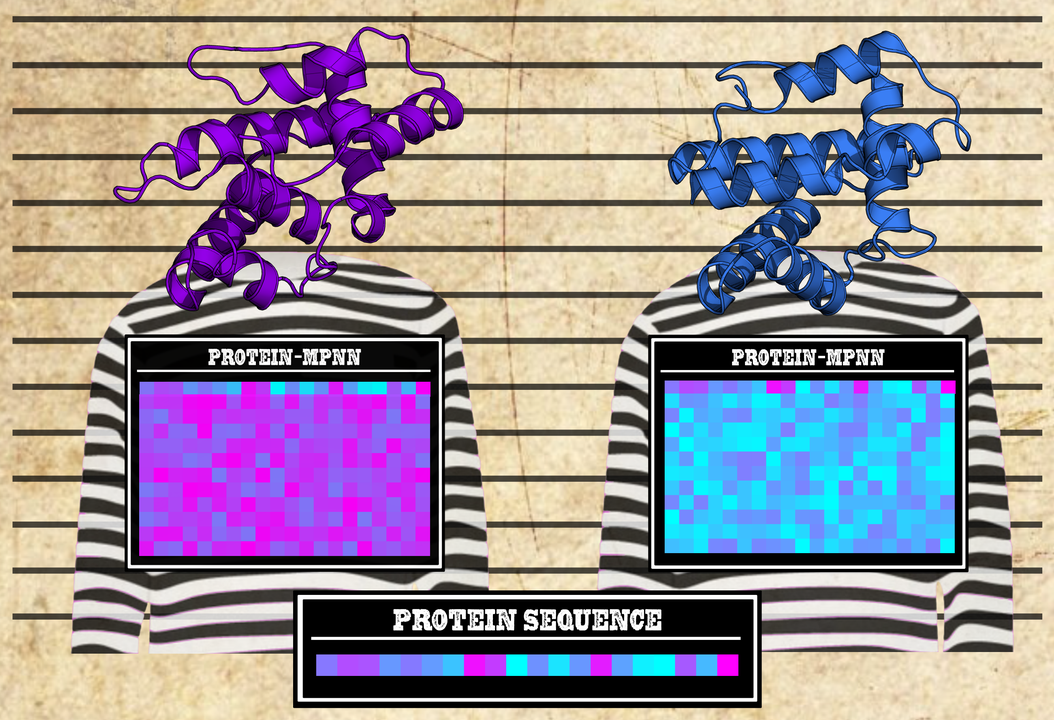
WANTED

SYNTHETIC OR NATURAL

MSAS TO IDENTIFY CONFORMATIONS

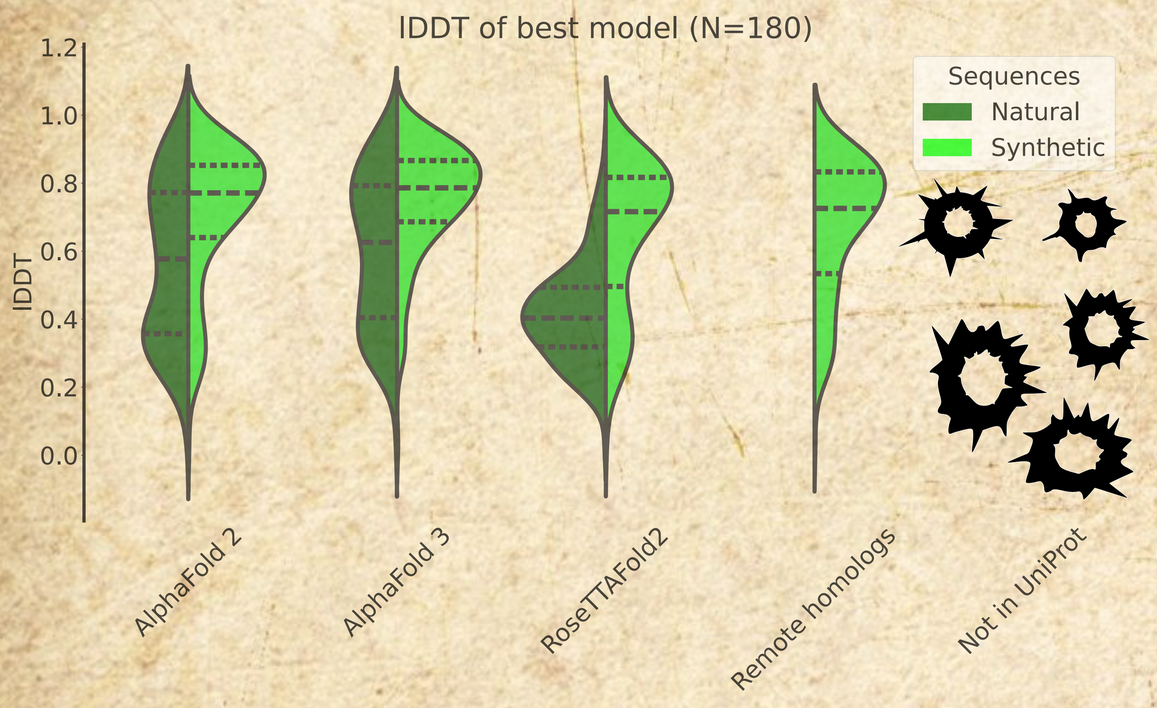
b
UNIVERSITÄT
BERN



OFFENCE

Proteins are not static objects, but can adapt to conditions and form different conformations with different three-dimensional structures. AlphaFold struggles to find these conformations, leading to the conclusion that "AlphaFold2 has more to learn about protein energy landscapes"[1]. We hypothesize that this weakness is not due to an inherent limitation of AlphaFold 2, but rather due to suboptimal multiple sequence alignment (MSA) input.

To test our hypothesis, we utilize the previously tested dataset[1], which consists of conformation pairs with different structures but similar sequences. We then generate synthetic MSAs tailored to the known structures using ProteinMPNN with a higher temperature. The designed MSA is then processed by protein folding models to predict structures, which are subsequently evaluated against the structures in the dataset.

