LOW N, HIGH N PROTEIN ENGINEERING.

Gabriel Abrahams, Harrison Steel
Department of Engineering
University of Oxford
Oxford, United Kingdom
{gabriel.abrahams,harrison.steel}@eng.ox.ac.uk

Carlos Outeiral, Charlotte Deane
Department of Statistics
University of Oxford
Oxford, United Kingdom
{carlos.outeiral,charlotte.deane}@stats.ox.ac.uk

ABSTRACT

Machine learning assisted directed evolution often involves experimentally collecting data from a relatively small number of variants to update a surrogate model, due to experimental limitations of characterisation and sequencing at high throughput. We propose an alternative approach, involving collecting high-throughput experimental data in a manner that results in a large number of characterised variants at the cost of reduced information: although the sequences and the measured fitness values are known, their correspondence is not. In particular we explore applying this method to the optimisation of a recently discovered phenomenon: magnetically sensitive fluorescent proteins.

1 INTRODUCTION

1.1 HIGH N, LOW N DIRECTED EVOLUTION

Routine optimisation of the properties of proteins, particularly with novel functions such as enzymatic activity, remains a challenging task. Directed evolution is a powerful experimental method for exploring the sequence search space (Arnold, 2018), however it can be burdensome to experimentally characterise a large number of variants. Recent approaches to high-throughput methods for cell characterisation promise to overcome this by enabling large-scale characterisation of protein function and selection using a combination of technologies such as microfluidics (Potvin-Trottier et al., 2018), automation (Chait et al., 2017; Yu et al., 2023) and in-vivo mutagenesis (Molina et al., 2022). In parallel, machine learning techniques are being developed that can learn from limited data, and make informed predictions that guide the experimental search, thus requiring fewer variants to be characterised. Machine learning directed evolution (MLDE) uses an algorithmic approach to reduce experimental burden by introducing in-silico exploration to the directed evolution loop; typically a continuously updated surrogate model predicts protein function, and an exploration function proposes a relatively small number of candidates for experimental study (Wu et al., 2019; Freschlin et al., 2022; Wang et al., 2023). While methods to perform high-throughput in-vitro sequencing are becoming available (Rodriguez-Mateos et al., 2020), here we develop an alternate approach; asking whether a high-throughput approach that enables large scale phenotype characterisation at the downside of reduced sequencing to function information can benefit from machine learning assistance.

1.2 MAGNETO-FLUORESCENT PROTEINS

There is still much to learn about the processes nature has evolved that take advantage of the many ways in which biomolecules can interact with physical processes and forces, such as the sensing of magnetic fields by birds and other species (Mouritsen, 2018). Furthermore, the ability to engineer biomolecular systems to interact with physical processes to sense and actuate biological functions has already revolutionised microbiology, for example with voltage sensing and fluorescent reporters like GFP (van Dijk et al., 2018; Tsien, 1998). Recently, Hayward et al. reported a magnetic field effect (MFE) response in the fluorescence of widely used fluorescent proteins including EGFP (Cormack et al., 1996) and mScarlet (Bindels et al., 2017). When these fluorescent proteins are mixed with flavin molecules, either in-vitro or in-vivo, the fluorescent signal was found to reversibly reduce in the presence of a ($\sim$ 10 mT) magnetic field. This effect is reminiscent of other flavoprotein systems (Evans et al., 2013; 2015; Déjean et al., 2020), though the mechanism is at present unknown. In-vivo magneto-responsive proteins could have important consequences in medical biology, for example magnetogenetic tools or drug delivery mechanisms (c.f. optogenetics which relies on light access, which is impeded by opacityPacker et al. (2013)), as well as applications in industrially focused synthetic biology as an additional means of applying control to a cellular system. A more immediate application could be to broaden the library of fluorescent reporters used to probe genetic functions or circuits. At present, fluorescent reporters of different colours can be used to monitor multiple genetic functions, however the number of signals which can be distinguished is quickly limited due to signal noise and spectrum overlap. Engineered magneto-fluorescent reporters could address this in two respects: firstly, modulation of the signal enables the implementation of lock-in amplification, greatly increasing signal integrity, perhaps allowing fluorescent reporters of significantly overlapping spectra to be separated. Building on this, a library

of magneto-fluorescent proteins of the same colour but with different dynamical responses could potentially be engineered (with better understanding of the effect rationally, or through the directed evolution method we propose), expanding the number of signals that can be multiplexed.

