Structural modeling of nanobodies: A benchmark of state-of-the-art artificial intelligence programs

Mario S. Valdés-Tresanco†§\*, Mario E. Valdés-Tresanco‡§, Daiver Jiménez† and Ernesto Moreno†\*

† Faculty of Basic Sciences, University of Medellin, Medellin, Colombia.

‡ Centre for Molecular Simulations and Department of Biological Sciences, University of Calgary, Alberta T2N 1N4, Canada.

§ These authors contributed equally to this work

\* To whom correspondence should be addressed:

Ernesto Moreno: [emoreno@udemedellin.edu.co](mailto:emoreno@udemedellin.edu.co)

Mario S. Valdés-Tresanco: [mariosergiovaldes145@gmail.com](mailto:mariosergiovaldes145@gmail.com)

**Table S1**. Nbs included in the dataset. All PDB files can be found at: <https://github.com/Valdes-Tresanco-MS/NbModelingBenchmark>

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PDB** | **PDB** | **PDB** | **PDB** | **PDB** |
| 7d8b\_B | 7p5y\_G | 7ndf\_C | 7lzp\_B | 7rby\_B |
| 7anq\_B | 7czd\_A | 7d6y\_B | 7apj\_B | 7e53\_B |
| 7r4r\_D | 7rnn\_C | 7fat\_B | 7nfr\_B | 7djy\_A |
| 7f5g\_B | 7vfa\_D | 6zwk\_A | 7nqa\_C | 7p5w\_G |
| 7r98\_D | 7s2r\_B | 7o06\_A | 7r74\_B | 7omn\_A |
| 7fau\_B | 7p5v\_G | 7r4q\_D | 7nft\_C | 7t5f\_B |
| 7m1h\_F | 7z1a\_E | 7r20\_B | 7lzp\_C | 7vfb\_B |
| 7nfq\_C | 7pqg\_B | 7qbg\_E | 7lzp\_G | 7olz\_B |
| 7p6k\_G | 7kd0\_C | 7oan\_F | 7na9\_D | 7ocy\_C |
| 7e6u\_B | 7p60\_G | 7r63\_A | 7djx\_A | 7z1c\_E |
| 7qiv\_C | 7l6v\_B | 7sl9\_B | 7b2p\_D | 7tpr\_F |
| 7qbe\_E | 7t5f\_C | 7php\_N | 7d4b\_B | 7r4i\_D |
| 7m1h\_E | 7te8\_A | 7ooi\_A | 7zfb\_M | 7l6v\_D |
| 7n9c\_D | 7nqk\_B | 7s7r\_B | 7l6v\_C | 7ar0\_B |
| 7ngh\_D | 7tpr\_D | 7cz0\_E | 7z1b\_E | 7te8\_C |

Chart, histogram

Description automatically generated

**Figure S1.** Resolution distribution of crystallographic structures of Nbs in the dataset. The vertical red dashed line represents the median resolution value.

