Fitness Landscapes of Orthologous Green Fluorescent Proteins

by
Louisa González Somermeyer
September, 2024

A thesis submitted to the
Graduate School of the
Institute of Science and Technology Austria
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Committee in charge:

Fyodor Kondrashov
Karen Sarkisyan
Gašper Tkačik

The thesis of Louisa González Somermeyer, titled Fluorescent Landscapes of Orthologous Green Fluorescent Proteins, is approved by:

Supervisor: **Fyodor Kondrashov**, Okinawa Institute of Science and Technology, Okinawa, Japan

Signature: _____

Committee Member: **Gašper Tkačik**, ISTA, Klosterneuburg, Austria

Signature: _____

Committee Member: **Karen Sarkisyan**, MRC Laboratory of Medical Sciences, London, UK

Signature: _____

Defense Chair: **Carrie Bernecky**, ISTA, Klosterneuburg, Austria

Signature: _____

Signed page is on file

ISTA Thesis, ISSN: 2663-337X

I hereby declare that this thesis is my own work and that it does not contain other people's work without this being so stated; this thesis does not contain my previous work without this being stated, and the bibliography contains all the literature that I used in writing the dissertation.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee, and that this thesis has not been submitted for a higher degree to any other university or institution.

I certify that any republication of materials presented in this thesis has been approved by the relevant publishers and co-authors.

Signature: _____

Louisa González Somermeyer

September, 2024

Signed page is on file

Abstract

Understanding the relationship between a given phenotype and its underlying genotype or genotypes is one of the most pressing challenges of biology, as it lies at the heart of not only basic understanding of evolutionary theory, but also of practical applications in medicine and bioengineering. Understanding this relationship is complicated by the ubiquitous phenomenon of epistasis, wherein mutation effects are dependent on their genetic context. Fitness landscapes — representations of phenotype as a function of genotype — are being increasingly used as a tool to study the effects and interactions of thousands of mutations, but are experimentally limited to exploring a small fraction of a protein's theoretical sequence space. Furthermore, not all regions of said sequence space are necessarily equally informative. Thus, gene selection for landscape surveys should be carefully considered in order to maximize the usable output of necessarily limited data.

In this work, we analyzed the fitness landscapes of orthologous green fluorescent proteins from four different species, by systematically measuring the phenotype, fluorescence, of tens of thousands of mutant genotypes from each protein. These landscapes were highly heterogeneous, with some genes being mutationally robust and displaying epistasis only rarely, and others being highly epistatic and mutationally fragile. We used this data to train machine learning models to predict fluorescence from genotype. Although the training data contained almost exclusively genotypes with less than 3% sequence divergence from the original wild-type sequences, we were able to create novel, functional genotypes with up to 20% sequence divergence. Counterintuitively however, genes with high mutational robustness and rare epistasis were *more* difficult to introduce large numbers of mutations into, not less. This represents the first study of large-scale fitness landscapes of a protein family, and provides insights into how to approach future landscape surveys and their applications in novel protein design.

Acknowledgements

I must start with the following disclaimer: if your name does not appear on this page, it does not mean I don't like you; it means my brain is no longer what it was before writing >50,000 words about GFP (I cannot even manage a proper joke about not being as bright as cgreGFP).

With that said...

I am grateful firstly to Fedya K, for his unfaltering support from the moment that I joined his lab, and to Karen (master of Zen) for being an amazing mentor through the years. Thank you to all past and present Kondrashov lab members who made moving to Maria Gugging into something fun (Pilar! Nastia L! Ana G! Katya P! Peter!) and who helped keep the lab as something more than just a workplace (Aygul! Mia! Katya M! Rodrigo! Catalin! Arina! Anna T!), and thank you to all the great people from other labs who did likewise (Elizabeth! Bor! Nicoló! Julian! Amika!). Thank you to everyone who saved this project with their critical contributions (Aubin! Nina! Sasha! Alex M, best A2P!). And a huge thank you to Calin for standing up for me against Certain Authorities and to all the rest of Guet group (Bryan! Mike! Katya K! Nathalie!) for adopting and welcoming me after Fedya moved most of the lab to Japan. A big thank you as well to my family and to everyone outside IST who helped maintain my sanity with their friendship, board game nights, and/or 12-session "one-shot" D&D campaigns (Andi! Lena! Sven! Misha V! Anya P! Fedya G!).

An especially big thank you to Ondra for generally putting up with me all the time!

Last but certainly not least, I want to thank all the fluffsters who ever got hair on my PhD student clothes. In particular, my grumpy little Fifi (R.I.P.), my grumpy little Juniper, and my sweet little Koblížek. But also: Amy & Nelli, Nola (and Kili, Fili, Nori, Dori, Merry, Sammy, and Bilbo), Odin & Villi, Roy boy, Murray, Finn, Byron & Güira, and Mr Cheese. You are all good boys and girls.

P.S.: Obligatory statement that this project was partially funded by the following grants: the *European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie COFUND Grant Agreement No. 665385* (2018-2020), and the *European Research Council 771209-CharFL "Characterizing the fitness landscape on population and global scales"* grant (for as long as this ERC was allowed to be hosted at IST.)

About the Author

Louisa completed a BSc in Biology and an MSc in Genetics and Genomics from the University of Barcelona, Spain. During that time, she spent almost 2 years interning at different research institutes. Freshly graduated but disillusioned with academia, she then joined Kondrashov Lab in at the Center for Genomic Regulation in Barcelona as a no-strings-attached research technician. She ended up liking the group enough to relocate to Maria Gugging and start an actual PhD, becoming the first Andorran citizen to join IST. She likes cats, spoon-billed sandpipers, and making things glow in the dark, whether they be fluorescent bacteria or chemiluminescent plants carrying fungal transgenes.

List of Collaborators and Publications

Collaborators

The following colleagues' contributions were crucial to this work:

Aubin Fleiss & Ekaterina Putintseva
Creation of neural net models trained on GFP genotype-to-phenotype library data. Implementation of genetic algorithm to generate novel protein sequences expected to be functional.

Anna Toidze
Protein purification, urea- and thermosensitivity assays of 12 cgreGFP-derived variants, as part of her Bachelor's thesis. FACS re-runs of cgreGFP-derived gene libraries at OIST.

Alexander Mishin
Scripts for rescaling and merging data from duplicate amacGFP/cgreGFP/ppluGFP2 FACS experiments, and for assigning fitness values to genotypes based on FACS data. Discussions and advice regarding data analysis and statistics.

Nina Bozhanova
Crystallization and 3D structure determination of amacGFP. Calculation of ddG values for single mutants of avGFP, amacGFP, cgreGFP, and ppluGFP2.

Karen Sarkisyan
Supervision and mentorship.

Fyodor Kondrashov
Supervision and mentorship.

Publications

Part of this work's findings have been published in the following article:

Louisa Gonzalez Somermeyer, Aubin Fleiss, Alexander S Mishin, Nina G Bozhanova, Anna A Igolkina, Jens Meiler, Maria-Elisenda Alaball Pujol, Ekaterina V Putintseva, Karen S Sarkisyan, Fyodor A Kondrashov (2022). **Heterogeneity of the GFP fitness landscape and data-driven protein design.** *eLife* 11:e75842

Table of Contents

List of Figures

List of Tables

List of Abbreviations

bp. Base pair(s).
BSA. Bovine serum albumin.
CDF. Cumulative distribution function.
CEITEC. Central European Institute of Technology.
CGSC. Coli Genetic Stock Center.
CRISPR. Clustered regularly interspaced short palindromic repeats.
DFE. Distribution of fitness effects.
DNA. Deoxyribonucleic acid.
    dsDNA. Double-stranded DNA.
    gDNA. Genomic DNA.
DSC. Differential scanning calorimetry.
DSF. Differential scanning fluorimetry.
EDTA. Ethylenediaminetetraacetic acid.
FACS. Fluorescence-activated cell sorting.
GFP. Green fluorescent protein.
    amacGFP. GFP derived from the hydrozoan jellyfish *Aequorea macrodactyla*.
    avGFP. GFP derived from the hydrozoan jellyfish *Aequorea victoria*.
    cgreGFP. GFP derived from the hydrozoan jellyfish *Clytia gregaria*.
    ppluGFP2. GFP derived from the copepod *Pontellina plumata*, equivalent to copGFP.
IDT. Integrated DNA Technologies.
IPTG. Isopropyl β-D-1-thiogalactopyranoside.
ISTA. Institute of Science and Technology Austria.
KDE. Kernel density estimate.
LD50. Lethal dose (enough to kill 50% of test sample).
LB. Lysogeny broth (not Luria Bertani).
ML. Machine learning.
NaCl. Sodium chloride.
NEB. New England Biolabs.
NGS. Next generation sequencing.
nt. Nucleotide.
OD. Optical density.
OIST. Okinawa Institute of Science and Technology.
TAE. Tris-acetate-EDTA buffer.
PBS. Phosphate buffered saline.
PCR. Polymerase chain reaction.
    qPCR. Quantitative PCR.
PDB. Protein Data Bank.
SEC-MALS. Size exclusion chromatography with multi-angle light scattering.
T_m. Melting temperature.
VBCF. Vienna Biocenter Core Facilities.
WT. Wild type.

1. Introduction

1.1. Understanding genotype-to-phenotype maps

1.1.1. Motivation

An organism's phenotype is shaped by its genotype, modulated by environmental factors. This is true whether the phenotype in question is the color of a cat's fur, how many eggs a spider lays, or how brightly a fluorescent protein glows. Some phenotypes are governed by interactions of many genes together, while others are monogenic. But even in the latter case, the precise relationship between a set of genotypes and their corresponding phenotypes is often obscure (de Visser et al., 2011), making it difficult or impossible to reliably predict one from the other.

This difficulty is a source of chagrin to researchers from all walks of science, as understanding the map between genotype and phenotype would have actionable implications for a broad range of fields. In medicine, it would inform predictions of disease progression (Huang, 2013) as well as patients' response to therapy (Bolton et al., 2020), in addition to monitoring the emergence of drug resistance in pathogens (Lozovsky et al., 2009; Palmer et al., 2015; Flynn et al., 2023). In evolutionary theory, it would aid interpretation of population dynamics (Watson et al., 2020), quantitative genetics (Mackay, 2014), and evolutionary arms races between co-evolving species (Gupta et al., 2022). Biotechnological applications include but are not limited to protein design (Wrenbeck et al., 2017; Ogden et al., 2019) and crop engineering (DeHaan & Van Tassel, 2014). Additionally, it is fun (Sarkisyan et al., 2016). This work is therefore dedicated to better understanding and exploiting the genotype-to-phenotype relationship, one of the most pressing challenges in biological research.

1.1.2. Fitness landscapes

A fitness landscape is a representation of fitness, or any other phenotype, as a function of genotype. The concept was originally introduced as a way to to picture the evolution of populations or genes (Wright, 1932), where said populations or genes can move through genotype space and acquire a higher or lower fitness (or other phenotypic value) depending on where in the space they move to. The nature of this space and the parameters used to quantify fitness can vary depending on the organismal scale and on the questions one seeks to address. For example, population studies might use reproduction or growth rate as a proxy for fitness as a function of allele combinations across the genome (Gupta et al., 2022), whereas genes or proteins can be thought of as moving through a sequence space consisting of nucleotide or amino acid sequences of a given length (Melamed et al, 2013; Sarkisyan et al., 2016; Chan et al., 2017), with expression level or molecular activity being the phenotype of interest.

If one then pictures the space being moved through as the set of possible genotypes, with the distance between any two points being roughly proportional to the number of mutational steps between them, then assigning each point a fitness value creates a "landscape" of high-fitness peaks and low-fitness valleys. The smaller the fitness changes between adjacent points, the smoother the landscape appears (Figure 1).

Such a visual representation of a genotype-to-phenotype map is naturally a vast oversimplification of its true properties, namely its massive dimensionality. For instance, the sequence space of a gene of length N is N-dimensional, with one axis for each position in the sequence, and becomes N+1 dimensional after assigning fitness values. Given the discrepancy between the number of dimensions an average human can visualize and the number of positions in an average gene, achieving any real understanding of a complete fitness landscape's true shape is literally unthinkable. Reducing such complex data to a low-dimensional projection of hills and valleys has therefore been occasionally criticized as misleading (Fragata et al., 2018; Kaplan, 2008), mainly because it may mask the true distance between genotypes and create the false

impression of isolated fitness peaks separated by valleys of death, when peaks might in fact be connected by a ridge along some higher dimension.

Nevertheless, the inherent intuitiveness of fitness landscapes means that they remain a useful abstraction in the study of genotype-to-phenotype maps (Fragata et al., 2018, Hartman & Tullman-Ercek, 2019), and their use has been expanding in recent years in fields ranging from synthetic biology (Davidson et al., 2012; Wrenbeck et al., 2017; Ogden et al., 2019) to evolution (Canale et al, 2018, Lozovsky et al., 2012) and everything in between (Huang, 2013; Bolton et al., 2020; DeHaan & Van Tassel, 2014).

While "fitness" has traditionally been defined in evolutionary biology as reproductive success (Orr, 2009) and therefore a property only applicable on the organismal scale, studies of protein fitness landscapes typically use the protein's function or activity as a proxy for fitness in cases where the molecular phenotype itself, and not its effect on overall organism survival or reproduction, is of interest. We ourselves will also use the term "fitness" in this work in such a manner.

The experimental approach in studies of protein fitness landscapes is typically to generate a library of mutant variants of the protein in question, and measure the phenotype of interest for each of those mutant genotypes. Of course, the sheer size of genotype space mentioned earlier is not only a problem for humans trying to imagine higher dimensions, but also for laboratory setups, as assaying all or even most genotypes of any given gene would be physically impossible. Sewall Wright gives the example of a sequence space of 1000 loci, each with 10 possible alleles; this corresponds to 10^{1000} combinations, which Wright correctly identifies as "a very large number" (Wright, 1932) and, indeed, higher than the number of atoms in the universe.

Due to these experimental constraints, studies of protein fitness landscapes must either limit their focus to specific functionally or evolutionarily relevant sites or domains (O'Maille et al., 2008; Hietpas et al., 2011; Olson et al., 2014; Podgornaia & Laub, 2015; Poelwijk et al., 2019; Pokusaeva et al., 2019; Johnston et al., 2024), or explore a library of random mutants in nearby sequence space (Bershtein et al., 2006; Fowler et al., 2010; Jacquier et al., 2013; Melamed et al., 2013; Sarkisyan et al., 2016), or analyze only single mutant effects while forgoing interactions (Sanjuan et al., 2004; Chan et al., 2017). The second approach is broad but shallow while the first is deeper but narrow, but in either case, only a small fraction of the theoretically possible genotypes ever make themselves available for study.

This raises the question: how many measured genotypes are actually necessary in order to construct a reasonable, and useful, genotype-to-phenotype map?

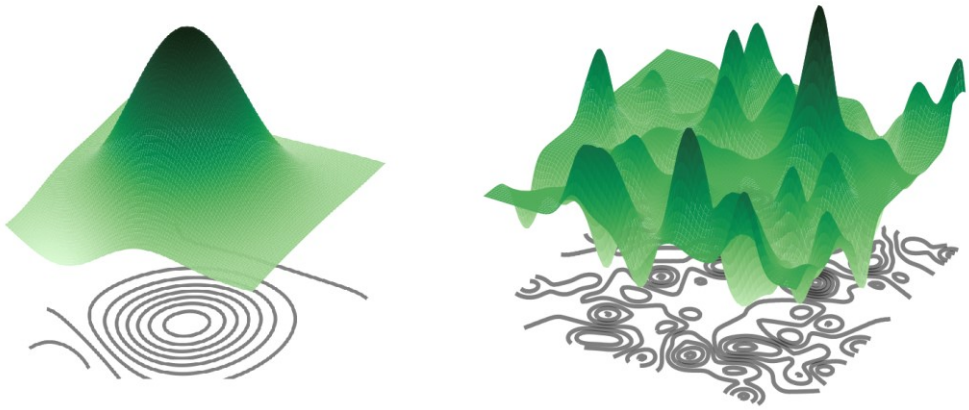

FIGURE 1. Conceptual illustration of smooth and rugged fitness landscapes. The X and Y axes represent a simplified projection of genotype space, where the distance between two points is indicative of the similarity of the corresponding genotypes. Genotype fitness is shown on the Z axis, in green. The more gradual and predictable are the fitness changes between similar genotypes, the smoother the landscape appears.

1.1.3. Epistasis

Even considering the harsh limitations in terms of the amount of experimentally feasible measurements, constructing an accurate genotype-to-phenotype map would be trivial if interactions between mutations never occurred. In such a world, the phenotype of a clone with multiple mutations AB could be readily extrapolated from the phenotypes of the individual mutants A and B. A protein of length 100 would thus require a mere $100 \cdot 19 = 1900$ measurements (19 being the number of alternative amino acids from the wild type), which is rather more manageable than the full theoretical protein sequence

space of 20^{100}.

However, experimental data has long shown that different mutations can and do interact with each other (Bateson et al., 1905; de Visser et al., 2011; Mackay, 2014; Hopf et al., 2017), such that a given mutation's phenotypic effect is dependent on the genetic context in which it occurs, i.e. the presence or absence of other particular mutations. This phenomenon is called epistasis. For qualitative traits, such as color markings or morphology, mutations may mask or modify each other's effects in a variety of surprising ways (Ishida et al., 2006; Schmidt-Küntzel et al., 2009). For quantitative traits, it can manifest as synergistic or antagonistic interactions (when mutations amplify or inhibit each other's effects, respectively) and be either positive or negative (when the final effect on fitness is more beneficial or more deleterious than expected, respectively); in some cases, it can even change the sign of a mutation's effect on phenotype (when the same mutation is beneficial in one context but deleterious in another) (Poelwijk et al., 2016) (Figure 2). One context — though not the only — in which sign epistasis arises is that of speciation, where mutations accumulate in diverging lineages to the point of being incompatible (Orr, 1995); pathogenic variants in one lineage may thus be neutral in another (Orr & Turelli, 2001; Kondrashov et al., 2002; Kulathinal et al., 2004). An interesting, if rare, subtype of sign epistasis is reciprocal sign epistasis, where two individually deleterious mutations create a neutral or even net positive effect (or vice versa) when they co-occur (Poelwijk et al., 2011). Epistasis may be intergenic, where the effects of one gene allele are modified by the existence of another allele at a different gene locus, or intragenic, where the interaction is between mutations at different sites within the same gene.

This non-independence of mutation contributions to phenotype makes intuitive sense, and was already illustrated in 1970 by John Maynard Smith (Smith, 1970) by a word game analogy wherein words are proteins, letters are amino acids, and the goal is to find a path from one word to another without generating any meaningless (low fitness) words in between. The shortest path from WORD to GENE is thus WORD—WORE—GORE—GONE—GENE, wherein epistasis can be readily observed: for example, the same change from O to E in the second position is permissible when it occurs in the GONE background, but not in the WORD background, where it would create an unfit WERD.

From a molecular point of view, epistasis can be explained and even expected in some cases when it occurs between physically close or interacting sites, such as between adjacent residues in a folded tertiary structure or subunits in a protein complex (Podgornaia & Laub, 2015; Kumar et al., 2017) or even between genes which do not interact directly on the molecular level but are part of the same metabolic pathway (de Visser et al., 2011). However, a structural or biophysical reason behind epistatic interactions is not always immediately apparent (Starr et al., 2016), making epistasis difficult to predict. In turn, this obfuscates the link between genotype and phenotype and increases the amount of data required in order to make sense of it: the more pervasive epistasis is, the more difficult it is to analyze and predict mutation effects. In fitness landscape terms, epistasis, particularly sign epistasis, renders the landscape more rugged (Figure 1) (Poelwijk et al., 2007; Poelwijk et al., 2011; Saona et al., 2022) by removing some of the direct paths between fitness peaks, even if it can also create indirect, otherwise inaccessible paths. Thus, an understanding of the rules governing epistatic interactions is key in understanding fitness landscapes and in leveraging genotype-phenotype data in downstream applications.

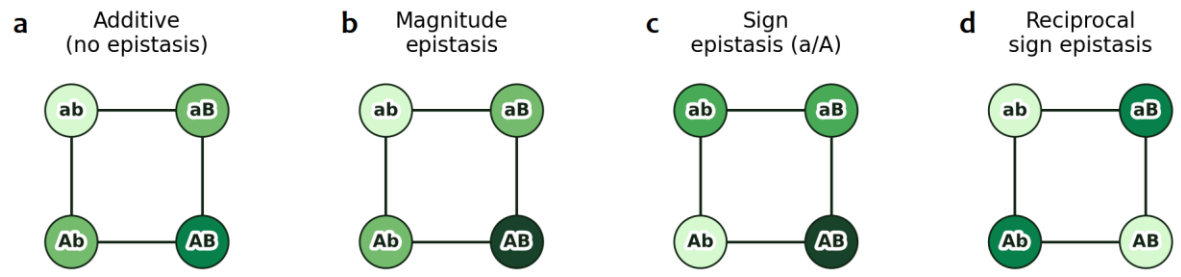

FIGURE 2. Pairwise epistasis affecting a quantitative trait. The trait's fitness, or phenotype, is represented by color, with darker greens being more fit. **(a)** Case of no epistasis: the mutations *A* and *B* do not modulate each other's effects. **(b)** Magnitude epistasis: *A* and *B* jointly result in a greater fitness change than expected (synergistic epistasis) from their individual effects on the *ab* background. Conversely, *a* and *b* jointly result in a smaller than expected change (antagonistic epistasis), when starting from the *AB* background. **(c)** Sign epistasis: *A* decreases fitness in the *b* background, but increases it in the *B* background. **(d)** Reciprocal sign epistasis: both *A* and *B* are either beneficial or deleterious depending on whether they co-occur or not.

1.2. GFP as a model for fitness landscape studies

As far as using a protein's function as a proxy for fitness, fluorescent proteins are attractive candidates for fitness landscape studies since light emission is a quantitative and readily measurable phenotype. The first such protein, green-glowing and isolated from the hydrozoan jellyfish *Aequorea victoria*, was discovered in the 1960s and eventually earned Osamu Shimomura and his colleagues the Nobel Prize in Chemistry. Fluorescent proteins produce light by absorbing higher energy (shorter wavelength) photons and then emitting lower energy (longer wavelength) photons upon relaxing to their ground state (Adan et al., 2017). Different fluorescent proteins may differ in their optimal excitation (absorption) and emission spectra, but all share similar tertiary structures consisting of eleven β-sheets arranged in a barrel around a light-emitting chromophore (**Figure 3a**) (Chudakov et al., 2010).

They are user-friendly molecules as a rule, being generally physically and structurally stable, non-toxic, relatively small (under 240 amino acids or 27 kDa), and not prone to interact with other cellular components (Chudakov et al., 2010). They also require no chaperones to fold correctly in either prokaryotic or eukaryotic cells, need no cofactors, and the only post-translational modification they require is self-catalyzed: the maturation of the chromophore, achieved by the cyclization of the serine, tyrosine and glycine residues in positions 65–67 in the presence of oxygen (Tsien, 1998; Chudakov et al., 2010). These features, which have so popularized GFP and other fluorescent proteins as reporters for gene expression and protein localization in model organisms from all domains of life, also make them ideal subjects for studying the more abstract question of genotype-to-phenotype maps.

Furthermore, fluorescent proteins are interesting in and of themselves due to how different members of this family can occupy vastly distant positions in sequence space and yet maintain such similar functionalities and 3D conformations (Chudakov et al., 2010). GFPs and GFP-like proteins have been documented in evolutionarily distant taxa ranging from jellyfish (Tsien,1998; Xia et al., 2002; Fourrage et al., 2014), corals (Alieva et al., 2008; Shagin et al., 2004), arthropods (Wilmann et al., 2006), and even cephalochordates (Baumann et al., 2008; Yue et al., 2016). Fluorescent proteins, including GFPs, are believed to share a single common origin (Shagin et al., 2004) rather than having arisen as a result of convergent evolution, yet sequence diversification over time has been such that many GFPs from different species are known to share under 20% amino acid identity (**Figure 3b,c**), while still exhibiting near-identical 3D β-barrel structures (**Figure 3a**) and light emission properties (**Figure 3d**). This makes GFPs a compelling group of proteins to use for fitness landscape studies, as it allows for a broad exploration of sequence space without requiring modifications to the type of phenotype data collected, thus allowing direct comparisons between distant sequences.

Indeed, the sequence space of GFP has already been explored at a local scale, focusing on around 50 thousand random genetic variants of *Aequorea victoria* GFP, or avGFP (Sarkisyan et al., 2016). This study found the fitness peak of avGFP to be narrow, meaning that protein fitness, measured as fluorescence output, tended to decrease dramatically in sequences only a few mutations away from the wild type. This sharp fitness loss was attributable to widespread negative synergistic epistasis among genotypes with multiple mutations, which meant that the overall effect of combinations of mutations was worse than the sum of their individual effects. Importantly, these mutation effects on fluorescence were found to be linked to their effects on protein structure and stability, according to their predicted effect on protein ΔΔG: individual mutations, each only slightly destabilizing on its own, together caused a sudden and sharp decline in fluorescence once their joint effects on protein structural integrity exceeded a critical threshold (Sarkisyan et al., 2016).

But, how far can these results on avGFP be extrapolated to the rest of GFP sequence space? Do other local landscapes, centered around other, distant GFP sequences, share similar properties? How do interactions between mutations change from one region of GFP space to another?

The fitness landscapes of various unrelated proteins — ranging from β-lactamase to Hsp90 to a WW domain and more — have been analyzed and published in the last two decades (Bershtein et al., 2006; Jacquier et al., 2013; Melamed et al., 2013; Olson et al., 2014; Hietpas et al., 2011; Fowler et al., 2010). However, only two surveys of multiple members of the same protein family are currently available to our knowledge. One study is that of orthologous His3 proteins in yeast (Pokusaeva et al., 2019), which focused exclusively on positions and amino acid states observed to be polymorphic across 21 natural yeast species. The other is that of three orthologous TIM barrel proteins (Chan et al., 2017), which assayed only single-mutant genotypes and as such did not study intramolecular epistasis. In this work, we will expand from the 2016 avGFP landscape by characterizing the fitness landscapes — including multi-mutant data — of GFP genes from three more species. This data will allow us to study the effects of mutations and their interactions within and across

genes, with the aim of understanding the broader sequence space of the GFP family as a model for fitness landscapes of protein families.

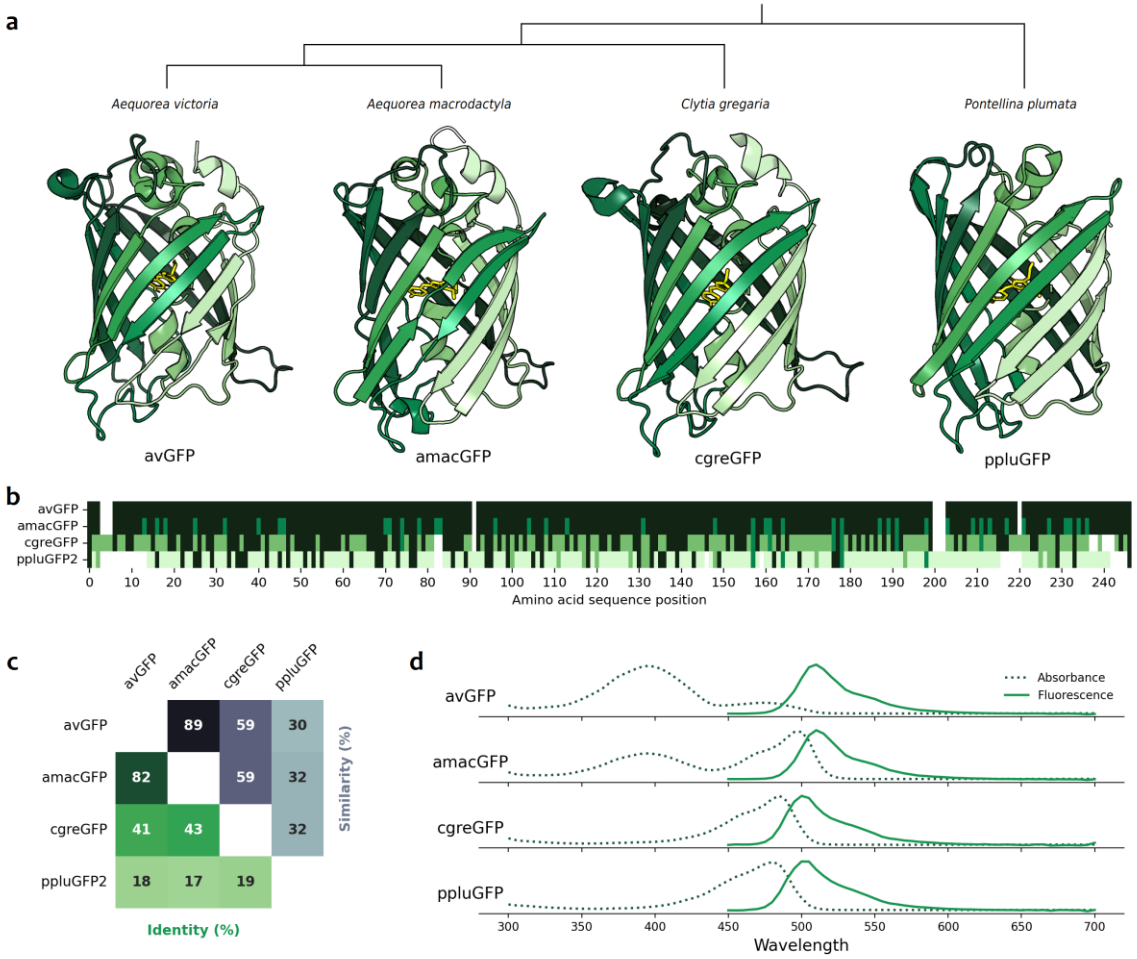

FIGURE 3. General properties and comparisons of four green fluorescent proteins. (a) Tertiary conformations of the four main GFPs used in this work, showing similar barrel structures (green gradient) surrounding the light-emitting chromophore (yellow). The phylogenetic relationship of the species from which these GFPs are derived is indicated by the cladogram above. **(b)** Structural alignment of the same GFPs, made with the T-Coffee Expresso alignment tool. Identical amino acids at any given position are labeled in the same color. Alignment gaps are left in white. **(c)** Amino acid identities (green) and similarities (grey) shared between each pair of proteins, structurally aligned as in (b). Aligned amino acids were counted towards the similarity score if they belonged to the same category, i.e.: aliphatic (M/L/A/I/V), aromatic (W/Y/F), negatively charged (D/E), positively charged (K/H/R), polar (S/T/N/Q), or special (P/G/C). For reference, randomly generated amino acid sequences of the same length average around 5% identity and 18% similarity (own simulations). **(d)** Absorbance and emission spectra of the four GFPs.

1.3. Experimental approach

We explored the sequence space of the green fluorescent protein family by focusing on multiple sequence-divergent GFPs from which to construct local fitness landscapes. An overview of the experimental design follows below; more detailed explanations are available in the methodology section (see: 4. Materials and Methods). Briefly: for each gene, we generated a mutant library comprising thousands of variants, each of which was barcoded with a unique molecular identifier. Libraries were expressed in *E. coli* and variants were separated by FACS according to their fluorescence intensity (fitness) (Figure 4). By sequencing the barcodes of sorted cells and analyzing their distribution across the designated green gates, we paired each variant's genotype to its corresponding fitness. These data were then used for analyses of mutation effects, epistasis, comparisons of local landscapes, etc., and ultimately used to train machine learning algorithms and generate novel artificial, functional GFP sequences.

1.3.1. Selection of genes

We selected documented GFP genes from different species based on several criteria. Firstly, we required candidate proteins to share similar optimal excitation and emission spectra (**Figure 3d**) in order to allow direct phenotype comparisons between genes without introducing unnecessary noise in the data due to shifting measurement parameters or settings. Related to this point, they should also be functional in *E. coli*, our chosen model organism, under standard culture conditions. Secondly, we wanted the selected genes to share varying degrees of sequence identity with each other, in order to be able to compare distant regions of sequence space. Thirdly, we favored proteins with known, resolved tertiary structures, in order to simplify any downstream structure-based analysis. Fourthly, many species are known to contain multiple, sometimes dozens of, GFP copies in their genome (Takahashi-Kariyazono et al., 2015; Baumann et al., 2008; Kashimoto et al., 2021), but we prioritized genes from species with fewer known copies on the basis that high gene redundancy may lead to weaker selection pressure (Nowak et al., 1997) and therefore to potentially less fit genes. (Note however: since the start of this work, emerging studies have shown that many species in fact contains more FP genes than previously believed: for example, multiple new FPs discovered in *Aequorea victoria* (Lambert et al., 2020) and *Clytia hemisphaerica* (Leclère et al., 2019).)

Based on the above considerations, we selected three GFP genes from which to construct local fitness landscapes (see: 4.2.1. Gene selection), referred to in this work as amacGFP, cgreGFP, and ppluGFP2. The former two are derived from hydrozoan jellyfishes, *Aequorea macrodactyla* and *Clytia gregaria* respectively, while ppluGFP2 is derived from the copepod *Pontellina plumata* and is sometimes referred to in other literature as copGFP. Together with avGFP, these four genes share between 17% and 82% amino acid identity with each other (**Figure 3c**).

Note: for simplicity, we will refer to these reference genotypes as the "WTs" of their respective local landscapes, even in cases where the reference amino acid sequences differ from the original, natural wild-type sequence by 1-3 point mutations (see: 4.2.1. Gene selection).

1.3.2. Creation of mutant libraries

E. coli codon-optimized sequences of the selected genes were used as the DNA template in mutagenic PCRs in order to create a library of mutant variants. Barcodes were introduced during this step thanks to a randomized 20N region in the reverse PCR primer (see: 4.2.1. Generation of mutant sequences). PCR reaction conditions were optimized to generate approximately four nucleotide mutations per variant, corresponding to an average of 1-2 amino acid substitutions per protein variant.

Plasmid-based mutant libraries were then obtained in a two-step process. First, PCR products were cloned into a promoter-less storage vector and transformed into *E. Coli*, recovering tens of thousands of colonies in each case. This step serves to estimate and control the size (number of variants) of the libraries, as the number of recovered colonies at this stage is roughly equivalent to the number of unique barcodes going forward, given that a) the probability of multiple variants sharing the same 20N nucleotide barcode is negligible (one chance in 4^{20}); b) double transformation is rare, so the vast majority of colonies contain only one variant; and c) transformed cells are not given enough recovery time to divide before being plated, so each transformed cell only gives rise to a single colony (see: 4.2.1. Cloning of mutants into storage vectors). (Of course, the total number of unique protein sequences in the final library is always lower than the number of barcodes, due to some variants containing no, or only synonymous, mutations.) At this stage, the plasmid library is also sequenced, in order to know which genotype is represented by which barcode (see: 4.6. Library sequencing: Coding regions).

Second, variants are shuttled from the storage vector into an expression vector (see: 4.2.1. Generation of final expression constructs). Final expression vectors contain a GFP variant cloned under a constitutive promoter and in-frame with mKate2, a red fluorescent protein, as in Sarkisyan et al., 2016. This fusion protein setup ensures a 1:1 mKate2:GFP ratio and thereby allows mKate2 signal to serve as a control for GFP expression level. mKate2 was originally chosen as a reference based on several key properties (Sarkisyan et al., 2016), primarily that a) it does not undergo a green-emitting stage during chromophore maturation, unlike many other red fluorescent proteins, and b) there is minimal overlap between the excitation and emission spectra of mKate2 and those of GFP (**Figure 5a**). Furthermore, in our setup, a rigid alpha-helical linker prevents direct physical interaction between mKate2 and GFP. These features help ensure that mKate2 fluorescence not affect or interfere with measured GFP signal.

While the mKate2-GFP fusion protein setup described above was also used previously for the avGFP fitness landscape (Sarkisyan et al., 2016), the libraries described in this work differ from the overall avGFP

setup in two main ways: genome integration of the constructs to limit copy number variation, and the use of secondary barcodes to allow for internal replicates in a single experiment.

Expression from a chromosomally integrated construct can be expected to be less noisy than that from a plasmid, due to the lack of copy number variation from cell to cell (Boyd et al., 2000). Whereas the avGFP library was expressed from a low/mid-copy number plasmid (Sarkisyan 2016), our current expression vectors contain 5' and 3' homology arms which flank the mKate2-GFP construct and map to a safe harbor on the *E. coli* chromosome, allowing genomic integration via homologous recombination as in Bassalo et al, 2016 (see: 4.3. Genome integration).

Furthermore, our expression vectors also contain a 10N barcode located after the GFP sequence (see: 4.2.1. Generation of destination vector). Each cloned molecule thus contains both a 20N identifier specific to a particular variant sequence, as well as an additional 10N identifier, henceforth referred to as "primary barcode" and "secondary barcode" respectively. This allows for the possibility of measuring biological replicates in a single experiment, as cells containing the same primary barcode but different secondary barcodes must necessarily have originated from independent cloning and genome integration events. We harvest 3-5 times more colonies at this stage than in the storage vector step, to maximize the number of primary barcodes associated to multiple secondaries in the final library.

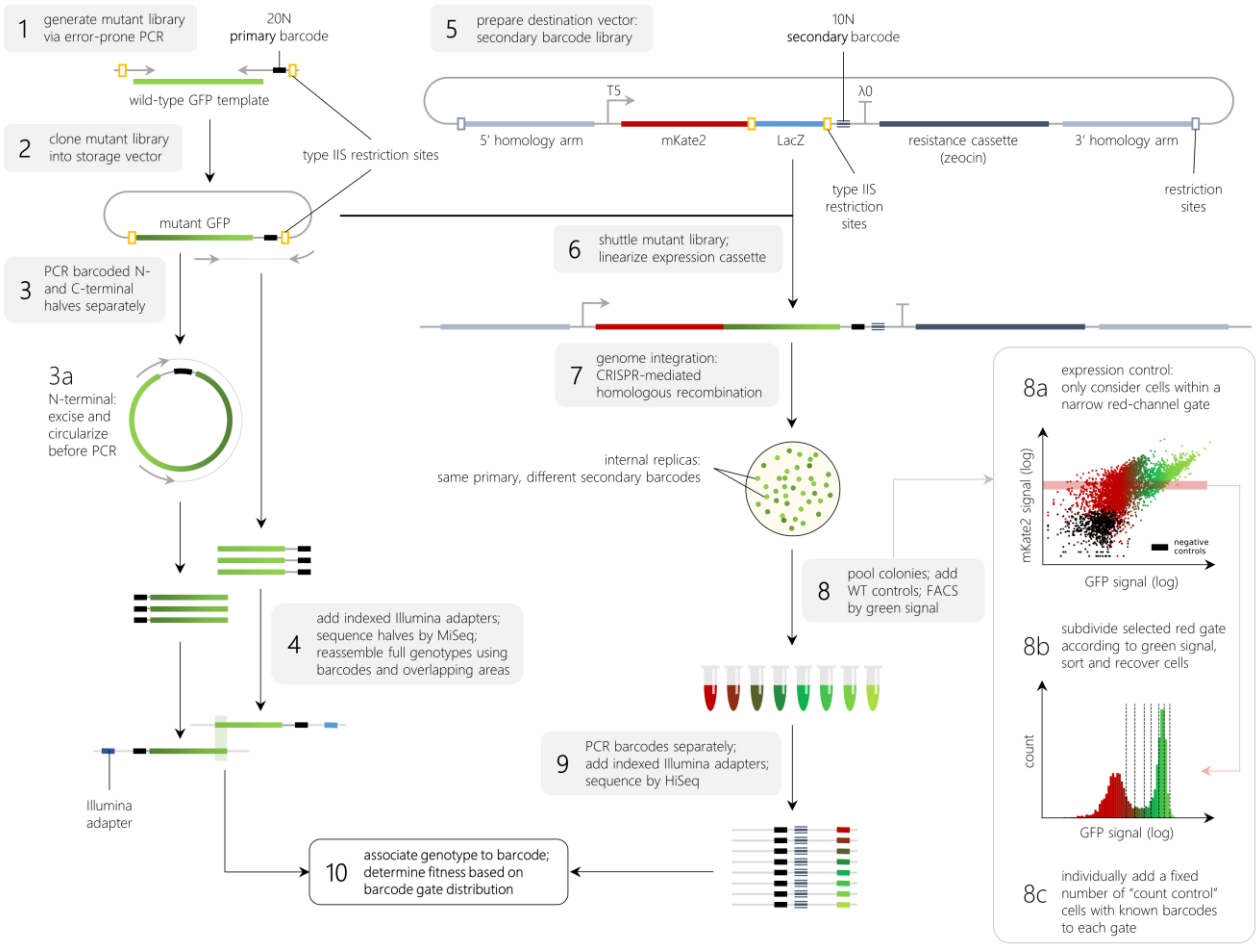

FIGURE 4. Experimental pipeline for the generation, sequencing, expression and sorting of mutant libraries. Figure adapted from Gonzalez Somermeyer et al., 2022, Figure 2.

1.3.3. Fitness measurements

Genome-integrated bacteria were processed by FACS (see: 4.4. Fluorescence-activated cell sorting). We defined a narrow gate in the red channel, limiting ourselves to cells with comparable mKate2 (and therefore GFP) expression levels, in order to minimize capturing variations in fluorescence caused by differences in gene expression (as opposed to caused by differences in genotype). The selected red gate was subdivided into eight gates according to the intensity of green fluorescence (**Figure 5b**), and cells falling into any of these eight green gates were physically sorted into separate tubes.

Note: the values reported by FACS machines are directly dependent on the laser voltage settings, as

well as on the machine model etc.; these values should not be interpreted as an absolute quantification of the fluorescence of sorted samples. Therefore, to facilitate comparisons across different gene libraries (processed by necessity on different days and/or different machines), all libraries had a small amount of wild-type control cells of other genes mixed in prior to sorting. These controls are sorted along with the rest of the library, and act as reference points to help compare different genes' relative fluorescences across experiments (see: 4.4.1. FACS sample preparation). In addition, we employ a number of "count controls", cells with known barcodes generated separately from the libraries. These cells are not mixed or sorted together with the libraries, but are added in fixed amounts to each tube after sorting, thereby serving as controls to convert the number of NGS reads to a cell count later on. I.e., if we know that we obtain X reads from Y control cells, we can calculate how many reads per cell to expect, and then determine for other barcodes how many cells were actually sorted during the experiment (see: 4.4.2. FACS setup).

The barcode regions of recovered sorted cells were amplified and sequenced. For each barcode, we used the distribution of cell counts across the eight green gates to assign a fluorescence (fitness) value (see: 4.7.1. Determination of fitness values). By combining the barcode-to-fitness and barcode-to-genotype data, we were able to reconstruct genotype-to-fitness maps containing data from tens of thousands of mutant variants for each of our selected starting genes. This data formed the basis for analyses of mutation effects and epistatic interactions within and across genes, and served as the training data for machine learning models used to generate novel, functional protein sequences with up to 20% sequence diverge from the original wild-types.

