[BIOINFO 2023]

Petabase-scale Homology Search for Structure Prediction

Sewon Lee[1,*], Gyuri Kim[1,*], Eli Levy Karin[2], Milot Mirdita[1], Sukhwan Park[3], Rayan Chikhi[4], Artem Babaian[5,6], Andriy Kryshtafovych[7], and Martin Steinegger[1,3,8,9,†]

[1]*School of Biological Sciences, Seoul National University*
[2]*ELKMO*
[3]*Interdisciplinary Program in Bioinformatics, Seoul National University*
[4]*Institut Pasteur, Université Paris Cité, G5 Sequence Bioinformatics*
[5]*Department of Molecular Genetics, University of Toronto, Toronto*
[6]*Donnelly Centre for Cellular and Biomolecular Research, University of Toronto*
[7]*Genome Center, University of California, Davis*
[8]*Artificial Intelligence Institute, Seoul National University*
[9]*Institute of Molecular Biology and Genetics, Seoul National University*
**These authors contributed equally*
†*Corresponding author: martin.steinegger@snu.ac.kr*

The CASP15 competition highlighted the critical role of multiple sequence alignments (MSAs) in protein structure prediction, as demonstrated by the success of the top AlphaFold2-based prediction methods. To push the boundaries of MSA utilization, we conducted a petabase-scale search of the Sequence Read Archive (SRA), resulting in gigabytes of aligned homologs for CASP15 targets. These were merged with default MSAs produced by ColabFold-search and provided to ColabFold-predict. By using SRA data, we achieved highly accurate predictions (GDT_TS > 70) for 66% of the non-easy targets, whereas using ColabFold-search default MSAs scored highly in only 52%. Next, we tested the effect of deep homology search and ColabFold's advanced features, such as more recycles, on prediction accuracy. While SRA homologs were most significant for improving ColabFold's CASP15 ranking from 11th to 3rd place, other strategies contributed too. We analyze these in the context of existing strategies to improve prediction.