**Chart

Description automatically generated**

**Figure S2.** Sequence identity among Nbs in the dataset. Only 4 Nbs have sequence identities greater than 95%.

**Table S2.** Results from Dunn’s test nonparametric pairwise multiple comparison procedure.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Method1** | **Method2** | **TM-score** | **GDT\_TS** | **GDT\_HA** | **Fw** | **CDR1** | **CDR2** | **CDR3** | **Global** |
| OF | AF2 | ns | ns | ns | ns | ns | ns | ns | ns |
| OF | IF | \* | \* | \* | ns | ns | ns | \* | ns |
| OF | NN | \*\* | \*\* | \*\* | ns | ns | ns | \*\*\* | \*\* |
| OF | ESM | ns | ns | ns | ns | ns | ns | ns | ns |
| OF | YS | \*\*\*\* | \*\*\*\* | \*\*\*\* | \*\*\*\* | \*\*\* | \*\*\*\* | \*\*\*\* | \*\*\*\* |
| AF2 | OF | ns | ns | ns | ns | ns | ns | ns | ns |
| AF2 | IF | ns | ns | ns | ns | ns | ns | ns | ns |
| AF2 | NN | \* | \* | \* | ns | ns | ns | \* | ns |
| AF2 | ESM | ns | ns | ns | ns | ns | ns | ns | ns |
| AF2 | YS | \*\*\*\* | \*\*\*\* | \*\*\*\* | \*\*\*\* | \*\*\* | \*\*\*\* | \*\*\* | \*\*\*\* |
| IF | OF | \* | \* | \* | ns | ns | ns | \* | ns |
| IF | AF2 | ns | ns | ns | ns | ns | ns | ns | ns |
| IF | NN | ns | ns | ns | ns | ns | ns | ns | ns |
| IF | ESM | ns | ns | ns | ns | ns | ns | ns | ns |
| IF | YS | \*\*\*\* | \*\*\*\* | \*\*\*\* | \*\*\*\* | \* | \*\*\* | \* | \*\*\*\* |
| NN | OF | \*\* | \*\* | \*\* | ns | ns | ns | \*\*\* | \*\* |
| NN | AF2 | \* | \* | \* | ns | ns | ns | \* | ns |
| NN | IF | ns | ns | ns | ns | ns | ns | ns | ns |
| NN | ESM | \* | \* | \* | ns | ns | ns | ns | ns |
| NN | YS | \*\*\*\* | \*\*\*\* | \*\*\*\* | \*\*\*\* | ns | \*\*\* | ns | \*\*\*\* |
| ESM | OF | ns | ns | ns | ns | ns | ns | ns | ns |
| ESM | AF2 | ns | ns | ns | ns | ns | ns | ns | ns |
| ESM | IF | ns | ns | ns | ns | ns | ns | ns | ns |
| ESM | NN | \* | \* | \* | ns | ns | ns | ns | ns |
| ESM | YS | \*\*\*\* | \*\*\*\* | \*\*\*\* | \*\*\*\* | \*\*\* | \*\*\*\* | \*\*\* | \*\*\*\* |
| YS | OF | \*\*\*\* | \*\*\*\* | \*\*\*\* | \*\*\*\* | \*\*\* | \*\*\*\* | \*\*\*\* | \*\*\*\* |
| YS | AF2 | \*\*\*\* | \*\*\*\* | \*\*\*\* | \*\*\*\* | \*\*\* | \*\*\*\* | \*\*\* | \*\*\*\* |
| YS | IF | \*\*\*\* | \*\*\*\* | \*\*\*\* | \*\*\*\* | \* | \*\*\* | \* | \*\*\*\* |
| YS | NN | \*\*\*\* | \*\*\*\* | \*\*\*\* | \*\*\*\* | ns | \*\*\* | ns | \*\*\*\* |
| YS | ESM | \*\*\*\* | \*\*\*\* | \*\*\*\* | \*\*\*\* | \*\*\* | \*\*\*\* | \*\*\* | \*\*\*\* |

**Table S3.** Number of occurrences of predicted structures with RMSD above 2.5 Å.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **CDR1** | **CDR2** | **CDR3** | **Fw** | **Global** |
| OF | 18 | 4 | 36 |  | 8 |
| AF2 | 18 | 4 | 42 |  | 10 |
| IF | 18 | 6 | 53 |  | 9 |
| NN | 20 | 3 | 56 |  | 14 |
| ESM | 20 | 3 | 49 |  | 11 |
| YS | 28 | 7 | 64 | 1 | 28 |

Chart

Description automatically generated

**Figure S3.** Distribution of heavy atoms RMSD values by position for the OmegaFold, AlphaFold2, IgFold, and Nanonet models. CDR1, CDR2, and CDR3 regions are colored pink, cyan, and brown, respectively. The RMSD distributions are represented by a boxplot. The lower and upper edges of the box represent the first (Q1) and third quartile (Q3), respectively. The difference Q3 - Q1 is known as the interquartile range (IQR). Whiskers extend to the minimum and maximum points within ±1.5 × IQR, respectively. The red line indicates the median value by position.

