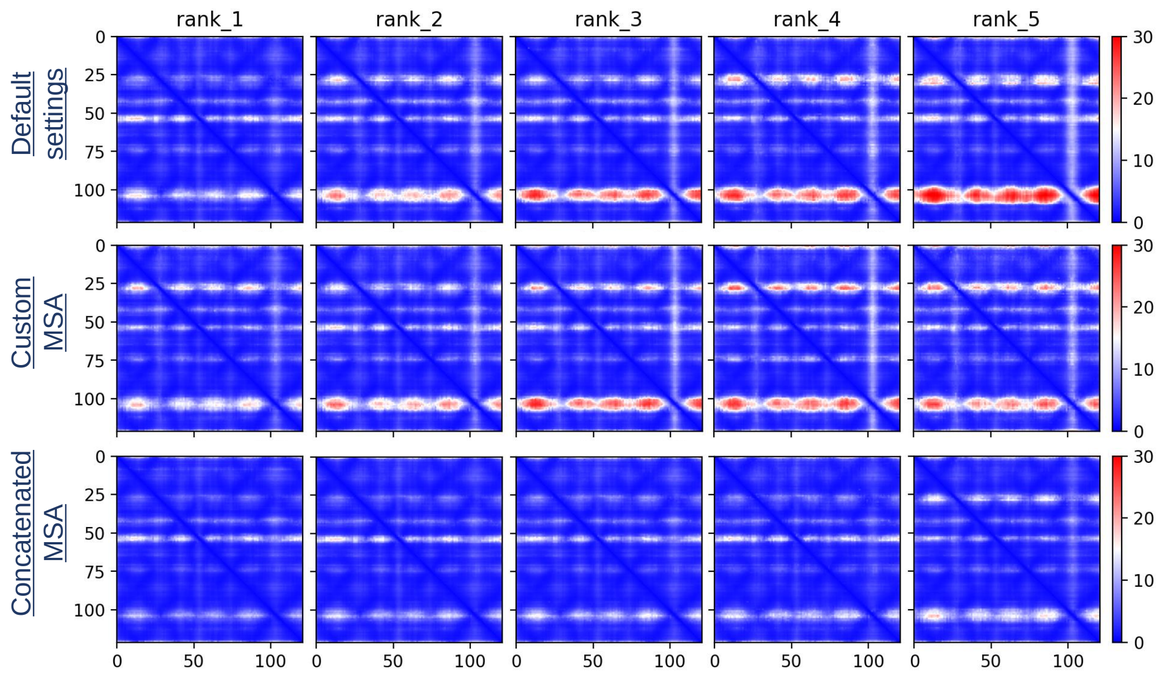
Dataset

Results

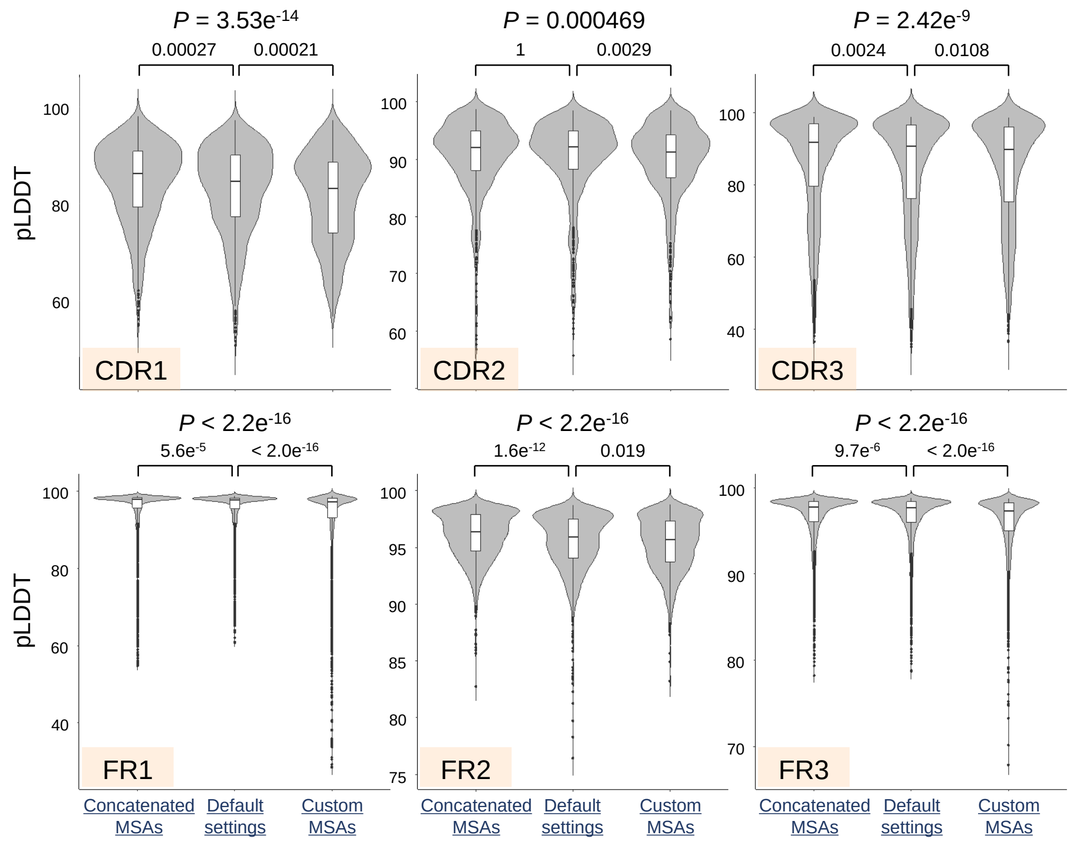
Prediction Outputs for Jessica ONU2



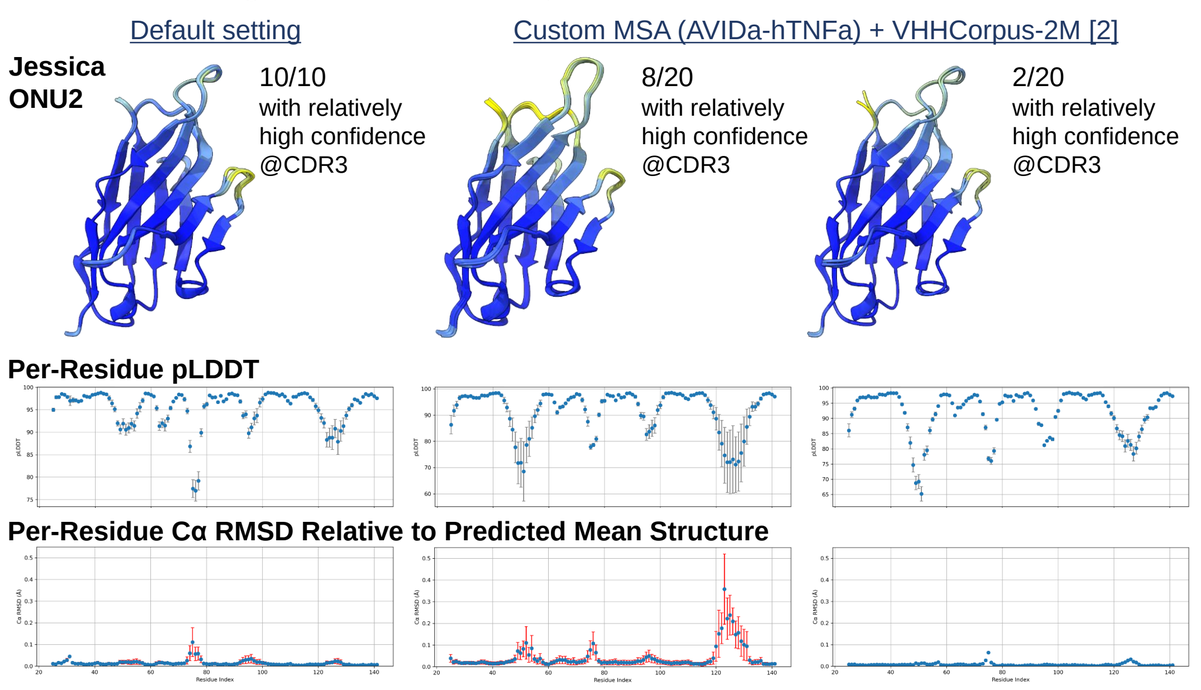
PAEs of All Predictions for Jessica ONU2



Statistical Analysis for the Per-Residue pLDDT of All Predictions for the 26 Clones



Exploring Structural Variability with Dataset-Informed Predictions



References

[1] Tsuruta et al., AVIDa-hIL6: A large-scale VHH dataset produced from an immunized alpaca for predicting antigen-antibody interactions (2023).
[2] Tsuruta et al., A SARS-CoV-2 interaction dataset and VHH sequence corpus for antibody language models (2024).
[3] Mirdita et al., ColabFold: making protein folding accessible to all (2022).
[4] Kumar et al., MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms (2018).
[5] Dunbar et al., ANARCI: antigen receptor numbering and receptor classification (2016).