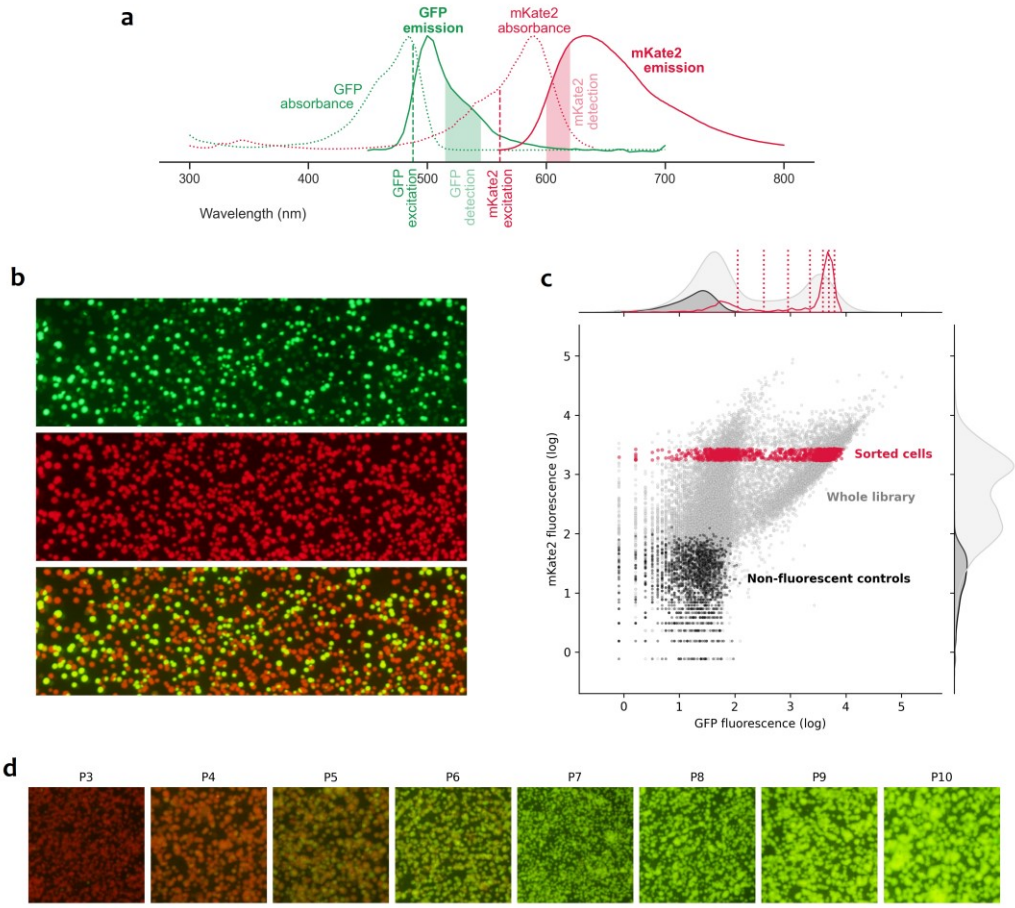

FIGURE 5. GFP library cell sorting. (a) Absorbance and emission spectra of GFP (green) and mKate2 (red). GFP spectra for amacGFP, ppluGFP2, and cgreGFP are comparable (Figure 3). FACS laser excitation wavelengths (488 nm for GFP, 561 nm for mKate2) are indicated by vertical dashed lines. The shaded areas represent fluorescence signal collection for GFP (515-545 nm) and mKate2 (600-620 nm) during FACS. **(b)** Representative example of *E. coli* colonies expressing an mKate2-GFP library. Note the variable green intensities (top) versus homogenous red intensities (middle). The bottom panel shows green and red channels merged. **(c)** Representative FACS setup for library sorting. Only cells within a narrow gate in the mKate2 channel (red) are sorted; these represent cells with comparable mKate2-GFP expression levels. This red gate is subdivided into 8 green gates which are then sorted; the darkest green gate is based on the distribution of GFP/mKate2-negative control cells. Fluorescence values are shown $\log_{10}$-transformed. **(d)** Representative example of colonies grown from sorted cells from the 8 green gates of a FACS run. Images all show merged red and green channels. Cells in (b) and (d) were grown on LB-zeocin agar overnight at 30°C then overnight again at room temperature, then photographed with a Canon EOS 600D SLR camera in (b) or Nikon SMZ25 stereo microscope in (d), under blue and yellow light; aside from cropping and merging red/green channels, photographs were not altered.

2. Results

We created and analyzed mutant libraries for amacGFP, cgreGFP, and ppluGFP2 independently (Figure 4). A large fraction (~34%) of the genotypes in the amacGFP library contained a V11L mutation (see: 4.2.1. Generation of mutant sequences), so for some downstream analyses we considered the V11L subset separately from the rest of amacGFP genotypes. A general summary of library statistics can be found in Table 1. Note: mutation indexing will generally refer to the structurally aligned position (equivalent across orthologous genes), with counting starting from Methionine = 0, unless otherwise specified.

Table 1. General dataset statistics. False positives refer to genotypes which contain mutations affecting chromophore or chromophore-interacting sites Y68, G69, R99, E229 (numbering refers to the structurally aligned positions, equivalent to Y66, G67, R96, E222 in traditional avGFP notation) but which were assigned a bright fluorescence value; as the chromophore is crucial for fluorescence, mutations in these sites can be expected to eliminate fluorescence, with very few exceptions. False negatives refer to nucleotide genotypes containing exclusively synonymous mutations which were assigned low fluorescence values (see: 4.7.4. Library data filtering). WT fitnesses and standard deviations are determined from the distribution of synonymous variants encoding WT proteins.

	amacGFP	cgreGFP	ppluGFP	avGFP (Sarkisyan et al., 2016)
Protein length (aa)	238	235	222	238
Chromophore	SYG	SYG	GYG	SYG
Number of assayed protein genotypes	35500 (of which, 12010 V11L)	26165	32260	51715
Average number of measured replicates per genotype	8.7	6.8	12	1.2
False positives (chromophore mutants assigned high fitness)	0.55% (9/1635)	0.75% (14/1860)	0.49% (11/2242)	0.24% (2/839)
False negatives (WT genotypes assigned low fitness)	0% (0/1084)	0% (0/1583)	0% (0/2744)	0.08% (2/2444)
WT fitness ($\log_{10}$) ± standard deviation	3.97 ± 0.031 (V11L: 3.96 ± 0.03)	4.5 ± 0.028	4.23 ± 0.027	3.72 ± 0.082

2.1. General features of local landscapes

2.1.1. Distributions of mutation effects on fluorescence are bimodal

For all new genes (amacGFP, amacGFP:V11L, cgreGFP, and ppluGFP2), as well as for the previously published avGFP, the distribution of fluorescence values of the mutant libraries was bimodal (Figure 6a-b). The majority of assayed genotypes were either relatively bright or non-fluorescent entirely, although the proportion of bright versus dark genotypes varied across genes. This pattern was echoed in the bimodal distribution of the effects of single mutations, with the majority of observed amino acid substitutions being either lethal or only slightly deleterious to fluorescence. This is in line with existing literature on the distribution of fitness effects (Wloch et al., 2001; Sanjuán et al., 2004; Eyre-Walker & Keightley, 2007; Wylie & Shakhnovich, 2011) and with other experimental studies of protein fitness landscapes (Hietpas et al., 2011; Jacquier et al., 2013; Chan et al., 2017), which consistently describe the effects of mutations on fitness as following a bimodal distribution.

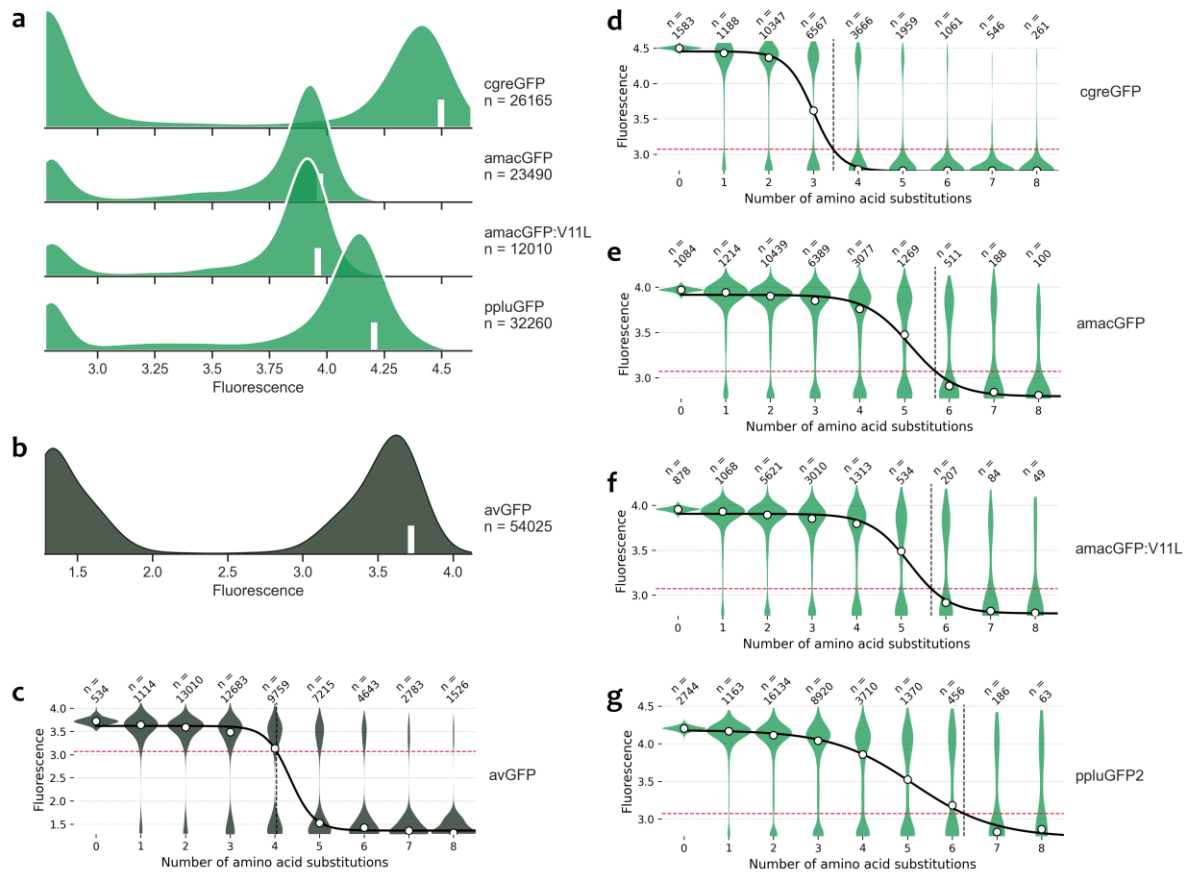

FIGURE 6. Fitness distributions of mutant libraries. Libraries created during this work are in green; publicly available data from avGFP (Sarkisyan et al., 2016) is in dark grey. **(a)** KDE plots showing the overall distributions of fluorescence of cgreGFP, amacGFP, amacGFP:V11L, and ppluGFP2 libraries, in green. WT values for each library are indicated by a white bar near the X axis. The total number of protein genotypes assayed is indicated on the right side of the figure. **(b)** As (a), but for avGFP. **(c)** Distribution of fluorescences of avGFP genotypes categorized by the number of amino acid mutations (X axis). The number of assayed genotypes is indicated at the top: for the WT, this number refers to the amount of distinct nucleotide genotypes containing exclusively synonymous mutations (see: Table 1 for WT distribution variance); for all other categories, it refers to the number of distinct protein sequences containing the specified amount of amino acid substitutions. The red horizontal dashed line represents the fluorescence cutoff below which genotypes were considered non-functional in the original publication. The median fluorescences of each category are shown in white and were used to fit a logistic curve (black line). The vertical dashed black line marks the average number of mutations necessary for the median fluorescence value to fall below the non-functionality cutoff. **(d,e,f,g)** As (c), but for cgreGFP, amacGFP, amacGFP:V11L, and ppluGFP2. Here, the non-fluorescence cutoff corresponds to the values of the upper border of the darkest FACS gate used during sorting, itself defined by the distribution of GFP-negative control cells.

2.1.2. Mutational thresholds mark fluorescence loss

For all genes, median fluorescence decreased as a function of the number of mutations (Figure 6c-g). This was expected, given that the more mutations are introduced, the greater the chances of one or more of those mutations being deleterious. This decrease was not linear: the bimodality of fluorescence values described above was largely maintained across data subsets irrespective of the number of mutations, with the proportion of bright versus dark populations shifting in favor of the latter as the number of mutations increased. This resulted in a threshold effect: a sharp decrease in median fluorescence coinciding with the dark population becoming the majority after a critical number of mutations was reached. It has been suggested that such loss of function patterns are attributable to the cumulative effect of mutations on protein stability, which, once past a certain threshold, render the protein thermodynamically unable to fold and therefore function (Bloom et al., 2005; Bershtein et al., 2006; Sarkisyan et al., 2016; Starr & Thornton, 2016). We will address the role of protein stability on GFP fluorescence in a later section of this work (see: 2.3. Structure, stability, and mutational robustness).

2.1.3. Mutational robustness varies across genes

While all GFP genes exhibited the threshold effect described above, the value of the threshold itself in

terms of number of mutations was highly variable. The amacGFP and ppluGFP2 libraries maintained high proportions of functional, bright genotypes containing up to 5-6 mutations, while cgreGFP already showed dramatic loss of fluorescence after only ~3 mutations (Figure 6c-g). This was consistent with the overall distribution of fluorescence values from the cgreGFP library showing a higher proportion of dark genotypes than that of amacGFP or ppluGFP (Figure 6a). The loss of fluorescence could be closely approximated by a logistic (sigmoid) curve fitted to the median fluorescence values for each number of mutations (Figure 6c-g), highlighting the threshold effect behind fluorescence loss.

The ability of a DNA sequence to accumulate mutations without manifesting significant changes in phenotype has been termed "mutational robustness" (de Visser et al., 2003; Bershtein et al., 2006), and we will use this term throughout this work to describe the degree to which different GFP genes can tolerate amino acid substitutions without losing fluorescence. In terms of fitness landscape imagery (Wright, 1932), mutationally robust genes such as amacGFP and ppluGFP2 can be thought of as having flatter fitness peaks, while mutationally fragile genes such as cgreGFP and avGFP have sharper or steeper fitness peaks.

We separately looked at genotypes which maintained WT-level fluorescence and genotypes which were non-functional, with increasing numbers of mutations. As a simple measure of mutational robustness, we determined the numbers of mutations necessary to either eliminate fluorescence, or cause it to fall below WT levels in 50% of genotypes, by fitting a curve to this data and solving for $f(x) = 0.5$ (Figure 7). We affectionately termed these values "mutational lethal dose 50", or $\text{MutLD50}_{(\text{Dark})}$ and $\text{MutLD50}_{(\text{WT})}$ respectively. MutLD50 values (WT/Dark) for the various GFP orthologues were as follows: 0.9/3.2 (cgreGFP), 2.2/4.1 (avGFP), 1.8/5.7 (amacGFP), 1.9/5.7 (amacGFP:V11L), 1.7/6 (ppluGFP2). (Note: these values differ marginally — between 0-0.2 — from those published in Gonzalez Somermeyer et al., 2022 due to the rescaling of amacGFP and ppluGFP2 datasets to the cgreGFP range of values which was done as part of this dissertation. This rescaling was done to improve direct comparability of fluorescence values across datasets after adding five new local landscapes in 2.6. Novel cgreGFP-Derived Genes).

Interestingly, mutational robustness did not appear to cluster according to the genes' sequence identity: for instance, amacGFP and avGFP, the two closest genes at 82% shared amino acid identity, exhibited opposite tendencies in this regard.

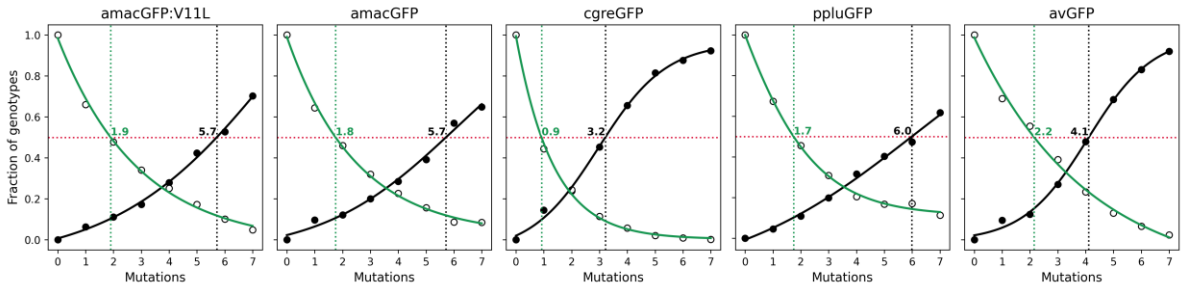

FIGURE 7. Number of mutations required to critically affect fluorescence in 50% of genotypes. For each gene and for increasing numbers of mutations, the fraction of assayed genotypes maintaining WT-level fitness (i.e. fluorescence values within 2 standard deviations of the WT) is shown in green, and the fraction of non-functional genotypes (i.e. fluorescence values within the darkest FACS gate) in black. Each set of points is fitted with a logistic curve, $f(x) = L/(1 - e^{-k(x-x0)})$. The number of mutations x at which the fraction of genotypes with WT-level fitness drops below 50% ($\text{MutLD50}_{(\text{WT})}$) was determined by calculating the inverse of the fitted function and solving for $f(x) = 0.5$, and is marked by a vertical dotted green line. The same procedure was done of the fraction of non-functional genotypes ($\text{MutLD50}_{(\text{Dark})}$), in black.

2.2. Intramolecular epistasis

We calculated epistasis as the difference between the observed fluorescence of a given genotype and its expected fluorescence in the absence of epistasis. In the absence of epistasis, the joint effect of all mutations is by definition equal to the sum of their individual effects (see: 4.7.6. Calculation of mutation effects and epistasis), with deviations from this rule being caused solely by measurement error.

To minimize false discovery of epistasis, we set a minimum threshold of |0.3|, as in Sarkisyan et al., 2016, and did not accept values under this threshold as necessarily indicative of real epistasis. For some analyses focused specifically on cases of strong epistasis, we set the threshold even higher, as indicated in the relevant sections. A difference of |0.3| is equivalent to a two-fold difference between measured and

expected fluorescence values. This value falls well outside our typical range of measurement errors (see: 4.7.7. Estimation of noise) and therefore may even lead to an underestimation of the actual amount of weaker epistatic interactions.

2.2.1. Epistasis is predominantly negative and a feature of sharper peaks

The majority of observed epistasis was negative, meaning that the measured fluorescence of multi-mutant genotypes was less than that expected under an additive model of non-interacting mutation effects (Figure 8a) Furthermore, epistasis was much more common overall in mutationally fragile libraries (cgreGFP, avGFP) than in robust ones (amacGFP, ppluGFP2) (Figure 8a,b). The latter finding is consistent with the observed threshold effects mentioned previously: if mutations are acting, even additively, upon an intermediate phenotype such as protein stability which results in sudden fluorescence loss past a certain threshold, then crossing this threshold will be detected as negative epistasis when measured in terms of fluorescence effects (Starr & Thornton, 2016). The lower the threshold, the more frequently epistasis will be detected.

In accordance with the above, if there exists a link between mutational robustness and the pervasiveness of epistasis due to the relationship of both with protein stability (or any other underlying phenotype), it may follow that negative epistasis in robust genes will be more likely to manifest when a higher number of mutations are in play. Epistatic interactions involving three or more mutations have been termed higher order epistasis (Weinreich et al., 2013). In our data, the distribution of epistasis as a function of the number of mutations also differed between libraries. In the case of the more mutationally robust amacGFP and ppluGFP2, genotypes with more (5-7) mutations were more likely to display negative epistasis than genotypes with fewer (3-4) mutations, while the opposite was true for the mutationally fragile cgreGFP. Indeed, with the exception of avGFP, negative epistasis tended to peak in genotypes with a similar number of mutations as the gene in question's $MutLD50_{(Dark)}$ (Figure 8c).

On the other hand, positive epistasis, while rare, tended also to be of higher order (Figure 8c). However, this may partially be a consequence of the experimental setup: the range of data measurements does not extend very far past WT bright values (Figure 6a), making the beneficial mutation effects more difficult to detect than deleterious effects on the WT. The detection of positive effects and positive epistasis therefore depends on the *expected* phenotype to be low. Because the expected fluorescence under the additive model tends to be inversely proportional to the number of mutations, positive epistasis may be more easily detected in genotypes with a higher mutation count, which may not necessarily reflect its true distribution.

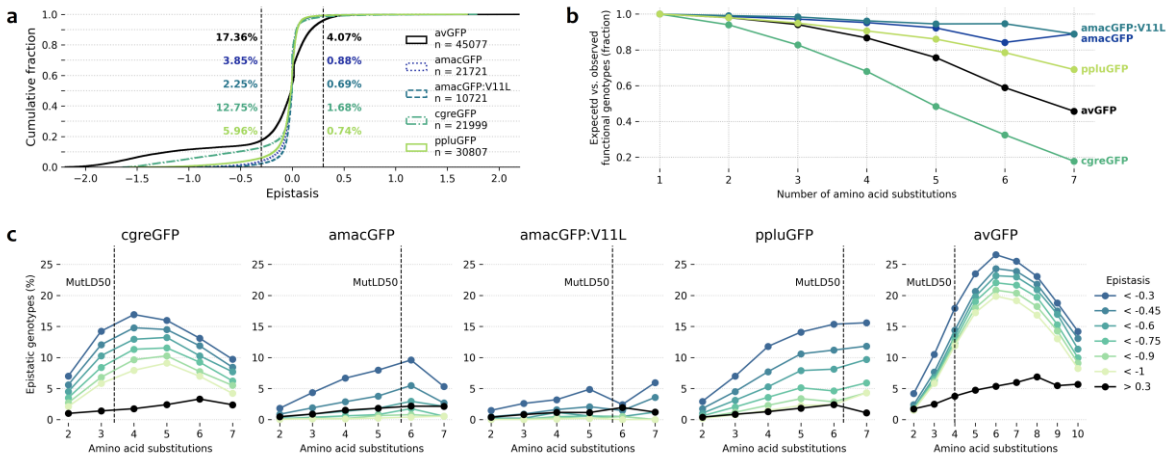

FIGURE 8. Overview of epistasis. All avGFP data shown in this figure is taken from Sarkisyan et al., 2016. **(a)** Cumulative distribution of observed epistasis in all datasets. Dashed vertical lines places at -0.3 and 0.3 indicate the minimal values for epistasis considered in this work; percentages indicate, for each library, the proportion of genotypes displaying epistasis below or above these limits, out of all genotypes where epistasis was calculable. **(b)** For *n* mutations (X axis), the fraction of genotypes observed to be functional, out of all those expected to be so under an additive model of mutation contributions. The cutoff for functionality was chosen to be the upper border value of the darkest FACS gate, except for avGFP, where the cutoff was set to 3 as in the original publication. In the absence of epistasis, Observed/Expected values are expected to equal 1. **(c)** Prevalence of higher order negative epistasis of varying magnitudes (color), and positive epistasis over 0.3 (black). Dashed vertical lines mark the number of mutations needed to eliminate fluorescence in 50% of genotypes ($MutLD50_{(Dark)}$).

2.2.2. Physically proximal residues are more likely to exhibit pairwise epistasis

In folded proteins, residues which are distant from each other in the primary structure (amino acid sequence) may be physically near in the tertiary one (3D conformation). Amino acids which are spatially proximal have greater opportunities to interact with one another, so pairwise epistatic interactions may be expected to be more prevalent between adjacent residues in a protein's folded structure (Melamed et al., 2013; Sarkisyan et al., 2016).

Using PyMOL and the proteins' solved 3D structures, we calculated the physical distance between all pairs of amino acids, defined as the minimal distance, in Ångströms, between any two atoms belonging to different residues (see: 4.7.8. Determination of physical distances between residues). For all proteins, the average distance between epistatic pairs was less than for non-epistatic pairs (**Figure 9**). This difference was particularly pronounced in the flatter landscapes where epistasis was rarer overall (amacGFP, ppluGFP2), consistent with the hypothesis of mutations affecting fluorescence through an intermediate phenotype exhibiting a threshold effect (e.g. protein stability): in mutationally fragile genes where just two mutations are often sufficient to cross this threshold, pairwise epistasis can be expected to be less dependent on specific (proximal) interactions between residue pairs.

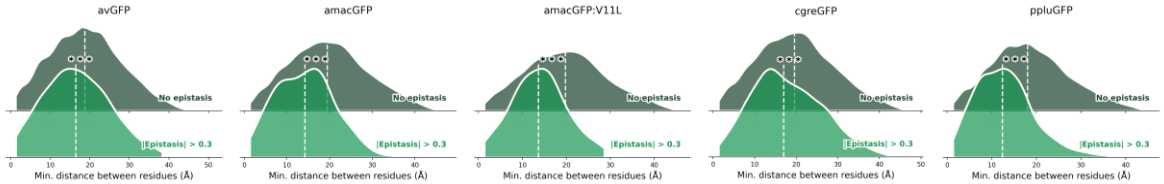

FIGURE 9. Distances between epistatic and non-epistatic pairs of residues. Distances in Ångström represent the minimal distances between amino acids (i.e. between any atom from one residue and any other atom from the other residue). Pairs of residues were considered epistatic if they displayed epistasis values under -0.3 or over 0.3 (representing a two-fold change in fluorescence compared to the non-epistatic expectation). KDE plots show the distributions of distances (Å) between amino acids of epistatic and non-epistatic pairs. Median values are indicated by vertical dashed lines; in all cases, the difference was highly significant (Mann-Whitney U test, all p-values < 10^{-10}).

2.2.3. Sign epistasis is rare but detectable

Sign epistasis refers to cases where a given mutation is observed to have a deleterious effect in one genetic background, but a neutral or even beneficial effect in another (de Visser et al., 2011; Starr et al., 2016; Poelwijk et al., 2016). As the majority of mutations were measured in multiple genetic contexts within the same library (see: 4.7.6. Calculation of mutation effects and epistasis), we were able to see how consistent mutation effects were in different contexts (variants of the same gene), and how frequently its sign changed. While many mutations displayed occasional severely deleterious effects even if they were near-neutral in the majority of backgrounds, mutations with effects ranging from significantly deleterious (-0.3, or a two-fold decrease in fitness) to significantly beneficial (+0.3, or a two-fold increase in fitness) were rare. AvGFP had the highest incidence of single mutations displaying effects below -0.3 and above 0.3, at 134/1431 (9.4%) of mutations assayed in multiple backgrounds, followed by cgreGFP (47/1412, 3.3%), ppluGFP2 (24/1312, 1.8%), amacGFP (10/1383, 0.7%), and amacGFP:V11L (5/1247, 0.4%). This was roughly proportional to the overall rates of epistasis in each dataset (**Figure 8a**).

A rare subset of sign epistasis is reciprocal sign epistasis (Kondrashov & Kondrashov, 2015), wherein individually deleterious mutations rescue each other's effects to result in a net positive joint effect (or, conversely, wherein individually beneficial mutations are mutually incompatible). While rare, we did detect instances of reciprocal sign epistasis in our data. For example: the cgreGFP dataset contained two triple-mutant genotypes, R96S:Q183R:R234C and R96S:T113M:Q183R, both of which were assigned nearly-WT fitness values. The mutations R96S and Q183R, present in both genotypes, were individually lethal to fluorescence; the R96 residue is known to play a role in chromophore maturation (Wood et al., 2005). The other two mutations, R234C and T113M, were individually near neutral. Therefore, the high fluorescence levels of the two triple-mutant genotypes — if accurate — were likely due to reciprocal sign epistasis between R96S and Q183R. Existing literature has reported that the "debilitating mutation R96A or R96M can be rescued by Q183R" (Banerjee et al., 2017; Wood et al., 2005), although we are not aware of existing references to R96S. To confirm that this was a genuine case of reciprocal epistasis, we individually expressed the R96S, Q183R, and double mutant genotypes in bacteria (**Figure 10**), which revealed that

while both single mutants were non-fluorescent, fluorescence in the double mutant was successfully rescued.

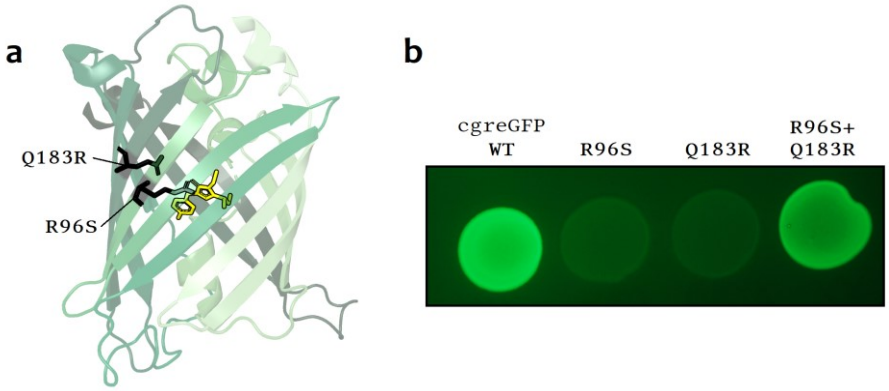

FIGURE 10. Example of reciprocal sign epistasis in cgreGFP. (a) 3D structure of cgreGFP, with the chromophore colored in yellow and positions R96 and Q183 labeled in black. **(b)** Spots of *E. coli* expressing WT cgreGFP, cgreGFP:R96S, cgreGFP:Q183R, and cgreGFP:R96S:Q183R. Mutant sequences were generated by amplifying WT cgreGFP with primers containing point mutations, then ligating the fragments into an expression vector. These cells confirm the non-functionality of both single R96S and Q183R mutants, as well as the fluorescence of the double mutant, proving the existence of reciprocal sign epistasis between these two mutations. Cells were grown on LB/ampicillin-agar, overnight at 30°C.

2.3. Structure, stability, and mutational robustness

Mutations affecting key locations of a protein's structure, such as an enzyme's catalytic site, are likely to be deleterious for obvious reasons. In the case of fluorescent proteins, mutations affecting the chromophore itself, or residues involved in chromophore maturation, are expected to be mostly lethal even if the overall structure of the final, folded protein is largely undisrupted (Banerjee et al., 2017; Wood et al., 2005). However, such mutations are only a small subset of all possible deleterious mutations. In general, the effect of a mutation on protein function is often believed to be a result of its underlying effect on protein stability and/or folding ability (Bloom et al., 2005; Bershtein et al., 2006; Zeldovich et al., 2007; Sarkisyan et al., 2016; Starr & Thornton, 2016). If a mutation causes misfolding, it follows that the resulting protein's functionality will be affected; similarly, even if a mutated protein does successfully fold, its function may be impaired if its structural integrity is more sensitive to perturbations in its molecular environment.

We analyzed the effects of mutations as a function of their position in the protein sequence to see the importance of protein structure on fluorescence, and compared mutations' observed effects with their computationally predicted effects on protein folding. We also tested WT, folded proteins for their thermostability and sensitivity to chemical denaturing agents, on the basis that less physically stable GFPs may also be less tolerant to mutations due to their structure being more easily disrupted.

2.3.1. Buried residues are more sensitive to mutations

The tertiary structure of fluorescent proteins consists of a barrel made of eleven β-sheets which surround the fluorescent chromophore inside. The side chains of the amino acid residues along the β-sheets are alternatingly inward- and outward-facing. Residues whose side chains are internally oriented ("buried" residues) are thus more likely to interact with and/or affect the immediate surroundings of the chromophore than residues with externally oriented side chains ("exposed" residues) (Figure 11). Mutations in buried sites can thus be expected, on average, to have greater fluorescence-dampening effects due to altering the chromophore environment; this was previously observed to be the case for avGFP (Sarkisyan et al., 2016).

Our data confirmed this was the case for amacGFP, amacGFP:V11L, cgreGFP, and ppluGFP2 as well. Comparisons of fitnesses of single-mutant genotypes showed drastic differences in overall distribution as well as median fluorescence depending on whether the mutation affected a buried residue or an exposed one (Figure 12a,b). This tendency was also clear in genotypes with multiple mutations, where maintenance of fluorescence was in general dependent on how many of those mutations affected buried sites (Figure 12b).

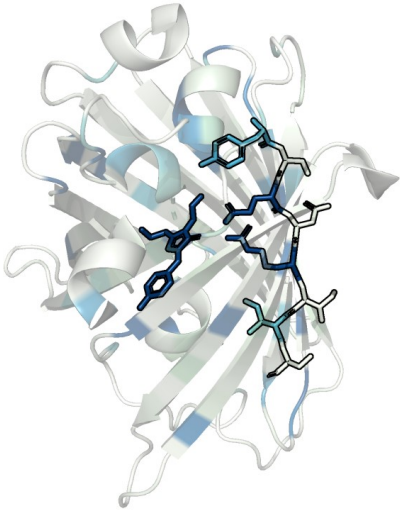

FIGURE 11. Close-up of buried and exposed residues. Tertiary structure of a representative GFP (cgreGFP), with residues colored according to the median effect of mutations in that position, as in Figure 12a, where darker colors signify more deleterious effects on fluorescence. The central chromophore, as well as residues along one of the barrel's β-sheets, are highlighted with their side chain structures visible.

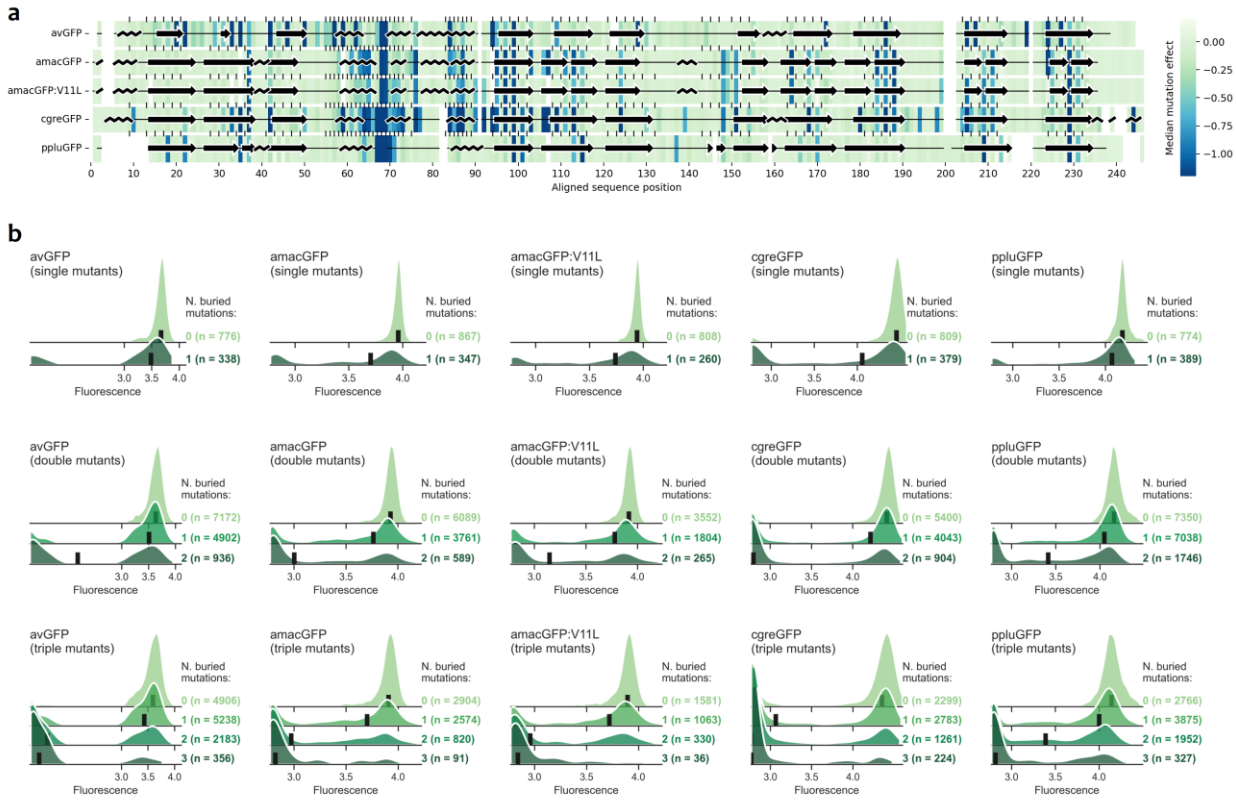

FIGURE 12. Effects of mutations in solvent-exposed and buried positions. (a) Median effects of mutations according to their position along the protein sequence. Only genotypes with a single mutation were used in this calculation. Sequence positions of orthologous GFPs are shown structurally aligned; white cells indicate alignment gaps or lack of data. Secondary structures were extracted from the proteins' PDB files: β-sheets are represented by arrows and α-helices by squiggly lines. **(b)** Distributions of fluorescence of genotypes with a total of one, two, or three mutations, split according to the proportion of those mutations affecting a buried site. KDE plots are scaled so that the area under each curve is equal to 1. The number of data points (genotypes) in each category is stated in color on the right.

2.3.2. Mutations predicted to be more destabilizing are more deleterious

A commonly used metric for predicting a mutation's effect on protein stability in terms of protein folding is ΔΔG (Kellogg et al., 2011; Zhang et al., 2012; Bigman & Levy, 2018), which measures the difference in the Gibbs free energy change (ΔG) of the protein with and without the mutation: $\Delta\Delta G = \Delta G_{(mutant)} - \Delta G_{(WT)}$. The ΔG itself refers to the energy change between a protein's unfolded and folded states: $\Delta G = G_{(unfolded)} - G_{(folded)}$, with subzero values being thermodynamically conducive to protein folding. The higher the ΔΔG value is, the

more destabilizing the mutation in question is expected to be, as the protein is predicted to fold more poorly with the mutation than without it. Conversely, negative ΔΔG values indicate the mutation is predicted to be stabilizing.

In avGFP, a moderate negative correlation (Spearman's ρ = -0.45) was found between mutations' predicted ΔΔG values and their effects on fluorescence (Sarkisyan et al., 2016), consistent with higher ΔΔG values being indicative of destabilizing mutations.

In order to calculate ΔΔG predictions, a solved crystal structure of the protein of interest is required. Those of cgreGFP and ppluGFP2 have been publicly available since 2011 and 2006 respectively (Malikova et al., 2011; Wilmann et al., 2006), while the crystal structure of amacGFP was determined as part of this work (see: 4.8.4. Crystallization and structure of amacGFP). Calculation of ΔΔG values for all libraries, as well as amacGFP protein crystallization, was performed by **Nina Bozhanova** (see: 4.9.2. Calculation of ΔΔG predictions).

Direct comparison of point mutations' predicted ΔΔGs versus observed fitness effects showed only weak negative correlations in the case of amacGFP and ppluGFP2, while that observed in the cgreGFP library was somewhat stronger (**Figure 13a**). It is worth noting that mutations in the more epistatic and less mutationally robust libraries (cgreGFP and avGFP) showed stronger, if still moderate, correlations between their predicted ΔΔG values and measured fitness effects, compared to the less epistatic and more mutationally robust libraries (amacGFP and ppluGFP2).

Nevertheless, a categorical comparison of neutral mutations (observed to maintain fluorescence levels within one standard deviation of the wild-type) and lethal ones (observed to eliminate fluorescence) showed a clear and significant tendency for lethal mutations, as a group, to have higher (more destabilizing) predicted ΔΔG values than neutral mutations (**Figure 13b**). This suggests that there is indeed a link between a mutation's effect on measurable protein fitness and its underlying effect on protein folding, even if any singular mutation's ΔΔG value is not always enough to confidently predict its actual effect.

Moreover, the difference between the ΔΔG distributions of neutral and lethal mutations was more pronounced in the mutationally fragile libraries than in the robust ones: in avGFP and cgreGFP, only ~1% of neutral mutations were associated to ΔΔG values higher than the median for lethal mutations, while for amacGFP and ppluGFP the value was closer to ~10% (**Figure 13b**). This trend suggests a possible link between a protein's mutational robustness and its physical stability — i.e., proteins who are better able to tolerate multiple co-occurring mutations may also better withstand individual mutations with greater destabilizing effects. This would be consistent with a robust protein's destabilization threshold being more difficult to reach, thus requiring either more mutations, or more highly deleterious ones.

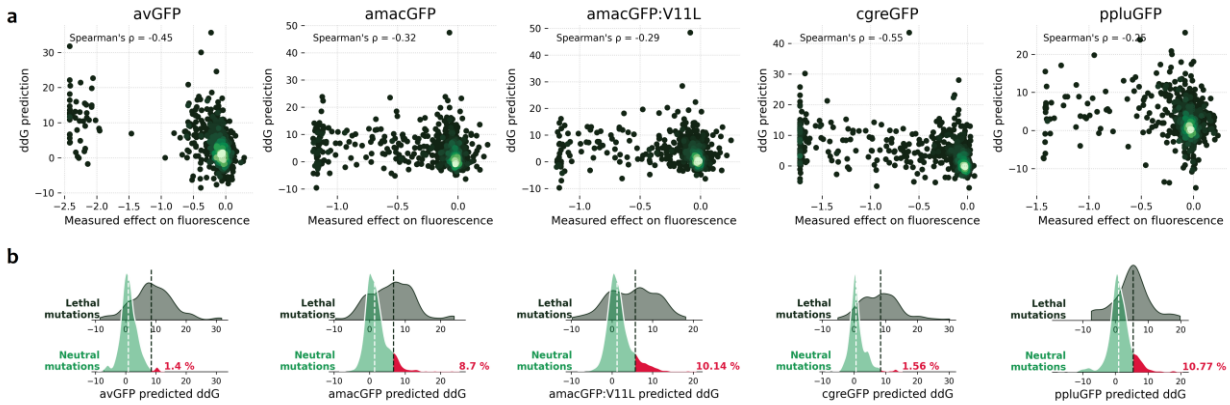

FIGURE 13. ΔΔG predictions versus observed effects of single mutations. (a) Scatterplot showing the relationship between a mutation's measured fitness effect and its predicted ΔΔG value. Each dot is a different mutation; lighter colors represent higher density of data points. Spearman's correlation coefficient is indicated in the top left corner of each plot; the p-value in all cases was under 10^{-15}. **(b)** Distribution of ΔΔG values of neutral mutations (causing a loss of fluorescence no greater than one standard deviation from the WT) and lethal mutations (causing fluorescence to drop to the values of the darkest FACS gate). Mann-Whitney U tests show a significant difference between the two distributions in all cases, with p-values under 10^{-5}. Dashed vertical lines represent the median ΔΔG value of each group. The fraction of neutral mutations associated with a ΔΔG higher than the median of the lethal group is highlighted in red.

2.3.3. Dark variants show greater propensity for aggregation

Misfolded proteins are highly prone to aggregation in general (Hartl & Hayer-Hartl, 2009). If deleterious

mutations affect fluorescence through their effect on protein stability and/or folding, it may follow that non-functional — likely misfolded — GFP variants should be expected to be more prone to forming protein aggregates than their functional, bright counterparts. We bulk tested bacteria grown from sorted cells from the darkest (P3) and brightest (P10) FACS gates (see: 4.4.2. FACS setup). Cells were lysed and centrifuged, and the insoluble (pellet) and soluble (supernatant) fractions were run on a protein gel (see: 4.9.10. SDS-PAGE and Western Blot). (Note: this was performed only for amacGFP, cgreGFP, and ppluGFP2 libraries, as avGFP cells from Sarkisyan et al., 2016 were not available.) The presence of mKate2-GFP fusion protein was detected by Western Blot using an anti-His-Tag antibody (our construct contains a 6H tag at mKate's N-terminus; see: 4.2.4. Generation of destination vector). Protein extract from all bright cells contained mKate2-GFP primarily in the supernatant, indicating the variants were soluble, as expected for correctly-folded GFP. On the other hand, protein extract from dark cells, i.e. containing GFP variants with highly deleterious mutations, showed much greater mKate2-GFP localization in the insoluble pellet fraction than was the case for bright variants. This tendency of forming insoluble aggregates is consistent with protein misfolding.

Interestingly, not all GFP orthologs were equally aggregation-prone. Dark amacGFP variants showed the highest tendency to aggregate, followed by dark ppuGFP2 and cgreGFP variants. We performed SEC-MALS analysis — size exclusion chromatography with multi-angle light scattering — to determine the natural oligomeric states of our WT orthologs (see: 4.9.9. SEC-MALS). At physiologically relevant concentrations, avGFP and cgreGFP are reportedly monomeric and dimeric, respectively (Malikova et al., 2011), and this was corroborated by SEC-MALS peak analysis (**Figure 15**). PpluGFP2 is reported to run as a dimer on gel-filtration chromatography (Wilmann et al., 2006), and TurboGFP, derived from ppluGFP2, is also described as dimeric in concentrations up to 5 mg/ml (though tetrametic in crystal form) (Evdokimov et al., 2006). However, SEC-MALS on ppluGFP2 (1 mg/ml) indicated it to be primarily tetrameric, while also forming high-molecular weight oligomers/aggregates. Finally, amacGFP and amacGFP:V11L appeared to be an approximately even mix of monomers and dimers. Given these results, there did not appear to be a correlation between a WT protein's usual oligomeric state and its propensity to aggregate when mutated, nor with its mutational robustness.

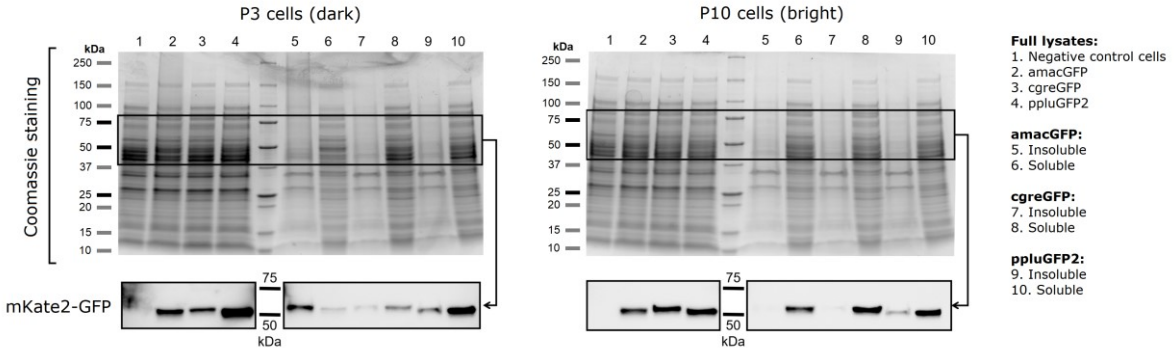

Figure 14. Protein gels showing aggregation of dark vs. bright GFP mutants. Figure adapted from Gonzalez Somermeyer et al., 2022, Figure 4-S2. AmacGFP, cgreGFP, and ppluGFP2 cells from the darkest (P3) and brightest (P10) sorted FACS gates were lysed. Full lysates (lanes 1-4), as well as the pellet (lanes 5, 7, 8) and supernatant (lanes 6, 8, 10) fractions derived from centrifuging the lysates, were stained with Coomassie (top) and Western Blotted (bottom) using an anti-His Tag antibody expected to label mKate2-GFP fusion proteins (~53-56 kDa, depending on the GFP variant). Bright protein variants (right) are detected primarily in the soluble fraction (supernatant) of the lysate, while dark protein variants (left) are more likely to be detected in the insoluble fraction (pellet).

