Leveraging Self-Supervised Contextual Language Models for Deep Neural Network Antibody CDR-H3 Loop Predictions

David Jimenez[1] **Nazim Medzhidov**[1]
david.jimenez@elix-inc.com nazim.medzhidov@elix-inc.com

[1] Elix Inc., Daini Togo Park Building 3F, 8-34 Yonbancho, Chiyoda-ku, Tokyo 102-0081 Japan

Keywords: Protein modeling, Unsupervised Learning, Antibody Structure Prediction, Transformers, Self-Supervision

Immunoglobulins take a structural conformation that is conserved in most parts, except for the antigen-binding fragment (Fab) that includes six complementarity-determining regions (CDRs). These CDRs are peptide loops on both antibody heavy and light chains that impact antigen recognition. Therefore accurate prediction of CDR structures from amino acid sequences is of great importance in antibody design. While existing methods are capable of predicting folding of five of these CDRs [1], determining the structure of the CDR H3 loop remains a challenge. This is due to the complex diversity in observed folding conformations in CDR H3 loop compared to the canonical folds in other five CDR loops. Deep neural networks strive at capturing complex patterns, which makes them a promising tool for protein structure prediction. However, the domain of antibody structure prediction has relatively scarce annotated data compared to general proteins, which usually limits the depth and complexity of the models that can be trained. To bypass this limitation, we propose to use a contextual language model trained unsupervised on a large general protein dataset using a proxy task, which is then joined with a H3-Loop predicting model. Here, the language model spans a representation space reflecting protein biochemical knowledge, which is exploited by the H3-Loop model for the antibody structure prediction. This results in a deeper and more expressive model that outperforms the prediction capabilities of the H3-Loop model alone.

[1]Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, Sheriff S, Padlan EA, Davies D, Tulip WR, et al. Conformations of immunoglobulin hypervariable regions. Nature. 1989 Dec 21-28;342(6252):877-83. doi: 10.1038/342877a0. PMID: 2687698.

[2]Jeffrey A Ruffolo, Carlos Guerra, Sai Pooja Mahajan, Jeremias Sulam, Jeffrey J Gray, Geometric potentials from deep learning improve prediction of CDR H3 loop structures, Bioinformatics, Volume 36, Issue Supplement_1, July 2020, Pages i268–i275

[3] Alexander Rives, Siddharth Goyal, Joshua Meier, Demi Guo, Myle Ott, C. LawrenceZitnick, Jerry Ma, Rob Fergus, Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences, *Proceedings of the National Academy of Sciences* doi: 10.1073/pnas.2016239118