**MS11-2-1 High-accuracy protein structure models in AlphaFold DB**
**#MS11-2-1**

**M. Varadi [1], M. Deshpande [1], S. Nair [1], S. Anyango [1], D. Bertoni [1], S. Velankar [1]**
**[1]EMBL-EBI - Hinxton (United Kingdom)**

**Abstract**
Proteins are essential macromolecules with vital biological functions. They are involved in a wide range of research activities and medical and biotechnological applications, from fighting infectious diseases to tackling environmental pollution. Knowledge of the three-dimensional (3D) arrangement of the atoms of a protein can provide essential clues to understanding the roles and mechanisms underpinning protein functions. However, while the Universal Protein Resource (UniProt) archives almost 230 million unique protein sequences, the Protein Data Bank (PDB) holds only over 190,000 3D structures (178,000 based on X-ray diffraction) for over 58,000 distinct proteins, thus severely limiting the use of structural information to support biomolecular research globally [1, 2].
One way to attempt to decrease this gap is to predict the structures of millions of proteins. Increasingly, researchers deploy Artificial Intelligence (AI) techniques to predict a protein's structure computationally from its amino-acid sequence alone [3, 4].
AlphaFold DB is an openly accessible, extensive database of high-accuracy protein-structure predictions [5]. Powered by AlphaFold v2.0 of DeepMind, it has enabled an unprecedented expansion of the structural coverage of the known protein-sequence space. While the AlphaFold AI system has certain limitations, especially in terms of modelling complexes, high accuracy models are helpful even in determining the crystal structure of more complex target proteins. AlphaFold DB provides programmatic access to and interactive visualisation of predicted atomic coordinates, per-residue and pairwise model-confidence estimates and indicated aligned errors. The current version of AlphaFold contains approximately 1 million predicted structures across 37 model-organism proteomes, with significant increases expected in terms of the data size sets.

**References**
[1] UniProt: the universal protein knowledgebase in 2021; UniProt Consortium; Nucleic Acids Res.; 2021 Jan 8;49(D1):D480-D489

[2] PDBe: improved findability of macromolecular structure data in the PDB; Armstrong D. et al; Nucleic Acids Res.; 2020 Jan 8;48(D1):D335-D343

[3] Highly accurate protein structure prediction with AlphaFold; Jumper J. et al; Nature; 2021 Aug;596(7873):583-589

[4] Accurate prediction of protein structures and interactions using a three-track neural network; Baek M. et al; Science; 2021 Aug 20;373(6557):871-876

[5] AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models; Varadi M. et al; Nucleic Acids Res.; 2022 Jan 7;50(D1):D439-D444