

## MACHINE LEARNING

# Lighting up protein design

Using a neural network to predict how green fluorescent proteins respond to genetic mutations illuminates properties that could help design new proteins.

GRZEGORZ KUDLA AND MARCIN PLECH

**Related research article** Gonzalez Somermeyer L, Fleiss A, Mishin AS, Bozhanova NG, Igolkina AA, Meiler J, Alaball Pujol M-E, Putintseva EV, Sarkisyan KS, Kondrashov FA. 2022. Heterogeneity of the GFP fitness landscape and data-driven protein design. *eLife* 11:e75842. doi: [10.7554/eLife.75842](https://doi.org/10.7554/eLife.75842)

Protein engineering is a growing area of research in which scientists use a variety of methods to design new proteins that can perform certain functions. For instance, enzymes that can biodegrade plastics, materials inspired by spider silk, or antibodies to neutralize viruses (Lu *et al.*, 2022; Shan *et al.*, 2022).

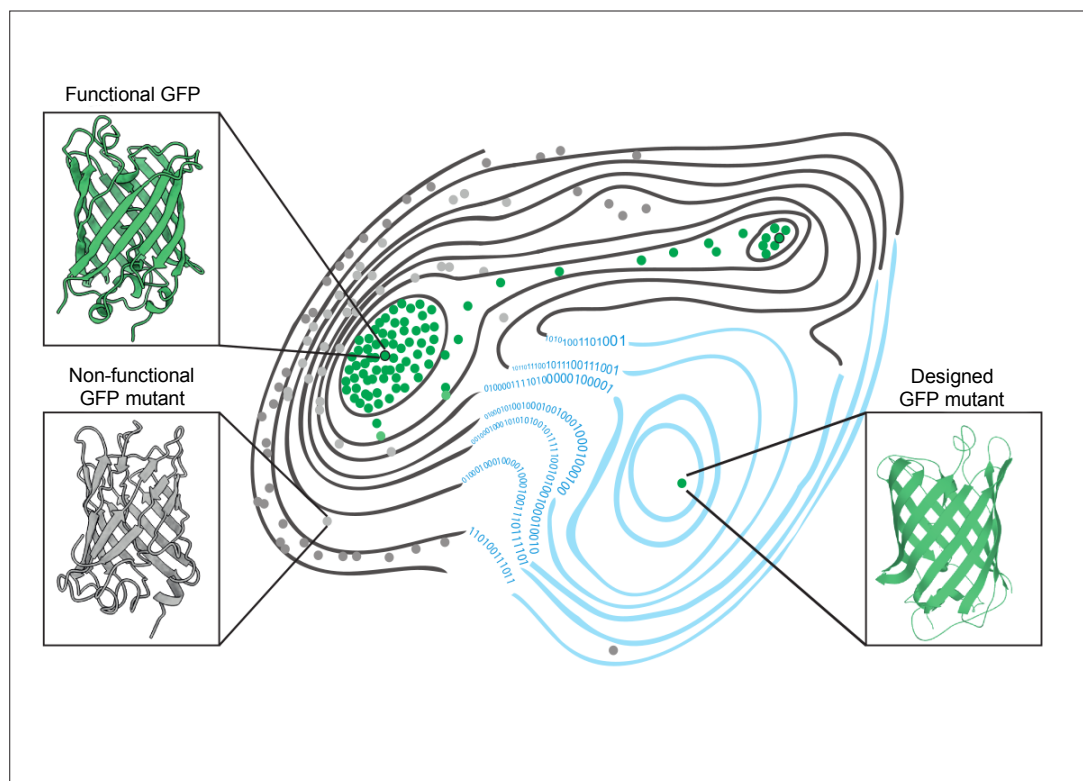
In the past, protein engineering has commonly relied on directed evolution, a laboratory procedure that mimics natural selection. This involves randomly mutating the genetic sequence of a naturally occurring protein to create multiple variants with slightly different amino acids. Various selection pressures are then applied to identify the ‘fittest’ variants that best carry out the desired role (Chen *et al.*, 2018). Alternatively, researchers can use a rational design approach, in which new proteins are built using principles learned from the study of known protein structures (Anishchenko *et al.*, 2021).