1.3 Optimising Magneto-Fluorescent Proteins

To measure the magnetic response of GFP in bacteria we employ a custom widefield microscope (Fig. 1), see Appendix A.1 that will have an integrated microfluidics device (Potvin-Trottier et al., 2018) (see Fig. 2). Genetically designed cells continuously replicate, providing a stream of clonal cells that flow in confined channels. By modulating the applied magnetic field, the fluorescence intensity is quantified following image segmentation into individual cells. Fluorescence is thus recorded over time and integrated over many identical clones to systematically phenotype variants (e.g. by the contrast in fluorescence, or temporal dynamics of the response). Due to the miniature scale of the microfluidics, traditional sequencing via collection of individual variants is not possible. However, it is possible to perform manual selection (i.e. terminate some proportion of lineages using ultra-violet light) and sequence the remaining variants in a batch. Thus a simple strategy is binary batching: collecting the sequences of the top 20% of variants and their phenotypes, and similarly (if the original sequence space is known) the sequences and phenotypes of the bottom 80% of variants though within each batch the sequence-to-function identity is lost.

2 Discussion

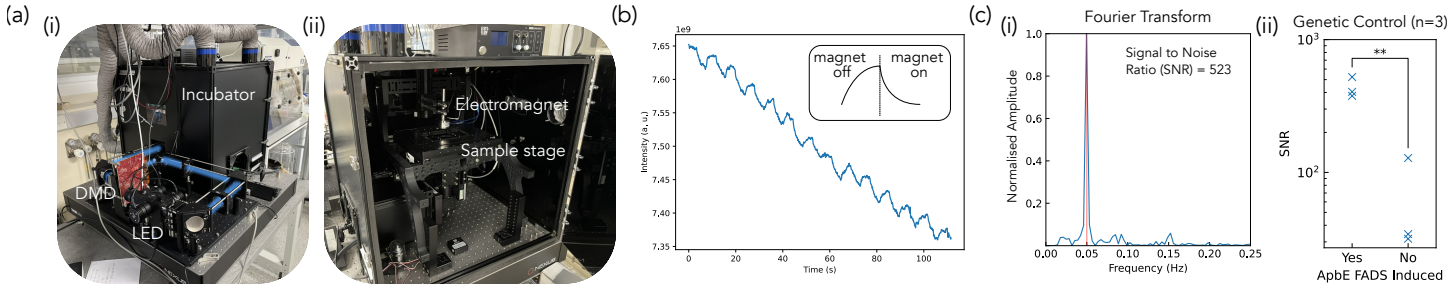

Figure 1: (a) Widefield microscope. (i) Optical system. (ii) Stage inside the incubator. (b) Magnetically modulated fluorescence of *E. Coli* expressing the MagnetoGFP proteins. (c) (i) Fourier transform (FT) of a similar fluorescent response (note: not the same data as (i)). SNR = signal/noise is extracted from the FT by integrating in (signal) and out of (noise) a small frequency band around the magnet on/off frequency (in this case 0.05 Hz). (ii) where EGFP was expressed, ApbE and FADS were either both induced or neither induced, demonstrating lack of magnetic response when flavins are not present (ApbE transfers flavin to FlavinTag, FADS synthesises flavin).