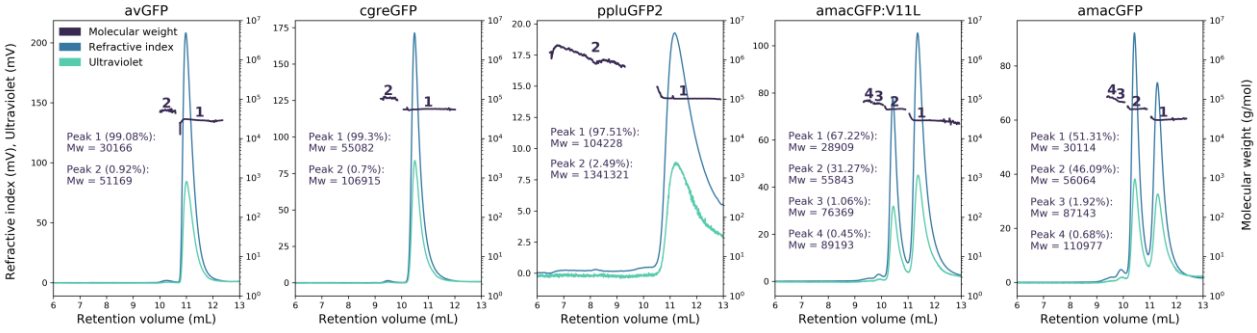

Figure 15. SEC-MALS analysis of WT GFPs. Figure adapted from Gonzalez Somermeyer et al., 2022, Figure 4-S2. Depending on the gene, the molecular weight of a GFP monomer is between 25-27 kDa. Peak analysis of different GFP orthologs, run at 1 mg/ml,

indicates primarily monomeric (avGFP), dimeric (cgreGFP), mixed monomeric and dimeric (amacGFP, amacGFP:V11L), or tetrameric (ppluGFP2) states. PpluGFP2 also displays a population of high-molecular weight aggregates. Mw/Mn ratios for all peaks were ideal, at between 1 and 1.002, except for the larger ppluGFP2 aggregates (Mw/Mn = 1.147) .

2.3.4. Thermostability correlates with mutational robustness with one exception

Assessing a protein's sensitivity to temperature is commonly part of studying its physical stability. We purified WT amacGFP, amacGFP:V11L, cgreGFP, ppluGFP2 and avGFP proteins (see: 4.8. His-tagged protein purification) and assayed their thermostability through a series of complementary tests in which protein samples were heated at a constant rate (see: 4.9.8. Thermosensitivity assays). Melting temperatures were determined by differential scanning fluorimetry, which measures the temperatures of protein unfolding and aggregation by monitoring changes in fluorescence emission of aromatic residues (**Figure 16a,c**); differential scanning calorimetry, which measures denaturation temperature by monitoring the difference in heat absorption between the sample and a reference (**Figure 16d**); circular dichroism, which measures loss of secondary structures by monitoring absorbance at a specified wavelength (**Figure 16e,f**); and by heating purified proteins in a qPCR machine, which directly measures loss of green fluorescence emission itself (**Figure 16b**).

Results from the various types of test were largely consistent and comparable. Specific melting temperatures for the same protein varied across methods by 1-2°C, which is not unexpected given that melting temperatures are sensitive to a wide variety of factors including temperature ramp rate, pH, and salt or buffer composition (Crowther et al., 2009), and different methods required slightly different sample treatments by necessity. The overall pattern, however, was the same regardless of the assay used: cgreGFP had the lowest melting temperature, followed, in order, by amacGFP:V11L, amacGFP, ppluGFP2, and avGFP. Furthermore, logistic (sigmoid) functions fitted to the CD data indicated that cgreGFP also had the steepest transition slope (followed by amacGFP:V11L, amacGFP, avGFP, and ppluGFP2 in order), indicating that it not only denatured at lower temperatures, but also did so more quickly.

With the exception of the mutationally fragile yet highly thermostable avGFP, the other four proteins indicated a correlation between tolerance to mutations and tolerance to heat. In order of most to least sensitive to both, they were: cgreGFP, amacGFP:V11L, amacGFP, and ppluGFP2. This suggests that lower physical stability of the WT state may be behind reduced mutational robustness: if a protein's starting point is already less stable, it may take fewer mutations, on average, to push it past its viability threshold and cause it to misfold.

AvGFP was an outlier in this regard, displaying the highest melting temperature of all tested GFPs despite being one of the least mutationally robust; indeed, a comparison of its CD spectra before and after heating suggest it retains some secondary structure even after heating to 98°C, whereas the spectra of all other proteins flattens out to zero after heating (**Figure 16e**). Possibly however, the shape of the avGFP landscape — a sharp peak featuring pervasive epistasis — was in part influenced by its gene expression levels during fluorescence measurements, which were higher than those of amacGFP, cgreGFP, or ppluGFP2 libraries (see: 3.1. On protein stability and selecting candidate genes for landscape surveys)

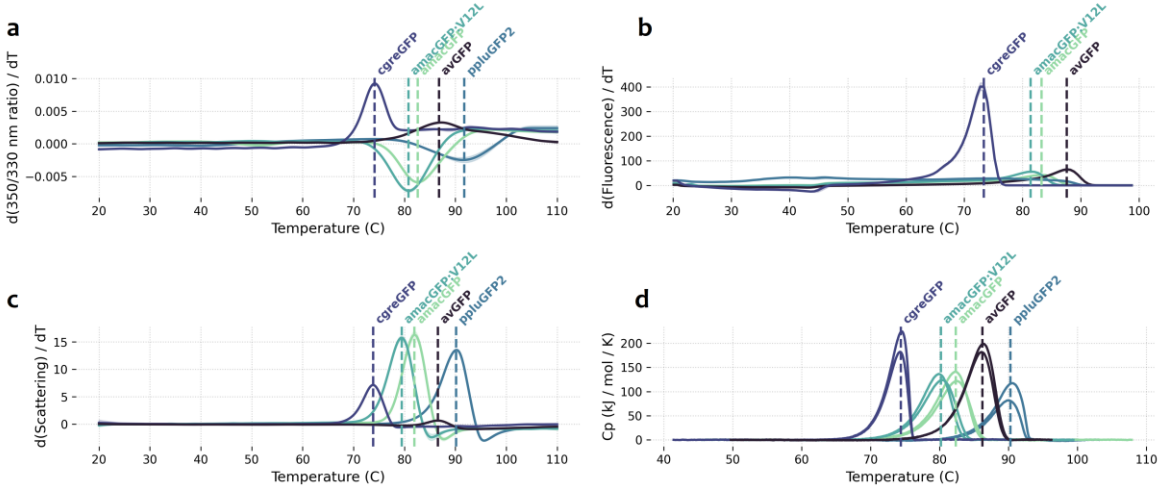

Figure continued on next page.

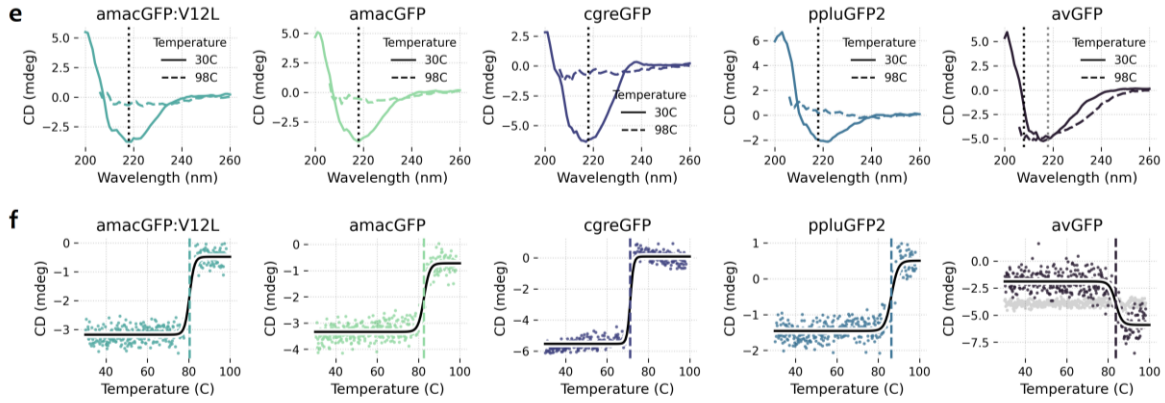

FIGURE 16. Thermostability of GFP orthologues. Figure adapted from Gonzalez Somermeyer et al., 2022, Figure 4. Vertical dashed lines in (a), (b), (c), (d), and (f) indicate melting temperatures of different GFPs. In (a), (b), (d), and (f), temperature was increased at a rate of 1°C per minute, in (c), at a rate of ~2°C per minute. **(a)** DSF-measured thermal unfolding. The first derivative 350/330 nm emission ratio is shown. Shaded areas indicate standard deviation of three replicates. **(b)** Melting curves of green fluorescence emission (510 nm) as a function of temperature measured on a qPCR machine. Shaded areas indicate standard deviations of eight technical replicates. **(c)** DSF-measured thermal aggregation. The first derivative of the light scattering is shown. Shaded areas indicate standard deviation of three replicates. **(d)** Specific heat capacities measured by DSC in duplicate. **(e)** Spectra measured by circular dichroism before and after heating. Vertical dotted lines mark the wavelength monitored during the melting curves represented in (f). **(f)** Circular dichroism melting curves monitored at 208 nm (avGFP) or 218 nm (all other genes), fitted with a logistic curve. For avGFP, monitoring at 218 nm (light grey) did not reveal a transition.

2.3.5. Protein sensitivity to urea does not correlate well with mutational robustness

Another widespread way to assess protein stability is by the use of chemical denaturing agents, such as urea or guanidinium chloride (Reddy et al., 2012). We subjected all WT GFPs to 9 M urea and scanned full absorbance and fluorescence spectra at regular intervals for a total period of ~60 hours (see: 4.9.7. Urea sensitivity assays). Protein denaturation is naturally expected to result in fluorescence loss as well as to changes in the overall absorbance spectra (Sarkisyan et al., 2012).

Different GFPs displayed very different sensitivity to urea, as was immediately noticeable in their spectral shifts over time (Figure 17a,b), with ppluGFP2 and avGFP being the least affected, and amacGFP:V11L the most, closely followed by cgreGFP. We tracked the signal loss at the wavelengths corresponding to the absorbance and fluorescence peaks of each protein in order to determine their half-lives, i.e. time points at which the signal fell below half of its initial value. Fluorescence half-life times in 9 M urea ranged from ~1 h (amacGFP:V11L) to ~2.5 h (amacGFP, cgreGFP), 16 h (ppluGFP2), and 57 h (avGFP) (Figure 17d). Absorbance peaks for avGFP and ppluGFP2 never fell below half of their initial values during the 60 hours of measurements, while half-lives for amacGFP, amacGFP:V11L, and cgreGFP were ~19 h, ~26 h, and ~14 h, respectively (Figure 17c).

While the mutationally robust ppluGFP2 was comparatively stable in 9 M urea and the mutationally fragile cgreGFP was degraded rapidly, the opposite pattern was true for the mutationally fragile yet urea-resistant avGFP and the mutationally robust yet urea-sensitive amacGFP:V11L. The sample size of only five data points, however, is likely insufficient to draw any conclusions about the correlation, or lack thereof, between the two attributes.

Furthermore, fluorescence loss curves for amacGFP, amacGFP:V11L, and ppluGFP2 could be fitted closely by the sum of two exponential decay functions, while this was not true of cgreGFP and avGFP. Indeed, cgreGFP in particular displayed a visibly different pattern of fluorescence loss over time than, for instance, amacGFP, despite the two proteins sharing nearly the same half-life (Figure 17d). This suggests that denaturation of different fluorescent proteins in urea is likely influenced by a variety of complex underlying factors.

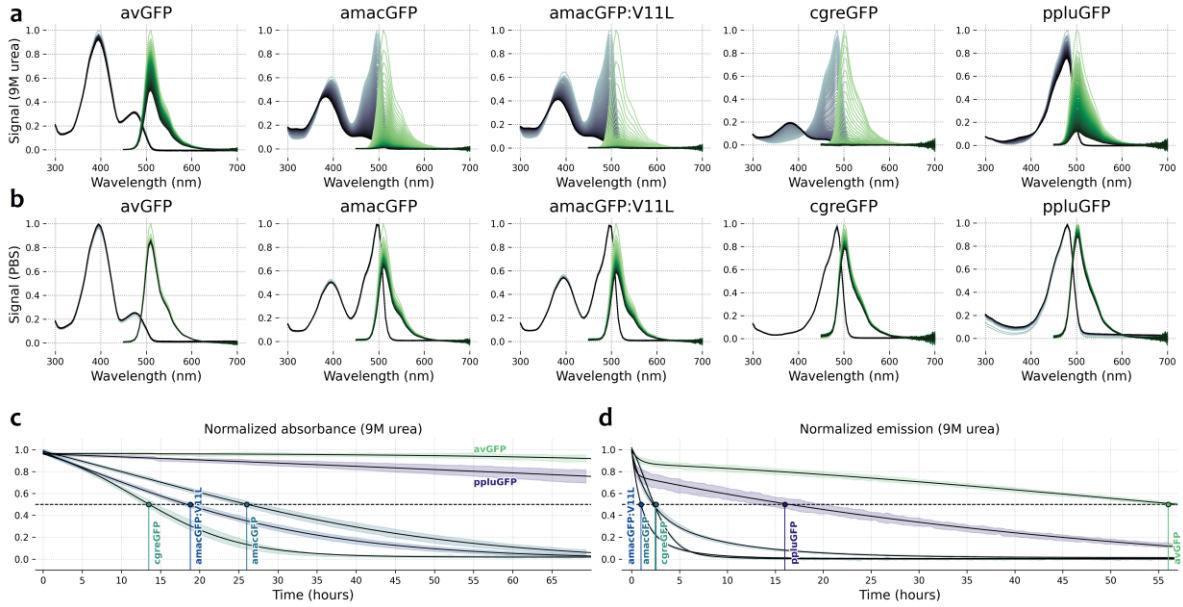

FIGURE 17. GFP sensitivity to urea. (a) Absorbance (blue) and emission spectra (green) of purified WT proteins in 9 M urea. Spectra were scanned for ~60 hours at regular intervals; the darker the line, the later the time point. Values are normalized such that the spectrum peak at time point zero equals 1. **(b)** As in (a), but in PBS instead of urea. **(c)** Loss of absorbance signal over time in 9 M urea, monitored at the wavelength corresponding to the absorbance peak for each protein. With a 5 nm resolution, absorbance peaks were: 495 nm (amacGFP, amacGFP:V11L), 485 nm (cgreGFP), 480 nm (ppluGFP2), or 395 nm (avGFP). Curves could be successfully fit with a logistic function (black). Half-lives are labeled where possible in colored vertical lines. **(d)** As in (c), but showing fluorescence emission instead of absorbance. With a 5 nm resolution, fluorescence peaks were: 510 nm (amacGFP, amacGFP:V11L, avGFP) or 500 nm (cgreGFP, ppluGFP2). Curves for amacGFP, amacGFP:V11L, and ppluGFP2 could be fit with a simple sum of two exponential decay functions, but cgreGFP and avGFP could not.

2.3.6. Case study: V11L alters amacGFP's sensitivity to mutations in a structure-dependent way

As mentioned previously, around one third of all amacGFP genotypes contained the V11L mutation (see: 4.2.2. Generation of mutant sequences, Table 1), which is enough data for amacGFP:V11L to be analyzed separately from amacGFP. As a standalone mutation in the WT amacGFP background, V11L may be considered neutral since its measured effect (approximately -0.01) falls within one standard deviation of the WT (approximately -0.03, as calculated from the distribution of WT-coding genotypes with synonymous mutations, Table 1). AmacGFP:V11L's mutational robustness is the same as amacGFP's (Figure 6e,f), and epistasis was similarly uncommon in both datasets (Figure 8).

The effects of nearly 1000 single amino acid substitutions were measured in both amacGFP and amacGFP:V11L backgrounds. We looked at the difference between each mutation's effect in one gene versus the other, as a function of its position along the protein sequence. The difference was calculated as $Effect_{(V11L)} - Effect_{(WT)}$, so negative values indicate a worse effect in the V11L background than in the WT, while positive values indicate the opposite. For the majority of positions, mutations had comparable effects in both genes (Figure 18c). However, ~10% of positions showed a trend towards harboring mutations more deleterious in one background than the other (Figure 18a,b). Notably, the sites at which mutations were most deleterious in amacGFP:V11L versus amacGFP were all clustered at one of the barrel lids (Figure 18d), suggesting that the V11L mutation, while not affecting amacGFP's mutational robustness overall, has an effect on amacGFP structure (and/or folding ability) such that tolerance to additional mutations is positionally dependent.

Furthermore, as shown previously, assays of thermal and urea sensitivity on WT proteins showed that amacGFP:V11L had a lower melting temperature (Figure 16) as well as lower stability in urea (Figure 17) than amacGFP. This shows that V11L affects the overall stability of the final folded protein (possibly by influencing the molecular environment of the barrel lid (Figure 18d)), even if its observable effects on mutational robustness and fluorescence in standard culture conditions are minimal.

Taken together, the case of amacGFP:V11L both supports the claim of mutations affecting a protein's underlying structure and stability (even if they may not immediately affect the phenotype of interest), while also highlighting the complicated nature of the relationship between proteins' physical properties and their mutational robustness.

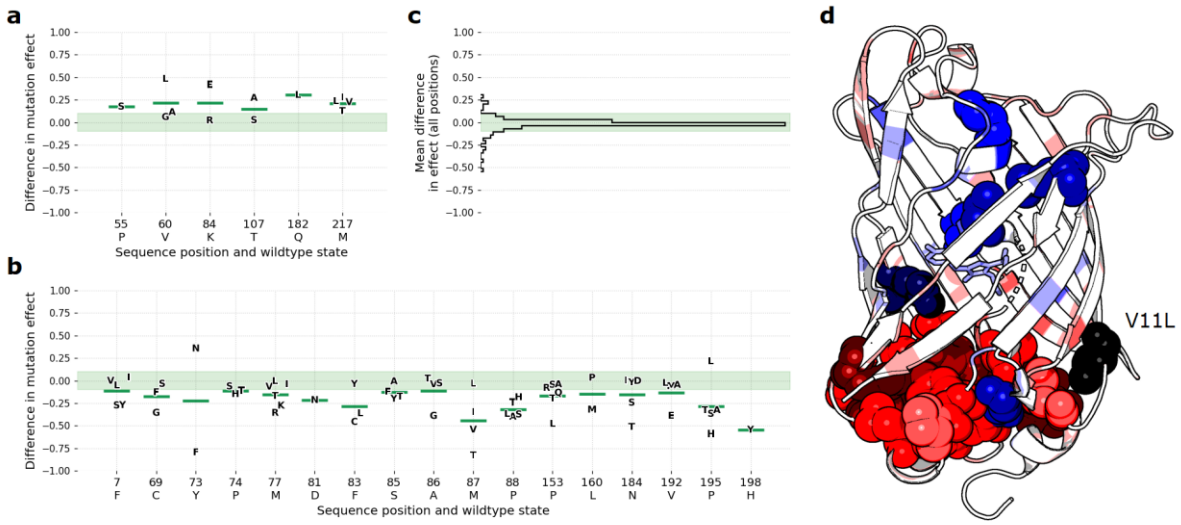

FIGURE 18. Positional differences of mutation effects in amacGFP and amacGFP:V11L. (a) Difference in effect, measured as $Effect_{(amacGFP:V11L)} - Effect_{(amacGFP)}$, of mutations measured in both backgrounds, at positions where mutations were on average less deleterious in amacGFP:V11L than in amacGFP. The average difference in effect at each position is marked by a green line. The light green shaded area shows the region into which the majority of average positional differences fell. **(b)** As (a), but showing positions where mutations were on average *more* deleterious in amacGFP:V11L than in amacGFP. **(c)** Distribution of the average difference in mutation effects of all 238 sequence positions. The shaded green area, from -0.1 to 0.1, includes 90% of positions. Only positions falling outside this region are shown in (a) and (b). **(d)** 3D structure of amacGFP, with residues colored according to the average difference in effect between amacGFP and amacGFP:V11L at each position. Red indicates sites where mutations are on average worse in amacGFP:V11L than in amacGFP; blue indicates the opposite. Positions from (a) and (b) are shown here in spheres representation. V11L itself is shown in black.

2.4. Epistasis across genes

After considering each local landscape individually, we asked how transferable the information from each gene was: how correlated are mutation effects in different gene contexts? How prevalent is sign epistasis across genes — and therefore, how rugged the "global" GFP fitness landscape? We already observed that different genes displayed different mutational robustness, frequency of epistasis, and even structural properties of the mature WT protein. To understand to what degree knowledge about one gene is useful for describing another, we directly compared mutation effects across genes and looked at the effect of extant mutations (known to have been observed in other, functional GFPs in nature).

2.4.1. Extant mutations are less likely to be deleterious

Drawing from UniProt and available literature describing the discovery of novel fluorescent proteins in nature (Alieva et al., 2008; Fourrage et al., 2014; Labas et al., 2002; Shagin et al., 2004; Baumann et al., 2008), we manually curated a list of 68 extant, confirmed-functional FP sequences with recorded emission spectra. We used the multiple sequence alignment tool T-Coffee Expresso (Armougon et al., 2006) to structurally align these 68 protein sequences along with avGFP, amacGFP, cgreGFP, and ppluGFP2. The majority of these proteins shared between 25-50% sequence identity with each other. We will refer to amino acid states observed in one or more wild GFPs, at a given position of the alignment, as "extant states" or "extant mutations".

We expected extant states to generally be, in the absence of epistasis, non-deleterious, based on the fact that they are observed in nature to be present in functional FP sequences. Indeed, in comparisons of genotypes with the same total number of mutations, genotypes containing solely extant states were overall brighter than those containing solely non-extant mutations; this was the case for all of our focal genes (Figure 19). However, even combinations of exclusively extant states were not universally neutral, consistent with the accumulation of individually small deleterious effects.

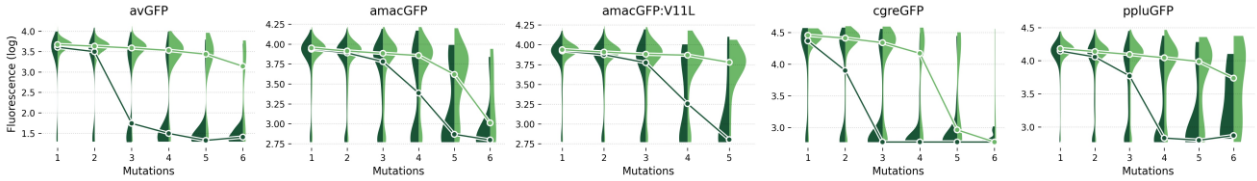

FIGURE 19. Effects of extant and non-extant mutations. For different total numbers of amino acid substitutions, fluorescence distributions of genotypes containing exclusively non-extant mutations (i.e. mutations not confirmed to exist in nature) are shown in dark green, and of those containing exclusively extant mutations (i.e. which have been documented in other functional GFPs) are shown in light green. Only categories containing at least 15 genotypes in total are shown. Median values are also plotted, in the respective color.

2.4.2. Changes in mutation effects across genes is not proportional to genes' sequence identity

For any given pair of genes, hundreds of mutations existed for which measurements had been carried out in both backgrounds. We defined a mutation as being the same in two different genes if it both a) occurred in the same position, according to the structurally aligned protein sequences (Figure 3b), and b) mutated to the same final amino acid state (regardless of the original WT state in that position). For each pair of genes, we measured the correlation between mutations' effects in one gene background and their effects in the other. The nearly identical amacGFP and amacGFP:V11L genes displayed a very strong correlation of mutation effects (Figure 20a), which is expected (Greenbury et al., 2016). Beyond that, we observed that overall, near-neutral mutations tended to remain near-neutral regardless of the gene background, and lethal mutations tended to remain lethal resulting in significant but moderate correlations seen in Figure 20a. However, correlations appeared independent of the sequence identity shared by the pair of genes under consideration (Figure 20b), even though mutation effects can be expected to change over the course of evolution (Orr, 1995; Starr & Thornton, 2016; Bazykin, 2015) and thus to correlate with phylogenetic or sequence distance.

In a similar vein, we looked at the rate of change of a mutation's sign, i.e. how frequently a neutral or beneficial mutation became deleterious in a different gene background. Frequency of sign epistasis is an indicator of the ruggedness of a fitness landscape, as it limits the viable mutational paths available between one sequence and another (Poelwijk et al., 2007; Saona et al., 2022). Furthermore, the probability of a mutation changing its sign from one background to another may be expected to increase as a function of the sequence divergence between the two backgrounds (Orr, 1995); this is the basis of, for instance, Dobzhansky-Muller incompatibilities, where speciation leads to increased genetic divergence, eventually to the point where mutations which are acceptable in one species are incompatible with the other (Orr & Turelli, 2001). For each pair of genes, we considered the fraction of mutations which were measured as neutral in one gene yet deleterious in the other, out of all the mutations which has been measured in both. Once again, sign epistasis was exceedingly rare between sequence neighbors amacGFP and amacGFP:V11L, but occurred in comparable amounts between other more distant pairs independent of their sequence distance (Figure 20c).

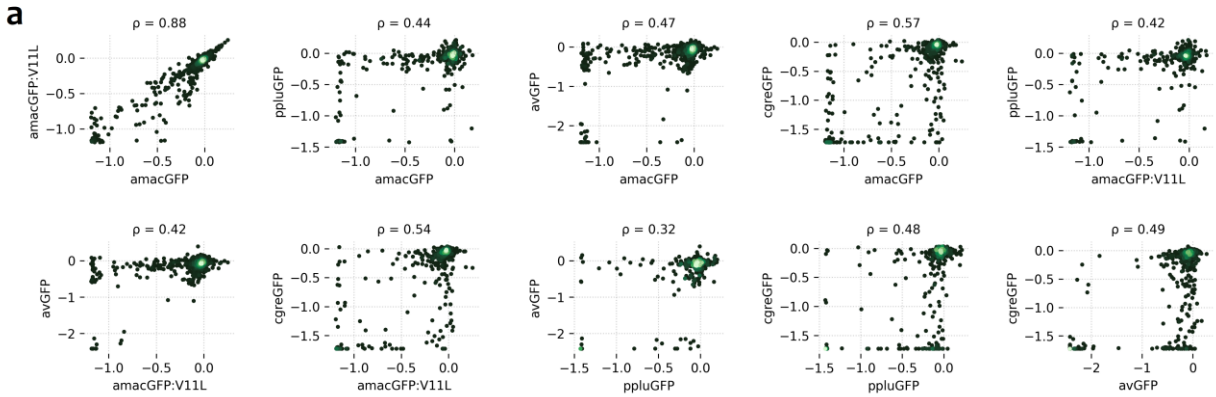

Figure continued on next page.

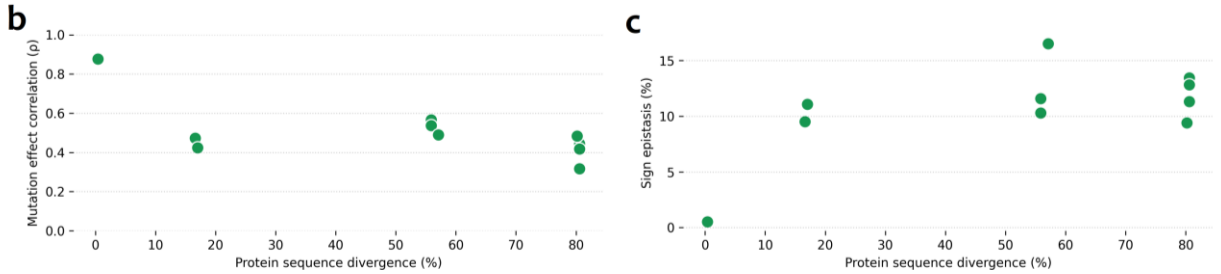

FIGURE 20. Effects of single mutations in different gene backgrounds. (a) Pairwise comparisons of the effects of single amino acid substitutions measured in multiple genes. The Spearman correlation coefficient is indicated; p-values were all below 10^{-8}. **(b)** Spearman correlations from (a) as a function of the sequence divergence (100 – amino acid identity, in percent) between each pair of genes. **(c)** Prevalence of sign epistasis. For each pair of genes, the percentage of mutations which are neutral in one (causing fluorescence loss of less than two standard deviations from WT brightness) but deleterious in the other (causing fluorescence loss of more than five standard deviations from WT brightness), out of all mutations measured in both backgrounds. X axis as in (b).

2.5. Machine learning-guided protein design

As described previously, in the absence of epistatic interactions between residues, the phenotype of any given multi-mutant genotype could be predicted accurately by simply summing up the individual effects of the relevant mutations. Extrapolating from this, novel and functional genotypes with untested combinations of mutations could be readily created by selecting and incorporating any number of mutations observed to be neutral or beneficial. However, epistasis is rampant (Figure 8), making the joint effect of groups of mutations unpredictable. In practice, identifying compatible combinations of mutations with no understanding of their potential interactions can be expected to be as successful as the idea, frequently misattributed to Charles Darwin, of a blind man in a dark room searching for a black cat who isn't there.

To leverage our datasets of tens of thousands of genotype-phenotype measurements, we used our library data to train machine learning models to predict fitness from genotype (see: 4.10. Machine learning). We expected that a successful understanding of the rules governing a protein's fitness landscape should enable not only describing the observed data, but also creating novel functional genotypes. To test this, we generated artificial protein sequences with up to 48 amino acid substitutions, predicted to fluoresce by ML models, and experimentally tested them. Furthermore, as "more data = more better" is widely accepted to apply to machine learning (as long as the data is of good quality), we tested whether data from different GFP genes could be jointly applied to produce better results than data from one library alone.

Our thanks go to **Aubin Fleiss** and **Katya Putintseva** for creating the machine learning models used in this section.

2.5.1. Neural networks with sigmoid activation layers can transform fitness potential into fluorescence

The amacGFP (including amacGFP:V11L), cgreGFP, ppluGFP2 and avGFP datasets were used separately to train a variety of neural net architectures to predict fitness from genotype (see: 4.10.1. Modeling of local landscapes), using 60% of the data for training, 20% for validation, and 20% for testing (Table 2). As expected, simple linear models performed notably better in predicting fluorescence in amacGFP and ppluGFP2 datasets, where epistasis was less abundant, than in cgreGFP and avGFP datasets, as judged by the models' R^2 (Figure 21a). Building from that basic linear model, the addition of an output node with sigmoid activation improved predictions for all genes, though only very moderately in the case of cgreGFP (Figure 21b), while the addition of an output subnetwork of ten sigmoid neurons followed by a final linear output node resulted in a substantial improvement in cgreGFP predictions and a small improvement in the already good performance for other genes (Figure 21c).

The final network architecture was based on that of the models with output subnetworks. Optimized models consisted of one input layer which received one-hot encoded protein sequence information, one hidden layer of neurons with linear activation functions, a second hidden layer of neurons with sigmoid activation functions, and one final linear node which output the predicted fluorescence of the input

genotype (see: 4.10.1. Modeling of local landscapes). The optimized architectures performed similarly well for the different genes, with R^2 values around 0.9 (Figure 21d).

The output from the network's linear layer can be understood as the fitness potential (P), a weighted sum of mutation effects (m_i) on fitness such that $P = \alpha_0 + \alpha_1 m_1 + \alpha_2 m_2 + \ldots + \alpha_n m_n$. However, due to epistatic interactions and/or effects of mutations on intermediate phenotypes, the fitness potential alone does not reliably predict final fluorescence (Figure 21a). The much more accurate fluorescence predictions after the addition of a sigmoid layer show that the underlying relationship between fitness potential and fluorescence can be captured by non-trivial sigmoid functions, which is consistent with the libraries' previously observed bimodal fluorescence distribution and mutational threshold effects (see: 2.1.1. Distributions of mutation effects on fluorescence are bimodal). The transformation of fitness potential into fluorescence output by the optimized architecture can be visualized in Figure 21e.

Table 2. ML models and gene predictions: numbers and statistics.

	avGFP	amacGFP	cgreGFP	ppluGFP2
R^2 (linear model)	0.68	0.77	0.6	0.82
R^2 (sigmoid model)	0.9	0.86	0.63	0.85
R^2 (output subnetwork)	0.94	0.88	0.86	0.9
R^2 (optimized architecture)	0.95	0.91	0.89	0.92
N. mutations used across all novel genotypes	—	359	243	427
N. mutations used across successful novel genotypes	—	210	218	225
N. conditionally deleterious mutations used in successful novel genotypes	—	28 (13.4%)	58 (26.7%)	33 (13.2%)

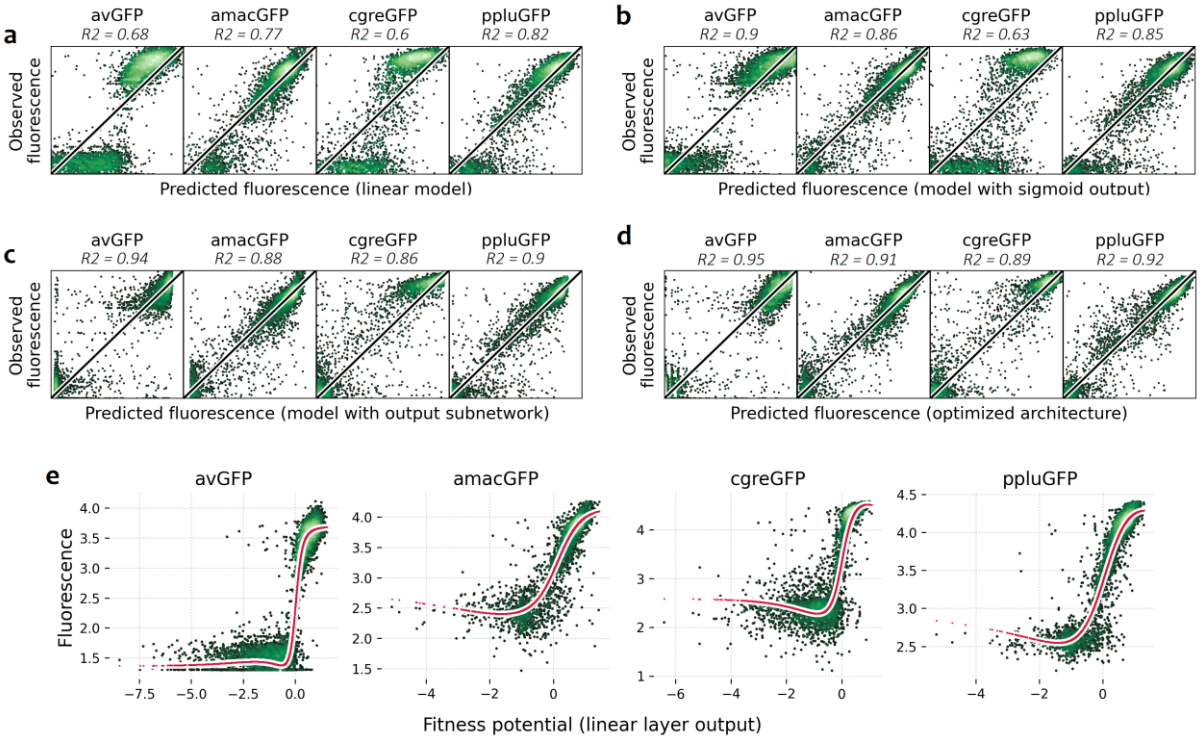

FIGURE 21. Fluorescence-predicting performance of machine learning models. In (a), (b), (c), and (d), each dot represents a genotype whose measured fluorescence is depicted on the Y axis; the corresponding fluorescence predictions output by the neural networks are shown on the X axis; lighter color represents higher density of data points; and the diagonal line represents a theoretical 1:1 perfect correlation between real (Y) and predicted (X) values. The models tested are as follows. **(a)** Linear model consisting of one input layer and one linear output node. **(b)** As (a) with the addition of one sigmoid output neuron. **(c)** Model with output subnetwork: one input layer, one linear node, one layer of 10 sigmoid nodes, and one linear output node. **(d)** Final,

optimized architecture: one input layer, one layer of linear nodes followed by a Monte Carlo dropout layer, one layer of sigmoid nodes followed by a Monte Carlo dropout layer, and one linear output node. **(e)** Observed fluorescence values (point cloud) and model prediction (red line) as a function of the fitness potential output by the network's linear activation layer.

2.5.2. Mutationally fragile libraries provide better training data for novel protein design

A genetic algorithm was used to generate artificial protein sequences with increasing numbers of mutations in combinations not observed in the data (see: 4.10.2. Generation of novel protein sequences predicted to fluoresce). The expected fluorescence of these artificial sequences was then predicted by the optimized neural network of the gene in question, and the top sequences were selected for experimental validation (see: 4.11. Experimental validation of novel gene sequences).

We focused initially on the mutationally robust amacGFP and ppluGFP2, on the basis that a) these genes were already observed to tolerate greater numbers of mutations, making them better candidates for the creation of novel multi-mutation genotypes, and b) they displayed comparatively little epistasis overall, thus decreasing the potential for unexpected interactions between mutations chosen by the genetic algorithm. We started by experimentally testing 12 sequences each of amacGFP and ppluGFP2 genotypes with 6, 12, 18, and 24 mutations. All tested genotypes were predicted to glow at least as brightly as the WT, but while the intensity of fluorescence was not always up to expectations, genotypes of either gene with 6-12 mutations were largely functional, as were half of ppluGFP2 genotypes with 24 mutations (Figure 22).

We then expanded the range of ppluGFP2 predictions to include 12 genotypes each with 30, 36, 42, and 48 mutations. We optimistically also tested amacGFP predictions with 30 mutations, nearly all of which were found to be non-functional, and did not push any further with amacGFP. Crucially, we decided to also begin testing predictions for cgreGFP, with 6 and 12 mutations, despite cgreGFP's mutational fragility and pervasive epistasis promising to make it a difficult gene to successfully mutate. The ppluGFP2 genotypes with 30+ mutations were only occasionally functional, but interestingly, all but one of the cgreGFP 6-12-mutant predictions were near WT fluorescence, which could not be said for either amacGFP or ppluGFP2 predictions with the same number of mutations. We then expanded cgreGFP predictions up to 48 mutations; while the average fluorescence intensity decreased with increasing numbers of mutations, the majority of sequences were functional, all the way up to 48 mutations — representing around 20% sequence divergence (Figure 22).

The fact that novel genotypes incorporating so many mutations were much more successful in the background of mutationally fragile and highly epistatic cgreGFP was a surprising result, particularly so because the algorithm was not limited to using so-called universally neutral mutations. In fact, all tested genotypes, successful or not, incorporated at least one conditionally deleterious mutation (i.e. observed to be deleterious, often highly so, in at least one context within the library, even if usually near-neutral) (Figure 22b). Furthermore, out of all mutations used across ML-predicted genotypes for a given gene, successful cgreGFP genotypes incorporated conditionally deleterious mutations around twice more frequently than amacGFP or ppluGFP2: 26.7% of mutations in functional cgreGFP predictions were observed to have deleterious effects of at least -0.3 (two-fold fluorescence decrease), versus 13.4% in amacGFP and 13.2% in ppluGFP2 (Table 2). This suggests that the model was able to capture epistasis to some extent and thereby avoid negative interactions. Indeed, out of all cgreGFP library genotypes with 6 mutations which were expected to be functional under an additive model, only ~30% actually were (Figure 8b); by comparison, ML-generated cgreGFP genotypes with 6, 12, and 18 mutations were universally functional, and frequently even maintained near-WT fluorescence levels (Figure 22b).

The greater success of the cgreGFP model was not an artifact of a larger training dataset, as the amacGFP and ppluGFP2 datasets contained 6,000-9,000 more assayed genotypes than cgreGFP (Table 1). However, the mutational fragility of cgreGFP may have resulted in a less "noisy" dataset for ML models to learn from: the fact of its fluorescence being so easy to kill may have made it easier for the models to understand what doesn't work. Conversely, the failed amacGFP and ppluGFP predictions may have suffered from higher order epistatic interactions which were not captured in the library data due to the scarcity of measured genotypes with 8 or more mutations; this would be consistent with the observed epistasis in amacGFP and ppluGFP datasets tending towards the higher order (Figure 8c).

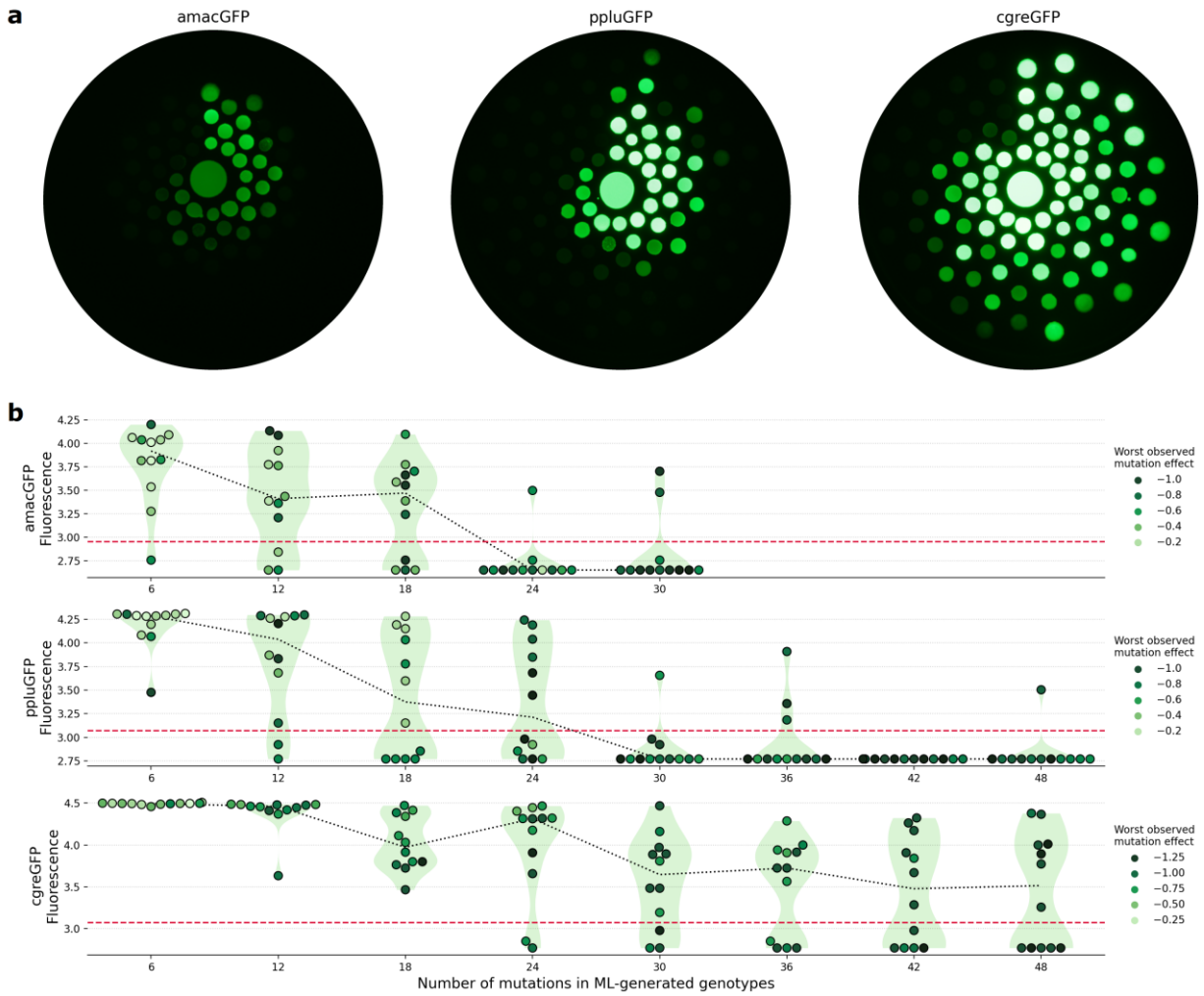

FIGURE 22. Experimental validation of ML-generated artificial genotypes. (a) Plated spots of *E. coli* expressing ML-generated protein sequences. The WT is represented in the middle of each plate; genotypes with increasing numbers of mutations (from 6, up to 30 or 48, in steps of 6) are arranged in concentric circles around the WT. **(b)** Quantified fluorescence of ML-generated genotypes from (a). To highlight the frequency of conditionally deleterious mutations even in functional sequences, the color of each dot (genotype) represents the worst observed effect (as measured in the corresponding gene's library data, across multiple backgrounds) of any mutation comprising the genotype in question. The horizontal dashed red line marks the non-functionality threshold (P3 gate border). The dotted black line tracks the median fluorescence of genotypes in each group.

