Machine Learning for Improved Directed Evolution Efficiency and Outcome

Bruce J. Wittmann[1], Zachary Wu[2], Frances Arnold[1,2]

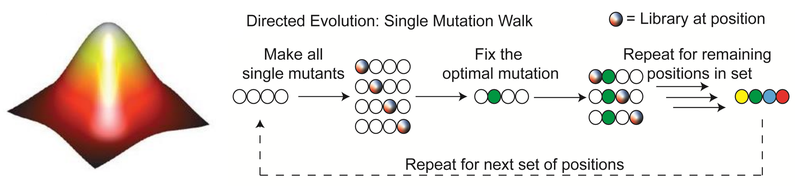
[1]Division of Biology and Biological Engineering, [2]Division of Chemistry and Chemical Engineering
California Institute of Technology, Pasadena, California, 91125

Abstract

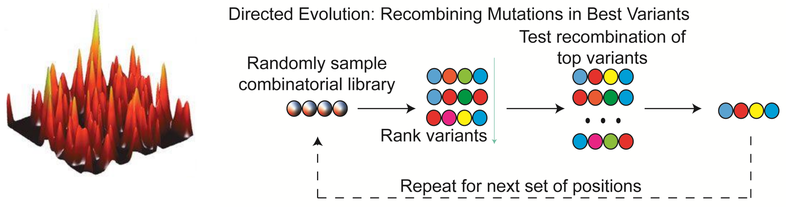
Enzymes inexpensively, efficiently, and selectively catalyze useful and challenging reactions under mild conditions. With an increasing drive to adopt greener practices, these biocatalysts are becoming more prevalent in industrial syntheses. Optimal enzymes are developed by searching the protein fitness landscape, the conceptual relationship between protein amino acid sequence and fitness. This landscape is beyond astronomical in size (for reference, there are 20^{100} unique 100-residue-long proteins and ~10^{80} atoms in the observable universe), and so cannot be exhaustively searched. Rather than randomly searching the fitness landscape, protein engineers improve enzyme activity using directed evolution (DE). Standard DE techniques are burdened by a limited ability to screen protein variants, and often arrive at local fitness optima rather than the global. By integrating machine learning into the directed evolution workflow, we were able to both lower the number of protein variants screened and increase the probability of finding the global optimum. In the future, we will encode substrates in our workflow to specify enzyme productivity with multiple substrates.

Background: Fitness Landscapes and Standard Directed Evolution Approaches

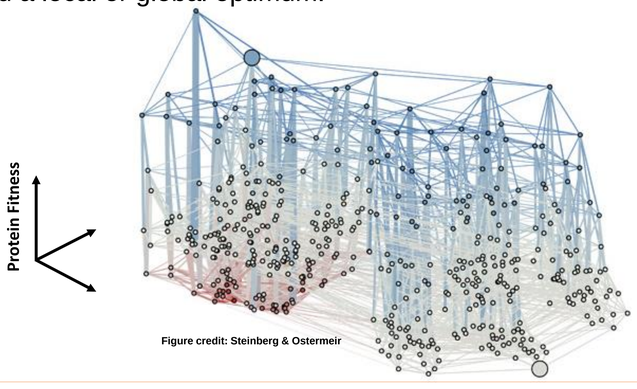
Single mutant walks climb smooth local fitness peaks.



Recombination might jump to a different local fitness peak.

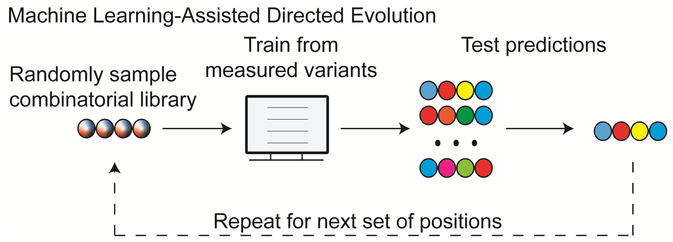


It is impossible to know with standard methods if you are climbing toward a local or global optimum.