2.1 Machine learning pipeline

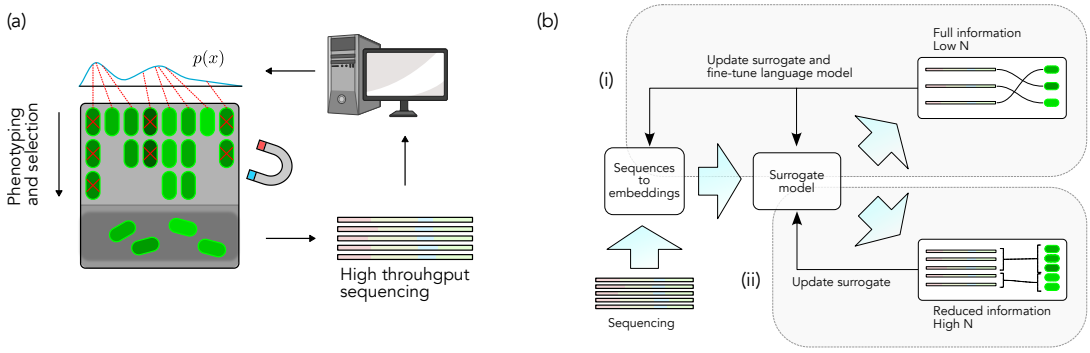

Figure 2: (a) High throughput directed evolution cycle: variants are produced probabilistic, characterised at scale in a microfluidics device, selected for, sequenced, and the data processed before repeating. (b) (i) Low N approach: rather than using the microfluidics device, a small number of variants are individually characterised, which is used to update surrogate model and fine-tune the embedding model. (ii) High N approach: a large number of variants are characterised, but the sequence to function relationship is not known. Only the surrogate model is updated, e.g. by classification loss on batched fitness classes.

We propose bringing together "low N" approaches to directed evolution (Biswas et al., 2021) in which foundation models encode general (presumably) biophysical properties of proteins in encodings (Unsal et al., 2022) that are used to train regression or more

complex surrogate models, with a "high N" approach where $10^3 - 10^6$ of variants are sequenced, but are batched into two or more fitness classes. A small scale experiment is initially performed that measures the fitness of ~ 100 variants. This data is used to train a surrogate model by conventional means. Subsequently, a reinforcement algorithm such as EvoPlay (Wang et al., 2023) proposes a variant library for experimental characterisation. Such a library can be produced at scale by designing a mutagenesis library that produces a distribution of sequences similar to that of the model output (Weinstein et al., 2021). The experimental outcome are classified batches of sequences and phenotypes. With this data there are a number of methods to refine the surrogate model. Firstly, the loss function can be formulated as a classification problem: whether the surrogate prediction on the sequences in a given class correctly fall within the fitness range of that class or not. Secondly, we use the distribution of fitness data within a class, i.e. calculate the loss between the distribution of measured fitness in a class and the distribution of finesses predicted by the surrogate on the sequences of that class. Such methods must be approached with care: for example, simply training a model to produce the correct output *distribution* could result in a model whose output depends nonsensically on its input. Optimal transportTai et al. (2021) offers a promising solution: formulating the problem as the optimal way in which to re-distribute the class sequence embeddings such that the surrogate model correctly predicts the class fitness distribution. For a minimal demonstration of these techniques, see Appendix A.2.

REFERENCES

Frances H. Arnold. Directed Evolution: Bringing New Chemistry to Life. *Angewandte Chemie International Edition*, 57(16):4143–4148, 2018. ISSN 1521-3773. doi: 10.1002/anie.201708408.

Daphne S Bindels, Lindsay Haarbosch, Laura Van Weeren, Marten Postma, Katrin E Wiese, Marieke Mastop, Sylvain Aumonier, Guillaume Gotthard, Antoine Royant, Mark A Hink, and Theodorus W J Gadella. mScarlet: A bright monomeric red fluorescent protein for cellular imaging. *Nature Methods*, 14(1):53–56, January 2017. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.4074.

Surojit Biswas, Grigory Khimulya, Ethan C. Alley, Kevin M. Esvelt, and George M. Church. Low-N protein engineering with data-efficient deep learning. *Nature Methods*, 18(4):389–396, April 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01100-y.

Remy Chait, Jakob Ruess, Tobias Bergmiller, Gašper Tkačik, and Călin C. Guet. Shaping bacterial population behavior through computer-interfaced control of individual cells. *Nature Communications*, 8(1):1535, November 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-01683-1.

Brendan P. Cormack, Raphael H. Valdivia, and Stanley Falkow. FACS-optimized mutants of the green fluorescent protein (GFP). *Gene*, 173(1):33–38, January 1996. ISSN 0378-1119. doi: 10.1016/0378-1119(95)00685-0.

Victoire Déjean, Marcin Konowalczyk, Jamie Gravell, Matthew J. Golesworthy, Catlin Gunn, Nils Pompe, Olivia Foster Vander Elst, Ke-Jie Tan, Mark Oxborrow, Dirk G. A. L. Aarts, Stuart R. Mackenzie, and Christiane R. Timmel. Detection of magnetic field effects by confocal microscopy. *Chemical Science*, 11(30):7772–7781, August 2020. ISSN 2041-6539. doi: 10.1039/D0SC01986K.

Emrys W. Evans, Charlotte A. Dodson, Kiminori Maeda, Till Biskup, C. J. Wedge, and Christiane R. Timmel. Magnetic field effects in flavoproteins and related systems. *Interface Focus*, 3(5):20130037, October 2013. doi: 10.1098/rsfs.2013.0037.