2.5.3. Handpicked combinations of beneficial mutations perform poorly

After observing that half of ppluGFP2 artificial genotypes with 24 mutations were functional, we wondered how well the ML models were learning about the underlying mutation interactions, as opposed to simply creating combinations of seemingly-neutral mutations. To this end, we generated 12 ppluGFP2-derived sequences, also with 24 mutations each, without the use of ML models (see: 4.11.2. Manual selection of top mutations in ppluGFP2). One of these genotypes consisted of the "top 24" mutations observed in the ppluGFP2 dataset, i.e. the 24 mutations with the most beneficial median effects, as measured over multiple ppluGFP2 backgrounds (see: 4.7.6. Calculation of mutation effects and epistasis). The remaining 11 genotypes consisted of random combinations of the "top 175" best mutations, all of which were observed to have median effects of 0 or higher; we chose a pool of 175 because this would provide comparable sequence diversity to the ML-generated genotypes.

Overall, the pool of mutations available to these non-ML genotypes had more beneficial effects on fluorescence than the group of mutations used in the ML-generated ones (Figure 23a). Nevertheless, these manually curated genotypes performed worse than their ML counterparts: only 3/12 were fluorescent (and the "top 24" candidate was not one of them), compared to 6/12 for ML genotypes, and none of the 3/12 were as bright as the WT or the top three ML genotypes (Figure 23b). This supports the notion of the optimized ML models being able to capture, and avoid, negative epistatic interactions.

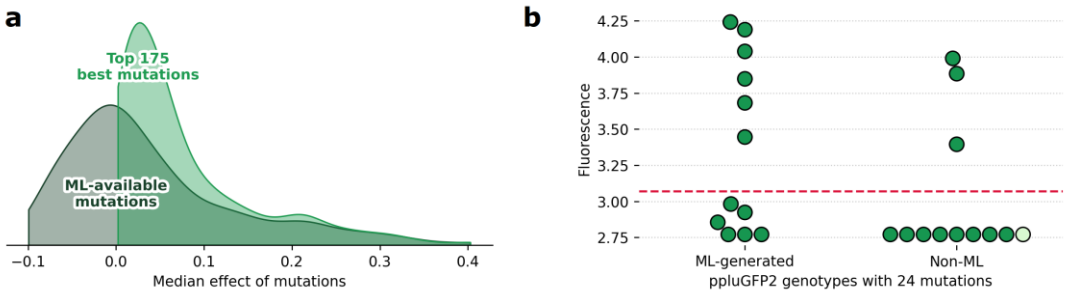

FIGURE 23. Comparison of artificial genotypes generated with and without ML. We generated 12 ppluGFP2-derived genotypes containing 24 mutations by randomly combining the most beneficial/neutral mutations, and compared their fitness with that of ML-generated ppluGFP2 genotypes with 24 mutations. **(a)** KDE plots showing the distribution of mutation effects of mutations used by the genetic algorithm in the 24-mutation ppluGFP2 ML genotypes (dark grey) and the mutations used to generate the non-ML genotypes (green). In either case, only mutations measured in at least 10 ppluGFP2 backgrounds were used; the X axis shows their median effect across all observed backgrounds. **(b)** Experimentally-measured fluorescence of ML-generated and non-ML-generated genotypes. The dashed horizontal line represents the cutoff for non-fluorescence (upper P3 border). The non-ML genotype consisting of the "top 24 best" mutations (as opposed to random combinations of the top 175) is represented by a lighter color.

2.5.4. Combining data from different local landscapes worsens performance

In addition to training ML models on single libraries and creating novel genotypes using a single gene as a starting point, we also trained models using combined data from multiple landscapes. While the different GFP orthologs differed from each other in terms of mutational robustness or physical stability of the mature WT protein, they also shared similarities — dependence of mutation effects on position and ΔΔG, threshold effects underlying fluorescence output — suggesting a common biophysical framework underlying mutation effects. This raises the possibility that information from one landscape may be partly extrapolated to others. While a large minority of specific mutations substantially changed effect across genes (Figure 20a), our ML models discussed above appeared capable of capturing and handling epistatic interactions to some degree. To what extent can data from distant local landscapes be combined in order to create an understanding of the global GFP sequence space? Can this data be generalized further, and be used to engineer novel, functional sequences that are distant from all assayed genes?

The range of fluorescence values of the different libraries varied considerably. Before combining them, the data were linearly rescaled to a common range by aligning the peaks of the fitness distributions. The combined dataset, consisting of rescaled genotype-phenotype data for amacGFP, cgreGFP, ppluGFP2, and avGFP, was used to train neural nets of the same general architecture as previously (one linear and one sigmoid hidden layers, one linear output node; aligned gene sequences were input under one-hot encoding). However, these models performed either worse or no better than models trained exclusively on data from one gene library, when predicting the fitness of genotypes from that library. A closer look at the neuron activation patterns revealed that given neurons were "specializing" in only one of the four genes. Although explicit information on library origin was not provided to the model, input sequences were processed by a different set of neurons according to which WT gene the network identified them as having originated from. Essentially, the global model consisted of four submodels, each dedicated to a different GFP (Figure 24a).

We speculated that this was due to the format of the input genotypes, which were one-hot encoded. As mutant genotypes typically contain only a few mutations, the vast majority of sequences from a given library will contain the same (WT) amino acid state at any given position. Input sequences can thus easily be clustered according to gene origin. To avoid this, new models were trained wherein only information on mutated positions was provided as input, and positions containing the WT state were not defined, precluding identification of the gene background by sequence alone. However, when models were trained on data from two or more libraries, they performed universally worse than when they were trained only on one (Figure 24b). Furthermore, models trained on one dataset performed very poorly on other gene datasets. This approach for input genotype formatting was abandoned.

Finally, we considered whether information could be transferred, or reused, from one dataset to another, by starting from the optimized single-gene models and partially re-training them on data from another library. However, models where either the linear or the sigmoid hidden layer was re-trained on data from a second library also resulted in a dramatic decrease in performance, with R^2 values dropping to below 0.5.

These results indicate that the rules underlying mutation effects and interactions are too different from one gene to the next for information pertaining to one gene to be relevant for another — at least for our particular panel of GFP orthologs.

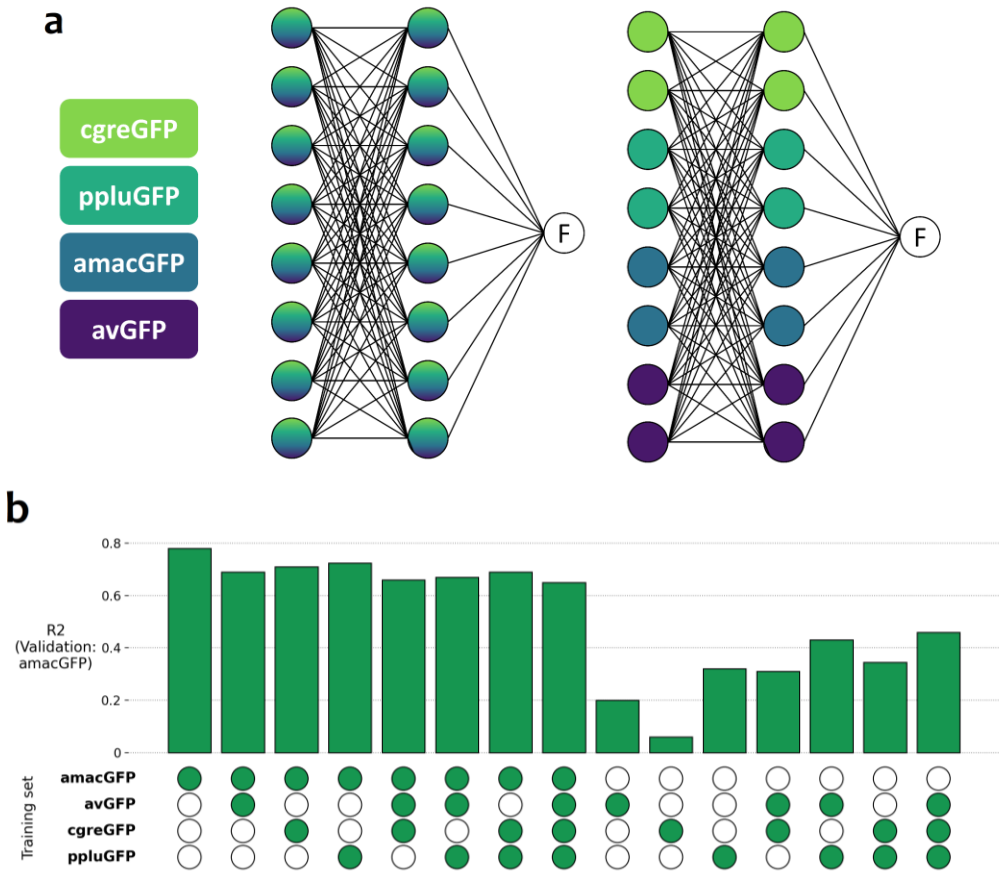

FIGURE 24. ML model training on multiple gene datasets. (a) Simplified schematic of an "expected" global model trained on multiple libraries (left) versus the observed reality of emerging gene-specific submodels (right). **(b)** Performance of ML models trained on different combinations of datasets. As an example, validation is shown for the amacGFP dataset. The addition of multi-gene training data universally worsened performance.

2.6. Landscapes of artificial cgreGFP-derived genes

Results from ML-guided creation of novel GFP sequences showed that the library of cgreGFP, the least mutationally robust gene and highly epistatic, yielded the best outcome in terms of designing functional, distant genotypes. This suggests that machine learning models may be able to capture negative epistatic interactions — and thereby avoid them during the generation of artificial sequences, and that cgreGFP's high sensitivity to mutations may have resulted in "cleaner" data, easier to interpret by ML models. Deleterious effects were more readily and consistently apparent in the cgreGFP library, whereas in amacGFP or ppluGFP2 they were often masked by the protein's general tolerance to perturbations.

On the other hand, combining data from multiple gene libraries resulted in a clear decline in ML models' ability to accurately predict fitness from genotype, possibly due to cross-gene epistasis causing conflicting information about mutation effects. In addition, changes in mutation effects from one gene background to another, while negligible in extremely close backgrounds (amacGFP vs. amacGFP:V11L), did not vary according to the sequence divergence of the genes under consideration (Figure 20). Key properties of local landscapes (shape, frequency of epistasis) and WT proteins (chemical/thermal stability) were highly variable, but these properties were not more similar in genes with higher sequence identity. As sequence divergence between pairs of genes ranged from 18% to over 80% (Figure 3b,c), these findings suggests that the underlying rules governing the emergence of epistatic interactions may change on a scale smaller than 18% sequence divergence. But how so? Is it a gradual change? Is there a threshold effect

and if so, what is the threshold?

To better understand this process of changing interactions, we sought to to fill this gap and analyze the behavior of mutations across pairs of genes with less than 18% sequence identity. For this, we expanded our datasets to include local landscapes of less sequence-divergent genes. However, wild GFPs are a highly varied gene family: the majority of members share only ~25-50% amino acid identity with each other, with only rare, isolated pairs sharing over 82% sequence identity (see: 2.4.1. Extant mutations are less likely to be deleterious). For our candidate genes for new landscapes, we therefore decided to draw from our collection of ML-generated, functionally-validated cgreGFP sequences, rather than from extant, natural genes. This would further allow a) with fewer tested genes, a greater number of pairwise comparisons in the 0-18% identity range, b) an opportunity to compare the changes emerging after departing in different "directions" from the same reference WT, and c) to assess the utility of multiple small local landscapes of highly similar sequences versus a single larger local landscape of a gene, for the training of machine learning models.

We selected two genes 6 mutations away from WT cgreGFP (cgreGFP:1338 and cgreGFP:132) and two genes 12 mutations away (cgreGFP:4111 and cgreGFP:9708). The "1338", "4111", and "9708" variants were chosen from amongst the most successful ML-generated genotypes, while cgreGFP:132 was created as a midway point between WT cgreGFP and cgreGFP:9708 (**Figure 25a,b**). We processed these four genes independently, creating and processing new mutant libraries for each. Furthermore, from among the ML-generated cgreGFP genotypes with 18, 24, 30, 36, 42, and 48 mutations, we selected the top two best (brightest) genes from each category and pooled them to create a fifth, "minis" library. Unfortunately, not all twelve variants were recovered in the final, filtered dataset, although we discovered an unexpected subset of "hybrid" genotypes — apparent crosses between different templates, likely caused by PCR chimeras. General statistics on these libraries can be found in **Table 3**.

Table 3. General statistics of new datasets. False positives and negatives are defined as in 4.7.4. Library data filtering. MutLD50 values were determined as in **Figure 7**. Values for WT cgreGFP are prvided for reference. The "1338", "132", "9708", and "4111" were processed independently; all other variants are part of the "minis" library (see: 4.2.1. Gene selection). Values labeled "—" were not calculable.

	Distance from cgreGFP	Number of assayed genotypes	False positives	False negatives	"WT" fitness	MutLD50 WT / Dark
cgreGFP	—	26165	0.75% (14/1860)	0% (0/1583)	4.5 ± 0.028	0.9 / 3.2
cgreGFP:1338	6	8934	0.32% (1/308)	0.43% (3/693)	4.41 ± 0.017	0.3 / 1.5
cgreGFP:132	6	4267	0.666% (1/150)	0% (0/504)	4.58 ± 0.012	0.9 / 6.6
cgreGFP:9708	12	4180	1.55% (2/129)	0.14% (1/701)	4.56 ± 0.026	1.6 / 7
cgreGFP:4111	12	8214	0.71% (2/280)	0.47% (3/643)	4.49 ± 0.013	0.5 / 2
cgreGFP:2880	18	1670	0% (0/96)	1.3% (1/77)	4.5 ± 0.017	0.9 / 2.9
cgreGFP:3224	18	150	0% (0/8)	0% (0/8)	4.47 ± 0.034	1.4 / 2.4
cgreGFP:900x	24	1106	0% (0/73)	0% (0/50)	4.46 ± 0.036	0.9 / 2.6
cgreGFP:575	24	38	0% (0/4)	0% (0/1)	4.44 ± NA	0.9 / 2.5
cgreGFP:83	30	265	0% (0/10)	0% (0/23)	4.36 ± 0.027	0.7 / 2.1
cgreGFP:626	30	182	7.1% (1/14)	0% (0/10)	4.44 ± 0.026	0.8 / —
cgreGFP:121	36	206	0% (0/13)	0% (0/7)	4.32 ± 0.013	0.3 / 2.3
cgreGFP:985	36	358	0% (0/15)	0% (0/14)	3.98 ± 0.124	1.3 / 2.4
cgreGFP:13	42	1	—	—	—	—
cgreGFP:911	42	—	—	—	—	—
cgreGFP:567	48	97	0% (0/5)	0% (0/7)	4.35 ± 0.046	1.4 / 2.3
cgreGFP:1414	48	45	25% (1/4)	0% (0/1)	4.5 ± NA	—
Hybrid	—	608	4.3% (1/23)	—	—	—

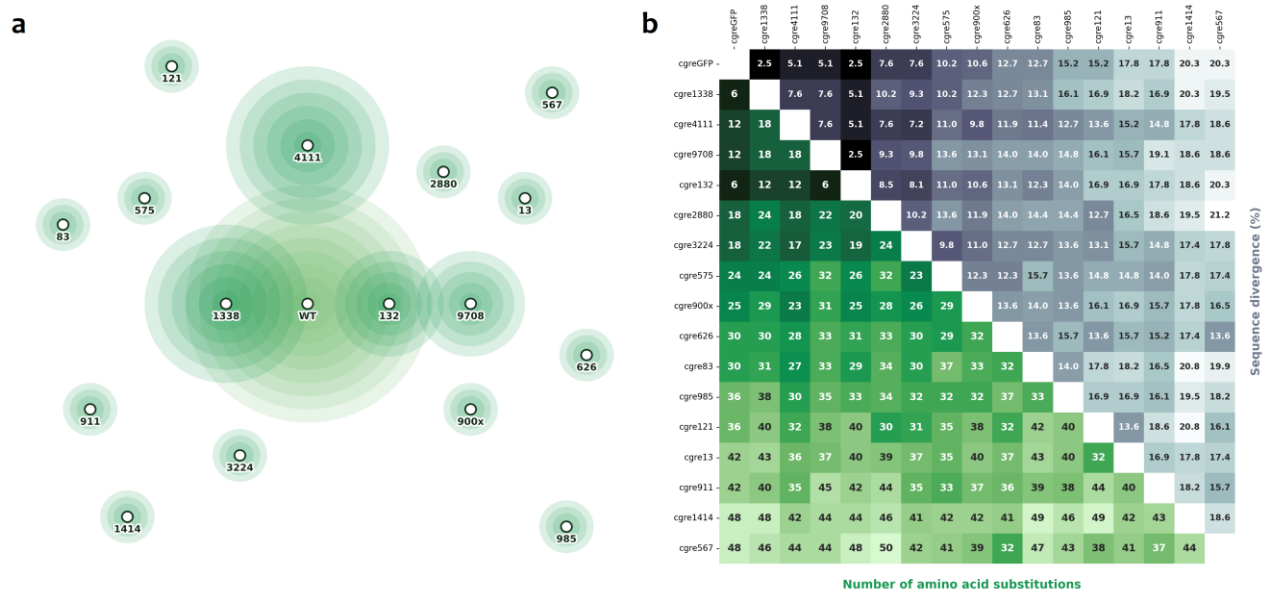

FIGURE 25. Novel cgreGFP-derived gene libraries. (a) Conceptual representation of landscapes of cgreGFP-derived genes. The original (WT) cgreGFP is in the middle, in lighter green. Distances (number of mutations separating each gene) are roughly proportional between cgreGFP, cgreGFP:1338, cgreGFP:4111, cgreGFP:132, and cgreGFP:9708; other distances are not to scale. Note: figure overlap between libraries does not imply actual overlap in terms of specific genotypes measured. **(b)** Hamming distances between all cgreGFP-derived genes: number of mutations in green, percent sequence divergence (100 – protein identity) in grey.

2.6.1. Few mutations are needed to drastically alter general properties

Consistent with previously described genes, all new libraries displayed bimodal distributions of fluorescence (Figure 26a), with peaks centered at near-WT and near-null fitness. Also consistent with previous findings and with general expectations based on GFP structure, mutations in the new genes were more deleterious overall if they affected buried, internally oriented residues compared to solvent-exposed ones (Figure 26b,c).

However, despite all being derived from the same mutationally fragile and highly epistatic cgreGFP wild-type, the new set of libraries displayed great variability in terms of mutational robustness and pervasiveness of epistasis. While cgreGFP:1338 and cgreGFP:4111 were even more mutationally fragile than WT cgreGFP, cgreGFP:132 and cgreGFP:9708 were the most robust genes of any analyzed so far (Figure 27, Table 3).

The previously observed link between low mutational robustness and high epistasis was supported by the new data, with a much higher proportion of variants of fragile genes ("1338", "4111") than robust genes ("132", "9708") losing functionality due to negative epistasis (Figure 28b). In particular, cgreGFP:4111 (more fragile and more epistatic overall than WT cgreGFP) and cgreGFP:9708 (highly robust with relatively little overall epistasis, similarly to ppluGFP2) highlighted the robustness/epistasis link (Figure 28a). On the other hand, cgreGFP:1338 ($MutLD50_{(Dark)}$ = 1.5, ~10% of genotypes with negative epistasis) displayed less epistasis overall than might be naively expected from comparing it to WT cgreGFP ($MutLD50_{(Dark)}$ = 3.4, ~12% epistasis) or cgreGFP:4111 ($MutLD50_{(Dark)}$ = 2, ~17% epistasis). This is likely due to the fact that, for negative epistasis to be detected, the fluorescence expectation under a non-epistatic model must be high enough; yet, cgreGFP:1338's extreme mutational fragility meant that the majority of multi-mutant genotypes were *expected* to be low-fitness even under an additive model of mutation effects. Indeed, the rate of negative epistasis in cgreGFP:1338 peaks in double-mutant genotypes and then steadily declines, while other genes show more higher-order epistasis (Figure 28c).

Both cgreGFP:1338 and cgreGFP:132 are only six mutations away (2.5% protein sequence divergence) from WT cgreGFP, and cgreGFP:4111 and cgreGFP:9708 are 12 mutations away (5% sequence divergence). Yet, the local landscapes of these genes vary considerably from each other and from the original WT from which they derive. This indicates that the general features of protein fitness landscapes can be readily altered through seemingly minor changes in the starting protein sequence. This is consistent with existing knowledge on, for example, specific point mutations causing significant changes in protein properties (Bloom et al., 2005; Jacquier et al, 2013).

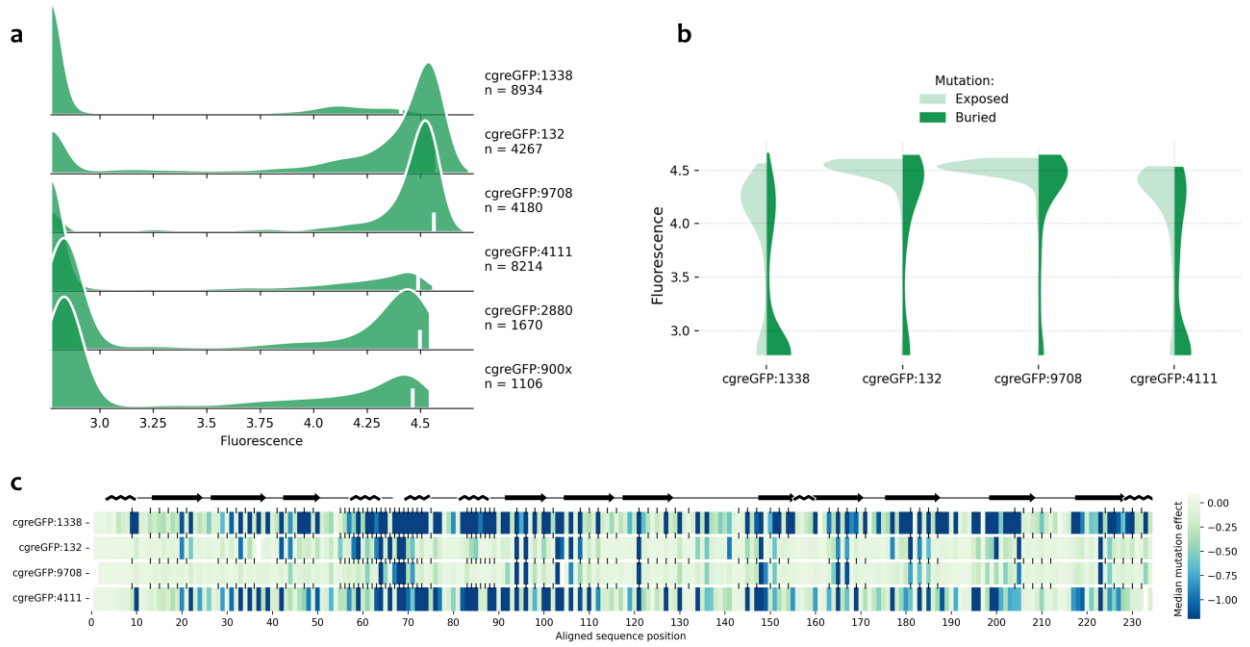

FIGURE 26. General features of cgreGFP-derived landscapes. (a) KDE plots showing the bimodal distribution of fluorescence values. Respective "WT" values are represented by white bars. The number of measured genotypes is indicated on the right; only genes with >1000 measured genotypes are included. **(b)** Distribution of fluorescences of genotypes with exactly one mutation, split according to whether the mutation affects a buried residue (side chains facing inward) or an exposed one (side chains facing outward). Only genes with >800 measured single-mutant genotypes are represented. **(c)** Median effects of single mutations based on their position in the protein sequence. Only genes with >800 measured single-mutant genotypes are represented. Buried positions are labeled with black tick marks. As a reference, the secondary structure of WT cgreGFP is displayed at the top.

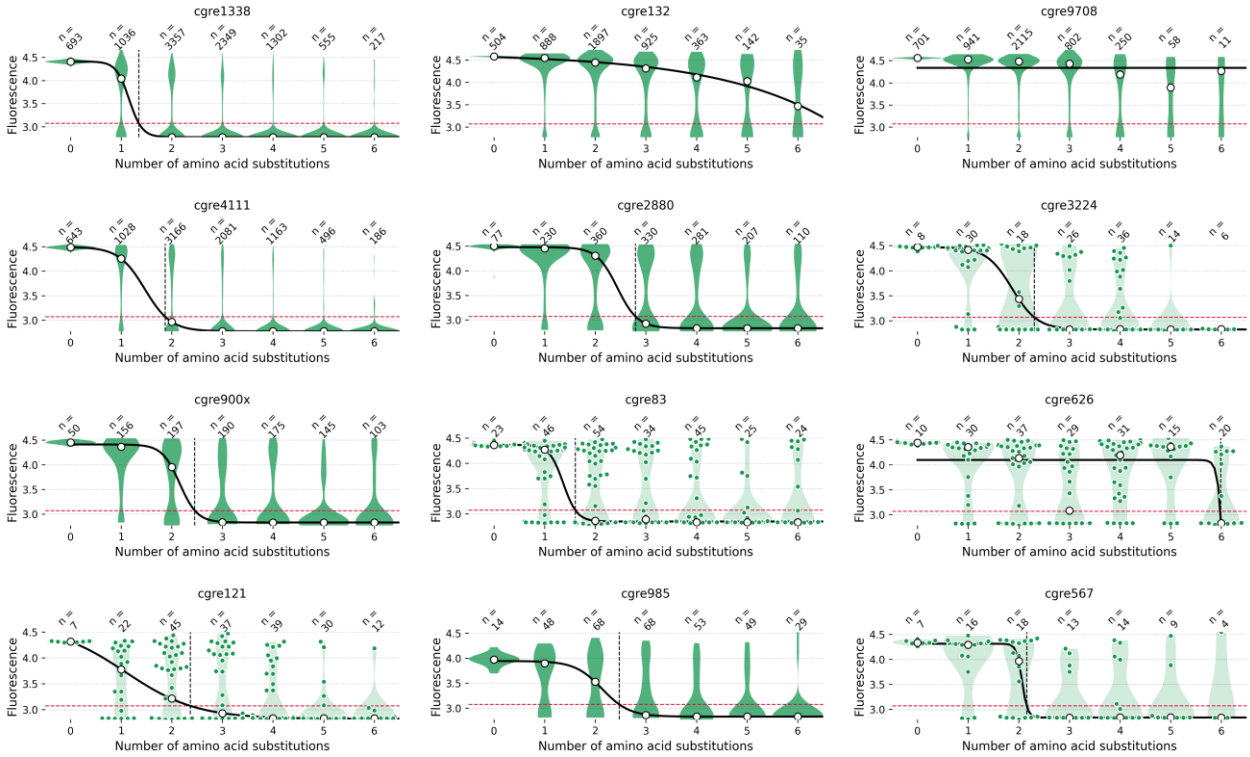

FIGURE 27. Distribution of fluorescences of cgreGFP-derived genes, according to number of mutations. Only genes with >50 measured genotypes are displayed; for genes with <300 measured genotypes, individual data points are shown in addition to the overall distribution. The number of measured genotypes in each category is indicated at the top. Median fluorescence values (white circles) were used to fit a logistic curve (black); horizontal red dashed lines represent the threshold for non-functionality. The cgreGFP:9708 data could not be fit by a logistic curve.

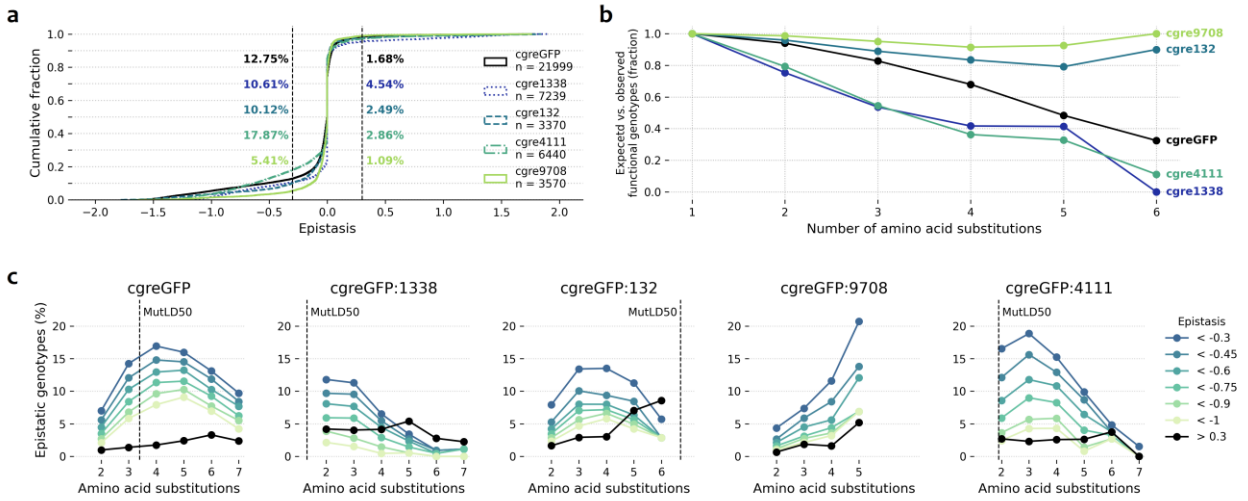

FIGURE 28. Overview of epistasis in new cgreGFP-derived libraries. Compare: Figure 8. Only libraries with >4000 measured genotypes (cgreGFP:1338, cgreGFP:132, cgreGFP:9708, cgreGFP:4111) are displayed. **(a)** Cumulative distribution of observed epistasis. WT cgreGFP data is provided in black, for comparison. **(b)** For *n* mutations, fraction of genotypes observed to be functional, out of all those expected to be so under a non-epistatic additive model. WT cgreGFP data is provided in black, for comparison. **(c)** Prevalence of negative epistasis of varying magnitudes (color) and positive epistasis over 0.3 (black). Where possible, dashed vertical lines mark the number of mutations needed to eliminate fluorescence in 50% of genotypes. Only categories with >10 epistasis-measured genotypes are displayed. WT cgreGFP data is shown for comparison.

2.6.2. Mutations change effect across genes depending on differences in sequence and robustness

As previously, we compared the effects of single mutations in different gene backgrounds. Having over a dozen small genotype-phenotype datasets centered around highly similar cgreGFP-derived genes allowed us to expand the number of pairwise comparisons from 10 (see: 2.4.2. Changes in mutation effects across genes is not proportional to genes' sequence identity) to 49 (considering only gene pairs with at least 100 mutations measured in both backgrounds).

The correlation between mutations' effects across two gene backgrounds was dependent on the sequence identity between the genes in question. While this dependence was not observable previously when comparing genes with a minimum sequence divergence of 18% (Figure 20), cgreGFP-derived gene pairs provided coverage of the 2-18% sequence divergence range and showed a steady decline in the correlation of mutation effects (Figure 29a). New data points in the >50% sequence divergence range were consistent with previous data.

The same was not the case for sign epistasis: two genes' sequence divergence was only moderately correlated to mutation effects changing sign from one to the other, and variability was very high (Figure 29b). The chance that a deleterious mutation will become neutral in a different background (or vice versa) can be more easily explained as a function of the difference in mutational robustness of the two genes in question. The rate of sign epistasis was moderately correlated with the difference in $MutLD50_{(Dark)}$ between the two backgrounds (Figure 29c), and highly correlated with the difference in $MutLD50_{(WT)}$ (Figure 29d). This makes intuitive sense, as an overall lower proportion of single mutations will be deleterious in a gene with high mutational robustness, while mutationally fragile genes will see a lower fraction of single mutations being neutral. Thus, a mutation which is neutral in a robust background is more likely to exhibit sign epistasis (become deleterious) if said robust background is compared to a fragile one than to another robust one. That said, the fact that genes with high sequence identity exhibit significant differences in local landscape shape (see: 2.6.1. Few mutations are needed to drastically alter general properties), resulting in frequently high rates of sign epistasis even between pairs of similar sequences (Figure 29b), suggest that the global GFP fitness landscape is substantially rugged.

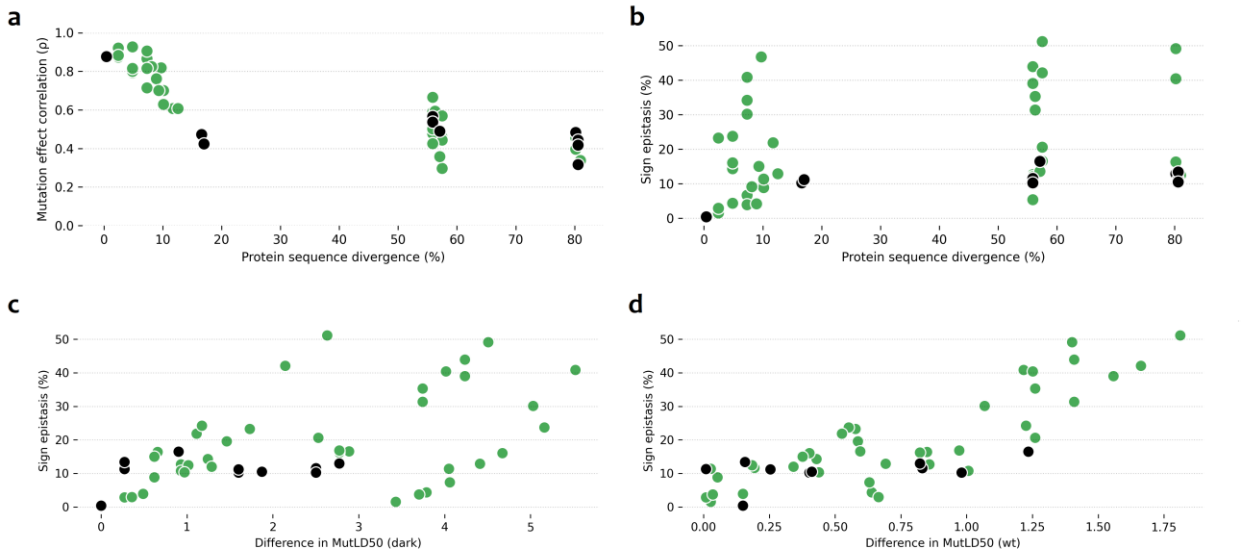

FIGURE 29. Effects of single mutations in different gene backgrounds, revisited. Compare: Figure 20b,c. **(a)** Spearman correlations between measured mutation effects in pairs of gene backgrounds, as a function of the sequence divergence between the genes. Only pairs with >100 mutations measured in both backgrounds are displayed. Data points from Figure 20b are in black, new data from cgreGFP-derived genes are in green. Overall, the negative correlation between sequence distance and mutation effects is high, at $\rho = -0.89$ ($p < 10^{-16}$). **(b)** For each pair of genes, the percentage of mutations which are neutral in one but deleterious in the other, out of all mutations measured in both backgrounds. X axis as in (a). Only pairs with >100 mutations measured in both backgrounds are displayed. Data points from *Figure 20c* are in black, new data are in green. Overall, the correlation between sequence distance and proportion of sign epistasis across genes is $\rho = 0.39$ ($p < 0.006$). **(c)** Sign epistasis calculated as in (b), but shown as a function of the difference in $MutLD50_{(Dark)}$ values of the two genes in each pairwise comparison. The correlation here is $\rho = 0.46$ ($p < 0.0009$). **(d)** As (c), but for $MutLD50_{(WT)}$. In this case, the positive correlation is higher, at $\rho = 0.73$ ($p < 10^{-9}$).

2.6.3. Protein stability and mutational robustness, revisited

As for amacGFP, ppluGFP, and original cgreGFP, we tested tested the stability of the new "WT" proteins in order to detect a potential relationship between physical stability and mutational robustness. As mentioned previously, a non-mutant reference protein starting from a low physical stability may be more vulnerable to mutations which would, on average, destabilize it even further (Bloom et al, 2005; Zeldovich et al., 2007; Bershtein et al., 2006). In the absence of resolved crystal structures of the 16 new cgreGFP-derived genes, we could not calculate the expected ΔΔG values of mutations. However, we performed urea and temperature sensitivity tests as previously (see: 4.9.7. Urea sensitivity assays, 4.9.8. Thermosensitivity assays).

We performed urea denaturation assays on all 16 new proteins. The initial shapes of absorbance and emission spectra were broadly similar to those of cgreGFP (Figure 30), though their responses to 9 M urea varied greatly in sensitivity. Some proteins displayed their own idiosyncrasies, such as cgreGFP:985 possessing a second absorbance peak around 390 nm (possibly reflecting differences in the anionic state of the chromophore (Chudakov et al., 2010)) or cgreGFP:121 in PBS increasing in fluorescence over time (possibly triggered by increased oxygenation during plate shaking, if cgreGFP:121 matured less readily than other variants?). Fluorescence emission and absorbance half-lives ranged from lower to much higher than those of original cgreGFP (Figure 31a,b). However, the proteins' half-lives appeared uncorrelated with their mutational robustness (Figure 32).

We also measured thermosensitivity of all proteins by performing melting curves in a qPCR machine (see: 4.9.8. Thermosensitivity assays). The melting temperatures of the "1338", "132", "9708" and "4111" variants were additionally measured by DSF and DSC (Figure 33a-d). However, the additional data points revealed that the tentative link between thermal stability and mutational robustness, suggested by amacGFP/cgreGFP/ppluGFP2 data, was not generalizable across other tested GFPs (Figure 32).

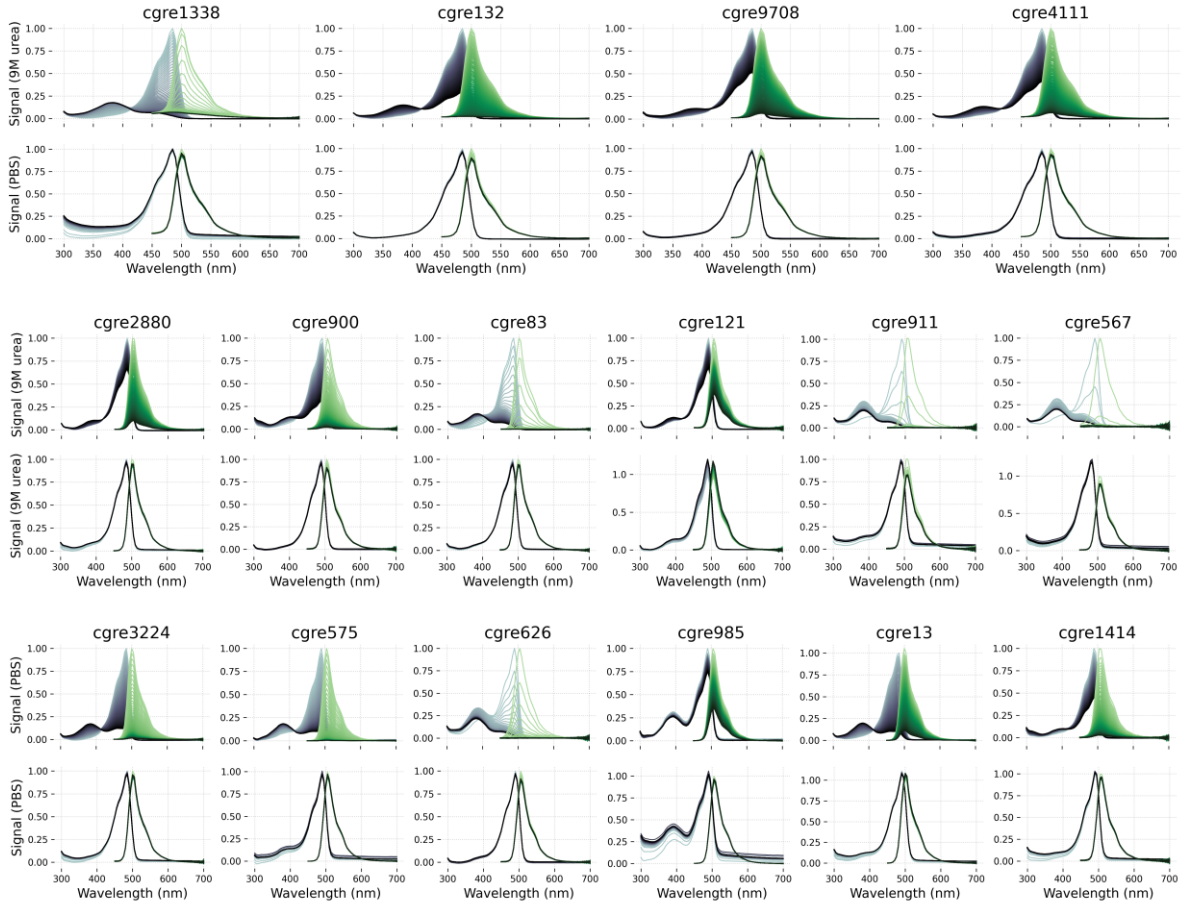

Figure 30. Absorbance and emission spectra of cgreGFP-derived genes in 9M urea and PBS. As in Figure 17a,b, spectra were scanned in 5 nm intervals, regularly over the course of ~60 hours: each line represents one time point (the darker, the later in the time series).

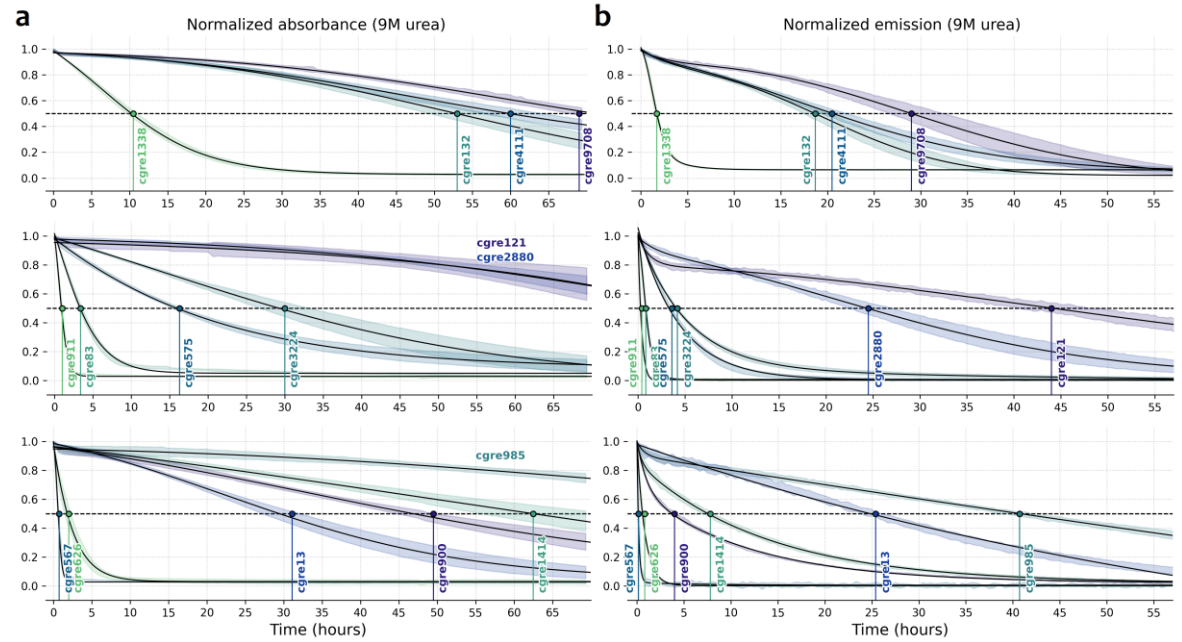

FIGURE 31. Decay of absorbance and fluorescence of cgreGP-derived genes in 9M urea. Compare: Figure 17c,d. The 16 genes pictured are split into three rows for ease of visualization. **(a)** Loss of absorbance over time, monitored at the wavelength corresponding to the absorbance peak of each protein. Absorbance peaks were at 485 nm for the "1338", "132", "4111", "9708", "2880", "13", "3224", and "83" genes, and 490 nm for all others. Curves were fitted with a logistic function. **(b)** Loss of fluorescence emission over time upon 420 nm excitation, monitored at the peak emission wavelength for each protein. Emission peaks were 500 nm for the "1338", "132", "4111", "9708", "2880", "3224", and "83" genes, 510 nm for cgreGFP:911, and 505 nm for all other genes. The fluorescence loss of all proteins except for the "1338", "132", "4111", "9708", "2880", "121", and "83" variants could be successfully fit with a sum of two exponential decay functions.

