TRANSFORMING PROTEIN ENGINEERING: ADVANCED INTEGRATION OF DEEP LEARNING AND 3DM TECHNOLOGY FOR SUPERIOR PROTEIN FUNCTION PREDICTIONS

Henk-Jan Joosten, Bio-Prodict BV
joosten@bio-prodict.nl
Bas Vroling, Bio-Prodict BV
Stephan Heijl, Bio-Prodict BV
Tom van den Bergh, Bio-Prodict BV

Key Words: Protein engineering, artificial intelligence, 3DM protein superfamily systems, Helix Engineering.

Transforming Protein Engineering: Advanced Integration of Deep Learning and 3DM Technology for Superior Protein Function Predictions

Deep learning has made a significant impact on the field of protein engineering, enabling the rapid exploration and optimization of protein sequences. Despite the availability of numerous benchmark datasets to evaluate model performance, the metrics used are often not suitable for measuring protein function and the proxy values used lead to inaccurate results. This reduces the ability to predict functional improvements in protein variants.

In this study, we assess the performance of five different protein language models. We measure spearman correlation for the top 40% variants across 56 datasets from ProteinGym[1], revealing an average performance of less than 0.1. We conclude that while these models can find non-functional variants, they are much less effective in discovering variants that enhance functionality. Within the sequence-structure-function' paradigm, AI is able to effectively map the relation between sequence and structure, but mapping the relation between sequence and structure to protein function remains elusive.

Recognizing these limitations, we have leveraged the powerful capabilities of Bio-Prodict's 3DM[2] protein family analysis technology and integrated it with multiple state-of-the-art AI techniques to develop enhanced predictors for protein engineering. We worked together with a large commercial partner to create a novel dataset and test our predictions. Our advanced approach, Helix Engineering, was trained on 165 single variants. We predicted a round with only 92 clones with one quintuple variant being 53 times more active than wildtype. . This experiment was replicated using brute force classic techniques, but while a similar result was achieved, obtaining a comparable increase in fitness required roughly 70 times as much experimental screening efforts. This shows that our approach efficiently identifies high fitness variants, which leads to significant cost savings.

[1] Notin, P., Dias, M., Frazer, J., Hurtado, J.M., Gomez, A.N., Marks, D. &amp; Gal, Y.. (2022). Tranception: Protein Fitness Prediction with Autoregressive Transformers and Inference-time Retrieval. *Proceedings of the 39th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* 162:16990-17017 Available from https://proceedings.mlr.press/v162/notin22a.html.
[2] R. Kuipers et al, 3DM: systematic analysis of heterogeneous superfamily data to discover protein functionalities. Proteins,
78(9):2101–2113, July 2010.