# BITS :: Call for Abstracts 2024 - Oral communication

| | |
|---|---|
| *Type* | Oral communication |
| *Session* | Bioinformatics AI, Models and Tools |
| *Title* | Development of an automated pipeline to predict protein structures. |
| *All Authors* | Gil Zuluaga FH(1), D'Arminio N(2), Bardozzo F(1), Tagliaferri R(1), Marabotti A(2) |

*Affiliation*

(1): Department of Management & Innovation Systems, University of Salerno, Fisciano (SA), Italy
(2): Department of Chemistry and Biology "A. Zambelli", University of Salerno, Fisciano (SA), Italy

*Motivation*

Protein structure prediction has long been a crucial goal. For decades, structural bioinformaticians have developed both template-based and template-free approaches for predicting 3D protein structures, with the aim of closing the gap between known protein sequences and protein structures determined by experimental methods, a gap of 3 orders of magnitude that until a few years ago seemed absolutely unbridgeable.
Things have changed with the advent of artificial intelligence approaches, and in particular with the development of AlphaFold2 (AF2), which released to the public the structural predictions of over 200 million proteins.
The accuracy of AF2's predictions is remarkable. However, it has been noted that it is not yet sufficient for applications such as virtual screening. Based on this consideration, we thought of improving the results of AF2 by applying a template-based modeling approach, MODELLER, to create a pipeline that can also give an evaluation of the obtained models based on independent metrics.

*Methods*

AlphaMod consists of 5 sequential steps (Figure) that allow: I. sequence homology and structural template search; II. protein structures predictions through AF2's. III. structure quality assessment of models, by a unified score (BORDAscore) including region confidence (pLDDT) and global and local geometrical aspects (QMEANDisCo). IV. the structures upon user selection undergo a further refinement step through MODELLER. V. a comprehensive evaluation of the final models using several unsupervised scores, and ranking them according to QMEANDisCo, which we found as the only metric strongly correlated with GDT_TS.