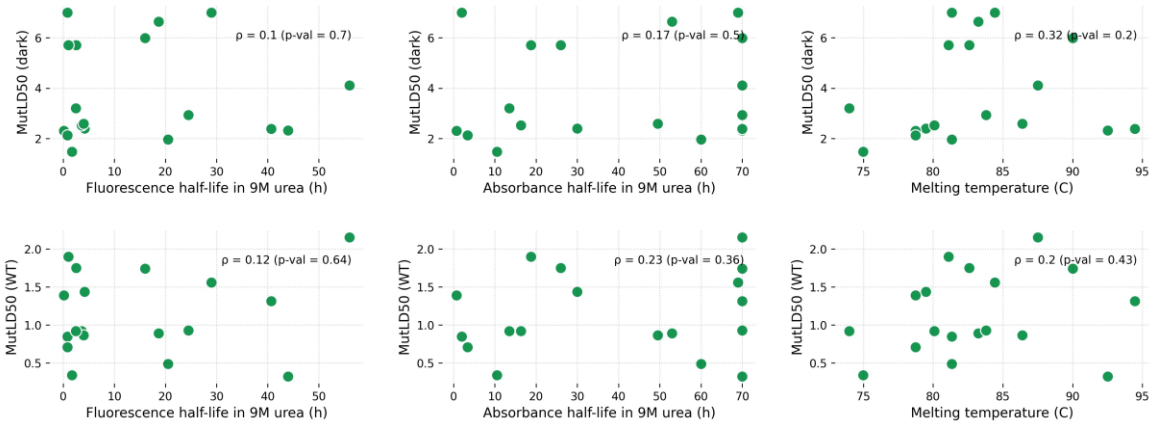

FIGURE 32. Mutational robustness and physical protein stability. Mutational robustness ($MutLD_{(Dark)}$, top row, and $MutLD50_{(WT)}$, bottom row) as a function of proteins' fluorescence half-lives in 9M urea (left), absorbance half-lives in 9M urea (center), or melting temperature as determined in a qPCR machine (right). Only genes with sufficient data to calculate MutLD50s are displayed. Absorbance half-lives half-lives were set to the maximum 70 hours for genes which had not yet lost 50% of their initial absorbance value by the end of the measurements.

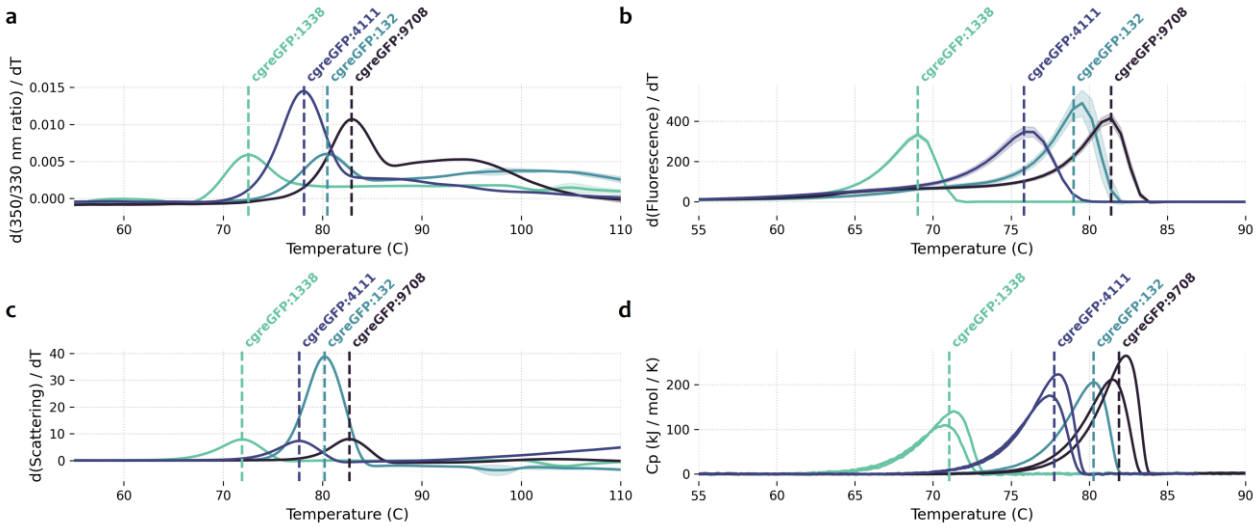

FIGURE 33. Thermostability of four cgreGFP-derived genes. Compare with Figure 16a-d. Only genes for which 4000+ genotypes were measured are shown. Melting temperatures are indicated by vertical dashed lines. **(a)** DSF-measured thermal unfolding, measured in triplicate. **(b)** Melting curves measured via qPCR machine, measured in eight replicates. **(c)** DSF-measured thermal aggregation, measured in triplicate. **(d)** DSC-measured specific heat capacities, measured in duplicate (due to a machine error during one of the cgreGFP:132 replicates, only one run is shown for that gene.)

2.6.4. Hybrid gene variants tend to maintain function

The cgreGFP:minis library, wherein 12 cgreGFP-derived variants were pooled at the beginning of library creation and mutagenesis, contained ~600 genotypes which could not be assigned a single origin, i.e. the coding sequence appeared not to be a version of one of the 12 original templates, but rather a hybrid or chimera molecule with more than one "parent". This was likely due to the formation of chimeras during the mutagenic PCR step (see: 4.6.5. NovaSeq PE250 data clean-up, **Figure 40**).

The calculation of a mutation's effect requires knowing the fitness of a reference (WT) sequence and that of the same sequence plus the mutation (see: 4.7.6. Calculation of mutation effects and epistasis). As these 600 genotypes are all unique hybrids between different combinations of the 12 "minis" origins, containing multiple (average: 12.8) mutations compared to the closest parent (plus any additional mutations introduced during mutagenesis), it was not possible to calculate specific mutation effects in this data.

However, we observed that fitnesses of the hybrid genotypes followed the usual bimodal distribution of all other libraries (**Figure 34**), with a substantial fraction being brightly fluorescent: ~40% of the genotypes overall fell within the four brightest FACS gates, even among hybrids with over 15 mutations

compared to the nearest parent. This seemed notable, as previous data showed that much fewer than 15 mutations are typically required to eliminate fluorescence in virtually all genotypes (Figure 6, Figure 27), even if the mutations in question represent extant states from functional wild FPs (Figure 19). This suggests that recombination between reasonably distant functional sequences (parent genotypes in the "minis" library all share ~80-85% sequence identity, see Figure 25b) may be a way to introduce large numbers of mutations at once while maintaining protein function. Indeed, recombination between homologous sequences is known to result in novel, functional genes: for instance, viruses famously recombine to give rise to medically relevant variants (Perez-Losada et al., 2015). Furthermore, the genetic algorithm employed during the generation of our artificial GFP sequences used a recombination approach (see: 4.10.2. Generation of novel protein sequences predicted to fluoresce), although it did not start from such divergent sequences. Hybridization between homologous genes may be another tool with potential to help generate novel distant, functional sequences.

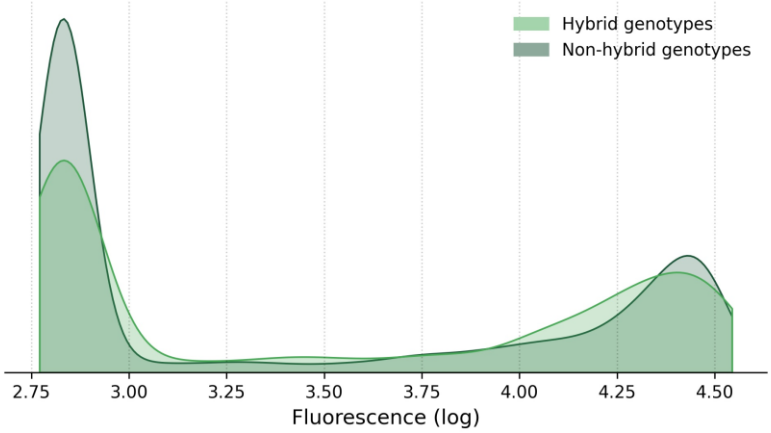

FIGURE 34. Fitness distribution of hybrid genotypes. The distribution of fluorescences of the ~600 hybrid genotypes from the "minis" library is is light green; that of the rest of the "minis" library is in dark green. KDE plots are normalized such that the area under the curve for each category sums to 1.

3. Discussion

Protein fitness landscapes remain a useful tool in the fields of protein design and bioengineering, as they help elucidate the effects and interactions of mutations which may be leveraged with the aim of creating novel sequences with particular functions. Indeed, experimental studies of protein landscapes are becoming more and more common as sequencing costs decrease and high-throughput experimental setups improve, even in the advent of emerging ML and other computational methods aimed at generating novel proteins by relying only on extant sequence data or biophysical protein folding models (Dou et al., 2018; Lisanza et al., 2023; Hayes et al., 2024).

However, given that epistatic interactions are rampant (Bershtein et al., 2006; Mackay, 2014; Olson et al., 2014; Poelwijk et al., 2019), experimental genotype-phenotype data must provide information on such interactions in order to be useful in predicting mutation effects or generating novel proteins. However, our data shows that not all proteins are equally amenable to providing such information, at least not within the constraints of an experimentally feasible amount of measurements: counter to intuition, mutationally robust proteins are not easier to introduce large numbers of mutations into, if the goal is to create very distant functional sequences.

This is partly a consequence of positive epistasis being rarer than negative — it is easier to break a thing than to fix it, so models are better served by learning the patterns of negative epistasis and avoiding what is known not to work. This appears to be more readily achieved with data from mutationally fragile proteins like cgreGFP, where such examples of failed mutational combinations are frequent and unambiguous. In contrast, data from mutationally robust proteins is arguably more confusing, inconsistent, and difficult to learn from, as slightly negative effects and interactions may remain undetected until it is too late (Figure 22a,b).

Thus, our results indicate that mutational robustness and frequency of epistasis are relevant factors to consider, when selecting a sequence to begin engineering. However, identifying good, low-robustness candidate genes is not necessarily straightforward: even with ever-improving technological and experimental advances, carrying out a fully-fledged landscape analysis is highly time- and resource-intensive.

Furthermore, the environment in which a protein of interest is studied should be carefully considered: the same collection of mutant genotypes may yield differently-shaped landscapes depending on the experimental conditions in which they are assayed. Different gene orthologs may have evolved in different habitats and thus require different conditions for optimal function. One should therefore carefully consider the parameters of the experimental setup in which the phenotype of interest will be measured, as well as the desired contexts in which the results will be applied.

3.1. On protein stability and selecting candidate genes for landscape surveys

Existing literature is replete with theory and examples suggesting that the effects of mutations in coding sequences are dependent on their effect on protein structure and/or stability. An important link between these two quantities has been shown in e.g. fitness landscape studies of the TEM-1 β-lactamase (Bershtein et al., 2006; Jacquier et al., 2013), the IgG-binding domain of protein G (GB1) (Olson et al., 2014), the chaperone protein Hsp90 (Hietpas et al., 2011), a poly(A)-RNA-binding protein Pab1 (Melamed et al., 2013), indole-3-glycerol phosphate synthase (TIM barrel) proteins (Chan et al, 2017), a human WW domain (Fowler et al., 2010), influenza nucleoprotein (Gong et al., 2013), and avGFP itself (Sarkisyan et al. 2016), as well as in more general, theoretical works (Zeldovich et al., 2007; Wylie & Shakhnovich et al., 2011; Starr & Thornton, 2016).

The ability to fold correctly and to maintain physical integrity in physiologically relevant molecular environments are general requirements that apply to the great majority of proteins, regardless of their function — intrinsically disordered proteins being a notable exception. So, are more physically unstable

variants typically also less mutationally robust, and therefore better candidates for fitness landscape studies? Intuitively, this makes sense: if mutations' effects on the phenotype of interest are mediated by their effect on an intermediate phenotype such as folding or stability, then it is plausible that more initially unstable variants are more vulnerable to further, small destabilizations which would more easily push them beyond the threshold of non-functionality (Bershtein et al., 2006). Under this model, a protein's mutational robustness is an emerging property of its physical stability and/or folding ability: the higher the latter, the more mutations are required, on average, before observing critically negative effects.

In this work, we have seen several factors which support a link between protein structure and/or stability and mutational robustness. Physically proximal pairs of residues are more likely to be epistatic (Figure 9). As a group, mutations with higher predicted ddG values — expected to have more destabilizing effects on protein folding — had overall worse effects on fluorescence (Figure 13). Buried residues with internally-oriented and potentially chromophore-interacting side-chains were more sensitive to mutation (Figure 12). Finally, mutation effects in amacGFP and amacGFP:V11L varied in a structurally-biased manner, with sites around the lower barrel lid being predominantly more sensitive in amacGFP:V11L (Figure 18). These features are all indicative of protein structure and/or stability playing an important role in the final phenotype, in this case fluorescence.

However, our assays on protein stability — in terms of folded proteins' sensitivity to urea and temperature — indicate that such straightforward measurements of a protein's physical properties are not necessarily a reliable indicator of its response to genetic mutations.

Our initial batch of genes — avGFP, amacGFP, ppluGFP2, and cgreGFP — suggested a link between melting temperature and mutational robustness, with avGFP being an outlier gene exhibiting great thermostability (as well as urea resistance) while being overall mutationally fragile (Figure 16). The case of avGFP may potentially be related, at least in part, to differences in experimental design between the 2016 avGFP landscape survey (Sarkisyan et al., 2016) and that of other GFPs in this work. AvGFP was expressed from a plasmid, with a copy number of ~30 per cell — protein expression levels were therefore at least an order of magnitude higher than those of subsequent GFPs which were expressed from a single, genome-integrated copy. Existing literature shows that some mutation effects do not manifest equally at different protein concentration levels (Jiang et al., 2013). While the use of mKate2 as an expression control allowed us to avoid such problems within any given experiment, libraries expressed from genomic and plasmid DNA may not be directly comparable in this regard. Furthermore, protein aggregation is a well-known consequence of protein misfolding (Hartl & Hayer-Hartl, 2009), but is also dependent protein concentration (Ignatova & Gierasch, 2004). Different mutant variants may be more prone to aggregate at different concentration levels, but higher expression levels may be more intolerant to slightly destabilizing mutations and lead to more widespread aggregation — affecting fluorescence readouts — overall. Thus, a given gene's fitness landscape may vary in shape — with a sharper peak meaning lower mutational robustness — depending on its expression level during experiments. In the case of avGFP, we cannot exclude the possibility that its local landscape may have appeared flatter and less epistatic if it had been expressed from a genomic copy.

Nevertheless, subsequent assays with additional proteins did not support a direct link between chemical and/or thermal sensitivity and mutational robustness of GFP variants (Figure 32). So while protein structure and folding ability do appear to underlie landscape shape, these features were not captured by assays measuring the denaturation of already folded proteins. Why was this the case? Proteins, even small ones, may pass through various structural intermediates before reaching their final native state (Brockwell & Radford, 2007); this has been shown to be the case for GFP as well (Andrews et al., 2007; Reddy et al., 2012). A proportion of slightly misfolded and/or partially folded variants may thus become trapped in local energy minima and fail to reach their intended final conformation — and while this is certainly exacerbated by the presence of destabilizing mutations, it is also the case even for sequences without such mutations (Reddy et al., 2012). Sequences with a higher propensity for becoming waylaid on the way to their native structure can be thought of as being more easily pushed past a critical stability threshold by the introduction of mutations, which in turn manifests as lower mutational robustness. However, the ability of successfully folded proteins — a subset of all starting molecules — to withstand temperature- or chemically-induced denaturation may not necessarily reflect the ability to fold correctly in the first place. A protein's ΔG — the change in energy between folded and unfolded states — has traditionally been used to describe folding ability, and this measure can be determined to some extent by controlled denaturation/refolding assays. However, recent literature casts doubt on the universal validity of this measurement as well (Sorokina et al, 2022).

What other measured could be used as a proxy for mutational robustness? A study in yeast has

indicated that mutations cause, on average, worse fitness costs when introduced into higher-fitness (i.e. faster-growing) strains (Johnson et al., 2019). This suggests that high-fitness genotypes may be more vulnerable to mutations in general and thus display lower mutational robustness than lower-fitness ones. However, this finding does not appear to generalize reliably across other phenotypes: our own data contains multiple examples of highly-fit (fluorescent) genotypes being more mutationally robust than their less-fit counterparts. For instance, cgreGFP:132 and cgreGFP:9708 are both brighter than cgreGFP and significantly more robust, while cgreGFP:1338 is both dimmer and more mutationally fragile (Table 3). Thus, simply surveying prospective proteins for the phenotype of interest and selecting the fittest variants may not reliably reveal the best candidates for fitness landscape studies in all cases.

Overall, more effective approaches for approximating a protein's mutational robustness are thus desirable, as this could be a deciding factor when choosing a starting sequence for a fitness landscape survey from among a pool of candidates.

For some proteins with easily measurable phenotypes, constructing small landscapes of several hundred mutants of different genes may suffice to grant reasonable insight into the most promising candidates, in terms of mutational robustness (Figure 27). Even without full library sequencing, the overall distribution of fitnesses is quite informative: while this distribution is bimodal across the board, more mutationally fragile genes display a higher proportion of genotypes in the "unfit" peak (Figure 26a): this can be ascertained even without any library sequencing.

3.2. On machine-learning methods in landscape data analysis

We have seen that datasets rich in negative mutation effects and interactions, i.e. those derived from mutationally fragile genes, are more conducive to learning the patterns required to create successful, novel combinations of mutations. High-fluorescence, ML-generated genotypes with 6-48 mutations were generated with the highest success in cgreGFP, the least robust gene, all of which contained at least one conditionally deleterious mutation — i.e. observed to be deleterious in at one or more backgrounds it was measured in (Figure 22b). Moreover, combinations of hand-picked, "best" (least deleterious) mutations were unsuccessful in ppluGFP2, a mutationally robust gene with low levels of detected epistasis (Figure 23b). This suggests that ML models were able to learn, to some degree, how to avoid genetic contexts or interactions which would trigger deleterious effects, but without limiting themselves to simply combining universally neutral mutations.

An alternate hypothesis may be that artificial cgreGFP variants were more successful due to WT cgreGFP being significantly brighter than WT ppluGFP2 or amacGFP (Figure 22a), thereby providing a larger buffer zone between functionality and non-functionality. However, this explanation is implausible on several levels. First, it would imply that brighter GFPs should be more mutationally robust, but our data has already shown that this is not the case: the majority of cgreGFP genotypes with >3 mutations are already non-functional (Figure 6d). Furthermore, if we accept the role of protein stability underlying mutational robustness, higher fluorescence values — or indeed many other phenotypes of interest — are not necessarily correlated with higher physical stability and/or folding ability, and therefore need not be linked to mutation tolerance.

It should also be noted that the differences between sharper and flatter landscapes can be thought of as a matter of degree and not a matter of kind. All assayed genes in this work display fluorescence loss following a threshold effect, with varying mutational robustness and epistasis levels defined by higher or lower thresholds. In the case of cgreGFP, negative interactions were detectable early on due to only few mutations being necessary to disrupt fluorescence. Whereas for amacGFP and ppluGFP2, slightly deleterious mutations were able to accumulate without immediate phenotypic consequence — however, once the (higher) threshold was crossed, negative effects and interactions manifested just as they did for cgreGFP. Thus, ML models' performance on flatter landscapes might improve if provided with more training data focused on this transition region (6+ mutations) and/or if training data prioritized genotypes with greater numbers of mutations. However, the size of the theoretical sequence space available increases exponentially with every step from the WT. An adequate sampling of e.g. 6-mutant space might require significantly more data points than a representative sampling of e.g. 3-mutant space, in order to achieve the same level of usefulness for ML. In that case, fitness landscape surveys of genes with low mutational robustness remain the more experimentally economical choice.

The new data from cgreGFP-derived genes (see: 2.6. Novel cgreGFP-derived genes) is currently being

analyzed by ML collaborators. Can models trained on one gene perform well on other genes as long as mutation effects are highly correlated in both? Does having a set of smaller landscapes located clustered in a small area of sequence space provide any benefit over a single, larger landscape centered on one gene — perhaps by improving generalizability without the performance drops observed when combining more distant landscapes? Future results will further inform the ideal approach to fitness landscape surveys aimed at ML methods for protein engineering.

3.3. On the global GFP landscape

In any study of multiple members of the same gene family, the question arises of how generalizable one gene's properties are to others. A study of >5,000 mutations in three orthologous TIM barrel proteins, sharing 30-40% sequence identity, found that their fitness landscapes were highly correlated, suggesting the possibility that "fitness landscapes can be translocated in sequence space" (Chan et al., 2017). Our work also showed significant, if moderate (ρ = ~0.5) correlations between mutation effects across genes (**Figure 29a**), the extent to which findings from one GFP landscape can be extrapolated to another appears limited, as indicated in particular by the poor performance of ML models trained on data from multiple orthologs (**Figure 24b**).

The correlation of mutations effects across pairs of GFPs was dependent on the sequence identity between the two genes, up to a certain point. Theory indicates that changes in mutations' effects across genes or species are expected to increase with phylogenetic distance (Orr, 1995); this explains phenomena such as hybrid incompatibilities, where mutations impact fitness according to the genetic context in which they occur (Orr & Turelli, 2001). However, beyond ~18% sequence divergence, we observed that correlations of mutation effects plateaued (**Figure 29a**). Similarly, changes in sign occur at a steady rate across such gene pairs, affecting ~10-15% of all mutations (**Figure 20c**); though variance is high, this rate does not appear to depend on sequence divergence, except for nearly-identical gene pairs (**Figure 29b**). This is in line with existing literature on sign epistasis across species, which have found that ~10% of pathogenic mutations in one species are neutral (compensated) in other backgrounds, and that this remains true regardless of the phylogenetic distance between the species being compared (Kondrashov et al., 2002; Kulathinal et al., 2004).

While mutation impacts are known and expected to change over the course of evolution (Starr & Thornton, 2016; Bazykin, 2015), the finding that sequence divergence beyond ~18% seems independent of the rate of change in mutation effects implies that the rules underlying epistatic interactions tend to shift on a smaller scale than this. This would be consistent with the observed heterogeneity of fitness landscapes of genes with high sequence identity (**Figure 27**, **Figure 28a,b**). However, this does not mean that the same rules cannot be maintained and applied across greater distances; indeed, successful ML-generated cgreGFP genotypes with 48 mutations (20% sequence divergence from the WT) likely were successful because the selected mutations interacted in ways which were still consistent with their behavior in the original cgreGFP library. Conversely, failed ML-generated genotypes may represent not only cases where the interactions of selected mutations were not fully understood, but also cases where the nature of these interactions had changed compared to the training data.

Our data points in the 2-18% sequence divergence range are limited to pairwise comparisons between cgreGFP-derived genes, all of which — except for the original WT cgreGFP itself — were ML-generated based on WT cgreGFP data. This raises the question of whether artificial sequences are biased, in terms of their landscape properties, compared to wild-evolved sequences with comparable levels of identity: wouldn't sequences purposefully derived from the same starting point be more likely to share similarities with each other? However, this argument may be applied to natural sequences as well. First, all fluorescent proteins are believed to have evolved from one common ancestor (Shagin et al, 2004); *a priori*, the genetic algorithm used to generate our artificial sequences (see: 4.10.2. Generation of novel protein sequences predicted to fluoresce) was as as just blind as naturally occurring mutations, prior to selection pressure. Second, empirical data showed that the closest near-natural sequences assayed, avGFP and amacGFP, did not share more similarities with each other than with cgreGFP and ppluGFP2, other natural sequences (**Figure 6**, **Figure 7**, **Figure 8**). As mentioned above, landscape features such as mutational robustness and pervasiveness of epistasis did not correlate with sequence distance past ~18%, and landscapes of artificial cgreGFP-derived sequences were themselves highly heterogeneous (**Figure 27**, **Figure 28**). Any biases in landscape features thus seem likelier to arise due to specific experimental measurement setups —

excitation at 488 nm, growth at 30°C, etc., which may differ from the optimal environment for any given variant — than from the fact of having derived form a common reference protein.

Overall, data from natural GFPs from four species as well as multiple artificially-generated GFP sequences indicate that the global fitness landscape of green fluorescent proteins is highly heterogeneous. While a common framework of protein stability and folding ability can be understood to underlie GFPs generally, there remains great variability across genes in terms of landscape shape as well as physical properties such as resistance to denaturation, and this variability may be modulated by the introduction of only a few mutations, such that sequence distance between two genes is not a reliable indicator for similarities in their behavior. At the same time, our data suggests that genes with lower mutational robustness yield useful training data for machine learning more readily than genes with higher mutational tolerance, where mutation impacts may be conditionally masked and more difficult to interpret. At a time when fitness landscape surveys are becoming ever more common and experimentally accessible (Fragata et al., 2019; Flynn et al., 2023), we hope that this study of orthologous versions of a model protein will impact the understanding of fitness landscapes of protein families and inform the selection of promising candidate genes for future studies.

4. Materials and Methods

4.1. General protocols

This section describes basic, general-use molecular biology techniques which were employed repeatedly as part of various other protocols. They are grouped here to avoid unnecessary repetition.

4.1.1. Golden Gate cloning

Golden Gate is a type of cloning where DNA digestion and ligation steps are carried out simultaneously in the same tube, eliminating the need for gel purification of specific digested DNA fragments prior to ligation. This is achieved through the use of Type IIS restriction enzymes, which cut DNA at a predictable location outside of their own recognition site. This provides two main benefits. Firstly, the recognition sites and cut sites can be placed such that, upon successful ligation of the desired molecules, the recognition sites are no longer present and the DNA can therefore not be cut again. This feature is what allows digestion and ligation to occur in a single reaction (when using restriction enzymes and ligase which are functional in same buffer) (Figure 35). Secondly, the user is able to control the exact sequence of nucleotides which will be cut into sticky ends, allowing easy customization and mixing and matching of modular cloning "blocks". The overhangs are typically designed to be non-palindromic, so there is only one way for all blocks to fit together; this reduces the incidence of incorrectly cloned constructs.

The Golden Gate protocol used in this work was adapted from Weber et al., 2011, with thermocycling parameters adapted from Iverson et al., 2016. The following parameters were used for all cloning reactions unless otherwise specified.

Reaction mix

- ✓ 50 ng insert DNA (if multiple inserts, 50 ng each)
- ✓ 50 ng destination vector
- ✓ 20U Type IIS restriction enzyme (BsaI or BpiI)
- ✓ 10U T4 DNA ligase buffer
- ✓ 2 µl 10X T4 ligase buffer
- ✓ H_2O up to 20 µl

Temperature cycling parameters

1. 10 min 37°C (restriction enzyme ideal temperature)
2. 25 cycles:
    2.1. 16°C, 90 s (T4 ligase ideal temperature)
    2.2. 37°C, 3 min
3. 5 min 50°C, 10 min 80°C (enzyme denaturation step)

Typically, the DNA of interest replaces a *lacZ* cassette in the destination vector, allowing for blue/white color screening of colonies. When it is important to minimize the total amount of *lacZ*-positive colonies on the plate, an extra digestion-only step may be performed after the above cycling protocol is complete, by adding fresh restriction enzyme and incubating at 37°C again.

Transformation of competent cells

10 µl of Golden Gate reaction were used to transform 50 µl of chemically competent cells. Cells were plated on LB agar plates containing the appropriate antibiotic, and incubated overnight at either 37°C or 30°C as needed. Colonies were first screened by PCR for the expected insert length and then sequence confirmed by Sanger sequencing.

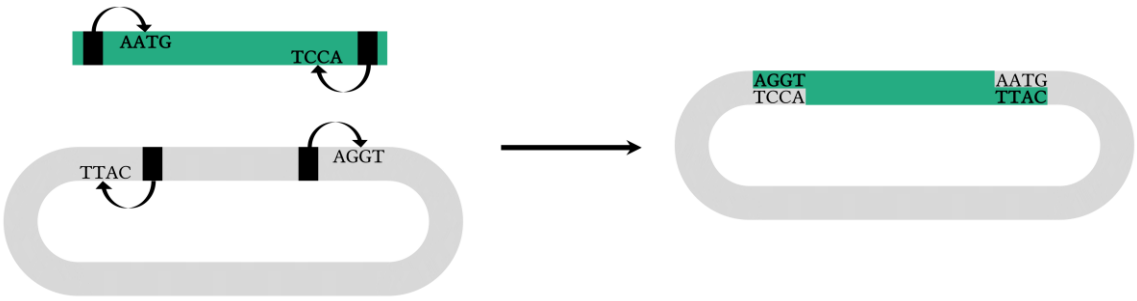

FIGURE 35. Principle of Golden Gate cloning. Type IIS restriction sites (black) cut outside of their recognition sites. After ligation of the desired insert into the destination vector, recognition sites are no longer available, allowing for one-tube simultaneous digestion/ligation reactions.

4.1.2. Colony screening and other PCRs

All PCRs were performed using Encyclo (Evrogen), OneTaq (NEB), or Q5 (NEB) (see: 4.12.1. List of consumables and services). General protocols are listed below; deviations will be specified in the relevant Methods sections.

Colony screening

The majority of our cloning setups featured blue/white color screening. Nevertheless, white colonies were screened by PCR for correct insert size wherever possible (i.e. *not* for all 100k variants in a library) and Sanger sequenced. Colony PCRs were performed in a total reaction volume of 10 µl, using bacterial cells directly as the DNA template by lightly touching a pipette tip to the colony and transferring the cells to the PCR reaction tube. Bacteria were not lysed separately prior to PCR, as the initial PCR step of ~3 minutes at 94-95°C was sufficient in that regard. Colony PCRs were performed using OneTaq or Encyclo polymerases for 25 cycles; see below for protocols.

OneTaq PCR

As a low-fidelity polymerase, OneTaq was exclusively used for colony screening. The reaction mix for 10 µl is shown below; scale up as needed.

- ✓ A tiny amount of cells, or 5+ ng of DNA
- ✓ 5 µl OneTaq 2X Master Mix
- ✓ 0.2 µl each of 5-10 µM forward and reverse primer
- ✓ H_2O up to 10 µl

Thermocycling parameters:

1. 3 min 94°C
2. Cycles (variable number):
    1. Denaturation: 30s 94°C
    2. Primer annealing: 20s T_m (primer-dependent)
    3. Extension: 68°C, 1 min per kb
3. 5 min 68°C; final extension

Encyclo PCR

Encyclo polymerase features ~20-fold higher fidelity than Taq, and higher processivity than Q5. It was used for occasional colony screening, and for amplification of some library DNA prior to sequencing (see: 4.6.1. Sample preparation: amacGFP, cgreGFP, ppluGFP2, 4.5.1. Barcodes sample preparation). The reaction mix for 10 µl is shown below; scale up as needed.

- ✓ A tiny amount of cells, or 5+ ng of DNA
- ✓ 1 µl 10X Encyclo buffer
- ✓ 0.2 µl 50X dNTP mix
- ✓ 0.2 µl 50X Encyclo polymerase

- ✓ 0.2 µl each of 5-10 µM forward and reverse primer
- ✓ H_2O up to 10 µl

Thermocycling parameters:

1. 3 min 95°C
2. Cycles (variable number):
    1. Denaturation: 10s 95°C
    2. Primer annealing: 10s T_m (primer-dependent)
    3. Extension: 72°C, 1 min per kb
3. 5 min 72°C; final extension

Q5 PCR

Q5 is one of the highest-fidelity polymerases available, but is a highly finicky creature (source: personal experience) especially compared to Encyclo. Q5 was used for amplification of some library DNA prior to sequencing (see: 4.6.3. Sample preparation: novel cgreGFP variants, 4.5.1. Barcodes sample preparation). The reaction mix for 50 µl was as follows:

- ✓ A tiny amount of cells, or 5+ ng of DNA
- ✓ 10 µl 5X reaction buffer
- ✓ 1 µl 10 mM dNTP mix
- ✓ 2.5 µl each of 10 µM forward and reverse primer
- ✓ 0.5 µl Q5 polymerase
- ✓ H_2O up to 50 µl

Note: the 2X Master Mix version tended to perform much more efficiently and reliably than the non-premixed Q5, unfortunately we only discovered it at the end of the project. Reaction mix for the 2X version was as follows:

- ✓ A tiny amount of cells, or 5+ ng of DNA
- ✓ 25 µl 2X master mix
- ✓ 2.5 µl each of 10 µM forward and reverse primer
- ✓ H_2O up to 50 µl

Thermocycling parameters:

1. 3 min 98°C
2. Cycles (variable number):
    1. Denaturation: 15s 98°C
    2. Primer annealing: 20s T_m (primer-dependent, but ~4°C higher than it would be for non-Q5 polymerases)
    3. Extension: 72°C, 30s per kb
3. 5 min 72°C; final extension

4.1.3. Electrocompetent cells and electroporation

E. coli cells were grown in liquid LB culture, with or without antibiotic as needed, in an air shaker at 37°C (or 30°C if containing pSIM5). Cells were grown in 50 ml aerated bioreactor tubes to an optical density of around OD_{600} = 0.8 (in our experience, this yielded better results than the often cited OD_{600} = 0.6), then washed with ice-cold water to render them electrocompetent.

Protocol (adapted from Sharam et al., 2009*):*

1. Chill cells on ice for 15 minutes, and cool centrifuges down to 4°C.
2. Centrifuge cells at 4500 g and 4°C for 7 minutes.
3. Discard supernatant and resuspend in 30 ml ice-cold distilled H_2O.
4. Centrifuge cells at 4500 g and 4°C for 7 minutes.
5. Discard supernatant and resuspend in 1 ml ice-cold distilled H_2O.
6. Transfer to 1.5 ml tubes.

7. Centrifuge at 10000 g and 4°C for 30 seconds.
8. Discard supernatant and resuspend in 1 ml ice-cold distilled H_2O.
9. Repeat steps 7-8.
10. Discard supernatant and resuspend in ice-cold 10% glycerol up to 400 µl.
11. Aliquot as needed (100 µl per reaction) and store at -80°C.

Before use, electrocompetent cells were thawed on ice. 100 µl of cells were mixed with 1-2 µl (5-50 ng) of DNA, then transferred to a chilled electroporation cuvette (1 mm electrode gap). Cells were electroshocked at standard settings for bacteria (1800 V, 25 uF, 200 Ω), paying attention to the time constants displayed after the pulse, as values below 4.6 ms are indicative of problems such as salt contamination which can lead to low transformation efficiency. 1 ml LB without antibiotic (supplemented with 0.2% arabinose if genome integrating) was added directly to the cuvette immediately afterwards, then cells were transferred to a 15 ml tube and incubated at 37°C with shaking for one hour (or two hours at 30°C if genome integrating).

After use, electroporation cuvettes were thoroughly rinsed with distilled water followed by 70% ethanol, three times, then given a final rinse with distilled water and left to air dry upside down. In our experience, cuvettes could be reused in this way indefinitely with no observed decrease in transformation efficiency.

4.1.4. Harvesting plated libraries

Cell libraries were plated on square plates. We used the number of colonies as a proxy for the number of variants in the library. To estimate that number, colonies were manually counted in several representative 2x2 cm sections and these counts were extrapolated to the full plated area.

Colonies were harvested from plates by adding 2-3 ml of M9 liquid media and using the long side of a generic glass microscopy slide (much faster and cheaper than using a cell scraper; credit to Bor Kavčič for this idea) to carefully scrape the cells off the surface of the agar. Plates were tilted to pool the liquid media into one corner and cells were recovered by pipetting.

If extracting plasmid DNA, cells were then pelleted. Cell pellets weighing up to 0.5 g were purified using Promega's PureYield Midiprep kit following the centrifugation protocol. Larger pellets (up to 2.5 g) were processed with Thermo Fisher's GeneJET Maxiprep kit following "Protocol A: Plasmid DNA purification using low speed centrifuges", or by splitting the sample across multiple PureYield Midiprep columns.

4.2. Creation of mutant GFP libraries

4.2.1. Gene selection

We selected an initial panel of eight GFP genes from six species to test:

- GFPxm191uv, derived from the jellyfish *Aequorea macrodactyla* (Luo et al., 2006); UniProt ref. Q8WTC7.
- CheGFP2 and CheGFP4 form the jellyfish *Clytia hemisphaerica* (Fourrage et al., 2014); UniProt refs. J9PGG2 and J9PJD5.
- cgreGFP derived from the jellyfish *Clytia gregaria* (Markova et al., 2010); UniProt ref. D7PM05; PDB structure 2HPW (Malikova et al., 2011).
- GFP509 from the jellyfish *Aldersladia magnificus* (unpublished); GenBank ref. ACC54354.1; UniProt ref. D3TI87.
- ppluGFP2, also known as copGFP, from the copepod (arthropod) *Pontellina plumata*; UniProt ref. Q6WV12; PDB structure 2G3O (Wilmann et al., 2006).
- GFP1 and GFP2 from *Asymmetron lucayanum*, a lancelet (cephalochordate) (Yue et al., 2016).

Protein-coding nucleotide sequences were codon optimized for *E. coli* expression (using known *Escherichia coli K12* codon usage tables downloaded from kazusa.co.jp/) and domesticated for Golden Gate

cloning (i.e. removing any restriction sites for BsaI, BpiI, and BsmBI) using custom Python scripts. For consistency, if multiple genes contained the same amino acid at a given position, the same codon was used for all genes at that site. Final sequences were ordered from Twist Bioscience as synthetic dsDNA and cloned as fusion proteins with mKate2, a red fluorescent protein used as a control of expression level in Sarkisyan et al., 2016, under a constitutive promoter. Constructs were transformed into *E. coli* DH5a cells and the resulting colonies were imaged under blue illumination to check for green light emission. GFPxm191uv, cgreGFP, and ppluGFP2 were strongly fluorescent under these conditions, while CheGFP2 was only dimly fluorescent and the remaining genes were not observed to fluoresce.

Note: the protein sequences of cgreGFP and ppluGFP2 used in this work were not modified from their wild type sequences, whereas GFPxm191uv, described in Luo et al., 2006, contains three amino acid substitutions compared to the true *A. macrodactyla* wild type: F64L (also present in the avGFP variant analyzed in Sarkisyan et al., 2016), Q69L, and T203C; these modifications are reported to improve the protein's fluorescence emission when expressed in standard laboratory conditions (Luo et al., 2006). For simplicity, we refer to this gene as "amacGFP" in this work. Also for simplicity, we refer to all reference sequences, cgreGFP, ppluGFP2, amacGFP, and avGFP as "wild-types" in the context of comparing them to mutant variants in their respective libraries, even though we acknowledge that this term is euphemistic in the case of the latter two genes.

Artificial cgreGFP-derived gene libraries

The cgreGFP-derived variants used to construct additional landscapes in the later sections of this work (see: 2.6. Novel cgreGFP-derived genes) were selected from among the successful artificial sequences generated by machine learning models trained on the cgreGFP data. The two top-performing sequences from each category (containing 12, 18, 24, 30, 36, 42, and 48 mutations) were selected, plus one of the top-performing 6-mutation genotypes (cgreGFP:1338), plus another 6-mutation genotype which was specifically generated for this part of the project (cgreGFP:132) to be a mid-way point between WT cgreGFP and one of the 12-mutant variants (cgreGFP:9708). Mutant libraries were subsequently generated separately for each of the following:

- cgreGFP:1338 (6 mutations)
- cgreGFP:132 (6 mutations)
- cgreGFP:9708 (12 mutations)
- cgreGFP:4111 (12 mutations)
- cgreGFP:minis, an equiproportional pool of the other 12 genes: cgreGFP:2880 and cgreGFP:3224 (18 mutations), cgreGFP:575 and cgreGFP:900x (24 mutations), cgreGFP:626 and cgreGFP:83 (30 mutations), cgreGFP:121 and cgreGFP:985 (36 mutations), cgreGFP:13 and cgreGFP:911 (42 mutations), cgreGFP:1414 and cgreGFP:567 (48 mutations)

4.2.2. Generation of mutant sequences

Randomly-mutated GFP sequences were generated via mutagenic PCR using Agilent's GeneMorph II Random Mutagenesis kit. This kit employs an error-prone polymerase which lacks proofreading capabilities. The PCR primers in this step (oligos 292 & 293, see: List of Oligos) included BpiI restriction sites for Golden Gate cloning of the PCR products into a storage vector; the reverse primer contained a degenerate 20N region which enabled the many resulting gene variants to be labeled with a unique barcode (**Figure 4**). Primers were ordered as PAGE-purified oligos, which is recommended for longer oligos in order to more efficiently remove aborted products and impurities.

The average number of mutations per molecule in a mutagenic PCR depends on the number of PCR cycles (with more cycles leading to more mutations) as well as on the initial quantity of DNA template (with more material leading to fewer mutations). A variety of different reaction conditions were tested. The average number of mutations was determined by cloning the PCR product into a plasmid (see: 4.2.3. Cloning of mutants into storage vectors), Sanger sequencing 20-25 clones and aligning the sequences with the wild-type template to count the mismatches.

We aimed for the GFP variants in each library to contain, on average, 4 mutations. Considering that around a third of nucleotide mutations are synonymous (see: any codon table), this corresponds on average to 1-2 amino acid substitutions per protein variant.

The official GeneMorph II protocol recommends the perplexingly high amount of 500-1000 nanograms of starting template material when aiming for 4.5 or fewer mutations per sequence. Please note that

"starting template" here refers only to the amount of DNA corresponding to the mutagenesis target, and does not include the rest of the plasmid. So, for an ~800 bp gene on a ~2500 bp plasmid as is our case, 500 ng of "starting template" would mean over 1500 ng of plasmid. The libraries in this work were all prepared using 75 ng of starting template (~240 ng of plasmid), because the PCRs with more than that consistently failed. Please see below for reaction details.

Reaction mix:

- ✓ 240 ng template plasmid*
- ✓ 5 µl 10X Mutazyme II reaction buffer
- ✓ 1 µl 40mM dNTP mix
- ✓ 1 µl each of 5µM forward and reverse primer
- ✓ 1 µl Mutazyme II polymerase
- ✓ H_2O up to 50 µl

* The template for the "minis" library consisted of an equimolar mix of twelve cgreGFP-derived template plasmids. For all other libraries, the template consisted of a single plasmid containing the appropriate wild type GFP sequence. Note: the use of a pool of different yet homologous sequences as the template in the "minis" reaction resulted in a subset of molecules consisting of hybridizations between different "parent" templates (see: 4.6.5. NovaSeq PE250 data clean-up, **Figure 40**).

Temperature cycling parameters:

1. 95°C, 2 min
2. (Variable number of) cycles:
   - 2.1. 95°C, 30 s
   - 2.2. 55°C, 30 s (the annealing temperature of 55°C was chosen based on gradient PCR tests with annealing temperatures ranging from 48°C to 60°C)
   - 2.3. 72°C, 60 s
3. 72°C, 10min

A note on amacGFP:V11L

We observed after the fact that around one third of genotypes in the final amacGFP library contained the V11L mutation. This was likely due to accidental contamination of the WT amacGFP DNA used as template during the mutagenic PCR, at some point during testing for the optimal mutagenesis conditions (early tests of sequencing ~20 clones to check mutation rate did not reveal an abundance of V11L genotypes, so the contamination must have occurred right before the final, optimized reaction). However, we decided to treat this as a good thing and an opportunity to study twin peaks in a single landscape: buy one, get one free.

A note on the number of mutagenic PCR cycles

In our experience, the GeneMorph II polymerase appears to lose activity over time, before even approaching its expiration date. amacGFP was the first library constructed and it was generated using 8 cycles of mutagenic PCR. Vexingly, after the amacGFP library was fully validated and it was time to proceed with cgreGFP and ppluGFP a few months later, 8 PCR cycles no longer yielded an average of ~4 mutations per clone, but closer to ~2. A few new tests showed that 16 PCR cycles was now the right amount. (The GeneMorph II kit had been stored correctly, at -20C, and was not expired.) Three years later, when preparing the new cgreGFP-derived libraries, the kit's mutagenic capabilities had mercifully not degraded very much more, and 20 PCR cycles were used on all the new libraries.

A note on certain degenerate primers

The reverse primer used in mutagenic PCRs contained twenty degenerate positions, meant to be occupied randomly by either A, T, C, or G. The primers used for the mutagenesis of amacGFP, cgreGFP, and ppluGFP were ordered as PAGE-purified oligos from Sigma (oligos 292 & 293, see: 4.12.2.List of Oligos). After sequencing the full libraries, we could observe that the twenty barcode positions were not, in fact, equally likely to be occupied by any of the four possible nucleotides. Cytosine and adenine were by far more

common than guanine or thymine, to varying degrees of extremity (Figure 36b).

