HELIX ENGINEERING: COMBINING THE POWER OF 3DM WITH AI TO DISRUPT PROTEIN ENGINEERING

Stephan Heijl, Bio-Prodict B.V. , the Netherlands
sheijl@bio-prodict.nl
Jeanine Boot, Bio-Prodict B.V. , the Netherlands
Bastiaan Brier, Bio-Prodict B.V. , the Netherlands
Tom van den Bergh, Bio-Prodict B.V. , the Netherlands
Bas Vroling, Bio-Prodict B.V. , the Netherlands
Henk-Jan Joosten, Bio-Prodict B.V, the Netherlands

Key Words: Protein Engineering, Machine Learning

The high dimensionality and practically infinite size of the sequence space requires effective techniques to explore, navigate and improve proteins. Machine learning methods have enabled the in silico screening of variants with vastly improved speed, but the current techniques are underwhelming in their accuracy and mainly resort to recombining variants present in the training data. This means that current techniques have a hard time finding novel promising mutations outside their training set.

With its 3DM technology[1] Bio-Prodict has long been at the forefront of providing protein engineering solutions. Recently we generated over 35.000 3DM databases that each contain large amounts of highly integrated protein related data for all protein superfamilies. Using this massive amount of information, we were able to develop Helix[2], a best-in-class AI pathogenicity predictor [3, 4]. Currently we are applying this signature innovation of Bio-Prodict to solve protein engineering problems.

We have automated conventional 3DM-based search strategies so that the program can now smartly pre-select positions to mutate initially. This step is designed to replace the need for broad alanine scanning or randomization of complete proteins often used by large pharmaceutical and biotech companies to find promising starting positions. It also removes the need for laborious bioinformatics analysis of the target protein to select positions manually.

For next rounds of evolution, we have developed a deep learning based ensemble architecture. Using multiple deep mutational scanning datasets we showed that this pipeline outperforms legacy machine learning methods[5] on average by 71.5% when mutations were selected randomly and with 115% when the 3DM-based initial selection step was used. Furthermore, we have shown that even as few as 50 initial mutations are needed to train a target specific AI network that already yields competitive hit rates.

In conclusion, we present an integrated methodology that combines the powerful 3DM technology with multiple state-of-the-art AI techniques to smartly optimize proteins. This can drastically decrease the number of rounds and samples required, thereby lowering costs and labtime. We expect that soon, once we have fully utilized all data inside 3DM to our AI methods, the Helix Engineering platform will become even more accurate, just like we did for Helix, making Helix Engineering the best-in-class tool to solve protein engineering problems.

[1] R. Kuipers et al, 3DM: systematic analysis of heterogeneous superfamily data to discover protein functionalities. Proteins, 78(9):2101–2113, July 2010.
[2] B. Vroling, S Heijl, White paper: The Helix Pathogenicity Prediction Platform, arXiv:2104.01033 [q-bio.GN], 2021
[3] L. Dorling et al, Breast cancer risks associated with missense variants in breast cancer susceptibility genes, https://doi.org/10.1101/2021.09.02.21262369, 2021
[4] R. Boonen, Functional analysis identifies damaging CHEK2 missense variants associated with increased cancer risk, Cancer Research, December 13, 2021
[5] Y. Xu et al., Deep Dive into Machine Learning Models for Protein Engineering, J. Chem. Inf. Model., vol. 60, no. 6, pp. 2773–2790, June, 2020