Now, in eLife, Fyodor A Kondrashov (from the Institute of Science and Technology Austria and the Okinawa Institute of Science and Technology Graduate University) and colleagues – including Louisa Gonzalez Somermeyer as first author – have combined elements of both approaches

to engineer new variants of naturally occurring green fluorescent proteins (GFP; Gonzalez Somermeyer *et al.*, 2022). First, the team (who are based at various institutes in Austria, Japan, the United States, the United Kingdom, Germany and Russia) generated tens of thousands of GFP variants that differed from each other by three to four mutations on average, and measured their fluorescence. This was used to create a ‘fitness landscape’ showing how the genetic sequence of each mutant relates to its performance (Figure 1). The data was then fed in to a neural network that can expand the landscape by predicting the performance of variants that were not observed experimentally.

Using this machine learning approach, Gonzalez Somermeyer *et al.* were able to design fluorescent proteins that differed from their closest natural relative by as many as 48 mutations. This is remarkable because in most protein mutagenesis experiments it only takes a few mutations before the function of the protein deteriorates. Evolution, on the other hand, can generate functional variants that differ by hundreds of mutations through a process of trial and error, which is akin to walking along a narrow ridge of high fitness one mutational step at a time. The neural network, however, appears to have jumped straight to a distant peak of high fitness (Figure 1). So, how did the network know where to take a leap?

To answer this, Gonzalez Somermeyer *et al.* experimented with three GFP proteins that originated from evolutionarily distant species. They found that machine learning was better at generating functional variants of cgGFP than its two homologues, amacGFP and ppluGFP2 (a fourth homologue, avGFP, was also studied, but



**Figure 1.** The fitness landscape of green fluorescent proteins. Fitness landscapes provide a graphical representation of how a protein's genetic sequence relates to its performance, leading to a multidimensional surface made up of peaks, ridges, and valleys. In the fitness landscape shown, horizontal distance represents the number of mutations that separate variants of a protein, while vertical elevation represented by contour lines indicates the fluorescence of each mutant. Two naturally occurring green fluorescent proteins (GFPs; dots outlined in black) reside on different peaks of the landscape (top left and top right) and are connected by a narrow ridge (area of high fitness). Mutant proteins at the peaks and ridges are all functional and able to fluoresce (green dots), whereas those in the valleys are non-functioning (grey dots). Application of a machine learning algorithm expanded the fitness landscape (right; blue contour lines) by including mutations that are not generated by evolution. This led to the creation of functional, synthetic variants (green dot, bottom right) that reside on different fitness peaks to variants that are naturally occurring.

Image credit: Marcin Plech & Grzegorz Kudla (CC BY 4.0).

not in the machine learning experiment). This allowed the team to look for properties within each protein's genetic sequence and fitness landscape which correlated with its machine learning performance.

Analysis of the fitness landscape revealed that the homologues differed in the number of mutations they could tolerate: it took on average three to four mutations until the fluorescence of cgreGFP and avGFP deteriorated, but seven to eight mutations were needed to compromise the function of amacGFP and pfluGFP2. The proteins also differed in their general sturdiness: pfluGFP2 was stable when exposed to high temperatures, whereas the structure of cgreGFP was more sensitive to changes in temperature.

Finally, Gonzalez Somermeyer et al. found that the increased mutational sensitivity of avGFP

and cgreGFP (and to a lesser degree pfluGFP2) was due to negative epistasis – that is, when an individual mutation is well tolerated, but has a negative effect on the protein's function when combined with other mutations (*Bershtein et al., 2006; Domingo et al., 2019*). The reduced fluorescence of amacGFP, however, could be ascribed almost entirely to additive effects, with each mutation incrementally making the protein less functional.

In order to generate functional variants, the network needs an opportunity to learn which properties of the fitness landscape are relevant from the data provided. The findings of Gonzalez Somermeyer et al. suggest that to predict a protein's function, the algorithm only requires data on the effects of single-site mutations and low-order epistasis (interactions between small


sets of mutations). This is good news for the protein engineering field as it suggests that prior knowledge of high-order interactions between large sets of mutations is not needed for protein design. Furthermore, it explains why the neural network is better at generating new variants of cgrGFP, which has a sharp fitness peak and high prevalence of epistasis.