For the new cgreGFP-derived libraries, we ordered new primers from Microsynth. In this case, we did not observe any extreme A/T/C/G bias (Figure 36c), although indels resulting in some barcodes being 19 or 21 nucleotides long seemed to be more common than in the Sigma primers, despite both being PAGE-purified. (We did not quantify the frequency of barcode indels, as there is no reason to expect them to interfere with experiments.)

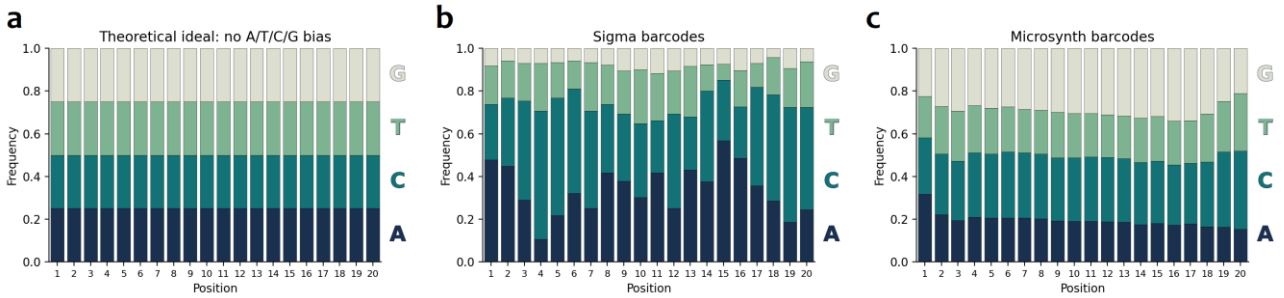

FIGURE 36. A/T/C/G bias from degenerate barcoding primers. (a) In the absence of bias, the four nucleotide types should be equally probably at any given position of the barcode. **(b)** The primary 20N barcodes from amacGFP, cgreGFP, and ppluGFP2 libraries were introduced with PAGE-purified oligos from Sigma containing a 20N degenerate region. However, cytosine and adenine were much more common than thymine or, especially, guanine, with the severity of the bias being dependent on the position in the sequence. **(c)** Novel cgreGFP-derived gene libraries were barcoded with identically-designed primers ordered from Microsynth, and showed much less nucleotide bias.

4.2.3. Cloning of mutants into storage vector

PCR products generated by mutagenic PCR were run on 1% agarose / 0.5% TAE gels, and the correct DNA band (~800bp) was excised and purified with whichever New England Biolabs, Thermo Fisher, or Blirt/DNA Gdansk (since assimilated by Qiagen) gel purification kit was within easiest reach.

Purified products were cloned into a storage (non-expression) vector using Golden Gate cloning with BpiI (see: 4.1.1. Golden Gate cloning). We used the plasmid pICH41258 a.k.a. "Level 0-SP" vector from the MoClo Toolkit (Addgene Kit #1000000044), which is high copy (pUC origin) and confers spectinomycin resistance, as a destination vector for this purpose.

10 µl of cloning reaction were used to transform 50 µl of chemically competent cells. Transformation protocols were followed according to the cell manufacturer's instructions, with the exception that post-heat-shock recovery time was kept to only ~20 minutes before plating (most protocols recommend 45-60 minutes). This protocol modification was done in order to minimize opportunity for cell division during recovery, which would interfere with colony-count-based quantification of the number of distinct clones (barcodes) obtained. Transformants were plated on LB agar plates containing 50 µg/ml spectinomycin and 40 µg/ml X-Gal, and incubated overnight at 37°C. Blue/white color screening allowed for an easy visual determination of cloning efficiency; the proportion of LacZ-positive colonies was typically between 0.5% and 5%.

For amacGFP, cgreGFP, and ppluGP2, Lucigen Ecloni 10G cells were used for transformation. A single 50 µl tube of these cells, transformed as described above, usually yielded above 100 thousand colonies. We harvested around 125k colonies for each of these three libraries. However, for the new cgreGFP-derived libraries in the second part of the project, Lucigen cells were not available so NEB 5-alpha High Efficiency chemically competent cells and ThermoFisher Library Efficiency DH5a competent cells were used instead. Sadly, both of these had a ~10-fold lower transformation efficiency than the Lucigen cells, so multiple vials had to be used per library. Only around 30k colonies (70k for cgreGFP:minis) were harvested for each of the new cgreGFP-derived libraries (see: 4.2.1. Gene selection), on the basis that this was a more manageable library size than the previously attempted 125k, where, after data filtering, we only recovered around 30k variants. (Though in the end, of course, we were not able to recover all 30k variants in the final data.)

4.2.4. Generation of destination vector

The destination vector for GFP mutant libraries consists of several parts (Figure 4):

✓ 600 bp 5' and 3' homology arms complementary to the *E. coli* chromosome, for use in downstream

genome integration; sequences taken from Bassalo et al. 2016.

- ✓ An mKate2-lacZ fusion protein under a constitutive promoter (T5) and lambda T0 terminator, as in the pQE30 vector where the avGFP library was previously expressed (Sarkisyan et al., 2016). The *lacZ* sequence is flanked by BsaI restriction sites in order to be easily replaced by GFP variants via Golden Gate cloning. Two restriction sites for BsmBI, another Type IIS enzyme, are located between *lacZ* and the terminator, oriented so as to cut away from each other. The mKate2 protein contains a 6H His-tag at the N-terminal end.
- ✓ A zeocin antibiotic resistance cassette under a constitutive promoter, to facilitate selection of cells with successful genome integration of libraries downstream.
- ✓ SpeI and NotI restriction sites, flanking the entire construct; these are used downstream to linearize the construct prior to genome integration.
- ✓ A plasmid backbone containing a high copy number origin and ampicillin resistance cassette, obtained by PCR on the CIDAR vector DVA_AH (Addgene #66043) with primers that removed unwanted BsaI sites (oligos 238 & 239, see: 4.12.2.List of Oligos).

The different parts were obtained by PCR where possible or ordered as synthetic dsDNA fragments, and assembled via Golden Gate cloning to form the final plasmid.

The BsmBI sites located after *lacZ* were then used to insert a library of 10N barcodes. Complimentary DNA oligos containing BsmBI sites and a degenerate 10N region were annealed by pooling them 1:1, heating them to 95°C, and slowly cooling to room temperature. The post-*lacZ* filler sequence was replaced by 10N barcodes in the destination vector via Golden Gate cloning, and the cloning mix was used to transform chemically competent cells. Around ten thousand colonies were recovered, pooled, and plasmid DNA was extracted using Thermo Fisher's GeneJET Maxiprep kit.

This library of ~10k plasmids, each labeled with a different 10N barcode (referred to as "secondary barcode" elsewhere in this work), was used as the destination vector for GFP libraries. In the final design, mKate2 and GFP are separated by a rigid alpha-helix linker (`GSLAEAAAKEAAAKEAAAKAAAARG`) to avoid potential interactions between them (Sarkisyan et al., 2016).

A note on antibiotic selection:

we chose the rarely-used zeocin antibiotic for this vector as it was sure to be compatible with all other cloning and genome integration steps in the project. The integration protocol results in cells which, at some point of their life stage, are already resistant to chloramphenicol (via the pSIM5 recombineering vector), kanamycin (via the pX2-Cas9 plasmid), and ampicillin (via the gRNA plasmid), so these antibiotics were all excluded as being inappropriate selection methods (see: 4.3. Genome integration). Furthermore, we also avoided spectinomycin as this antibiotic is used in various Golden Gate plasmids (Weber et al., 2011) including our chosen library storage vector and we wished to avoid any potential incompatibilities there as well.

4.2.5. Generation of final expression constructs

Mutant GFP libraries were shuttled from their non-expression storage vectors to the destination vector (containing secondary barcodes) via Golden Gate cloning, wherein the lacZ fragment is replaced by a barcoded GFP variant, in frame with mKate2, to form an mKate2-GFP fusion protein.

For each library, we harvested 3-5 times more colonies in this step than we had in the initial library generation step (over 500k colonies for amacGFP, ppluGFP2, and cgreGFP; 100k colonies for cgreGFP:1338, cgreGFP:132, cgreGFP:4111, and cgreGFP:9708; and 220k colonies for cgreGFP:minis). Therefore, each primary barcode can be expected to be associated with 3-5 secondary barcodes in the final expression library, on average.

4.3. Genome integration

We adapted protocols from Bassalo et al., 2016 and Sharam et al., 2009 to integrate our constructs into the *E. coli* genome via CRISPR-Cas9 mediated homologous recombination. This method combines the λ Red recombineering system with CRISPR targeting of the desired integration spot, and uses the following DNA:

- pSIM5 (Sharam et al., 2009), kindly provided by the Court lab, contains λ Red system genes under a heat-inducible promoter. This plasmid is chloramphenicol-selectable.
- pX2-Cas9 (Addgene #85811) (Bassalo et al., 2016), contains the Cas9 enzyme under an arabinose-inducible promoter. This plasmid is kanamycin-selectable.
- SS9_RNA (Addgene #71656) (Bassalo et al., 2016), contains the Cas9 guide RNA targeting the sequence "TCTGGCGCAGTTGATATGTA". This plasmid is ampicillin-selectable.
- Our mKate2-GFP library, linearized and flanked by homology arms to the target integration spot (see: 4.3.1. Preparation of DNA insert).

The λ Red system used here is based on the use of three genes derived from the λ bacteriophage (Sharam et al., 2009): *gam*, *bet*, and *exo*, encoding the proteins Gam, Beta and Exo respectively. The first of these, Gam, prevents linear DNA from being immediately degraded in the cytoplasm after transformation. The other two are responsible for inserting DNA at the target location: Beta facilitates the annealing of complementary ssDNA (such as the chromosome and our transformed DNA), and Exo is a 5' to 3' dsDNA exonuclease. The λ Red genes on pSIM5 are controlled by a heat-inducible promoter which activates expression from temperatures of 34°C and up. Cells transformed with pSIM5 must therefore be grown at lower temperatures, both to avoid unintentional recombination events between repetitive DNA regions, and because, perhaps more importantly, long-term *gam* expression is lethally toxic (Sharam et al., 2009).

The CRISPR/Cas9 guide RNA (SS9_RNA) targets an intergenic safe harbor where exogenous DNA can be integrated without disrupting native genes (Bassalo et al., 2016). After Cas9 induces a double stranded break, the DNA will be repaired via homologous recombination, using the λ Red machinery to integrate the GFP library construct at the site. Furthermore, the protospacer-adjacent motif (PAM) (a short sequence of nucleotides requires for Cas9 to cleave DNA) is mutated in the homology arms flanking the GFP construct, ensuring that the site cannot be cut again after integration.

4.3.1. Preparation of DNA insert

Linear constructs as in **Figure 4** were prepared by digesting the plasmid library with SpeI and NotI restriction enzymes (New England Biolabs) and gel purifying the relevant ~3.6 kbp band. Reactions were incubated at 37°C for one hour before being run on 1% agarose / 0.5% TAE gels at 120V, and DNA was extracted using gel purification kits from New England Biolabs (Monarch) or Thermo Fisher (GeneJET) according to the manufacturer's instructions.

Reaction mix (40 µl total volume):

- ✓ 32 µl (3-5 µg) plasmid DNA
- ✓ 4 µl CutSmart 10X Buffer
- ✓ 2 µl (40 U) SpeI-HF
- ✓ 2 µl (40 U) NotI-HF

The size of the construct containing mKate2-GFP, zeocin resistance cassette, and both homology arms is around 3.6 kbp. The full plasmid including the backbone is 5.3 kbp, but circular supercoiled DNA migrates faster than relaxed linear DNA on agarose gels (Lee at al., 2012). In this case, the uncut, likely supercoiled plasmid dares to migrate at a similar rate as the linear 3.6 band of interest. It is therefore crucial to run an uncut sample alongside the digest in order to monitor their migration, and only cut out the linear fragment when it has been sufficiently resolved (**Figure 37**). For 1% agarose gels running at 120V, at least one hour is required. Failure to exert rigor during this step results in the purified linear fragment being contaminated with plasmid DNA; in turn, this results in a high fraction of recombineering cells being transformed with plasmid instead of (or in addition to) the linear construct during the genome integration step. (While the vast majority of plasmid-carrying cells can be excluded during FACS thanks to their high fluorescent signal, a high proportion of such cells in the library implies a lower proportion of cells of interest, which is inefficient.)

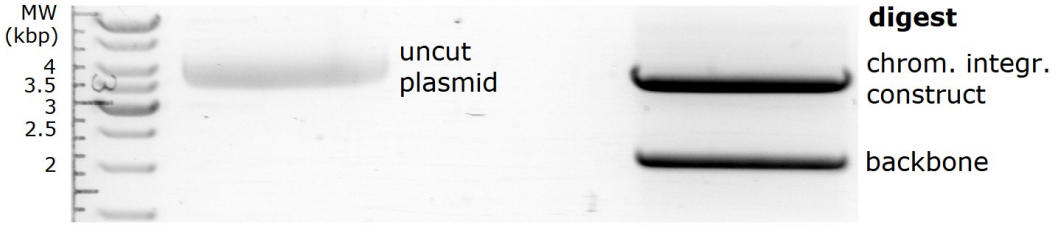

FIGURE 37. Agarose gel visualization of uncut vs. digested expression vector. After digestion with SpeI/NotI, the ~3.6 kbp linear DNA required for genome integration (containing homology arms, mKate2-GFP, and resistance cassette) migrates at a similar rate as the undigested plasmid. The image shows the gel after running for ~30 minutes at 120 V. At least one hour is required to reliably resolve the two bands before excision and purification, in order to avoid contaminating the linear construct with residual, uncut plasmid.

4.3.2. Preparation of recombineering cells

We used the *E. coli* strain BW29655 (CGSG #7934) (Zhou et al., 2003) for genome integration. Bassalo et al., 2009, whose method we adapted, employ the strain BW25113 (CGSC #7636), which is also the parent strain of the Keio Collection, a set of around 4000 single gene knockout strains (Baba et al., 2006). We did not have this strain on hand at the start of this work, but did have strain BW29655 in the lab. BW29655 is genotypically similar to BW25133 except for lacking the *rph-1* frameshift mutation and for having deleted the two-component system of EnvZ/OmpR porins (Zhou et al., 2003); the *Δ(envZ-ompR)520* deletion does not inhibit growth in standard culture conditions nor interfere with expression of fluorescent markers. As the most similar strain to BW25113 that we had immediately on hand, we decided to test the genome integration protocol with BW29655 right away. To our pleasant surprise, we were very quickly able to optimize a highly efficient genome integration protocol with our library constructs and this strain, so we simply kept using it and never switched to another strain.

Several steps are required in order to generate cells with recombineering capacity. We first generated electrocompetent BW29655 cells, which we co-transformed with pSIM5 and pX2-Cas9 plasmids and grew overnight at 30°C on LB-agar plates containing 25 µg/ml chloramphenicol and 50 µg/ml kanamycin. Resulting colonies were used to inoculate overnight liquid cultures in LB plus antibiotic. On the third day, recombineering cells were prepared according to the following protocol, adapted from Sharam et al., 2009:

1. ***Growth of cells to exponential phase.*** We diluted overnight cultures seventy-fold and incubated at 30°C with shaking. We used 500 µl of overnight culture per 35 ml of LB supplemented with chloramphenicol, kanamycin, and 0.2% arabinose. We grew cells in 50 ml aerated bioreactor tubes, and used multiple tubes rather than larger flasks when scaling up, as flask size and shape was observed to affect the efficiency of heat shock in the next step. Although Sharam et al., 2009 do not include arabinose in this step, we observed that its use here improved subsequent genome integration efficiency, compared to its use solely during the recover step. We also observed better efficiency when cells were grown to an OD_{600} of 0.8, rather than 0.6. This step takes around four hours.
2. ***Induction of λ Red system by heat shock.*** We exposed cultures to 42°C for 15 minutes to induce expression of λ Red genes. (We skipped this step when preparing non-induced cells for negative controls.) Although the literature and common sense both recommend using water baths for this step in order to optimize heat transfer, in our experience cells died *en masse* when using 42°C water baths, though we experimented with various lengths of heat shock time and in various water baths and air shakers. For best results, heat shock was performed in an Innova 44 air shaker.
3. ***Make electrocompetent cells.*** After heat shocking, cells were placed immediately on ice and subjected to several rounds of washing with ice-cold distilled water to render them electrocompetent (see: 4.1.3. Electrocompetent cells and electroporation). Each initial 35 ml culture yielded enough cells for four genomic integration reactions downstream; each reaction typically generated >200,000 genome-integrated colonies. Induced and non-induced electrocompetent cells were stored at -80°C until use.

4.3.3. Genome integration

Induced cells were transformed (see: 4.1.3. Electrocompetent cells and electroporation) with ~10 ng linearized

GFP library DNA and with ~50 ng plasmid encoding a guide RNA targeting a safe harbor in the *E. Coli* chromosome (Bassalo et al., 2016). Immediately after electroporation, cells were recovered in 1 ml LB/0.2% arabinose media at 30°C for two hours, then plated on LB agar plates supplemented with 50 µg/ml zeocin.

Plated cells were grown overnight at 30°C, then left at room temperature for one more day. We observed that this treatment improved the fluorescence intensity displayed by the cells, compared to only incubating for a single day at 30°C. (We speculate this may be due to some variants having an especially long maturation time.)

4.3.4. Sanity checks

Successful genome integration was confirmed by PCR using a primer pair complementary to the chromosomal sequences flanking the integration landing site (oligos 327 & 329, see: 4.12.2.List of Oligos) as well as a primer pair where one primer is located inside the insert and the other is chromosomal (oligos 321 & 329, see: 4.12.2.List of Oligos). Around two dozen colonies were screened, with chromosomal integration being confirmed in all of them. The absence of plasmid DNA in genome-integrated cells could also be confirmed, using primers specific to the plasmid backbone (oligos 356 & 359, see: 4.12.2.List of Oligos). Note: Aside from plasmid-carrying cells being generally undesirable, it is especially so for cells containing an integrated construct to also contain plasmid, as this could result in the same cell expressing different GFP variants at the same time. Care should be taken to ensure the linear dsDNA used for integration be as plasmid-free as possible (see: 4.3.1. Preparation of DNA insert.)

As a negative control, non-induced cells were also transformed, in the same manner as induced cells described above. As non-induced cells never have the chance to express λ Red genes, they are not capable of integrating DNA via homologous recombination. They are also not capable of maintaining linear DNA within the cytoplasm; to do so would require expression of λ Red *gam*, which in any case would not save them in the long term because a) the Gam protein is toxic, and b) the linear construct cannot be replicated and transmitted to future generations. Thus, any colonies arising from transformed non-induced cells must be the result of a) genomic integration of the linear construct via non-homologous end-joining, or b) contamination of the linear construct with undigested plasmid. The machinery for traditional non-homologous end-joining is absent in *E. coli*, although an alternative, rare mechanism has been reported in some strains (Chayot et al., 2010). Plasmid contamination is thus the only plausible culprit for the existence of colonies on negative control plates; indeed, this can be confirmed by the naked eye, as the difference in copy number variation between genome-integrated cells and plasmid-carrying cells is such that colonies of the latter appear visibly colored even under ambient light. PCR screening with plasmid-specific primers (oligos 356 & 359, see: 4.12.2.List of Oligos) also confirmed the presence of plasmid in all screened colonies growing on negative control plates. Therefore, transformation of non-induced cells was used in order to check that the purified linear dsDNA intended for genomic integration was not significantly contaminated with plasmid DNA.

4.3.5. Generation of wild-type and count controls

For each wild-type GFP gene, a pool of barcoded wild-type sequences was generated by PCR using a high-fidelity polymerase (Q5 or Encyclo) and barcoded primers (oligos 292 & 293, see: 4.12.2.List of Oligos). This pool was cloned into the mKate2 destination vector (see: 4.2.1. Generation of final expression constructs), and plasmid DNA was extracted from a few hundred pooled colonies and used for genome integration according to the protocol described above.

After integration, up to 6 colonies were selected from each gene and the GFP-barcode region was amplified and sent for Sanger sequencing to document the barcode and to check the integrity of the GFP coding sequence. Verified clones were stored as glycerol stocks. A limited number of clones which did not pass this check (due to having incorporated some mutation in the GFP region) were also stored as glycerol stocks and used as “count controls”.

4.4. Fluorescence-activated cell sorting

4.4.1. FACS sample preparation

Genome-integrated libraries were plated two days before FACS experiments (see: 4.3.3. Genome integration). In parallel, wild-type control cells and count control cells (see: 4.3.5. Generation of wild-type and count controls) were plated individually from glycerol stocks.

After overnight incubations at 30°C and room temperature, colonies were washed from plates (see: 4.1.4. Harvesting plated libraries), resuspended in 0.22 µm-filtered 1X M9 liquid medium and thoroughly mixed. Approximately 5 million colonies were recovered for each of the amacGFP, cgreGFP, and ppluGFP2 libraries, and around 2 million colonies for each of the novel cgreGFP-derived gene libraries. (In the case of FACS experiments repeated at OIST (Table 4), the libraries were plated from glycerol stocks made at ISTA, due to difficulties with genome integration efficiency in the new laboratory location. In all other cases, sorted libraries were the result of fresh genome integrations two days prior.)

Wild-type controls were also resuspended in M9 medium, and added to the library to a final controls:library ratio of approximately 0.5:100. The libraries of amacGFP, cgreGFP, and ppluGFP2 were all sorted alongside wild-type controls from all three genes as well as avGFP, each represented by 3-5 barcodes. The novel cgreGFP-derived gene libraries contained wild-type controls from 17 cgreGFP-derived genes, including the original WT cgreGFP, each represented by 2-6 barcodes.

Library mixes were generally diluted ~1:1000 in filtered M9 media before being allowed near the FACS machine; concentrations were then adjusted to achieve an event rate of ~15k-25k events/s at a flow rate of 1 (the minimum).

4.4.2. FACS setup

All FACS experiments were performed on various BD Aria III cell sorters, using a 70 micron nozzle. The amacGFP, cgreGFP, and ppluGFP2 libraries were sorted at the Institute of Molecular Pathology (IMP) in Vienna, together with IMP cytometry staff, with each library being simultaneously sorted on two machines in parallel. Novel cgreGFP-derived gene libraries were sorted at ISTA and/or OIST (Table 4).

Sort precision

In order to minimize data noise caused by cells being accidentally sorted into the wrong tubes, the Sort Precision Mode was always set to “Single Cell” for all FACS experiments. This is the highest possible precision setting, and implies that droplets will only be sorted if a) their leading and trailing droplets are empty (i.e. maximum Purity Mask setting), and b) the target cell is located near the middle of the droplet (i.e. Phase mask set to half the maximum). The former minimizes the chances of a sorted droplet merging with a non-empty adjacent droplet, causing it to be contaminated; the latter improves droplet trajectory accuracy (which can be affected by particles located at the edges of the droplet) and thus minimizes the chances of a sorted cell falling outside of the target tube.

Furthermore, the flow rate during sorting was always set to 1, the minimum value. (Possible values range from 1 to *11*, possibly as a nod to Spinal Tap.) A lower flow rate means a narrower stream, which increases resolution by making cells more likely to pass by the lasers in single file, as opposed to side by side.

Thresholding

The FSC (forward scatter) measures a particle’s size by collecting diffracted light along the same path as the laser beam, while the SSC (side scatter) measures internal complexity or granularity by collecting refracted and reflected light perpendicularly to the beam (Adan et al., 2017). Thresholds in the FSC and/or SSC channels are commonly used in order to filter out electronic noise during sorting; events smaller than the threshold will not be displayed or analyzed by the machine. We used a fluorescence-negative control bacteria to set the voltages for the FSC and SSC channels such that the bacterial cell population was easily identifiable, and set an SSC threshold of 1000, which in our case was comfortably below the minimum SSC values of the cell population. We observed no improvement in the overall sorting experience when using

an FSC threshold in addition to SSC, and an FSC threshold on its own was less effective than SSC.

Note: setting the threshold value too high will result in widespread contamination of all sorted gates with random cells: real events (cells) which fall below the threshold will not be displayed to the user, will be ignored entirely by the machine, and will therefore be free to contaminate sorted droplets without being filtered out according to the Sort Precision settings. Setting SSC voltage and threshold values intelligently is vital.

Fluorescence channels and compensation

We used two fluorescence channels: "FITC" for detection of GFPs (488 nm blue laser excitation, 530/30 nm band pass detection filter) and "PE-TexasRed" for detection of mKate2 (561 nm yellow laser excitation, 610/20 nm band pass detection) (Figure 5a).

The use of compensation controls is recommended when using two or more fluorescent markers in the same experiment, due to the possibility of different markers having overlapping spectra and therefore contaminating each other's signals. Despite the joint use of two fluorescent markers in our setup (mKate2 and GFP), compensation was not performed during the sorting of amacGFP, cgreGFP, and ppluGFP2 libraries at VBCF because it was deemed unnecessary due to the known lack of overlap between mKate2 and GFP spectra in the wavelengths used during FACS. Nevertheless and in the interest of procedural rigor, for all FACS runs of novel cgreGFP-derived genes, GFP-only and mKate2-only cells were used as controls to calculate compensation values according to the standard BD Aria III protocol.

Gate selection

Cell doublets can be detected by looking at the relationship between the measured height (H) and total area (A) in either the FSC or SSC channels. For single cells which are normally all the same shape, these values should be proportional, so plotting H as a function of A is a simple diagonal and events deviating from this ratio likely represent multiple cells stuck together. We looked at SSC-H versus SSC-A and defined a polygon gate encompassing the diagonal. This population (P1) was our population of single bacterial cells.

Within P1, we then selected an interval gate in the mKate2 (red) channel centered roughly around the peak of the mKate2 fluorescence distribution and encompassing ~10% of the total library cell population (Figure 5c). (The narrower this gate, the better our control for mKate2-GFP protein expression levels, but too narrow a gate would exclude too many cells and slow the experiment to the point of unfeasibility.) Only cells falling within this red gate (P2) were sorted; all others were discarded.

The selected red P2 gate was subdivided into eight interval gates (P3-P10) according to fluorescence intensity in the green channel. The darkest of these, P3, was defined based on the distribution of GFP-negative control cells and meant to capture cells with totally non-functional GFP mutants. Gates P4-P10 were spaced unevenly across the rest of the green value range with the intent of maximizing fluorescence resolution while taking into account cell population density (Figure 5c). This setup of gates P3-P10 was in contrast to the previous setup for avGFP, which defined gates in the green channel at equal intervals on the logarithmic scale (Sarkisyan et al., 2016).

Sorting rounds

Libraries were sorted at room temperature in multiple rounds of one hour each. Four green gates were sorted simultaneously at any given time (p3-P6, or P7-P10), with cells being recovered in room-temperature SOC or LB medium. After one hour of sorting, a fixed number of count control cells were sorted into each recovery tube (5000 per count control barcode in the case of amacGFP, cgreGFP, and ppluGFP2; 2000 or 1000 in the case of novel cgreGFP-derived libraries sorted at ISTA or OIST respectively).

The full volume of recovered cells from a given round was plated onto multiple large square LB/zeocin agar plates (up to 300k sorted cells per plate) and incubated overnight at 30°C then overnight again at room temperature. Colonies were then imaged with a Canon EOS 600D SLR camera under blue light to check the overall green brightness of different gate outputs, as a sanity check to ensure sorting was successful. Colonies derived from different FACS gates were stored as glycerol stocks and used to extract DNA for barcode sequencing (see: 4.5. Library sequencing: Barcodes).

Note: six separate FACS experiments were discarded entirely, and individual rounds were discarded from three other runs, due to heavy cross contamination caused by cells falling into the wrong gates during sorting. The four sorting streams were sometimes observed to fluctuate during sorting or even

dissolve into an untargeted spray (particularly on the ISTA machine), resulting in all recovery tubes being contaminated with random library cells.

Table 4. Controls and quantities of sorted GFP libraries. OIST-sorted libraries were prepared by growing cells from the glycerol stocks of unsorted aliquots of the corresponding libraries previously sorted at ISTA. In the case of the "minis" library, the initial ISTA sorting was unsuccessful (large amount of missorted cells apparent by microscope observation of colonies derived from sorted cells), but its unsorted glycerol stock was used for both OIST runs.

Library	Location	Sample	WT controls	Total number of sorted cells	Events/s
amacGFP	IMP	~5M colonies	avGFP, amacGFP, cgreGFP, ppluGFP2	~11,114,000	25k
amacGFP	IMP			~16,740,500	25k – 30k
cgreGFP	IMP	~5M colonies		~11,866,000	25k
cgreGFP	IMP			~16,390,000	25k – 30k
ppluGFP2	IMP	~5M colonies		~16,246,000	25k
ppluGFP2	IMP			~17,649,000	25k – 30k
cgreGFP:1338	ISTA	~1.3M colonies	cgreGFP and cgreGFP-derived genes ("1338", "132", "4111", "9708", "2880", "3224", "900x", "626", "575", "121", "83", "13", "911", "985", "567", "1414")	5,142,365	15k – 20k
cgreGFP:1338	OIST	~2M colonies (from ISTA glycerol stock)		7,557,122	15k – 20k
cgreGFP:132	ISTA	~1.5M colonies		6,385,511	15k – 20k
cgreGFP:132	OIST	~2M colonies (from ISTA glycerol stock)		14,580,691	15k – 20k
cgreGFP:9708	ISTA	~2M colonies		8,865,010	15k – 20k
cgreGFP:9708	OIST	~2M colonies (from ISTA glycerol stock)		11,020,876	15k – 20k
cgreGFP:4111	ISTA	~1.8M colonies		5,265,272	15k – 20k
cgreGFP:4111	ISTA	~2M colonies			
cgreGFP:minis	OIST	~2M colonies (from ISTA glycerol stock)		2,923,905	10k – 15k
cgreGFP:minis	OIST	~2M colonies (from ISTA glycerol stock)		6,395,383	15k – 20k

4.5. Library sequencing: barcodes

4.5.1. Barcodes sample preparation

Sorted cells from different gates (see: 4.4.2. FACS setup) were washed from plates and collected separately (see: 4.1.4. Harvesting plated libraries). Barcode regions were extracted from all samples by PCR and Illumina adapter sequences were added, in the following steps:

1. ***Extraction of genomic DNA (optional).*** For amacGFP, cgreGFP, and ppluGFP2, no gDNA was extracted, and the first PCR step was performed directly on cells. For novel cgreGFP genes, gDNA was extracted from an aliquot of cells (~0.05 g) from each gate from each sorting experiment, following the "Gram Negative Bacteria" protocol from Promega's Wizard gDNA Purification Kit. Extracted gDNA was checked by NanoDrop for purity and run on a 1% agarose gel to check for DNA integrity. In practice, we did not observe any improvement in PCR output when using gDNA instead of cells, making this step optional (or pointless).
2. ***First PCR: amplification of barcodes.*** In this step, primers containing partial NG adapter sequences were used (olgios 373-375 and 387, see: 4.12.2.List of Oligos). The forward primers (a pool of N-shifted oligos, to increase sequence complexity for NGS) were designed to anneal directly upstream of the primary barcode, while the reverse was located in the zeocin resistance cassette in order to generate a ~350 bp fragment of ideal size for Illumina sequencing. In the case of amacGFP, cgreGFP, and ppluGFP2, 15 PCR cycles with Encyclo polymerase were used; for novel cgreGFP-derived genes, it was 18 PCR cycles with Q5 (an even higher fidelity polymerase than Encyclo, but seemingly less processive). In all cases, PCR products were gel purified and eluted in 10 µl H_2O.
3. ***Second PCR: addition of NGS adapters.*** 1 µl of purified PCR from the previous step was used as the template for a second PCR, using primers corresponding to NGS adapters. For amacGFP, cgreGFP,

and ppluGFP2, TruSeq Illumina adapters were used (oligos 366 & 379-86/427-30, see: 4.12.2.List of Oligos) and the PCR consisted of 9 cycles with Encyclo polymerase. For novel cgreGFP-derived genes, dual-index primer pairs provided by VBCF were used (oligos DI5 & DI7, see: 4.12.2.List of Oligos) and the PCR consisted of 14 cycles with Q5 polymerase. PCR products were gel purified and submitted for NGS sequencing.

4. ***Sequencing***. As the full region of interest (primary and secondary barcodes, plus technical sequences) was around 65 bp, the above samples were sent for single-end, 100 bp-length sequencing (HiSeqV4 SR100 or NovaSeq SP/S1 SR100, as available). Between 10-15% PhiX DNA was added per run to increase sample sequence complexity.

4.5.2. Illumina SR100 data processing

Sequencing data files were converted from .bam to .fastq format using Bamtools as needed (see: 4.6.2. MiSeq PE300 data processing). Overall read quality was checked by FastQC. As a general rule, barcode data from amacGFP/cgreGFP/ppluGFP2 and from the novel cgreGFP-derived genes was processed in the same way, except that the efficiency of scripts was improved for the latter. The following data processing steps were performed using custom Python scripts:

1. ***Processing of individual samples (cells from one gate)***.
   - 1.1. ***Identification of barcode-adjacent constant sequence***. Reads were expected to consist of a 20N primary barcode and 10N secondary barcode, separated by an invariant "AGGTGCTAG" sequence, plus flanking technical sequences. Reads found not to contain the constant between-barcodes sequence were discarded.
   - 1.2. ***Barcode extraction***. The secondary barcode was assigned the 10 bp sequence immediately after the mentioned invariant region. The primary barcode was assigned to the 20 bp sequence immediately preceding it (for amacGFP, cgreGFP, and ppluGFP2 libraries), or to the 20 bp sequence immediately following the forward primer-annealing region (for novel cgreGFP-derived genes, due to the observation that these contained more frequent indels in the barcode region, such that counting 20 bp from the end instead of the beginning would cause a reading shift compared to barcodes extracted in the full-length gene sequencing).
   - 1.3. ***Barcode quantification***. The total number of reads corresponding to a given primary/secondary barcode combination in the sample was determined.
2. ***Unification of gate samples from a single FACS run***.
   - 2.1. ***Determination of barcode distribution across gates***. The data from all eight green gates from the same FACS experiment were combined. Each primary/secondary barcode combination was assigned an array of eight numbers corresponding to its total read counts across the different FACS gates.
   - 2.2. ***Correction of sequencing errors: primary barcodes***. Similar primary barcodes were merged together if a) their sequences differed by only one nucleotide, and b) they shared the same secondary barcodes (or subset of). Given the improbability of meeting both criteria, the less abundant primary was assumed to be a sequencing error of the more abundant one. For each secondary barcode of the less abundant primary, the read counts were added to the corresponding counts of the more abundant primary. (This step was not performed for amacGFP, cgreGFP, and ppluGFP2 libraries due to concerns about lower barcode diversity caused by C/A bias in the primers; see: 4.2.2. Generation of mutant sequences.)
   - 2.3. ***Correction of sequencing errors: secondary barcodes***. Following the same reasoning as above, two secondary barcodes of the same primary were merged together if a) their sequences differed by only one nucleotide, and b) the less abundant secondary had a ten-fold lower read count than the more abundant one.
3. ***Read count normalization***.
   - 3.1. ***Determination of the number of reads originating from a single cell***. Between 3 and 5 count control barcodes were added in fixed amounts to each gate (see: 4.4.2. FACS setup). For each gate, the median read count per cell was determined by dividing the read counts of each count control barcode by the corresponding number of count control cells, and taking the median.

3.2. ***Normalization of all read counts***. For each gate, the read counts of each primary/secondary barcode combination were normalized by dividing by the median read count per cell determined above. This approximation of cell counts were termed "pseudo-cell counts" going forward.

4. ***Association of barcodes to genotypes***. Primary barcode sequences were used to assign nucleotide and protein genotypes, referencing full-gene sequencing of the appropriate libraries (see: 4.6. Library sequencing: Coding regions). Barcodes with missing genotype data were discarded.
5. ***Determination of fitness values and downstream processing***. See: 4.7.1. Determination of fitness values, 4.7.2. Combining data from multiple experiments, and 4.7.4. Library data filtering.

4.6. Library sequencing: coding regions

Full-length coding sequences and their associated barcodes were sequenced by high-throughput Illumina NGS. The longest possible read lengths produced by Illumina platforms are paired-end reads of 250 or 300 bp, allowing for full coverage of sequences up to around 500 bp. However, as the full length of barcoded GFP gene sequences is closer to 800 bp, we sequenced the N-terminal and C-terminal sections separately. We used different approaches for the first and second parts of the project.

4.6.1. Sample preparation: amacGFP, cgreGFP, ppluGFP2

The mutant libraries of these three genes were first circularized, which allowed the N- and C-terminal halves to be amplified separately while still incorporating the barcode in region in each case (**Figure 4**). Illumina "TruSeq" adapters were then added, and the halves were sequenced with Illumina's MiSeq PE300 platform:

1. ***Circularization***.
   1.1. ***Excision of gene library from storage vector***. Mutant libraries in storage vectors (see: 4.2.3. Cloning of mutants into storage vectors) were digested with BsaI, and the GFP fragment (~750 bp) was isolated by agarose gel purification.
   1.2. ***Preparation of oligo bridge***. Complementary primers (oligos 388 & 389, see: 4.12.2.List of Oligos) were annealed to each other by pooling them 1:1, then heating at 95^{0}C and cooling to room temperature in a thermocycler at a rate of ~0.1^{o}C per second. These oligos contained BsaI restriction sites designed to leave compatible overhangs with those of the GFP fragment from the previous step, flanking a short filler sequence.
   1.3. ***Circularization***. The annealed dsDNA oligos and the linear GFP fragment were ligated together to form a circular molecule via a modified Golden Gate reaction. Initial reactions contained 100 U BsaI, 60 U T4 ligase, and 50 ng each of GFP fragment and oligo filler, in a final volume of 500 µl in 1X T4 ligase buffer. The reaction was performed at room temperature. Every 30 minutes, another 50 ng (~1 µl) of each DNA partner was added to the mix, up to a combined total of 2 µg DNA. The large reaction volume and the gradual and limited DNA addition helped minimize the concentration of free, unligated DNA (once ligated, circular molecules could not be cut again due to the loss of restriction sites); we observed during tests that the higher the DNA concentration, the more prevalent was the formation of circular multimers (**Figure 38a,b**). Naturally, multimers should be religiously avoided, as using them as templates in downstream steps would cause coding regions to be associated to the wrong barcodes.
   1.4. ***Purification of circular monomers***. The full reaction volume was processed through a column-based DNA clean-up kit in order to concentrate the sample into a volume that would fit into an agarose gel well. Whole samples were then run on 1% agarose/0.5% TAE gels and the band corresponding to the circular monomer was excised and purified. Circular molecules were observed to migrate faster than their linear counterparts on a gel (**Figure 38b**). Successful circularization was further confirmed by PCR using primers that would fail to amplify linear fragments (**Figure 38a**).

2. ***First PCR: amplification of gene halves.*** Barcoded N-terminal halves of GFP were amplified using the circularized libraries as template (oligos 373-5 & 376-8/421-3/431-3, see: 4.12.2.List of Oligos) (**Figure 4**). As the barcode is originally located after the stop codons, C-terminal halves were amplified directly from the libraries in the original storage plasmid (oligos 370-2 & 367-9/424-6/434-6, see: 4.12.2.List of Oligos). In both cases, 10 PCR cycles with Encyclo polymerase were employed. Primers included a template-specific region as well as a partial NGS adapter sequence. Furthermore, for each gene half, three separate, staggered primer pairs were employed which differed from each other only by being shifted by 1-3 nucleotides; this increase in sequence complexity is desirable for NGS (see note below). PCR products were run on a 1% agarose gel and gel purified, eluting in 10 µl H_2O.
3. ***Second PCR: addition of NGS adapters.*** 1 µl of purified PCR products from the previous step were used as the template DNA in a second PCR (also Encyclo, 10 cycles) (oligos 366 & 379-86/427-30, see: 4.12.2.List of Oligos). The primers in this case corresponded to the full-length TruSeq Universal Illumina adapters, initially annealing to the partial adapter sequence already incorporated during the previous step. The forward primer was the same for all samples; a different reverse primer, differing by a unique 6N index sequence, was used for each sample in order to be able to pool samples together in the same NGS run and demultiplex them afterwards.
4. ***Sequencing.*** All samples described above were sequenced at VBCF on the Illumina MiSeq platform, generating paired-end reads of length 300 bp each. A total of 5 MiSeq lanes were used, yielding ~100 million reads altogether. As amplicon libraries have low sequence complexity even when using N-shifted primers, ~20% of PhiX DNA was included per run.

A note on N-shifted PCR primers:

Introducing reading shifts is recommended for amplicon libraries or other samples where nucleotide diversity per position is expected to be low. This is the case for our mutant libraries, as only a minority of variants will contain a point mutation any given site. During NGS, starter molecules are amplified into clusters, and each cycle, clusters are imaged and their nucleotide states are determined by their fluorescent signal. If nucleotide diversity is low, resolving individual adjacent clusters becomes more difficult, as the majority signal may mask that from alternative nucleotides. This results in miscalled bases and sequencing errors. Using staggered primers, also called phased primers, increases the nucleotide diversity at any given site, minimizing this problem.

Sample sequence complexity may also be increased by the addition of PhiX DNA (typically 10-20%) prior to sequencing.

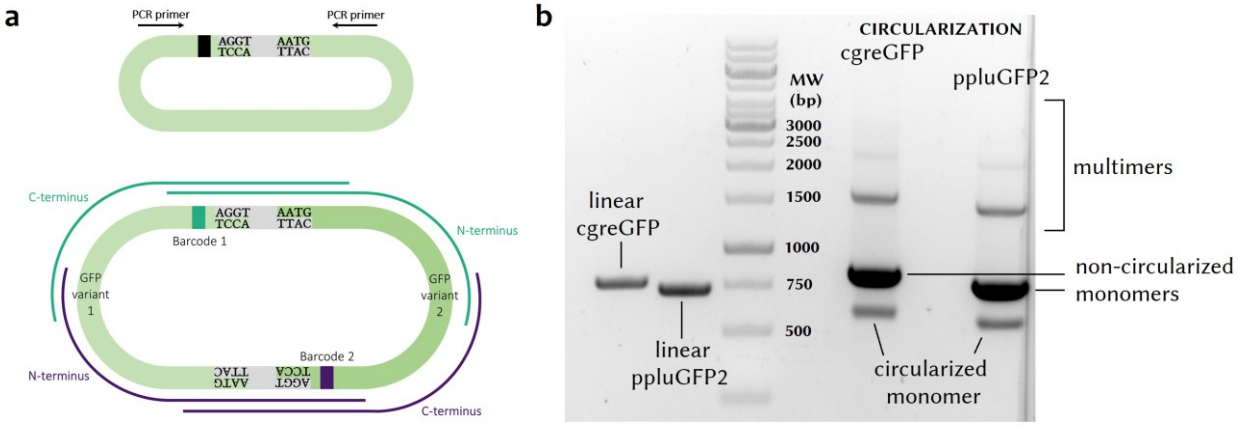

FIGURE 38. GFP library circularization. (a) Schematic of the desired circular molecule (top) and of an undesired multimer (dimer, bottom), showing the circulatizing oligo filler in grey and GFP sequence in green. Placement of PCR primers used to confirm circularization are shown on the monomer. In the dimer, PCR of the N-terminal half would lead to the amplified coding region being associated with the incorrect barcode (represented by green vs. purple). **(b)** Agarose gel visualization of pre- and post-circularization products. Purified, linear dsDNA of the cgreGFP and slightly shorter ppluGFP gene libraries are shown on the left. On the right, the circularization products are shown (see: 4.6.1. Sample preparation: amacGFP, cgreGFP, ppluGFP2). Circular monomers migrate faster than linear molecules of the same size. Multimers are also visible.

4.6.2. MiSeq PE300 data processing

Demultiplexed sequence data were downloaded from VBCF in .bam format and converted to .fastq using Bamtools:

```
bamtools convert -in sample.bam -format fastq -out sample.fastq
```

FastQC files provided by VBCF indicated overall good sequencing quality, with quality decreasing towards the ends of the reads, particularly Read 2; this is standard for MiSeq. No quality-based read trimming was performed.