Emrys W. Evans, Jing Li, Jonathan G. Storey, Kiminori Maeda, Kevin B. Henbest, Charlotte A. Dodson, P. J. Hore, Stuart R. Mackenzie, and Christiane R. Timmel. Sensitive fluorescence-based detection of magnetic field effects in photoreactions of flavins. *Physical Chemistry Chemical Physics*, 17(28):18456–18463, July 2015. ISSN 1463-9084. doi: 10.1039/C5CP00723B.

Chase R. Freschlin, Sarah A. Fahlberg, and Philip A. Romero. Machine learning to navigate fitness landscapes for protein engineering. *Current opinion in biotechnology*, 75:102713, June 2022. ISSN 0958-1669. doi: 10.1016/j.copbio.2022.102713.

Rebecca Frank Hayward, Julia R. Lazzari-Dean, Andrew G. York, and Maria Ingaramo. GFP Magnetofluorescence. https://andrewgyork.github.io/gfp_magnetofluorescence/.

Rosana S. Molina, Gordon Rix, Amanuella A. Mengiste, Beatriz Álvarez, Daeje Seo, Haiqi Chen, Juan E. Hurtado, Qiong Zhang, Jorge Donato García-García, Zachary J. Heins, Patrick J. Almhjell, Frances H. Arnold, Ahmad S. Khalil, Andrew D. Hanson, John E. Dueber, David V. Schaffer, Fei Chen, Seokhee Kim, Luis Ángel Fernández, Matthew D. Shoulders, and Chang C. Liu. In vivo hypermutation and continuous evolution. *Nature Reviews Methods Primers*, 2(1):1–22, May 2022. ISSN 2662-8449. doi: 10.1038/s43586-022-00119-5.

Henrik Mouritsen. Long-distance navigation and magnetoreception in migratory animals. *Nature*, 558(7708):50–59, June 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0176-1.

Adam M. Packer, Botond Roska, and Michael Häusser. Targeting neurons and photons for optogenetics. *Nature neuroscience*, 16(7):805–815, July 2013. ISSN 1097-6256. doi: 10.1038/nn.3427.

Gabriel Peyré and Marco Cuturi. Computational Optimal Transport, March 2020.

Laurent Potvin-Trottier, Scott Luro, and Johan Paulsson. Microfluidics and single-cell microscopy to study stochastic processes in bacteria. *Current Opinion in Microbiology*, 43:186–192, June 2018. ISSN 1369-5274. doi: 10.1016/j.mib.2017.12.004.

Pablo Rodriguez-Mateos, Nuno Filipe Azevedo, Carina Almeida, and Nicole Pamme. FISH and chips: A review of microfluidic platforms for FISH analysis. *Medical Microbiology and Immunology*, 209(3):373–391, June 2020. ISSN 1432-1831. doi: 10.1007/s00430-019-00654-1.

Kai Sheng Tai, Peter Bailis, and Gregory Valiant. Sinkhorn Label Allocation: Semi-Supervised Classification via Annealed Self-Training, June 2021.

Roger Y. Tsien. THE GREEN FLUORESCENT PROTEIN. *Annual Review of Biochemistry*, 67(1):509–544, June 1998. ISSN 0066-4154, 1545-4509. doi: 10.1146/annurev.biochem.67.1.509.

Serbulent Unsal, Heval Atas, Muammer Albayrak, Kemal Turhan, Aybar C. Acar, and Tunca Doğan. Learning functional properties of proteins with language models. *Nature Machine Intelligence*, 4(3):227–245, March 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00457-9.

Erwin L. van Dijk, Yan Jaszczyszyn, Delphine Naquin, and Claude Thermes. The Third Revolution in Sequencing Technology. *Trends in Genetics*, 34(9):666–681, September 2018. ISSN 0168-9525. doi: 10.1016/j.tig.2018.05.008.

Yi Wang, Hui Tang, Lichao Huang, Lulu Pan, Lixiang Yang, Huanming Yang, Feng Mu, and Meng Yang. Self-play reinforcement learning guides protein engineering. *Nature Machine Intelligence*, 5(8):845–860, July 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00691-9.

Eli N. Weinstein, Alan N. Amin, Will Grathwohl, Daniel Kassler, Jean Disset, and Debora S. Marks. Optimal Design of Stochastic DNA Synthesis Protocols based on Generative Sequence Models. Preprint, Synthetic Biology, October 2021.