**Chart

Description automatically generated**

**Figure S4.** Representation by positions of the Nbs in the dataset using the Aho numbering scheme.

Chart, histogram

Description automatically generated

**Figure S5.** CDR3 length distribution of non-redundant Nbs structures in the Protein Data Bank.

Imagen que contiene collar

Descripción generada automáticamente

**Figure S6.** Particular cases of structural variations in the crystallographic structures are used as references. A) PDB: 7tpr. Nb (white) in complex with the antigen (deep salmon), which sits below CDR3 (cyan) stretching it away from the framework. B) Comparison between the engineered human heavy chain variable domain (PDB: 7db8) without canonical disulfide bridge (framework in white and CDR3 in cyan) and a Nb of equal length (PDB: 5lhn) with canonical disulfide bridge (framework in dark gray and CDR3 in hot pink). The arrow indicates the displacement of the N-terminal portion of the VH CDR3 with respect to the Nb.

Chart, scatter chart

Description automatically generated

**Figure S7.** Correlation between the RMSD values and the average predicted confidences by AI models for the CDR regions in the 75 Nbs conforming the dataset. Regression lines relating the RMSD and the average predicted confidence variables are shown in different colors depending on the AI model used. Translucent bands around the regression lines indicate the 95% confidence interval for the regression estimates.

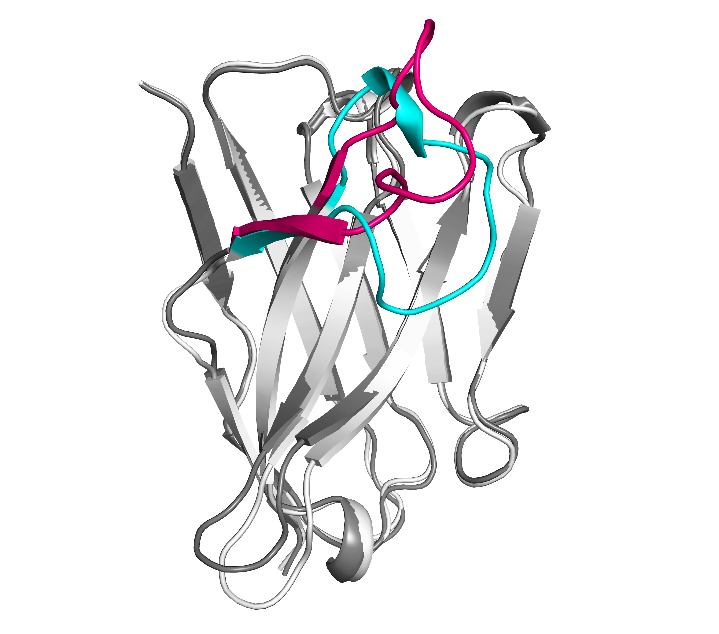
**Table S4.** Correlation between the RMSD values and the average predicted confidences by AI models for the CDR regions in the 75 Nbs conforming the dataset.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Region** | **Pearson** | **Pearson p-value** | **Spearman** | **Spearman p-value** | **Slope** | **Intercept** |
| AF2 | CDR1 | -0.62 | 2.10E-09 | -0.68 | 1.49E-11 | -0.11 | 11.34 |
| AF2 | CDR2 | -0.54 | 7.46E-07 | -0.42 | 2.12E-04 | -0.11 | 11.50 |
| AF2 | CDR3 | -0.59 | 3.02E-08 | -0.63 | 1.14E-09 | -0.17 | 16.65 |
| OF | CDR1 | -0.59 | 2.57E-08 | -0.62 | 3.55E-09 | -0.08 | 8.68 |
| OF | CDR2 | -0.44 | 6.55E-05 | -0.47 | 2.08E-05 | -0.07 | 7.44 |
| OF | CDR3 | -0.62 | 4.01E-09 | -0.71 | 7.02E-13 | -0.13 | 12.87 |
| IF | CDR1 | 0.54 | 5.91E-07 | 0.60 | 1.42E-08 | 1.35 | 0.73 |
| IF | CDR2 | 0.55 | 3.13E-07 | 0.47 | 2.21E-05 | 1.81 | -0.02 |
| IF | CDR3 | 0.46 | 3.61E-05 | 0.45 | 4.86E-05 | 2.65 | -0.26 |
| ESM | CDR1 | -0.48 | 1.05E-05 | -0.54 | 7.34E-07 | -0.08 | 8.81 |
| ESM | CDR2 | -0.54 | 6.66E-07 | -0.49 | 7.73E-06 | -0.09 | 8.66 |
| ESM | CDR3 | -0.51 | 2.45E-06 | -0.56 | 2.23E-07 | -0.14 | 13.52 |
| YS | CDR1 | -0.13 | 2.81E-01 | -0.21 | 7.02E-02 | -0.03 | 4.69 |
| YS | CDR2 | -0.22 | 5.91E-02 | -0.20 | 8.72E-02 | -0.04 | 4.72 |
| YS | CDR3 | -0.23 | 5.01E-02 | -0.24 | 3.55E-02 | -0.09 | 12.18 |