Raw sequencing reads were processed using custom Python scripts. Code is available on the `Orthologous_GFP_Fitness_Peaks` GitHub page. The data was processed in the following steps:

1. *Processing of individual samples (gene halves)*:
   - 1.1. *Identification of barcode-adjacent constant sequence*. All constructs included an invariable 22 bp sequence between the stop codon and the barcode. This was used as a primer-annealing region during mutagenic PCR and during post-FACS barcode amplification. Any sequencing reads found not to contain this sequence were discarded.
   - 1.2. *Barcode extraction*. Barcodes corresponded to the 20 nucleotides immediately adjacent to the above invariable sequence.
   - 1.3. *Trimming of primer sequences*. Sequences corresponding to primer-annealing regions used during sample preparation were discarded, as leaving them in would interfere with the discovery of real point mutations in those regions.
   - 1.4. *Pooling of reads with matching barcodes*. Reads with the same barcode were pooled, and barcodes with fewer than 5 reads to their name were discarded as having too low coverage. (Note: due to concerns about reduced barcode diversity caused by nucleotide bias in the barcoding primers (see: 4.2.2. Generation of mutant sequences), no attempts were made to correct for sequencing errors in the barcode region during this step.)
   - 1.5. *Making of consensus sequences*. Aligned reads were merged into a consensus sequence by taking the most abundant nucleotide at each position. (Note: by default, sequences are already aligned after primer trimming, so no external sequence aligner was used in this case.) This process was done separately for forward (Read 1) and reverse (Read 2) sequences. Sequences with ambiguous positions, i.e. less than 80% agreement among all reads for all positions, were discarded. This high threshold improved data quality downstream compared to our initial simple threshold of 50%.
   - 1.6. *Merging of Read 1 and Read 2 consensus*. For each barcode, the Read 2 consensus sequence was reverse-complemented and the forward and reverse consensus sequences were merged together. The overlap between Read 1 and Read 2 varied between samples but was always over 100 bp long. If the overlap region was not a 100% match between both consensus sequences, the barcode was discarded.
2. *Merging of N- and C-terminal gene halves*: For each surviving barcode, the N-terminal and C-terminal consensus sequences were merged. The overlap between gene halves was only 6 bp long for amacGFP, but 71 bp and 53 bp for cgreGFP and ppluGFP2, respectively. Barcodes where the overlap was not a 100% match between gene halves were discarded.
3. *Genotype determination*:
   - 3.1. *Nucleotide mutation extraction*. Global pairwise alignments were made between every full-length consensus sequence and the relevant wild-type reference, using Biopython (Cock et al., 2009). Nucleotide genotypes were determined by extracting the mutations (mismatches with the reference).
   - 3.2. *Protein translation*. Nucleotide coding sequences were translated to protein sequences and amino acid mutations were extracted by comparing with the wild-type reference.

4.6.3. Sample preparation: novel cgreGFP variants

For these genes, circularization was not performed. Instead, three PCRs were performed on each library such that the different-length products spanned the entire gene length, with the barcode always being included (**Figure 39**). Primers were designed to anneal to regions which were identical in all 16 of the novel cgreGFP-derived genes, in order to avoid any bias resulting from some variants being amplified

more efficiently than others.

Aside from the lack of circularization, the sample preparation steps were essentially the same as for the previous libraries (i.e., NGS adapters added via two PCR steps; see: 4.6.1. Sample preparation: amacGFP, cgreGFP, ppluGFP2), with the following modifications:

- ***Primer sequences***. For the first PCR, we used N-shifted oligos binding to different regions of the gene sequence (**Figure 39**). For the second PCR, we used premixed dual-index Illumina adapters provided by VBCF (DI5 & DI7, see: 4.12.2. List of Oligos).
- ***PCR conditions***. We used Q5 polymerase, 10 cycles for the first PCR and 13 cycles for the second.
- ***NGS run***. Samples were sequenced by Microsynth on a single NovaSeq SP PE250 flowcell, instead of using several MiSeq PE300 lanes.

4.6.4. NovaSeq PE250 data processing

Demultiplexed sequence data were downloaded from Microsynth in .fastq format.

Processing of raw read data was performed using custom Python scripts adapted from those used for MiSeq PE300 data (see: 4.6.2. MiSeq PE300 data processing):

1. ***Processing of individual samples (gene fragments)***:
   - 1.1. ***Identification of barcode-adjacent constant sequence***. As in 4.6.2. MiSeq PE300 data processing.
   - 1.2. ***Barcode extraction***. As in 4.6.2. MiSeq PE300 data processing.
   - 1.3. ***Trimming of primer sequences***. As in 4.6.2. MiSeq PE300 data processing.
   - 1.4. ***Pooling of reads with matching barcodes and correction of barcode sequencing errors***. Reads with the same barcode were pooled, taking into account potential sequencing errors. Error rates for Illumina NGS have been reported to range from 0.1% to 0.6% per base (Stoler & Nekrutenko, 2021), depending on the platform and on the individual run. If we pessimistically assume an error rate of 0.6%, then the probability of the 20N barcode being sequenced perfectly in any given read is $(1 - 0.006)^{20} = 0.89$, meaning that up to 11% of reads may contain miscalled bases in the barcode region. We therefore considered two barcodes to be equivalent, if a) they differed by only one base, and b) the less abundant barcode was associated to fewer than 10 reads.
   - 1.5. ***Alignment of reads***. Reads belonging to the same barcode were aligned using the multiple sequence alignment tool MUSCLE v3.8.1551 (Edgar, 2004), alongside the corresponding WT sequence for reference. This was done separately for the forward (Read 1) reads from each of the three fragments (**Figure 39**) and for the pooled reverse (Read 2) reads from all fragments.
   - 1.6. ***Making of consensus sequences***. From each multiple read alignment, a consensus sequence was generated by taking the most abundant nucleotide at each position. Any given position was considered ambiguous if less than 75% of reads were in agreement about nucleotide identity. Initially, nearly all consensus sequences contained ambiguous positions; see 4.6.5. NovaSeq PE250 data clean-up for details on this problem and how it was fixed. Any barcodes whose final, post-clean-up consensus sequences still contained ambiguities were discarded.
2. ***Merging of gene fragments***. For each surviving barcode, the four consensus sequences were merged together by comparing the their overlapping regions (**Figure 39**). Barcodes were discarded if the overlap regions did not match 100%.
3. ***Genotype determination***:
   - 3.1. ***Nucleotide mutation extraction***. As in 4.6.2. MiSeq PE300 data processing.
   - 3.2. ***Protein translation***. As in 4.6.2. MiSeq PE300 data processing.

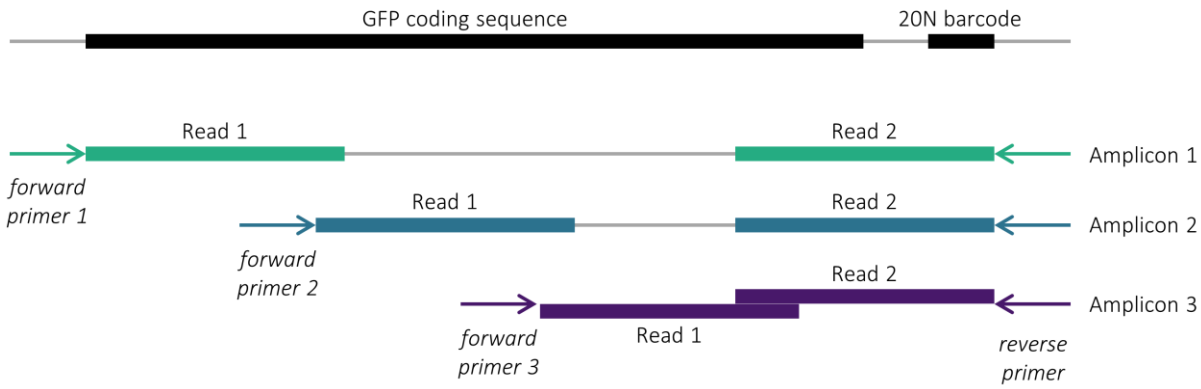

FIGURE 39. PCR setup for NovaSeq PE250. For NGS of new cgreGFP-derived gene libraries, three fragments of different lengths were amplified, with overlapping sections between them to allow for downstream sequence merging. The barcode was included in the Read2 of all fragments.

4.6.5. NovaSeq PE250 data clean-up

When combining reads to create consensus sequences and/or when merging consensus sequences from different amplicons (see: 4.6.6. NovaSeq PE250 data processing), the majority of coding sequences failed to assemble without ambiguities. Upon inspection, in nearly all cases, ambiguous positions were due to conflict between the WT state and a mutated nucleotide, with neither state achieving a 75% majority (our threshold for consensus building). That is, the same barcode was associated to different sequences at a rate much higher than could ever be expected from random one-off sequencing errors.

Note: considered individually, the conflict at any given position under dispute is of course always between the WT and a non-WT nucleotide, and not between two non-WTs (discounting rare sequencing errors). For example: if the same barcode is associated to two sequences with mutations at different sites, *nAnnn* and *nnnBn*, where n represents the WT state, the conflicts are necessarily *n/A* and *B/n*. However, we did not observe such cases in our data: taking into account the full read sequences, all conflicts were between 100% WT sequences and sequences with one or more mutations. We will hereafter refer to a read's joint nucleotide states across all conflicting positions as its "haplotype". So, for a barcode associated to a consensus sequence with ambiguities in positions X, Y, and Z, we observed reads with fully WT haplotypes (WT states at sites X, Y, and Z) and reads with fully mutant haplotypes (mutations at all sites X, Y, and Z). We did not observe ambiguities caused by conflicts between three or four nucleotide states.

In order to determine which sequence was the correct one, we performed long-read sequencing of full-length mKate2-GFP-barcode sequences from all cgreGFP-derived libraries, and compared the data to the Illumina reads. Long-read sequencing was done in-house on a MinION Mk1C device (Nanopore Technologies), using one R9.4.1 flowcell. Nanopore reads were processed as follows, using custom Python scripts:

1. ***Read filtering***. 20N primary barcode sequences were extracted from all Nanopore reads by scanning for the constant barcode-adjacent motif (see: 4.6.2. MiSeq PE300 data processing). Reads without this motif, or which were too short to span the full GFP sequence, were discarded.
2. ***Generation of consensus sequences***. Reads with the same primary barcode were grouped and aligned with the appropriate reference sequence (i.e. according to which library the barcode belonged to, which was known from the Illumina data; with 4^{20} theoretically possible barcodes, there was no barcode overlap between libraries). Consensus sequences were created for barcodes with at least 3 reads, by assigning the simple majority nucleotide at each aligned position. Barcodes with fewer than 3 reads were discarded. (Note: average sequence quality was overall low (Phred scores between 7 and 25) across the full read length, and indels were rampant, with most reads exhibiting multiple short deletions compared to reference GFPs. However, these deletions appeared randomly spaced, such that they were canceled out in alignments of multiple reads.)
3. ***Extraction of mutations***. Ambiguous positions in consensus sequences of Illumina reads were compared with their corresponding nucleotide states in the Nanopore consensus reads.

The Nanopore data confirmed that in virtually all cases where a conflict existed between a WT haplotype and a mutant haplotype, the mutant one represented the true genotype. Over 10k barcodes

from each library were captured by long-read sequencing:

- cgreGFP:1338: 14,189 barcodes, 99.3% in support of the fully mutant haplotype.
- cgreGFP:132: 12,783 barcodes, 99.5% in support of the fully mutant haplotype.
- cgreGFP:9708: 20,857 barcodes, 99.3% in support of the fully mutant haplotype.
- cgreGFP:4111: 23,010 barcodes, 99.4% in support of the fully mutant haplotype.

Based on this, we returned to the Illumina data for libraries "1338", "132", "9708", and "4111", and resolved all conflicts by assigning the mutant haplotype and discarding reads which supported the WT state. We then proceeded with merging gene fragments as described in 4.6.6. NovaSeq PE250 data processing.

Source of WT contamination

The abundance of false WT sequences in NGS data was likely caused by insufficient sequence diversity, despite the addition of PhiX DNA and the use of staggered primers during sample amplification. The majority of mutant variants have only a handful of mutated positions spread across the full ~700 bp gene length. Thus, for any given position, the overwhelming majority of sequences will contain the WT state. The genuine signal from a mutated position in one cluster may then be lost among the WT calls of surrounding clusters. NGS flowcells may use 4-channel chemistry (where A, T, C, and G are each labeled a different color) or 2-channel chemistry (only two fluorophores are used, and nucleotides are labeled with one or the other, or both, or neither) for base calling. It is possible that this problem may occur more readily in NovaSeq runs, which use 2-channel chemistry, than in the previously employed MiSeq runs, which use 4-channel chemistry [personal communication with sequencing facility staff].

Detection of hybrid sequences in the "minis" library

In the case of the cgreGFP:minis library, derived from a mix of 12 templates (see: 4.2.2. Generation of mutant sequences), data processing was more complex. Before extracting mutations and determining whether there is a conflict between WT and non-WT reads, we needed to first determine which of the 12 "WTs" is the correct one for each barcode. (When using a generic "cgreGFP consensus" sequence as the WT reference, containing the majority amino acid states from among all cgreGFP "WTs", only 39.9% of the 33,095 Nanopore consensus sequences supported the "fully mutant" haplotype.)

Out of the full 705 bp coding sequence, 195 sites are variable across the different cgreGFP "WT" references (called "origins" or "parents" hereafter). For each barcode, we looked at the nucleotide states of these sites to determine the most likely origin. This was done independently for each amplicon fragment. We observed that in a significant fraction of cases (~30%), different fragments appeared to originate from different origins. To investigate this, we looked at the Nanopore reads covering the full gene length. These confirmed the existence of apparently hybrid sequences consisting of segments of different origins (see **Figure 40** for an example). Once identified, we allowed these hybrid genotypes to remain in the dataset (see: 2.6.4. Hybrid gene variants tend to maintain function).

These hybrid or chimeric sequences were likely an artifact of the simultaneous use of 12 templates in a single PCR reaction during the mutagenesis step. The 12 different origins are all cgreGFP variants and therefore share high sequence identity, and thus may complementarily anneal to one another. Any molecule prematurely aborted during PCR extension may serve as a primer during the next cycle and bind to a template molecule from a different origin. This results in a subset of PCR products consisting of single molecule chimeras originating from multiple templates (Haas et al., 2011).

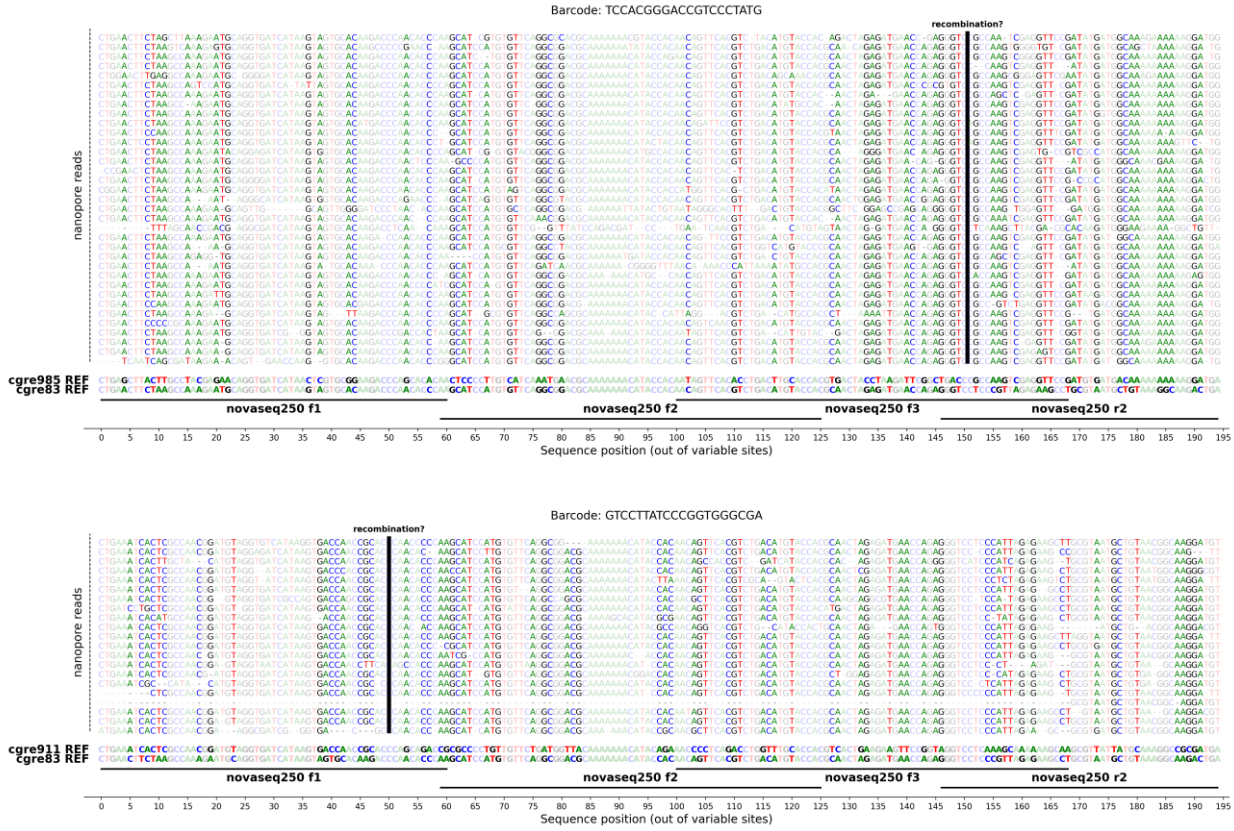

FIGURE 40. Chimeric GFP sequences. Two examples of barcodes associated with hybrid cgreGFP sequences from the "minis" library. Individual, aligned Nanopore reads are shown above the WT reference sequences corresponding to the genotype's identified original templates. Sequence positions on the X axis are limited to the 195 sites which are variable across different origins, but only those positions which vary between the immediately relevant origins are highlighted. The segments amplified for NovaSeq PE250 sequencing are labeled below, and the position at which the template switch occurs is marked by a vertical black line.

4.7. Library data processing and analyses

4.7.1. Determination of fitness values

A fitness was assigned to each primary/secondary barcode combination from each FACS run, using the cumulative distribution of pseudo-cell counts across gates (normalized to sum to 1) and the signal values (i.e. fluorescence intensities, non-log-transformed) corresponding to the borders of the sorting gates, as reported by the FACS machine during sorting. A cumulative density function of normal distribution was fitted to this data, and the fitted mean was assigned as the fitness; see **Figure 41** for a panel of example barcodes.

For amacGFP, cgreGFP, and ppluGFP2, fitting was performed using custom Python scripts by **Alexander Mishin** (code available on the `Orthologous_GFP_Fitness_Peaks` public GitHub page). The fitting algorithm was not bounded by the minimum observable experimental values (i.e. the middle of the darkest FACS gate, P3, see: 4.4.2. FACS setup), so barcodes with 100% of reads in the P3 gate were sometimes assigned fitnesses below the P3 midpoint. This data was left as-is for machine learning (see: 4.10.1. Modeling of local landscapes), but for all other analyses, fitness values under the P3 midpoint were set to the P3 midpoint itself, on the basis that a) it is unjustified to assign different outputs to barcodes with the same input information (i.e. all reads exclusively in P3), and b) the P3 gate was selected during FACS to encompass the distribution of GFP-negative control cells, so any within-P3 variation is stochastic and does not reflect true differences in fluorescence anyway.

For the novel cgreGFP-derived libraries, Mishin's code was modified to ensure the output was bounded by the minimum and maximum observed gate values, but otherwise untouched.

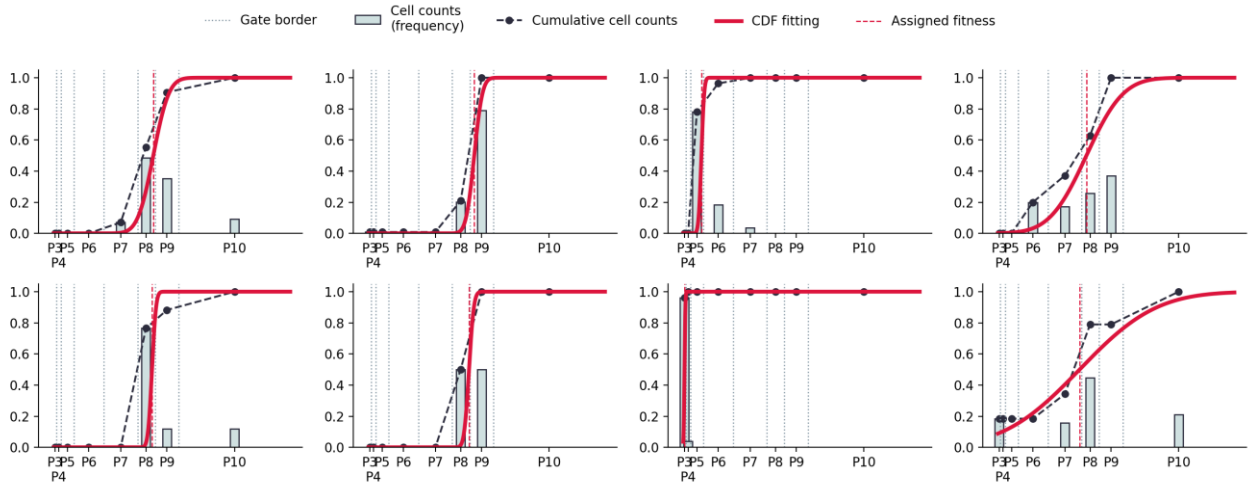

FIGURE 41. Representative examples of CDF fitting of FACS data. Each panel shows, for a different barcode, the CDF fitting (red curve) of the cumulative cell counts (black dashed curve) across green FACS gates. The X axis represents raw fluorescence values as output by the FACS machine (not log-transformed); for simplicity, only gate IDs are labeled (P3 – P10). Cell counts in each gate, normalized to sum to 1, are represented by grey bars. The fitness (fluorescence) value determined by the CDF fitting is shown as a vertical dashed red line.

4.7.2. Combining data from multiple experiments

After assigning fitness values, data from different FACS runs of the same library was pooled. In order to minimize any noise caused by inherent differences between machines, data from different FACS runs was scaled as follows.

For amacGFP, cgreGFP, and ppluGFP2, the values of wild-type controls included during sorting were used to match brightnesses by linear regression (see: Alexander Mishin's code available on the `Orthologous_GFP_Fitness_Peaks` GitHub page) in order to merge data from the two parallel FACS runs on different machines.

For novel cgreGFP-derived libraries, barcodes (i.e. primary/secondary pairs) measured in both machines and with a minimum cell count (at least 15 pseudo-cells) in each were fitted using Python's `scipy.optimize.curve_fit` module. The fitness values (non-log-transformed) from one experiment were then transformed to the scale of the other experiment using the fitted parameters. For cgreGFP:minis, the data from the two FACS runs were, linearly, largely in agreement, so the function $f(x) = a \cdot x$ was sufficient to transform one to the range of the other. However, for cgreGFP:1338, cgreGFP:4111, cgreGFP:9708, and cgreGFP:132, a more complex function was necessary to fit the data (Figure 42). For these libraries, first- and second-degree polynomials were poor fits; third-degree polynomials were good fits, but the function $f(x) = a \cdot x^b + c \cdot \sqrt{x}$ seemed marginally better than that, so we used this. (Note: this fit does not attempt to capture or explain any biological phenomena, merely to adjust data ranges from separate experiments so that they are comparable.) Once on the same scale, the transformed data were further adjusted slightly in order to align the minimum and WT fluorescence values of both experiments, according to the linear formula:

$$Min_{(OIST)} + (x - Min_{(IST)}) \cdot (WT_{(OIST)} - Min_{(OIST)}) \,/\, (WT_{(IST)} - Min_{(IST)}).$$

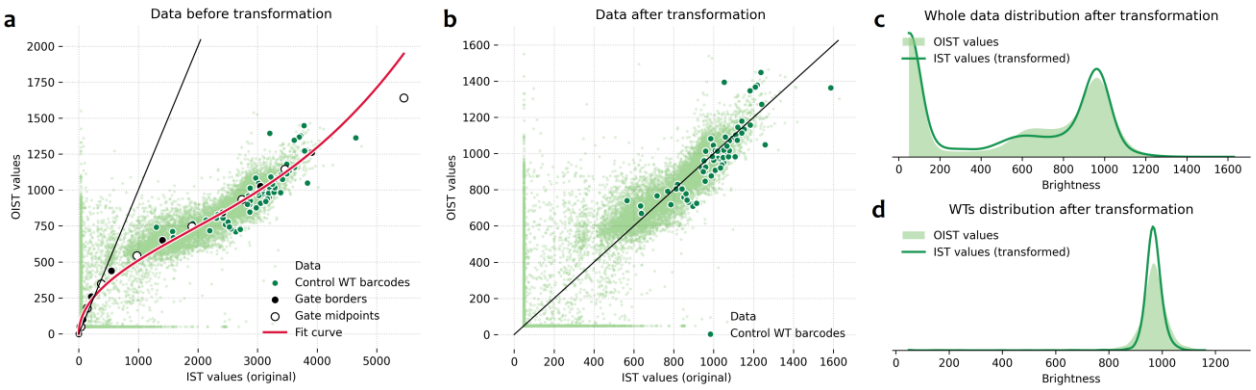

FIGURE 42. Merging of data from two FACS runs. Data from cgregFP:1338 is shown, as an example. **(a)** Data from IST and OIST runs, prior to rescaling; X and Y axes show respective fluorescence values. Light green points represent all barcodes with cell counts over 15 measured in both experiments. FACS gate borders (black dots), gate midpoints (white dots), and control barcodes from 17 cgreGFP-derived "WTs" (dark green dots) are also shown. For reference, the black line, $f(x) = x$, shows ideal fluorescence agreement between the two experiments. All data points represented were used to fit the red curve, $f(x) = a \cdot x^b + c \cdot \sqrt{x}$. **(b)** Data points as in (a), after transforming the IST values according to the parameters of the fitted curve and applying a linear rescaling to align the minimal and WT fluorescence values. Again for reference, the black line shows $f(x) = x$. **(c)** Distribution of the IST (dark green outline) and OIST (light green fill) library data, after rescaling. **(d)** As (c), but showing only WT genotypes.

4.7.3. The non-effect of synonymous mutations

Synonymous mutations are widely considered to have no phenotypic effects because they do not affect a protein's final amino acid sequence. Nonetheless, in some instances, synonymous mutations may have a fitness cost (Bailey et al., 2021), due to e.g. codon bias (mutating to a rare codon may affect translation by causing ribosomal stalling), or by affecting RNA stability or structure. Therefore, before merging data from genotypes containing different synonymous mutations, we checked for differences between the distributions of WT nucleotide genotypes vs genotypes containing synonymous substitutions. We did not see evidence of synonymous mutations affecting fluorescence in our data, and therefore barcode data pertaining to the same protein sequence was ultimately pooled together regardless of the presence of any synonymous mutations. It is likely that, if any effect caused by synonymous mutations existed, it would not be captured by our FACS setup, as RNA anomalies would likely affect mKate2 translation as well, and such cells would fall outside the sorted red gate (Figure 5c).

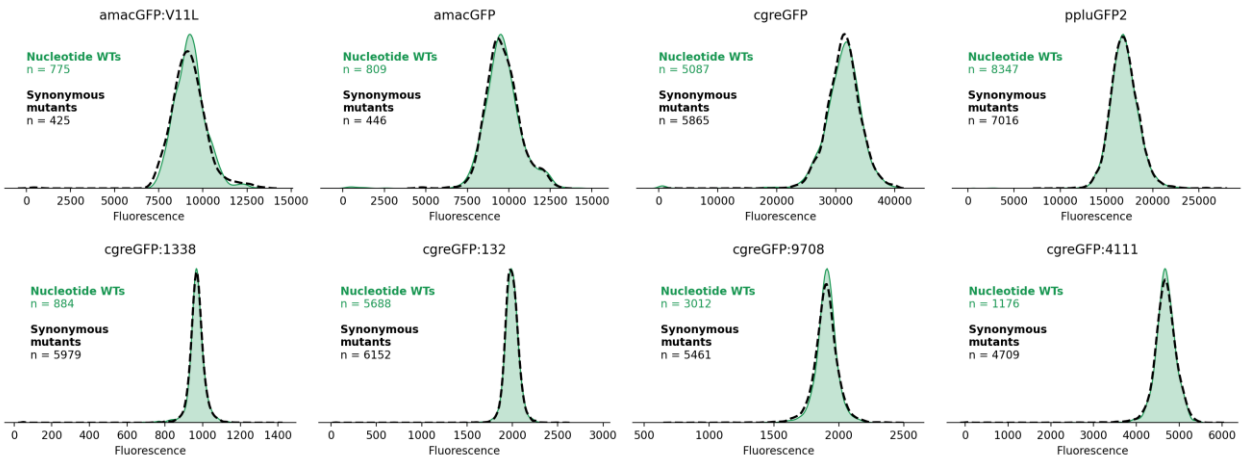

FIGURE 43. Effects of synonymous mutations. Fluorescence distributions of barcodes associated to WT nucleotide sequences (green) versus sequences encoding WT proteins but containing synonymous substitutions (black outline) are shown for all main libraries. These distributions were not significantly different (Mann-Whitney U test, $p > 0.15$) for all genes except cgreGFP:9708 ($p < 0.01$), which is likely a statistical fluke. Only quality barcodes with a minimum cell count of 50 were considered, as in Gonzalez Somermeyer et al., 2022, Figure 1-S1. Differences in cgreGFP:9708 distributions fluctuate into non-significance with higher or lower cell count thresholds, but changing the figure after the fact felt like p-hacking. Note: different libraries were run on different FACS machines with different voltage etc., so the X axis here is *not* comparable across genes.

4.7.4. Library data filtering

After performing fitness determinations for each primary/secondary barcode pair and merging data from replicate experiments, barcode-to-fitness datasets were quality filtered and barcodes/nucleotide genotypes corresponding to the same protein sequence were ultimately merged. ("Nucleotide genotypes" refer to DNA mutations; "protein genotypes" refer to amino acid substitutions in the final proteins. Due to synonymous mutations, the same protein genotype may be represented by multiple nucleotide genotypes.) Filtering steps all take place prior to $\log_{10}$-transforming the fitness values. Python code relating to amacGFP, cgreGFP, and ppluGFP2 datasets is publicly available on the `Orthologous_GFP_Fitness_Peaks` GitHub repository; this code was modified/improved slightly for new cgreGFP-derived genes.

Data processing steps were briefly as follows, with more detail below:

1. Discard low-quality barcodes (i.e. given primary/secondary barcode pairs measured on a given machine), based on low cell count or high spread across FACS gates.
2. Merge surviving barcodes according to their nucleotide genotype. This means grouping all replicates representing the same nucleotide genotype, summing their cell counts, and assigning the genotype a fluorescence corresponding to the average fitness of its replicates weighted by their individual cell counts.
3. Discard low-quality nucleotide genotypes, based on low total cell count, high variance, or few replicates. This filtered dataset is stored as the final, "Nucleotide Genotypes to Fitness" dataset for each gene (subject to downstream rescaling, see: 4.7.5. Scaling of library values), used for some downstream analyses.
4. Merge surviving nucleotide genotypes according to their protein genotypes. As in Step 2, where replicates are now variants with synonymous mutations encoding the same protein. This dataset is stored as the final, "Amino Acid Genotypes to Fitness" dataset for each gene (subject to downstream rescaling, see: 4.7.5. Scaling of library values), used for most downstream analyses.

Parameters subject to filtering

Data can widely be considered unreliable if it is based on few measurements/replicates or if those replicates contradict each other. In our data, these aspects can be captured in the following parameters, on either the barcode or genotype level:

- *Individual cell count*: for a given primary/secondary barcode ID, the number of physical cells of that ID which were FACS sorted (or normalized "pseudo-cell" counts, accounting for average number of NGS reads per cell).
- *Within-barcode variability*: for a given primary/secondary barcode ID, the spread of cell counts across FACS gates. While it is normal for a given genotype's fluorescence to span multiple gates, reads spread across dark and bright gates can be due to sorting errors.
- *Number of replicates*: for a given genotype, the number of unique barcodes measured: either same primary but different secondary barcode, or same primary/secondary measured on different machines, or different primary barcodes representing the same protein genotype (with or without synonymous mutations).
- *Global cell count*: for a given genotype, the total number of cells of all barcodes replicates representing that genotype.
- *Variability across replicates*: for a given genotype, the index of dispersion or coefficient of variation (respectively variance or standard deviation, normalized to the mean) in assigned fitnesses of different barcodes representing that genotype.

For amacGFP, cgreGFP, and ppluGFP2, no filtering was applied to individual barcodes, but genotypes with low global cell count, low number of replicates, or high variability across replicates were discarded. For new cgreGFP-derived libraries, barcodes were additionally filtered prior to nucleotide genotype merging. Exact threshold values for filter parameters can be found in (**Table 5**). Thresholds were determined independently for each library because, *a priori*, we cannot assume that all FACS runs to have had identical sorting precision in practice (despite machine settings), nor all genes to naturally exhibit the same variability in fluorescence.

Thresholds for each filter parameter were determined with the goal of maximizing the total number of surviving protein genotypes while minimizing the number of genotypes with incorrectly-assigned fitnesses (as estimated by quality-control genotypes described below). Thousands of combinations of

threshold parameters were applied to the entirely of each dataset uniformly and blindly.

Quality control genotypes

Each dataset contained many nucleotide genotypes encoding WT proteins; as the WTs are, naturally, known to be bright, we can assume that any apparently-WT-encoding genotypes assigned a low fitness in the dataset must be victims of a) FACS technical error: droplets being accidentally sorted into the wrong tube, or b) sequencing error: they are not truly WT, but actual mutations in the coding sequence were not detected. The latter possibility is minimized by imposing higher minimum read counts and overall read agreement for consensus sequences (see: 4.6. Library sequencing: Coding regions). WT genotypes therefore serve as internal controls for estimating the rate of "false negatives" — bright genotypes incorrectly classified as dark or dim.

Conversely, datasets also contain many genotypes encoding a mutation in the chromophore (Y66 and G67, in traditional avGFP numbering) or in other residues involved in chromophore maturation (R96 and E222, in traditional avGFP numbering), hereafter called chromophore mutants for simplicity. As a functional chromophore is a physical requirement for fluorescence, any chromophore mutants assigned a high brightness are also likely due to (a) or (b) described above. Note that the probability of (a), FACS mis-sorting, is likely higher in this case, due to how the machine works: it is easier for a dark cell to go undetected and tag along with a bright cell into a bright-gate tube, than it is for a bright cell to go undetected and be sorted into a dark-gate tube (**Figure 44**). Note also that in the case of chromophore mutants, a rare option (c) exists: That the mutation genuinely doesn't eliminate fluorescence. Rare mutations affecting chromophore maturation have been documented which maintain fluorescence (mainly, the substitution of tyrosine with another aromatic amino acid) (Chudakov et al., 2010; Sarkisyan et al., 2012), or which are rescued by other mutations elsewhere in the sequence (Wood et al., 2005); see **Figure 10** for an example. Nevertheless, such mutations are rare, so chromophore mutants remain a good internal control for "false positives" — dark genotypes incorrectly classified as dim.

Note: we did not use nonsense mutations as "false positive" checks, as stop codon read-through is frequent (Poole et al., 1995): multiple avGFP genotypes containing premature stop codons were cloned independently and confirmed to be fluorescent (Sarkisyan et al., 2016).

Table 5. Dataset filtering parameters and statistics. Chromophore sites are numbered starting from Met=0 for each gene. We stopped considering E223 in new datasets due to mutations at that site being frequently only slightly deleterious, making it a bad control for lethal mutations. WT genotypes contributed to the False Negative rate if they were assigned fitnesses falling within any of the four dimmest gates (P3 – P6), while chromophore mutants contributed to the False Positives rate if they were assigned fitnesses falling within any of the four brightest gates (P6 – P10).

	amacGFP	ppluGFP2	cgreGFP	cgreGFP: 1338	cgreGFP: 132	cgreGFP: 4111	cgreGFP: 9708	cgreGFP: minis
Chromo-phore sites	Y65, G66, R95, E221	Y57, G58, R86, E209	Y68, G69, R99, E223	Y68, G69, R99	Y68, G69, R99	Y68, G69, R99	Y68, G69, R99	Y68, G69, R99
Barcode cell count	—	—	—	4	10.8	7.8	2.7	7
Barcode spread (gates)	—	—	—	1.7	1.11	1.3	1.25	1
Minimum replicates	2	3	3	—	—	—	2	—
Global cell count	26	23	14	4.5	45.1	—	40.2	16.5
Replicate variance	525 (D)	1000 (D)	575 (D)	1 (CV)	—	1.2 (CV)	0.25 (CV)	0.32 (CV)
% data discarded	~78%	~68%	~80%	~16%	~44%	~23%	~57%	~56%
Surviving nucleotide genotypes	46173	47045	34758	11307	5894	10499	6477	4987
False negatives	0/1084	0/2744	0/1583	3/693 (0.43%)	0/504	3/643 (0.47%)	1/701 (0.14%)	1/184 (0.54%)
Surviving protein genotypes	35500	32260	25165	8934	4267	8214	4180	4597
False positives	9/1635 (0.55%)	11/2242 (0.49%)	14/1860 (0.75%)	1/308 (0.32%)	1/150 (0.666%)	1/280 (0.36%)	2/129 (1.55%)	2/242 (0.82%)

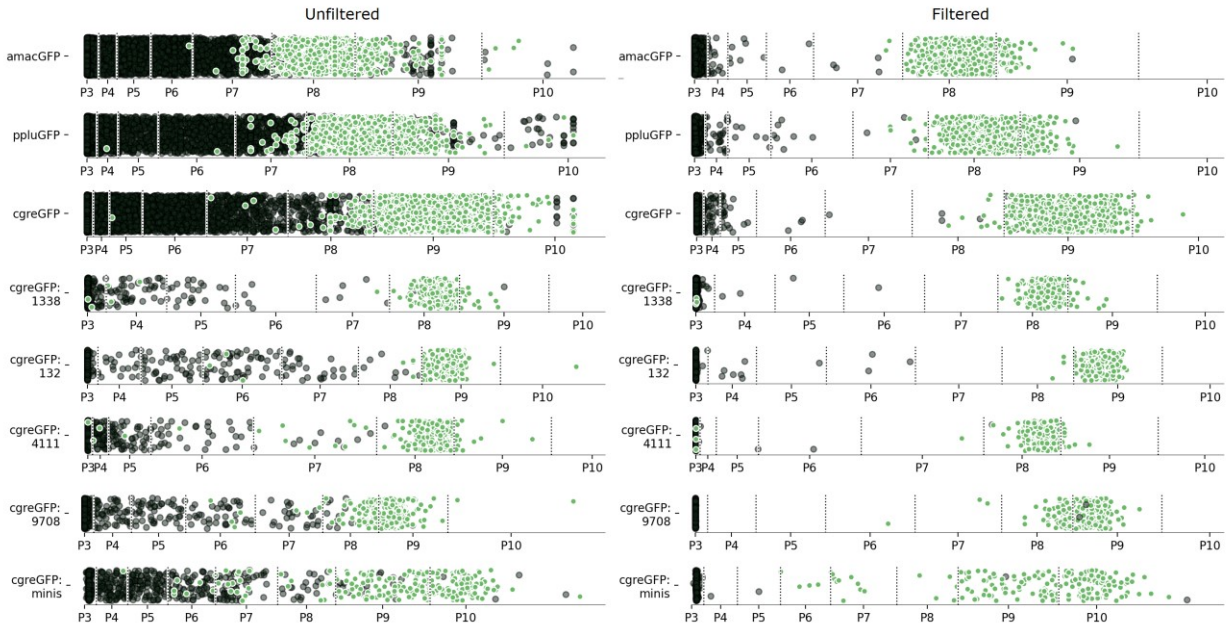

FIGURE 44. Brightnesses of WT and chromophore-mutant genotypes before and after data filtering. Each dot represents a unique nucleotide genotype, either WT-encoding (green) or containing a mutation affecting chromophore maturation (black). The X axis represents fluorescence (non-$\log_{10}$-transformed) across the 8 sorted FACS gates. Vertical dotted lines mark gate borders. Note: cgreGFP:minis WTs appear to have an excessively broad distribution because the 12 different WT genes of the "minis" library vary greatly in intensity from each other.

4.7.5. Scaling libraries to the same value range

FACS runs were performed on four different machines over the course of several years, and it was not always possible to maintain identical voltage and other settings. The range of fluorescence values of different libraries are therefore not always comparable. To be able to directly compare fitness values across libraries, we rescaled the filtered data by matching the brightnesses of the WT control genotypes, which were included in all sorted libraries for this purpose (see: 4.4.1. FACS sample preparation). This method was essentially the same as for merging data from separate FACS runs of the same library (see: 4.7.2. Combining data from multiple experiments), but using only WT controls (since no other genotypes would overlap between libraries).

Note: this rescaling was not performed on amacGFP, cgreGFP, or ppluGFP2 datasets as analyzed in Gonzalez Somermeyer et al., 2022, and machine learning models were trained on non-rescaled datasets (see: 4.10. Machine learning). However, the data ranges of new cgreGFP-derived gene libraries were very different from that of the original cgreGFP, so to allow direct comparisons, new libraries were rescaled to the original cgreGFP range. For consistency, the amacGFP and ppluGFP2 libraries were then also rescaled to this range for this dissertation, and non-ML analyses were repeated (to no difference in results from Gonzalez Somermeyer et al., 2022).

Datasets were rescaled such that the minimum fluorescence values (P3 midpoint) and the fitness value of WT cgreGFP (which was measured in all experiments) were matched. For amacGFP and ppluGFP2, a second degree polynomial $f(x) = a \cdot x + b \cdot x^2$ was fitted to the fluorescence values of control WT genotypes (avGFP, amacGFP, cgreGFP, ppluGFP2), and the fitted parameters were used to transform amacGFP and ppluGFP2 data to the cgreGFP range (**Figure 45a,b**). For new cgreGFP-derived libraries, which did not include WT avGFP, amacGFP, or ppluGFP2 control barcodes, the FACS gate border values and WT cgreGFP fitness values were used instead. This was possible because in all experiments, the FACS gates were selected in a similar fashion in the sense of which section or proportion of the library population fell into each gate. Finally, the data were further linearly rescaled (as in 4.7.2. Combining data from multiple experiments) to align the minimum and WT cgreGFP values (**Figure 45c,d**).

Note: avGFP data from Sarkisyan et al., 2016 was not rescaled in any way, due to a) no controls of any other genes were available in the avGFP data to use as references for fitting; b) FACS gates were also unsuitable for this purpose, since gates in the avGFP experiments were selected differently (fixed intervals) from those of all other libraries (P3 as fully dark, other gates relative to population size); c) avGFP's absorbance spectrum differs from that of the other genes (**Figure 3d**), thus the 405 nm laser was

used for excitation in avGFP experiments, while the 488 nm laser was used for other GFPs, making the values of the WT avGFP controls in those libraries unrepresentative of its actual relative brightness.

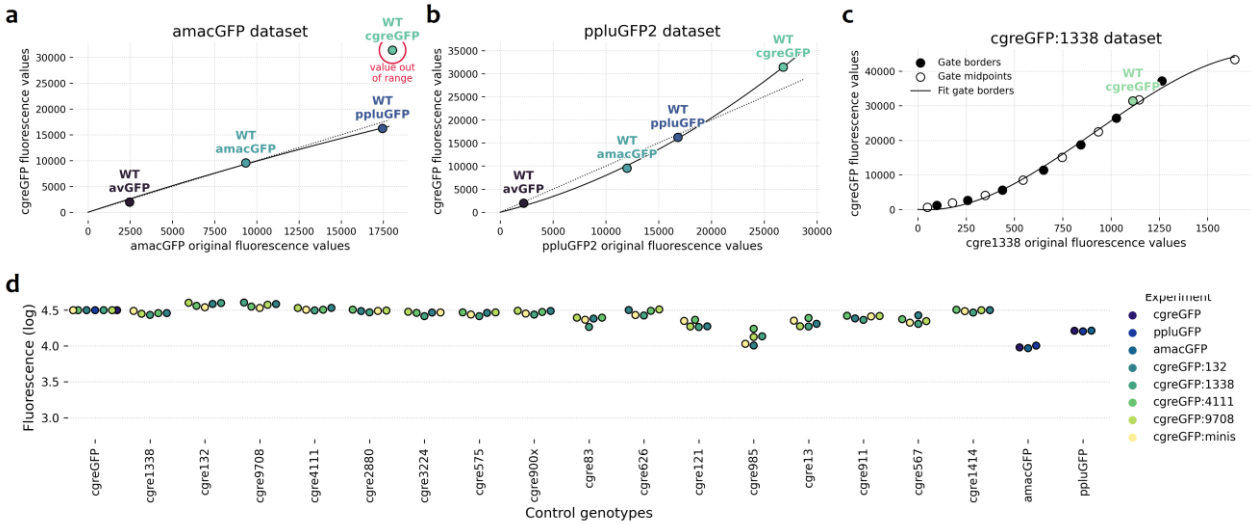

FIGURE 45. Rescaling of datasets to the WT cgreGFP data range. (a) Fitting of second-degree polynomial to amacGFP and cgreGFP control WT genotypes. The fitness value of the very bright WT cgreGFP was not included, as its intensity clearly fell beyond the measurement limits of the amacGFP experiment. **(b)** As (a), but for ppluGFP2. **(c)** Example (cgreGFP:1338) of fitting a third-degree polynomial to gate values in new cgreGFP-derived gene libraries. **(d)** Fluorescence values (Y axis) of different control WT genotypes (X axis) after rescaling of different libraries (color).