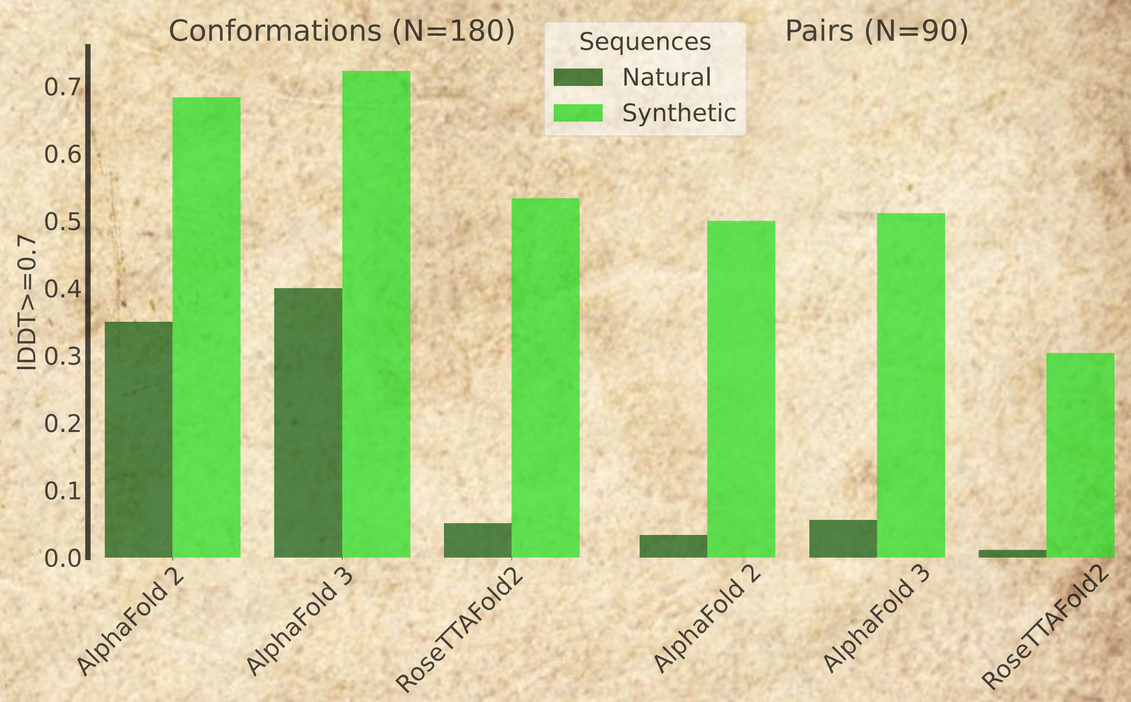


CAUGHT OFFENDERS

Synthetic MSAs perform better, not just for the AlphaFold family, but also for RoseTTAFold2, indicating this trend should persist for future folding models. The results stay good when taking away sequences that are more than 30% identical to the query sequence. No generated sequence matched UniProt entries above 70% identity, suggesting that folding does not rely on naturally occurring sequences.

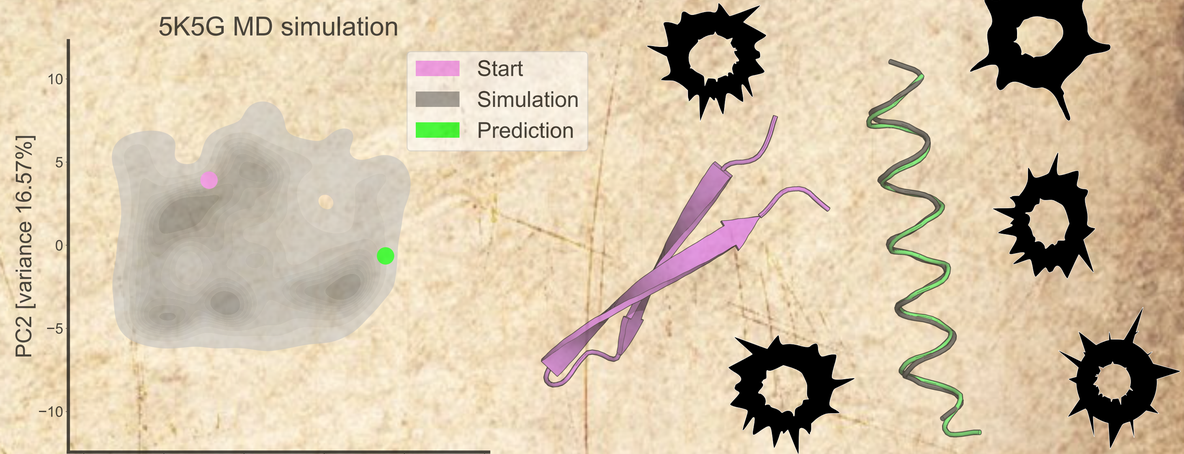
Thanks to Noah Kleinschmidt for MD simulations and the FrankenFold logo!

[1] Chakravarty, Devlina, et al. "AlphaFold2 has more to learn about protein energy landscapes." BioRxiv (2023).
[2] Wang, Jinan, et al. "Gaussian accelerated molecular dynamics: Principles and applications." Wiley Interdisciplinary Reviews: Computational Molecular Science 11.5 (2021): e1521.
[3] Joosten, Robbie P., et al. "A series of PDB related databases for everyday needs." Nucleic acids research 39.suppl_1 (2010): D411-D419.
[4] Lindorff-Larsen, Kresten, et al. "How fast-folding proteins fold." Science 334.6055 (2011): 517-520.
[AlphaFold 2] Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." nature 596.7873 (2021): 583-589.
[AlphaFold 3] Abramson, Josh, et al. "Accurate structure prediction of biomolecular interactions with AlphaFold 3." Nature 630.8016 (2024): 493-500.
[lDDT] Mariani, Valerio, et al. "lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests." Bioinformatics 29.21 (2013): 2722-2728.
[MMseqs2] Steinegger, Martin, and Johannes Söding. "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets." Nature biotechnology 35.11 (2017): 1026-1028.
[ProteinMPNN] Dauparas, Justas, et al. "Robust deep learning–based protein sequence design using ProteinMPNN." Science 378.6615 (2022): 49-56.
[RoseTTAFold2] Baek, Minkyung, et al. "Efficient and accurate prediction of protein structure using RoseTTAFold2." BioRxiv (2023): 2023-05.
[UniProt] UniProt Consortium. "UniProt: a worldwide hub of protein knowledge." Nucleic acids research 47.D1 (2019): D506-D515.