Zachary Wu, S. B. Jennifer Kan, Russell D. Lewis, Bruce J. Wittmann, and Frances H. Arnold. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proceedings of the National Academy of Sciences*, 116(18):8852–8858, April 2019. doi: 10.1073/pnas.1901979116.

Tianhao Yu, Aashutosh Girish Boob, Nilmani Singh, Yufeng Su, and Huimin Zhao. In vitro continuous protein evolution empowered by machine learning and automation. *Cell Systems*, 14(8):633–644, August 2023. ISSN 2405-4712. doi: 10.1016/j.cels.2023.04.006.

A Appendix

A.1 Microscope

We use a custom widefield microscope developed on the ASI RAMM platform (see Fig. 1). The system has a 0.5×0.5 mm^2 FOV using a Kinetix sCMOS camera and Nikon 40x/0.95 NA objective. The imaging setup is contained within a temperature controlled incubator. Initial experiments have been performed with cells immobilised on an Agar pad, high-throughput experiments will be performed using a microfluidics device that confines unique variants to channels. The magnetic field is applied using an electromagnet connected to a computer controlled current-set power supply. LED illumination for fluorescence imaging is performed using the ASI Tiger system with a 450nm LED at ~ 1 W/cm^2. Cells can be selectively illuminated by selecting switching pixels on and off using the Texas Instruments DLP660 4K digital micromirror device (DMD)- in future this will be used to perform selection.

A.2 Learning with Classification and Optimal Transport

Here we demonstrate a simple proof of principal using classification and optimal transport on a reduced dataset to learn to fit a trivial function $y = x^2$. Training points are chosen by uniformly sampling x values (x_{train}). A small multilayer-perceptron with the number of parameters in each layer being [12, 8, 4, 1] is trained using mean-squared error (MSE) loss on the interval $x \in X_1 = [0, 1]$. As seen in Fig. 3(a), in this region the model fits the target perfectly, but understandably fails to generalise. A second model is trained by alternatively updating with MSE loss on region X_1 as before, and also on the region $x \in X_2 = [1, 2]$: the data available in X_2 is now only a binary classification: 1 if the data point is in the top 20% of values in X_2 and -1 otherwise (x-values are sampled randomly to generate the training set). This improves the model, but also introduces an anomalous shape to the prediction function. Finally, the model is alternately trained with MSE loss on X_1, classification loss on X_2, and also Sinkhorn divergencePeyré & Cuturi (2020) on X_2, where the x_{train} points have been shuffled relative to the y_{train} to ensure the model does not have access to the direct correspondence between x_{train} and y_{train}. This appears to further improve the model, leading to better fitting in the region X_2 and better generalisation nearby. In Fig. 3(b), it can be seen that the inclusion of optimal transport derived loss indeed leads to the output distribution being shifted towards that of the target distribution.

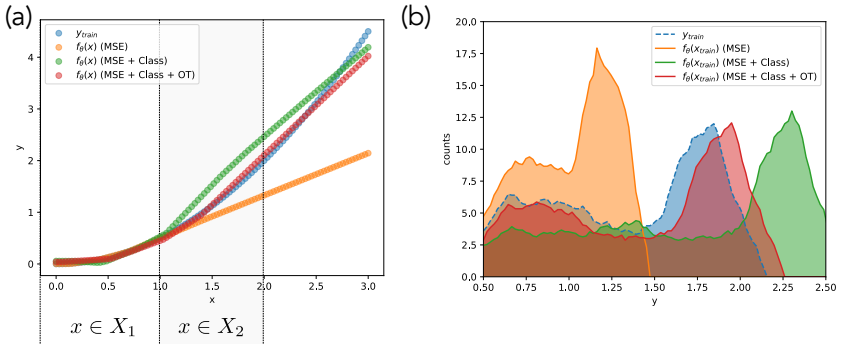

Figure 3: f_θ is the predictive model parameterised by θ. (a) Target function and model predictions after training on region X_1 with MSE loss, or X_1 with MSE loss and X_2 with classification (hinge loss) and optimal transport (OT) using Sinkhorn divergence. (b) Histogram of y values sampled on uniformly chosen $x \in X_2$, for the target function and models. Note the data has been balanced, i.e. points are uniformly removed from the lower 80% such that there are the same number of training points as in the upper 20%.