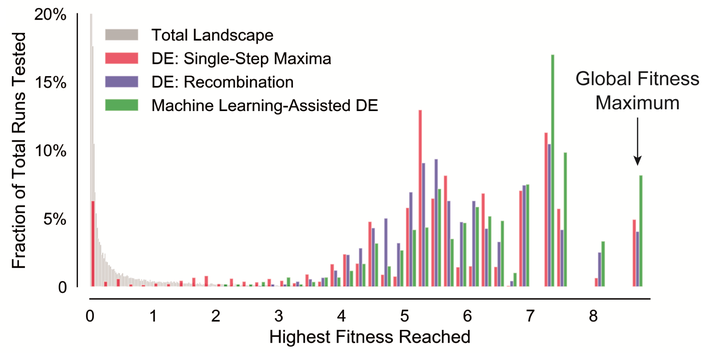


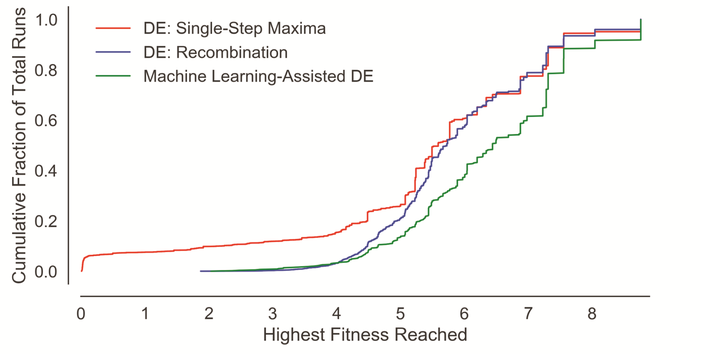
Machine Learning-Assisted Directed Evolution (ML-DE): Procedure and Validation

Machine learning removes some randomness from directed evolution by learning a function that describes the fitness landscape.



Simulated directed evolution studies using a published fitness landscape of human GB1 showed that ML-DE is more likely to reach the global fitness optimum than standard DE techniques.

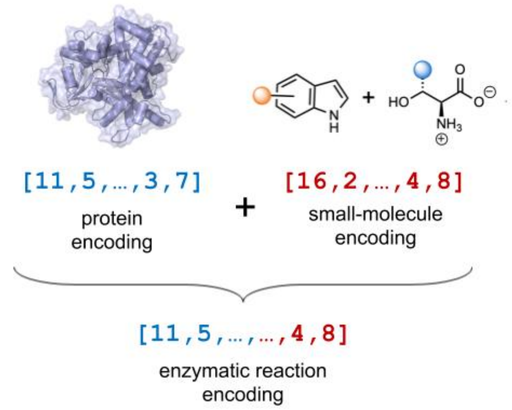




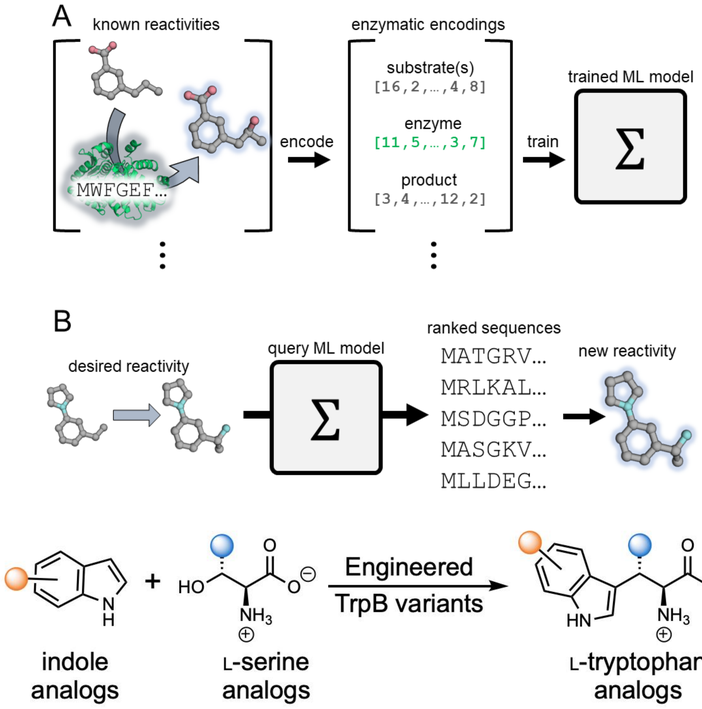
Approach	Expected Fitness Reached	Fraction of Runs that Reach the Maximum
ProSAR	3.00	0.20%
Recombining 3 top mutations at each position	4.07	1.18%
1000 random combinatorial sequences	5.04	0.40%
Single step mutation walk	5.41	4.91%
DE+ML (300 total sequences)	5.46	3.50%
Testing random sequences; recombining the top 3	5.93	4.03%
DE+ML (570 total sequences)	6.42	8.17%

Future Directions: Substrate Encoding for Rapid Access to New Substrate Diversity

Our previous ML-DE work only uses protein encodings. Future iterations will involve both protein and substrate encodings.



Including substrate encodings will allow us to identify protein sequences that are optimal for activity against a substrate.



References

Kan, S. B. J. et al. *Science* **354**, 1048–1051 (2016).
Romero, P. A. et al. *Proc. Natl. Acad. Sci.* **110**, E193–E201 (2013).
Steinberg, B. & Ostermeier, M. *Sci. Adv.* **2**, e1500921 (2016).
Wu, N. C. et al. *Elife* **5**, 1–21 (2016).
Wu, Z. et al. *Proc. Natl. Acad. Sci.* **116**, 8852–8858 (2019).

Acknowledgements

This work is supported by the U.S. Army Research Office Institute for Collaborative Biotechnologies [W911F-09-0001 to FHA] and the National Science Foundation [GRF2017227007 to ZW],