CAUGHT GANGS

Filtering conformer pairs to include only those with IDDT scores above 0.7 for both structures resulted in a success rate of 142 out of 180.

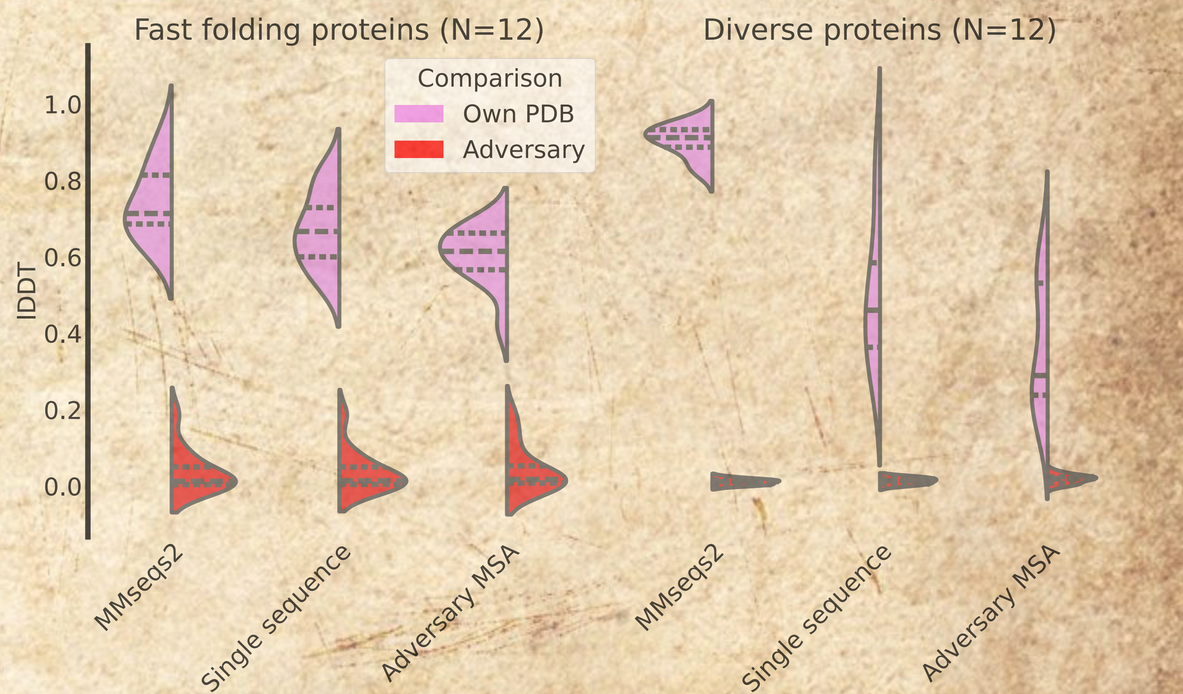


WHEREABOUTS OF UNCAUGHT

To catch the remaining conformations, we employed molecular dynamics (MD) simulations [2] to better characterize the conformational space. However, the predicted structures were consistent with the MD simulations in only a limited number of cases.

STILL ON THE LOOSE

Visual inspection of the remaining proteins revealed similar structural features despite low scoring metrics. A telling feature is that 24/38 missing proteins have more than half of the residues without secondary structures[3], making them hard to score with traditional scoring methods. We are heavily working on catching the last missing structures also with the help of the simulations.



FRAMING OF ADVERSARY

We tried our method on proteins that only have one conformation and give the MSA of a different, adversary protein with a different secondary structure. For small, fast-folding proteins[4], the MSA hardly makes an impact, for more complicated proteins, the MSA can lead AlphaFold 2 to a disordered prediction.

REWARD

- ★ MSA level combination of known template structure with uncertain parts in FrankenFold
- ★ Current upper bound for protein folding prediction performance
- ★ Precise tools for targeted folding model experiments

COURTESY OF

Jannik Adrian Gut, Thomas Lemmin
Institute of Biochemistry and Molecular Medicine, University of Bern
Graduate School for Cellular and Biomedical Sciences
jannik.gut@unibe.ch, thomas.lemmin@unibe.ch