BIOTRANS
LA ROCHELLE 2023

N°844 / OC
TOPIC(s) : Artifical intelligence / computational methods / Enzyme discovery and engineering

# Data-driven protein engineering through machine learning models

## AUTHORS

Mehdi D. DAVARI / LEIBNIZ INSTITUTE OF PLANT BIOCHEMISTRY HALLE, WEINBERG 3, HALLE (SAALE)
Alexander-Maurice ILLIG / RWTH AACHEN UNIVERSITY, WEINBERG 3, AACHEN
Niklas E. SIEDHOFF / RWTH AACHEN UNIVERSITY, WORRINGERWEG 3, AACHEN
Ulrich SCHWANEBERG / RWTH AACHEN UNIVERSITY, WEINBERG 3, AACHEN

## PURPOSE OF THE ABSTRACT

Protein engineering is nowadays used routinely in developing biocatalysts for biotechnological, biomedicine, and life sciences applications. Recently, data-driven strategies attracted attention in protein design and engineering because of advances in large experimental databanks of proteins, next-generation sequencing (NGS), high-throughput screening (HTS) methods, and the development of artificial intelligence algorithms [1-3]. However, the reliable prediction of beneficial amino acid substitutions, their combination, and effect on functional properties still pose significant challenges in machine learning assisted protein engineering [4]. In this presentation, we describe a general-purpose framework (PyPEF: Pythonic Protein Engineering Framework[5-6]) for data-driven protein engineering by combination of machine learning methods with signal processing and statistical physics techniques. PyPEF assist in the identification and selection of beneficial proteins of a given sequence space by either systematically or randomly exploring the fitness of variants and by sampling random evolution pathways. The predictive accuracy and throughput performance of PyPEF was evaluated based on four public protein and enzyme datasets using common regression models. It turns out that the PyPEF could efficiently predict the fitness of protein sequences for different target properties (predictive models with coefficient of determination values ranging from 0.58 to 0.92). By combining machine learning and protein evolution, PyPEF enabled the screening of proteins with various functions reaching a screening capacity of more than 500,000 protein sequence variants in the timeframe of only a few minutes on a standard PC. PyPEF exhibited significant accuracies on four public datasets (different proteins and properties) and underlined the potential of integrating data-driven technologies for covering different philosophies by either predicting the fitness of the variants to the highest accuracy or capturing the general trend of introduced mutations on the fitness in directed protein evolution campaigns. In essence, PyPEF provide a powerful solution to current sequence exploration and combinatorial problems present in protein engineering through exhaustive in silico screening of the protein sequence space.

## FIGURES

## FIGURE 1

## FIGURE 2

## KEYWORDS

Protein engineering | machine learning | Data-driven | enzyme engineering

## BIBLIOGRAPHY

[1] N. E. Siedhoff, U. Schwaneberg and M. D. Davari, Methods in Enzymology 2020, 643, 281-315.

[2] B. J. Wittmann, K. E. Johnston, Z. Wu and F. H. Arnold, Current Opinion in Structural Biology 2021, 69, 11-18.

[3] K. K. Yang, Z. Wu and F. H. Arnold, Nature methods 2019, 16, 687-694.

[4] S. Mazurenko, Z. Prokop and J. Damborsky, ACS Catalysis 2019, 10, 1210-1223.

[5] N. E. Siedhoff, A.-M. Illig, U. Schwaneberg and M. D. Davari, Journal of Chemical Information and Modeling 2021.

[6] A.-M. Illig, N. E. Siedhoff, U. Schwaneberg and M. D. Davari, bioRxiv 2022, 2022.2006. 2007.495081.