Diagram

Description automatically generated with low confidence

**Figure S8.** Comparison of the RMSD distribution for the models obtained varying the number of recycles of AlphaFold2 and OmegaFold. The RMSD distributions are represented by a violin plot. It combines the kernel density plot (outer) to show the distribution of values and a boxplot (inner) that summarizes the distribution statistics. The boxplot is composed of the white dot representing the median, the thick gray bar in the center representing the interquartile range, and the thin gray line representing the rest of the distribution. No statistically significant differences were found between the models obtained with the default parameters and those obtained with varying the number of recycles of AlphaFold2 and OmegaFold.



**Figure S9**. Structural variation of CDR3 of the same Nb in the asymmetric unit. Both CDR3 are highlighted in cyan and hotpink, respectively. PDB ID: 7omn

**Table S5.** Nanobodies selected for modeling Nb-antigen complexes. All PDB files can be found at https://github.com/Valdes-Tresanco-MS/NbModelingBenchmark

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PDB** | **PDB** | **PDB** | **PDB** | **PDB** |
| 7d8b\_B | 7l6v\_D | 7ocy\_C | 7m1h\_F | 7z1b\_E |
| 7nft\_C | 7rby\_B | 7d4b\_B | 7vfa\_D | 7fau\_B |
| 7nfr\_B | 7lzp\_C | 7l6v\_B | 7php\_N | 7anq\_B |
| 7m1h\_E | 7z1c\_E | 7olz\_B | 7d6y\_B | 7fat\_B |
| 7sl9\_B | 7o06\_A | 7ar0\_B | 7qbg\_E | 7qbe\_E |
| 7t5f\_B | 7lzp\_B | 7na9\_D | 7s2r\_B | 7r74\_B |
| 7r98\_D | 7e53\_B | 7f5g\_B | 7zfb\_M | 7ndf\_C |
| 7t5f\_C | 7ngh\_D | 7rnn\_C | 7lzp\_G | 7nfq\_C |
| 7pqg\_B |  |  |  |  |

Chart, bar chart

Description automatically generated

**Figure S10.** Comparison between the RMSD of models generated with AlphaFold2 (AF2) and AlphaFold-multimer (AF2m) clustered by CDR3 length.

**Diagram

Description automatically generated**

**Figure S11.** Comparison of the RMSD distribution between the default generated model and its minimized version for OmegaFold, AlphaFold2, IgFold, Nanonet, ESMFold and Yang-server. The RMSD distributions are represented by a violin plot. It combines the kernel density plot (outer) to show the distribution of values and a boxplot (inner) that summarizes the distribution statistics. The boxplot is composed of the white dot representing the median, the thick gray bar in the center representing the interquartile range, and the thin gray line representing the rest of the distribution. No statistically significant differences were found between the models obtained with the default parameters and their minimized version.