In sum, these experiments provide a successful case study in protein engineering. An interesting extension would be to analyse the three-dimensional structures of the variants using AlphaFold, an algorithm which can predict a protein's structure based on its amino acid sequence (**Jumper et al., 2021**). This would reveal if data from AlphaFold improves the prediction of functional variants, and help to identify structural features that rendered some of the variants non-fluorescent despite them being predicted to work. In the near future, assessing a new variant's structure before it is synthesized could become a standard validation step in the design of new proteins. Furthermore, studying the fitness landscapes of multiple related variants, as done by Gonzalez Somermeyer et al., could reveal how a protein's genetic sequence and structure changed over the course of evolution (**Hochberg and Thornton, 2017; Mascotti, 2022**). A better understanding of the evolution of proteins will help scientists to engineer synthetic molecules that carry out specific roles.

**Grzegorz Kudla** is in the MRC Human Genetics Unit, The University of Edinburgh, Edinburgh, United Kingdom  
gkudla@gmail.com

 <http://orcid.org/0000-0002-7924-2744>

**Marcin Plech** is in the MRC Human Genetics Unit, The University of Edinburgh, Edinburgh, United Kingdom

 <http://orcid.org/0000-0002-2040-3750>

**Competing interests:** The authors declare that no competing interests exist.

**Published** 19 May 2022

## References

**Anishchenko I**, Pellock SJ, Chidyausiku TM, Ramelot TA, Ovchinnikov S, Hao J, Bafna K, Norn C,

Kang A, Bera AK, DiMaio F, Carter L, Chow CM, Montelione GT, Baker D. 2021. De novo protein design by deep network hallucination. *Nature* **600**:547–552. DOI: <https://doi.org/10.1038/s41586-021-04184-w>, PMID: 34853475

**Bershtein S**, Segal M, Bekerman R, Tokuriki N, Tawfik DS. 2006. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**:929–932. DOI: <https://doi.org/10.1038/nature05385>, PMID: 17122770

**Chen K**, Huang X, Kan SBJ, Zhang RK, Arnold FH. 2018. Enzymatic construction of highly strained carbocycles. *Science* **360**:71–75. DOI: <https://doi.org/10.1126/science.aar4239>, PMID: 29622650

**Domingo J**, Baeza-Centurion P, Lehner B. 2019. The causes and consequences of genetic interactions (epistasis). *Annual Review of Genomics and Human Genetics* **20**:433–460. DOI: <https://doi.org/10.1146/annurev-genom-083118-014857>, PMID: 31082279

**Gonzalez Somermeyer L**, Fleiss A, Mishin AS, Bozhanova NG, Igolkina AA, Meiler J, Alaball Pujol M-E, Putintseva EV, Sarkisyan KS, Kondrashov FA. 2022. Heterogeneity of the GFP fitness landscape and data-driven protein design. *eLife* **11**:e75842. DOI: <https://doi.org/10.7554/eLife.75842>, PMID: 35510622

**Hochberg GKA**, Thornton JW. 2017. Reconstructing ancient proteins to understand the causes of structure and function. *Annual Review of Biophysics* **46**:247–269. DOI: <https://doi.org/10.1146/annurev-biophys-070816-033631>, PMID: 28301769

**Jumper J**, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Židek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**:583–589. DOI: <https://doi.org/10.1038/s41586-021-03819-2>, PMID: 34265844

**Lu H**, Diaz DJ, Czarnecki NJ, Zhu C, Kim W, Shroff R, Acosta DJ, Alexander BR, Cole HO, Zhang Y, Lynd NA, Ellington AD, Alper HS. 2022. Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* **604**:662–667. DOI: <https://doi.org/10.1038/s41586-022-04599-z>, PMID: 35478237

**Mascotti ML**. 2022. Resurrecting enzymes by ancestral sequence reconstruction. *Methods in Molecular Biology* **2397**:111–136. DOI: [https://doi.org/10.1007/978-1-0716-1826-4\\_7](https://doi.org/10.1007/978-1-0716-1826-4_7), PMID: 34813062

**Shan S**, Luo S, Yang Z, Hong J, Su Y, Ding F, Fu L, Li C, Chen P, Ma J, Shi X, Zhang Q, Berger B, Zhang L, Peng J. 2022. Deep learning guided optimization of human antibody against SARS-CoV-2 variants with broad neutralization. *PNAS* **119**:e2122954119. DOI: <https://doi.org/10.1073/pnas.2122954119>, PMID: 35238654