*Results*

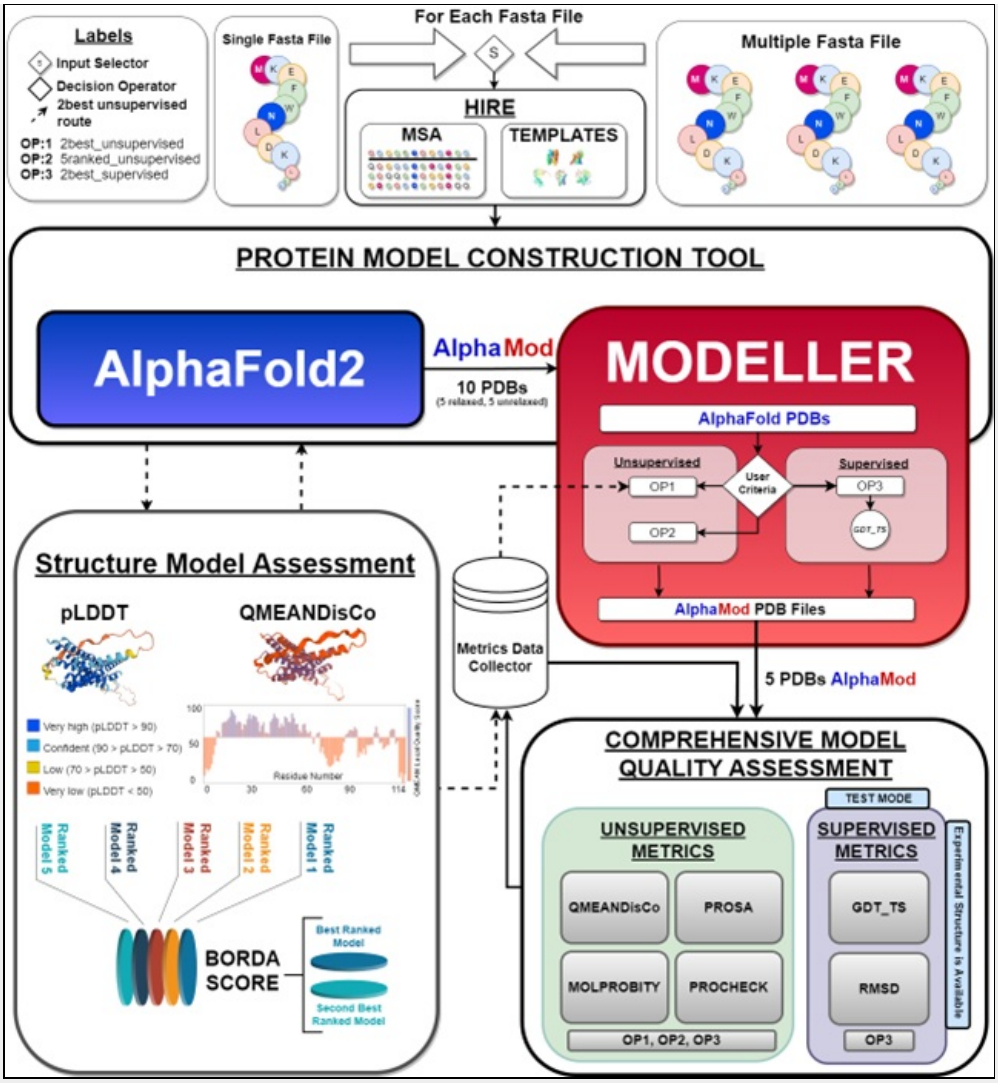
We tested AlphaMod pipeline on two different datasets. Test set A is formed by 43 structures extracted from CASP14 database, for which AF2 performance was assessed and both the structures of the proteins and the GDT_TS results are publicly available. Test set B is formed by 25 protein structures selected by Terwilliger and colleagues in order to assess the procedure they developed to improve AF2 performance by including implicitly experimental information (doi: https://doi.org/10.1038/s41592-022-01645-6). We performed 3 different simulations: OP1, MODELLER is launched using as input the two proteins with the highest structural accuracy as explained in Methods step III and IV; OP2, MODELLER is launched with all relaxed structures produced in step II; OP3 (test mode only), MODELLER is launched using as input the two proteins with the highest structural accuracy as per GDT_TS analysis.
Comparing the resulting protein structures obtained through OP1, OP2 and OP3 with those proteins predicted by AF2 alone, we found a statistically significant improvement for AlphaMod's structures when assessed on GDT_TS for both test sets. The differences between protein distributions were evaluated by means of Kullback-Leibler Divergence. Additionally, we found that our pipeline exhibits greater robustness with respect to the results of Terwilliger. Therefore, our results highlight AlphaMod's potential for quality enhancement in protein structure prediction.
In conclusion, the integration of deep learning and conventional protein modelling tools could lead to a significant improvement of the quality of a protein's structure. Moreover, our pipeline shortens the time complexity of manual tasks by coupling different protein quality assessment tools into an integrated and automated solution. In the future we aim to develop a tool to improve specifically for subsets of proteins exhibiting non-common structural features.

*Info*

Reference:
Gil Zuluaga FH, D'Arminio N, Bardozzo F, Tagliaferri R, Marabotti A. An automated pipeline integrating AlphaFold 2 and MODELLER for protein structure prediction. Comput Struct Biotechnol J. 2023 Nov 3;21:5620-5629.

| | |
|---|---|
| *filename* | AlphaModpipeline.png |

*Figure*




| | |
|---|---|
| *Availability* | AlphaMod is freely available at: https://github.com/Fabio-Gil-Z/AlphaMod |

**Dissemination Material**

*Social*

-

*Summary*

-

**Corresponding Author**

| | |
|---|---|
| *Name, Surname* | Anna, Marabotti |
| *Email* | amarabotti@unisa.it |
| *Submitted on* | 01.05.2024 |