4.7.6. Calculation of mutation effects and epistasis

All calculations of mutation effects were performed on $\log_{10}$-transformed fitness values unless otherwise stated.

Effects of individual mutations

The effect of any given mutation M was calculated as the difference between the measured fluorescences of the genotype containing the single mutation M and the wild-type:

$$Effect_{(M)} = Fluorescence_{(M)} - Fluorescence_{(WT)}$$

Expected joint effects without epistasis

The effects of single mutations were used to calculate the expected fitnesses of genotypes with multiple mutations under the assumption of no epistasis. Note that expected fitnesses could only be calculated for genotypes composed solely of mutations which had also been measured individually. Under this assumption, the individual mutations comprising any given *n*-mutant genotype contribute additively to the final fitness of said genotype:

$$Fluorescence_{(expected)} = Fluorescence_{(WT)} + Effect_{(M1)} + Effect_{(M2)} + ... + Effect_{(Mn)}$$

Note: the possible values of "expected" fluorescence were bounded by the minimum and maximum measurable values, for each dataset. For example, if two separate mutations were each observed to reduce fluorescence to zero, the expected fitness of the double-mutant genotype was still just zero, not the actual mathematical sum of both lethal effects.

Epistasis

Epistasis was calculated as the difference between the observed (measured) fitness and the expected fitness under the above additive model:

$$Epistasis = Fluorescence_{(observed)} - Fluorescence_{(expected)}$$

Background-calibrated mutation effects

Furthermore, within each library, many mutations were observed in multiple backgrounds. For

example, if the fitnesses of genotypes A, B, and AB were all measured, then the individual effect of A could be calculated not only in the WT background, but also in the background of B; likewise for the effect of B in the background of A. More generally, for any pair of genotypes differing by only one mutation M, the effect of M can be calculated as above:

$$Effect_{(M)} = Fluorescence_{(background + M)} - Fluorescence_{(background)}$$

Having multiple observations of the same mutation's effects in different backgrounds allowed for computing its average and/or median effect across backgrounds, as well as see how variable, or prone to interacting epistatically, the mutation was. These "background-calibrated" effects were used alongside "simple" (as measured in the WT) effects in various analyses, in particular in the generation of ML-guided design of novel sequences.

4.7.7. Estimation of noise

Measurement errors, caused by e.g. cells falling into the wrong FACS tubes and skewing fitness estimates downstream, can affect not only the genotype in question, but also calculations of "expected" (additive) fluorescence of other genotypes and thereby calculations of epistasis. Therefore, we should consider the possibility of spurious epistasis arising from measurement errors, and set an appropriate threshold for what we consider to be true epistasis such that false positives are reliably avoided. For this, we estimated our measurement error and simulated epistasis discovery due to it.

Under ideal circumstances (perfect measurements) and in the absence of epistasis, the following holds true for any two co-occurring mutations A and B (subject to the minimum and maximum of the experimentally measurable range):

$$Effect_{(AB)} = Effect_{(A)} + Effect_{(B)}$$

Epistasis, measured as $Effect_{(observed)} - Effect_{(expected)}$, should be equal to zero in this case:

$$Epistasis = Effect_{(AB)} - (Effect_{(A)} + Effect_{(B)}) = 0$$

However, if there exists measurement error, we have the following, where "$noise_{(x)}$" is the error in the measurement of genotype x:

$$MeasuredEpistasis_{(AB)} = MeasuredEffect_{(AB)} - (MeasuredEffect_{(A)} + MeasuredEffect_{(B)})$$
$$MeasuredEpistasis_{(AB)} = (Effect_{(AB)} + noise_{(AB)}) - (Effect_{(A)} + noise_{(A)} + Effect_{(B)} + noise_{(B)})$$

In the hypothetical case of no real epistasis, any detected epistasis will be due solely to measurement error:

$$MeasuredEpistasis_{(AB)} = (Effect_{(A)} + Effect_{(B)} + noise_{(AB)}) - (Effect_{(A)} + noise_{(A)} + Effect_{(B)} + noise_{(B)})$$
$$MeasuredEpistasis_{(AB)} = noise_{(AB)} - noise_{(A)} - noise_{(B)}$$

Furthermore, the chances of spurious epistasis discovery increase as the number of mutations in the genotype increase:

$$MeasuredEpistasis_{(ABCD)} = noise_{(ABCD)} - noise_{(A)} - noise_{(B)} - noise_{(C)} - noise_{(D)}$$

Note: measurement error (noise) can take either positive of negative values.

We estimated the measurement error in our data by looking at the standard deviations of fluorescences of genotypes represented by multiple replicates. Since amacGFP, cgreGFP, and ppluGFP2 datasets were filtered at the nucleotide genotype level but not at the barcode level (see: 4.7.4. Library data filtering), we used the nucleotide genotype datasets for this, as unfiltered barcodes would not be representative of the final, in-use datasets. Distributions of standard deviations were well fit by a skew-normal curve (**Figure 46a**).

We then simulated spurious epistasis for genotypes with increasing numbers of mutations (2–8) by simulating measurement error (noise). For a hypothetical genotype with n mutations, the simulated epistasis was equal to the sum of $n+1$ measurement errors drawn from the skew normal distribution which was fitted to the real data. Spurious epistasis values between -0.2 and 0.2 was common, but values beyond -0.3 or 0.3 — the thresholds set in this work for accepting epistasis as genuine — were negligible (**Figure 46b**).

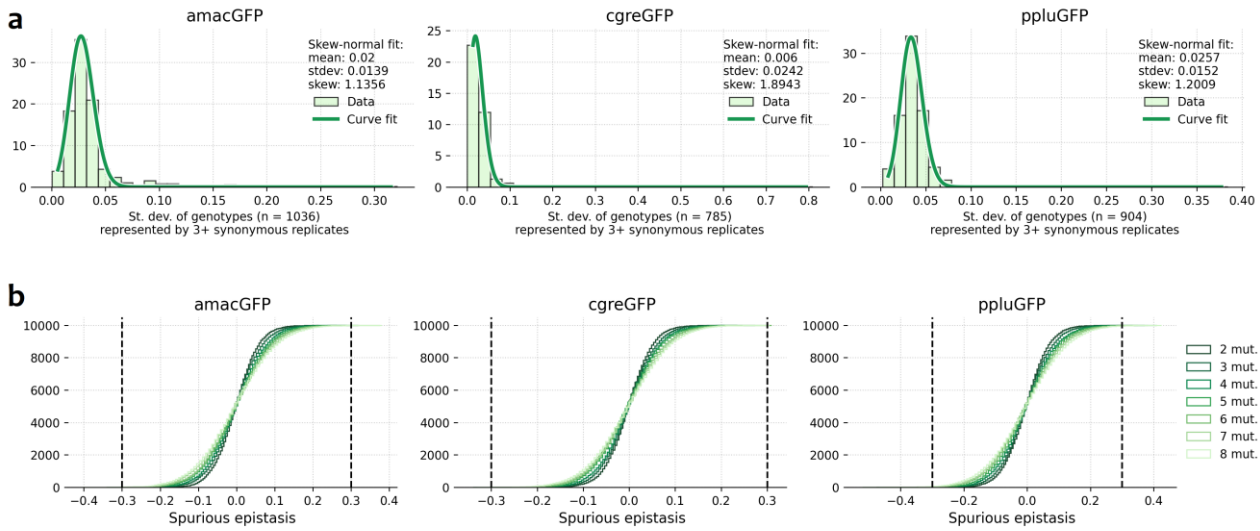

FIGURE 46. Estimation of noise in genotype-phenotype data. (a) Distributions of standard deviations of protein genotypes represented by at least three replicates (nucleotide genotypes with synonymous mutations) (bars) and skew normal curve fit to this distribution (green curve). **(b)** Simulations of spurious epistasis for genotypes containing 2-8 mutations, with noise values drawn from the fitted distributions in (a). 10,000 simulations were performed for each category (gene/number of mutations). Plots show the cumulative distribution of false epistasis detected under these conditions. Vertical dashed lines mark the thresholds (-0.3 and 0.3) used in this work for accepting epistasis as genuine.

4.7.8. Determination of physical distances between residues

Distance measurements (in Angstroms) between pairs of residues on the folded protein structure were determined as in Sarkisyan et al., 2016. Briefly, the coordinates of all atoms in all residues of protein PBD structures (2WUR, 2HPW, 2G3O, 7LG4) were extracted using PyMOL. The distances between pairs of atoms belonging to different amino acid residues was calculated using custom Python scripts (available on the `Orthologous_GFP_Fitness_Peaks` GitHub repository). For any given pair of residues, the minimum distance between any two atoms (one from each residue) was assigned as the distance between residues (in Ångströms). The structural similarity of sequence-divergent GFP orthologs can be appreciated in Figure 47.

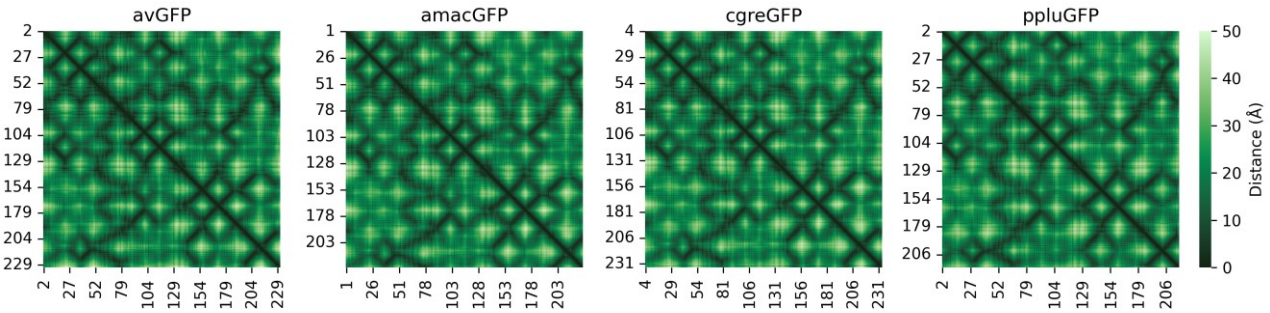

FIGURE 47. Physical distances between pairs of residues in folded GFPs. Mimimum distances between any two atoms belonging to two residues are displayed in Ångströms. Amino acid positions are represented on the X and Y axes.

4.8. His-tagged protein purification

4.8.1. Protein Expression

Coding sequences of wild-type GFPs, codon-optimized for *E. Coli* and flanked by Type IIS restriction sites for easy cloning (see: 4.1.1. Golden Gate cloning), were obtained by PCR where possible or ordered as synthetic DNA from Twist Bioscience. Genes were cloned in-frame with an N-terminal 6H polyhistidine

tag, under a T7 promoter, in a homemade low-copy (p15A origin) vector conferring kanamycin resistance.

Purified, sequence-verified plasmids were then used to transform BL21(DE3) cells. Unlike most laboratory strains, BL21(DE3) cells contain an inducible T7 polymerase, allowing the specific over-expression of proteins of interest under the T7 promoter. Transformed BL21(DE3) cells were plated on LB agar supplemented with 50 µg/ml kanamycin and 20 µM IPTG to induce T7 expression, and incubated overnight at 30°C and then overnight again at room temperature. We chose to use plated cells rather than liquid cultures, and constant rather than time-limited IPTG exposure, in order to maximize the amount of fully matured GFP (toxicity from extreme over-expression was not an issue).

4.8.2. Protein Extraction

His-tagged proteins were purified using sepharose beads coated with nickel ions (Ni^{2+}), which interact and bind to imidazole rings. As histidine contains an imidazole ring, proteins with exposed histidine residues are preferentially bound. Once the protein of interest is captured and other proteins and molecules have been washed away, the His-tagged protein can be released by flooding with high concentrations of imidazole, which competes with and takes the place of the proteins on the nickel-sepharose beads.

The following homemade buffers were used in this step:

- ***Binding buffer***: 150 mM NaCl, 20 mM Tris-HCl, 25 mM imidazole, pH 8 (note: a low but nonzero concentration of imidazole is recommended even during initial binding and washing, to minimize nonspecific binding of other histidine-containing proteins to the beads)
- ***Elution buffer***: 150 mM NaCl, 20 mM Tris-HCl, 500 mM imidazole, pH 8

Protocol:

1. ***Cell recovery***: Wash cells (densely grown colonies from 10-15 plates) and resuspend in 35-40 ml of binding buffer inside a 50 ml Falcon tube.
2. ***Cell lysis***. We sonicated cells in a Qsonica Q700 device, using the flat 1 cm probe nozzle and the following parameters: 20 kHz, amplitude 10, 1s on/4s off (20 minutes active sonication time in total). After sonication, the solution is visibly less murky, due to the bacterial cell walls being compromised.
3. ***Pellet cell debris***. Centrifuge sonicated cells for 30 minutes at 21,000 g, at 4-6°C.
4. ***Resin equilibration***. Wash 3 ml of nickel-sepharose beads with 20 ml of distilled water followed by 20 ml of binding buffer. The easiest way to do this is by adding the beads to a chromatography column and letting the washes flow through.
5. ***Incubation of beads and protein***. After centrifugation, collect the green supernatant and add the washed beads to it. Incubate with rotation for one hour at 4-6°C to promote capture of His-tagged GFP by the beads. Discard the pelleted cell debris.
6. ***Column set-up***. Add the supernatant/resin solution to the chromatography column. Allow the liquid to pass through to a waste bucket by gravity, or if you don't have all day, connect a silicone tube to the column's nozzle and use a peristaltic pump instead, but be careful to not allow the resin beads to dry out.
7. ***Washing***. Add 20 ml of binding buffer to the column and allow it to pass through. Repeat this step a total of three times.
8. ***Elution***. Remove the silicone tubing and peristaltic pump, and ready a collection tube. Add ~5 ml of elution buffer to the column and allow it to pass through. Ideally, place a new collection tube for every 0.5-1 ml eluted; the eluted protein is much less concentrated at the beginning and end of this process, so using multiple collection vials avoids diluting the more concentrated elute. Eluted GFP can be stored at 4°C indefinitely.

A note on bacterial sonication:

Sonication was performed in a cold room, and the tubes with cells were, additionally, on ice; nevertheless, the heat produced was enough to partially melt that ice, so it is important to ensure that the tube is properly secured and immobile during the whole process so that the tube walls don't hit the sonicator probe. (We never suffered any accidents in this regard, however, we were very annoyed the day somebody broke the integrated tube fastener.)

Extended sonication with the Q700 was the only successful method we were able to test. For instance, no amount of water bath sonication was effective at lysing the cells; even hours-long programs harsh enough to heat the water and the protein past its denaturing point (evidenced by the visible loss of green color) did not seem to succeed in actually lysing the cell walls.

Reuse of Nickel-Sepharose Beads

After use, nickel-sepharose beads were washed with 20 ml of resin stripping buffer (20 mM Na_2HPO_4, 0.5 M NaCl, 50 mM EDTA, pH 7.4) to remove residual protein, followed by 20 ml of binding buffer and 20 ml of distilled water to remove residual EDTA. As an extra precaution and to remove any stubborn precipitated protein, beads were also incubated in 1 M NaOH for one hour, then washed with 40 ml of binding buffer and 40 ml of distilled water.

Clean beads were then recharged with 5 ml 0.1M $NiSO_4$, rinsed with 20 ml of distilled water followed by 20 ml of binding buffer to adjust the pH, and finally stored in 20% ethanol for future use.

The above steps were performed with the beads resting in a chromatography column and the solutions passing through with the help of a peristaltic pump. Chromatography columns were rinsed thoroughly with distilled water and also reused.

4.8.3. Protein Quantification and Storage

Protein concentration was estimated via NanoDrop using the Protein A280 protocol, taking into account the extinction coefficient for each protein. Extinction coefficients were determined using ProtParam (https://web.expasy.org/protparam/), based on the amino acid sequence of each protein.

As part of Anna Toidze's bachelor thesis project, we also estimated protein concentrations by Bradford assay using the Pierce Coomassie Protein Assay Kit (Thermo Fisher) according to the manufacturer's instructions. The calibration curve was generated by fitting a linear function $f(x) = ax + b$ to the measurements of BSA samples of known concentration (0, 25, 125, 250, 500, 750, 1000, 1500, and 2000 mg/ml). Bradford-estimated protein concentrations largely coincided with those estimated by NanoDrop for the same samples.

4.8.4. Crystallization and structure of amacGFP

Purified amacGFP protein (obtained as in 4.8. His-tagged protein purification) was crystallized and used to resolve the 3D protein structure. Crystallization and structure determination were performed by **Nina Bozhanova**. Full methodology is described in Gonzalez Somermeyer et al., 2022.

Briefly: amacGFP was crystallized at 21°C, following the Hampton Research Additive Screen protocol and using the sitting drop vapor diffusion technique. Crystals were grown for 1 week and flash frozen in liquid nitrogen with mother liquor/20% PEG400 as cryoprotectant. Diffraction data was collected on a D8 Venture system and crystal structures were solved by molecular replacement using MOLREP.

The amacGFP structure can be found on PDB under the accession code 7LG4.

4.9. Measures of protein structure and stability

All assays were performed on purified GFP proteins (see: 4.8. His-tagged protein purification). All protein samples used in the assays below were diluted starting from 20 mg/ml stock solutions in elution buffer (see: 4.8.2. Protein extraction). Unless otherwise stated, stocks were diluted with imidazole-free elution buffer (20mM Tris-HCl, 150mM NaCl, pH 8).

4.9.1. Urea sensitivity assays

The urea sensitivity of all "wild-type" reference GFPs was assayed by measuring the absorbance and fluorescence (emission) spectra with and without 9M urea. 20 mg/ml stocks were diluted in either 1X PBS or 1X PBS/9M urea to final protein concentrations of 18.5 µM (for absorbance) or 0.15 µM (for fluorescence), in a final volume of 200 µl per technical replicate. Samples were measured in 96-well plates with flat- and clear-bottomed wells (and, in the case of fluorescence measurements, black walls).

Absorbance spectra were measured from 300 nm to 700 nm and fluorescence spectra from 420 nm to 700 nm, in 5 nm intervals, on a Biotek SynergyH1 plate reader at 42°C. For fluorescence, an excitation wavelength of 420 nm was used, which is suboptimal for all of our proteins but allowed us to measure a fuller emission spectrum than excitation at 488 nm would have. Measurements were repeated at fixed time intervals (~40 and ~25 minutes for absorbance and fluorescence respectively, due to the time necessary to measure full spectra in all 96 wells), for a total of around 60 hours. Data from 3–8 technical replicates was kept for each gene and each condition (some replicates were discarded due to aberrations caused by evaporation or bubbles in the well).

As blanks, 8 wells containing the corresponding amount of protein-free elution buffer were measured alongside the samples. These values were subtracted from those of the protein measurements, in order to avoid interference from the imidazole present in the protein elution buffer (which absorbs significantly around 300 nm).

The excitation and emission peaks for each gene were determined by detecting the highest values in PBS for absorbance and fluorescence, respectively, within the 5 nm resolution allowed by the measurement setup. Urea sensitivity was determined by tracking the decrease in absorbance and fluorescence of these peaks over time in 9M urea.

4.9.2. Thermosensitivity assays

The sensitivity to increasing temperatures of different GFPs was measured in a variety of ways. Simple melting curves generated in a qPCR machine were done at ISTA; DSF and CD measurements were performed at the Central European Institute of Technology (CEITEC) in Brno, Czech Republic, under the supervision of facility staff Josef Houser and Eva Fujdiarová; DSC runs were performed at CEITEC by Josef Houser.

qPCR melting curves

Samples were diluted to 0.1 mg/ml. 6-8 replicates of 20 µl were loaded onto white qPCR 96-well qPCR plates and monitored for fluorescence loss in a Roche LightCycler 480 qPCR machine, using a melting curve protocol of heating from 20°C to 99°C at a ramp rate of ~2°C/min (the slowest we could force the machine to go). The SYBR Green channel (excitation at 465 nm and detection at 510 nm) was chosen for monitoring as it was the most similar to GFP absorbance and emission parameters. Melting curve temperatures were calculated automatically, corresponding to the peak of the first derivative of the fluorescence emission loss.

Circular dichroism

200 µl of 0.1 mg/ml protein samples were run on a Jasco J-815 CD spectropolarimeter, in cuvettes of 1 mm thickness. Proteins' initial spectra between 200 and 260 nm were measured at 30°C; the spectrum of protein-free buffer was also measured and subtracted from those of the proteins. Spectra were not measured beyond 200 nm, as measurements because unreliable due to voltage increases of over 700 V; we observed that higher concentrations of imidazole in the protein buffer decreased the usable range of measurements. Initial spectra were measured at a scanning speed of 100 nm/min, a data pitch of 1 nm, a digital integration time of 2s with 1 nm bandwidth, and 10 accumulations. The wavelength corresponding to the spectrum peak was then selected (208 nm for avGFP, 218 nm for amacGFP, cgreGFP, and ppluGFP), and samples were heated to 98°C at a ramp rate of 1°C/min with constant monitoring of the selected peak wavelength. The full spectra were subsequently measured again post-denaturation, using the same settings as for the initial spectra.

The melting curves (monitored at 218 or 208 nm during heating) were fitted with a logistic function, $f(x) = L \,/\, (1+e^{-k(x-x0)})$, using custom Python scripts (`scipy.optimize.curve_fit`) in order to obtain the melting temperature (x_0).

Differential scanning fluorimetry

DSF measures the changes in the fluorescence emission at 330 nm and 350 nm of aromatic amino acids (primarily tryptophan, but also tyrosine), as proteins are heated. As proteins are denatured, buried residues become exposed. Because the fluorescence of aromatic amino acids depends on their immediate surroundings, this results in a detectable change in emission, and the ratio of 350/330 nm fluorescence as

a function of increasing temperature can be used to determine the temperature at which the protein unfolds.

Protein samples were diluted to 1 mg/ml and run in triplicate (10 µl per replicate) on a Prometheus NT.48 (NanoTemper Technologies) device, set to 100% excitation power. Samples were heated from 20°C to 110°C at a rate of 1°C/min. The melting temperatures for protein unfolding and protein aggregation were determined from the peaks of the first derivatives of either the 350/330 nm emission ratio (for unfolding) or the light scattering (for aggregation).

None of our GFPs' chromophores fluoresce in the range of 330-350 nm, which could interfere with interpretation of DSF measurements (Figure 3d). And while our GFPs all had a low tryptophan (W) content, tyrosine (Y) content was sufficient to generate a good signal. The aromatic amino acid content of our DSF-tested GFPs was:

- **avGFP**: 1 W, 11 Y
- **amacGFP** and **amacGFP:V11L**: 1 W, 10 Y
- **cgreGFP** and its derivatives (1338, 132, 9708, 4111): 3 W, 14 Y
- **ppluGFP**: 0 W, 12 Y

Differential scanning calorimetry

DSC measures a protein's specific heat capacity (C_p, measured in calories per mole), i.e. the amount of heat required to increase sample temperature, which changes as the protein denatures. The protein sample and a known reference (buffer solution) are heated simultaneously at a constant rate; when the protein unfolds, heat is absorbed, causing a temperature difference between the sample and the reference which can be converted into a C_p value. The enthalpy of unfolding (ΔH) is the area under the C_p curve, and the peak corresponds to the protein's melting temperature.

Samples were diluted to 1 mg/ml and run in duplicate on a MicroCal PEAQ-DSC machine (Malvern Analytical). Temperature settings were set to increase from 20°C to 110°C at a ramp rate of 1°C/min, in order to be able to compare with DSF data (because the ramp rate affects the detected melting temperature: the faster the temperature increases, the higher the detected T_m will be). Melting temperatures were calculated automatically as the peak of the C_p curve.

4.9.3. SEC-MALS

SEC-MALS consists of first separating proteins or other macromolecules in solution based on their hydrodynamic volume (size exclusion chromatography), then measuring the intensity of light scattering as the sample elutes from the chromatography column (multi-angle light scattering). The light scattering is proportional to the macromolecules' weight-averaged mass. This method therefore allows the molecular masses of different macromolecules in solution (for example, different oligomeric states of a protein) to be determined.

SEC-MALS analysis was performed at CEITEC, Brno, by Josef Houser. 1 mg/ml GFP samples were run on an OmniSEC system (Malvern Panalaytical). 50 µl were used as the injection volume, and samples were measured at 30°C with a flow rate of 0.7 ml/min.

4.9.4. SDS-PAGE and Western Blot

Cells from P3 and P10 FACS gates were grown from glycerol stocks on LB/zeocin agar. Colonies (>10k per condition) were recovered (see: 4.1.4. Harvesting plated libraries) and cells were pelleted. Pellets (0.25 g) were then resuspended in 30 ml of Lysis Buffer (1X PBS pH 7.4, 150 NaCl, supplemented with 50 µl protease inhibitor cocktail (Merck #P8340)) and sonicated on a QSonica Q700 (20 khz, amplitude 10, 10 minutes of active sonication at 1s ON/4s OFF). (30 *ml* is not a typo. Unfortunately, sonication of smaller volumes on various other devices was not efficient.)

15 µl of lysed cells were centrifuged for 10 minutes at 20,000 g. The supernatant was isolated and the pellet was resuspended in 15 µl of Lysis Buffer. Supernatants, resuspended pellets, and full lysates (15 µl) were mixed separately with 5 µl 4X Laemmli loading buffer (BioRad) and boiled for 5 minutes at 95°C. Samples were then run in 4-20% polyacrilamide Mini-Protean gels (BioRad) for one hour at 100 V alongside a protein ladder (Protein Precision Plus Standard, BioRad). The coomassie-based ReadyBlue dye (Sigma) was used to stain the gels, overnight at room temperature.

Coomassie-stained gels were imaged on a ChemiDoc MP system (BioRad), then transfered to PVDF

membranes using a Trans-Blot Turbo Transfer system (BioRad). Membranes were blocked with EveryBlot buffer (15 min, room temperature), then incubated overnight at 4°C with the primary anti-His-Tag antibody (Abcam) diluted 1:1000 in blocking buffer. Membranes were subjected to five washes of five minutes each in 1X PBS/0.05% Tween-20, then incubated for two hours at room temperature with anti-mouse HRP secondary antibody (Cell Signal), diluted 1:1000 in blocking buffer. Finally, membranes were washed five times again and incubated with SuperSignal West Pico-Plus ECL substrate according to the manufacturer's instructions. Membranes were imaged on a ChemiDoc MP system.

4.9.5. Calculation of ΔΔG predictions

ΔΔG predictions in this work were performed by **Nina Bozhanova**, using the resolved proteins structures 2WUR (avGFP) (Shinobu et al., 2010), 2HPW (cgreGFP) (Malikova et al., 2011), 2G3O (ppluGFP2) (Wilmann et al., 2006), and 7LG4 (amacGFP and amacGFP:V11L) (this work). Full methodology is described in Gonzalez Somermeyer et al., 2022.

Briefly: the first chain from each structure was minimized using Rosetta Relax with constrains to starting coordinates. 50 models were generated per protein, and the one with the lowest total score was selected. Chromophores were treated as non-canonical amino acids and their geometry was optimized using Gaussian density functional theory. ΔΔG predictions were calculated using Rosetta's `ddg_monomer`, for all mutations except nonsense mutations, chromophore mutations, and mutations at sites absent from the resolved crystal structure. Rosetta version 3.10 was used throughout.

4.10. Machine learning

All machine learning work described in this section was performed by **Aubin Fleiss** and **Ekaterina Putintseva**. Full details on the methods used can be found in Gonzalez Somermeyer et al., 2022, and the code is publicly available on the `Orthologous_GFP_Fitness_Peaks` GitHub repository.

4.10.1. Modeling of local landscapes

Separately for amacGFP (including amacGFP:V11L), cgreGFP, and ppluGFP2, data was split randomly into sets for training (60%), validation (20%), and testing (20%). Neural networks were trained on protein sequences represented by one-hot encoding (a binary encoding consisting of, for each position in the protein, a vector of length 20 where the relevant amino acid is represented by a 1 and the nineteen absent amino acids are represented by zeros). Protein sequences in the training set were paired with their respective fitness (fluorescence) values, and the neural networks were tasked with predicting fluorescence from input sequences. The Keras software package was used to build all ML models, and model goodness was judged by the coefficient of determination (R^2) between known and predicted fitness values of genotypes in the validation set.

The following models of increasing complexity were trained on each dataset:

1. ***Linear models***. Neural networks with one input layer, and one layer consisting of a single neuron with a linear activation function. These were trained for 30 epochs with the task of minimizing the difference between real and predicted fluorescence values, with loss being measured by the mean square error. To prevent overfitting, the validation loss was monitored with a patience of 10 epochs. These models output a sum of individual mutation effects on fitness (i.e. fitness potential).
2. ***Models with sigmoid output node***. As above, but with the addition of a single neuron with a sigmoid activation function.
3. ***Models with output subnetworks***. As the linear models in (1), with the addition of an output subnetwork consisting of ten neurons with sigmoid activation functions, followed by a final linear output node. The hidden layer of ten sigmoid nodes were able to effectively transform the fitness potential (i.e. the output from the first, linear neuron layer) into accurate fluorescence values.
4. ***Final optimized architecture***. The final neural nets consisted of the input layer, a first hidden layer of neurons with linear activation, a second hidden layer of neurons with sigmoid activation, and one linear output node. A wide array of networks with this general architecture were tested, differing from each other by the number of neurons (1, 10, 20, 50, 100, or 200, selected randomly) in the two hidden layers. These networks also contained a Monte Carlo dropout layer after each hidden layer,

with the aim of minimizing overfitting as well as allowing the models to output fluorescence values with uncertainty estimates. The various networks were trained for 10 epochs, after which the best-performing one for each gene (i.e. the one with the smallest mean square error on the validation data) was trained for up to 30 epochs more. The training and validation sets used on the optimized models were filtered in order to ensure fairness: genotypes were only included in the training data if all their constituent mutations were present in, at minimum, a total of 10 genotypes (this ensured that enough data on each mutation was available for the models to make an informed opinion), and the validation data was filtered so as to not contain any mutations not present in the training data.

Furthermore, independent models for each gene with 1, 10, and 100 leaky ReLu nodes were also generated, using 90% of the genotype-to-phenotype data for training and 10% for validation. These models were trained for up to 500 epochs while minimizing mean square error loss. Overfitting was avoided in the same way as for the linear models described above. These models were used as independent *a posteriori* fluorescence predictors, to double-check artificially generated sequences predicted to be fluorescent (see below).

4.10.2. Generation of novel protein sequences predicted to fluoresce

Artificial protein sequences were generated following a genetic algorithm approach, and the optimized neural networks described above were queried as to the predicted fluorescence value of the generated sequence. Candidate sequences with a given target number of mutations were created as follows:

1. ***Initialization***. The first round begins with 50 wild-type protein sequences.
2. ***Recombination and mutation events***. Half of the population of sequences is left as is. The other half is subjected to crossing over and/or mutation events:
    - ***2a. Crossing over***. Recombination (sequence exchange) occurred between pairs of sequences with a probability of 0.7. If and when it happened, the number of crossings (between 1 and 5) and their position along the sequence were chosen randomly.
    - ***2b. Mutagenesis***. Mutations were drawn randomly from a pool containing both mutant amino acid states and WT states (to allow for potential mutation reversion). The probability of suffering a mutation was 0.01–0.015 per amino acid, and if the target number of mutations (defined by the user) was exceeded, then a previously integrated mutation was reverted. The pool of potential mutant states was limited to mutations which a) had been observed at least 10 times in the neural net training data, and b) were observed from the data to have a median effect on fluorescence no worse than -0.1 (see: 4.7.6. Calculation of mutation effects and epistasis). This approach therefore excluded universally negative mutations, but did not limit the pool of potential mutations to those seemingly universally neutral. Furthermore, the pool was enriched in non-extant mutations (see: 2.4.1. Extant mutations are less likely to be deleterious), to avoid the algorithm converging toward sequences already known to be functional; non-extant states comprised 60% of states available to the algorithm.
3. ***Selection***. All sequences in the population above (mutated and/or recombined or not) were one-hot encoded and their fluorescence predicted by the optimized neural network (see: 4.10.1. Modeling of local landscapes). Their predicted fitness was set as the median of 20 calculations performed by the model. Protein sequences were sorted in descending order of predicted fluorescence, and the lower ranking sequences were culled. New WT sequences were added in order to keep the population size constant.
4. ***Rinse and repeat***. Steps (2) and (3) above were repeated for several generations. In order to avoid a loss of sequence diversity due to excessive number of generations, any given algorithm run was stopped once the median predicted fluorescence level of the population reached a plateau. Furthermore, the entire algorithm was repeated independently a minimum of 10 times, or until all the mutations available in the pool had been sampled.
5. ***Final candidate genotypes***. From among the final sequences generated, those with the desired number of mutations and whose fluorescence was predicted to be greater than the initial WT by both the optimized neural net and the ReLu *a posteriori* neural net (see: 4.10.1. Modeling of local

landscapes) were submitted as candidates for experimental validation.

Note: the numerical values listed above for the recombination rate and per-amino-acid mutation rate, as well as the ratio of WT to mutant states available in the pool and the number of generations, were adjusted empirically with the aim of reaching the targeted number of mutations per sequence while also maintaining high predicted fluorescence levels.

4.11. Experimental validation of novel gene sequences

4.11.1. Selection of ML-generated test sequences

Novel protein sequences for amacGFP, cgreGFP, and ppluGFP2 were computationally generated and predicted by machine learning models to fluoresce brightly (see: 4.10.2. Generation of novel protein sequences predicted to fluoresce). For each gene, sets of candidate sequences containing increasing numbers of mutations were created (6, 12, 18, 24, 30, 36, 42, and 48 mutations in the case of ppluGFP2 and cgreGFP; only up to 30 mutations for amacGFP; and eventually up to 84 mutations for cgreGFP).

For each gene and each category (number of mutations), we experimentally tested 12 protein sequences. In cases where more than 12 candidate sequences were available to choose from, we chose the 12 which were as different from each other as possible in order to maximize the tested sequence diversity. To this end, the hamming distance between each pair of sequences was calculated, and the top 12 candidates with the highest overall distance scores were selected.

Coding sequences corresponding to the chosen novel proteins were generated using custom Python scripts. All nucleotide sequences were codon-optimized for *E. coli* and free of any internal restriction sites that would make them incompatible with Golden Gate cloning (BsaI, BpiI, BsmBI) or even genome integration (SpeI, NotI). Finally, BsaI restriction sites were added to flank the coding sequences, for downstream Golden Gate cloning into expression vectors. Final nucleotide sequences were ordered as synthetic DNA fragments in 96-well plates from Twist Bioscience and IDT.

A note on synthetic DNA fragments

We ordered multiple 96-well plates of synthetic dsDNA gene fragments from Twist Bioscience and one plate of gBlocks from IDT. Genes from Twist tended to give a 100% correct sequence with the first tested clone in ~90% of cases, and in the remaining cases, the second sequenced clone was correct. In contrast, only ~68% of initial clones from IDT genes were correct, with a further ~12% being correct in the second clone, another ~12% requiring 3-5 clones, and the remaining requiring 6+ clones to be sequenced before identifying one with a perfect sequence. Out of all the clones with incorrect IDT sequences, 88% of them were incorrect due to an indel, while 12% contained point substitutions. The difference in error rates between the two companies, and the pervasiveness of indels in particular may be due to differences in (proprietary) synthesis methods.

4.11.2. Manual selection of top mutations in ppluGFP2

For the non-ML-generated ppluGFP2 test sequences containing 24 mutations, we first checked the diversity of mutations used in the ML-generated sequences; 175 unique mutations were observed to have been employed across the 12 ML ppluGFP2 24-mutation genotypes. To make comparisons fairer between genotypes constructed manually versus by ML in terms of available choice of mutations, we decided to also select 175 mutations from the dataset to serve as the pool to draw from. For this, we ordered all mutations measured in at least 5 backgrounds in the ppluGFP2 dataset according to their median observed effect across backgrounds, and selected the top 175 mutations. None of the "top 175" mutations had a negative median effect on fluorescence (**Figure 23a**). (However, it was not possible to limit the pool to "universally neutral" mutations, i.e. those never observed, in any background, to have a deleterious effect causing loss of fluorescence of over one standard deviations from the WT fluorescence; only 5 mutations fit such a stringent criteria.)

We generated 12 non-ML genotypes, each incorporating 24 mutations from the "top 175" pool. One of

the genotypes was manually constructed to contain precisely the top 24 mutations of the pool, the "best of the best" out of all mutations ever tested in the ppluGFP2 dataset. The remaining 11 genotypes were constructed by taking 24 mutations at random from the "top 175" pool.

Nucleotide sequences were generated and ordered as synthetic DNA, as described in section 4.11.1. Selection of ML-generated test sequences.

4.11.3. Fluorescence measurements of novel genes

All novel GFP sequences were cloned via Golden Gate cloning into a homemade expression vector under a constitutive T5 promoter and lambda T0 terminator. The homemade vector conferred zeocin resistance and contained an origin of replication derived from pBR322, resulting in a medium/low copy number of ~20 copies per cell.

Golden Gate reactions were carried out in 96-well PCR plates. Chemically competent XL10-Gold cells were transformed in chilled 96-well PCR plates and the heat shock was performed in a thermocycler. 96-well deep well plates were used for post-heat shock recovery. Cells were plated on LB agar/zeocin plates supplemented with autoclaved black drawing ink (Higgins) in an ink:media ratio of 1:100, in order to improve the contrast between the fluorescent signal from GFP-positive colonies and the LB itself (which autofluoresces green).

Plates were incubated overnight at 30°C and then overnight again at room temperature, as during library expression prior to FACS. On the second day, all genes could be visually confirmed to be functional or not. Either way, at least one colony per construct was sequence-confirmed via Sanger sequencing. Sequence-confirmed clones were stored as glycerol stocks, and also streaked side by side alongside WT constructs on fresh LB agar/zeocin/ink plates which were then incubated as above. Those plates were were then imaged with a Canon EOS 600D SLR camera under identical conditions (aperture 2.8, ISO 100, 0.8s exposure time); settings were selected to maximize visible fluorescent signal without resulting in any saturated pixels. FIJI was used to convert images to 8-bit and to extract the median pixel values for each streak; no other image parameters such as brightness or contrast were modified. These values were log-transformed and scaled to the range of values of the sorted libraries, in order to allow direct comparisons.

A note on the use of T5 promoter

The constitutive T5 promoter sequence, as in the commercial pQE30 vectors, is `AAATCATAAAAAATTTATTTGCTTTGTGAGCGGATAACAATTATAATAGATTCAATTGTGAGCGGATAACAATTTCACACA`. In ~3% of sequenced constructs, we observed a partial promoter deletion, resulting in `AAATCATAAAAAATTTATTTGCT----------TTGTGAGCGGATAACAATTTCACACA`.

Notably, if a promoter deletion occurred, it was always identical: the missing section was always `TTGTGAGCGGATAACAATTATAATAGATTCAA`. This exact T5 promoter deletion has been described in Kawe et al., 2011, and appears to occur as a result of lack of repression during key time periods of plasmid establishment. We strongly recommend avoiding headaches by avoiding this promoter in cases where medium-level, constitutive expression is desired. Note: while the T5 promoter was also used in mKate2-GFP sorted libraries, we do not expect this phenomenon to affect sorting results: any individual cells having suffered a spontaneous promoter deletion should be unable to express mKate2 properly, and will therefore fall outside the gate of interest.

4.12. List of materials

4.12.1. List of consumables and services

Antibiotics:

- **Zeocin** (InvivoGen #ant-zn-1p); 50 mg/ml stocks obtained by dissolving 1 g powder in 20 ml HEPES buffer; working concentration 50 µg/ml
- **Carbenicillin** (...); working concentration 100 µg/ml, interchangeable with ampicillin
- **Ampicillin** (ISTA media kitchen); working concentration 100 µg/ml
- **Kanamycin** (ISTA media kitchen); working concentration 50 µg/ml

- **Chloramphenicol** (ISTA media kitchen); working concentration 25 µg/ml
- **Spectinomycin** (ISTA media kitchen); working concentration 50 µg/ml

Antibodies:
- Anti-6X His-Tag Antibody HIS.H8 (Abcam #ab18184)
- Anti-Mouse IgG, HRP-Linked Antibody (Cell Signalling #7076)

Biological materials:
- *E. coli* strain **BW29655** (CGSC #7934)
- The **pSIM5** plasmid was kindly provided by Court lab
- E. cloni 10G Chemically Competent Cells (1 x 109 cfu/µg) (Lucigen #60107)
- Addgene plasmids and kits: **pX2-Cas9** (#85811), **SS9_RNA** (#71656), **MoClo Toolkit** (#1000000044), **CIDAR MoClo Parts Kit** (#1000000059)
- **BL21(DE3)** Competent *E. coli* (NEB #C2527I)
- Chemically competent **DH5α** (10^6 cfu/µg) and **XL10-Gold** (10^9 cfu/µg) prepared by the ISTA media kitchen

Buffers & media:
- **General solutions**: 50X TAE, 10X PBS, 5M NaCl, 50% Glycerol, Milli-Q water, 0.1M $NiSO_4$, 10M urea, were prepared by the ISTA media kitchen
- **Growth media**: LB agar, LB liquid media, and M9 liquid media were prepared by the ISTA media kitchen
- Any other solutions were homemade from other ingredients listed in this section

Chemicals:
- TopVision Agarose Tablets (Thermo Fisher #R2802)
- X-Gal Solution, ready-to-use (Thermo Fisher #R0941)
- Imidazole (Sigma Aldrich #12399-100G)
- L-(+)-Arabinose (Sigma Aldrich #A3256-25G)
- IPTG Molecular Biology Grade (Applichem #A4773,0005)
- Waterproof Drawing Ink, Black India (Higgins #44204)
- Protease Inhibitor Cocktail (Merck/Sigma Aldrich #P8340)

DNA oligos: sigma + microsynth
- NEB Unique Dual-Index Illumina Adapter primers obtained from VBCF

DNA purification, gel:
- EXTRACTME DNA Gel-Out Kit (Blirt #EM08, sadly discontinued)
- Monarch DNA Gel Extraction Kit (NEB #T1020S)
- GeneJET Gel Extraction Kit (Thermo Fisher #K0691)

DNA purification, genomic:
- Wizard Genomic DNA Purification Kit (Promega #A1125)

DNA purification, plasmid:
- EXTRACTME Plasmid Mini Kit (Blirt #Em01.1, sadly discontinued)
- ZymoPURE Plasmid Miniprep Kit (Zymo Research #D4211)
- PureYield Plasmid Midiprep System (Promega #A2492)
- GeneJET Plasmid Maxiprep Kit (Thermo Fisher #K0492)

Electroporation:
- Electroporation Cuvettes 1 mm Gap (VWR #732-1135)

Enzymes, polymerase:
- Encyclo Plus PCR Kit (Evrogen #PK101)
- Q5 High-Fidelity 2X Master Mix (NEB #M0492S)
- Q5 High-Fidelity DNA Polymerase (NEB #M0491S)
- OneTaq Quick-Load 2X Master Mix with Standard Buffer (NEB #M0486L)

- GeneMorph II Random Mutagenesis Kit (Agilent #200550)

Enzymes, restriction:
- Eco31I (BsaI) (Thermo Fisher #ER0291)
- BpiI (BbsI) (Thermo Fisher #ER1011)
- SpeI-HF (NEB #R3133S)
- NotI-HF (NEB #R3189S)

Enzymes, other:
- T4 DNA Ligase (Thermo Fisher #EL0011)

Molecular weight markers:
- GeneRuler 1 kb DNA Ladder (Thermo Fisher #SM0311)
- Precision Plus Protein Dual Color Standard (BioRad #1610374)

Petri dishes and other plates:
- 500 µl 96-Well V-Bottom Sterile Clear Assay Plates (Axygen #P-96-450V-C-S)
- Petri dish, square, 12x12 cm (Greiner Bio-One #688102)
- Petri dish, 92x16 mm (Sarstedt #82.1473)
- Generic 96-well clear flat-bottom plates
- Generic 96-well white qPCR plates

Pipette tips:
- All filter and non-filter tips (1250 µl, 200 µl, 12.5 µl) by Starlab

Protein gels and Western Blot:
- 4–20% Mini-PROTEAN TGX Precast Protein Gels (Biorad #4561095)
- ReadyBlue Protein Gel Stain (Sigma Aldrich #RSB-1L)
- 4X Laemmli Sample Buffer (BioRad #1610747)
- Immun-Blot PVDF Membrane (BioRad #1620174)
- EveryBlot Blocking Buffer (BioRad #12010020)
- SuperSignal West Pico PLUS Chemiluminescent Substrate (Thermo Scientific, #34579)

Protein purification:
- Ni Sepharose High Performance (GE Healthcare/Cytiva #17-5268-01)
- Econo-Pac Chromatography Columns (BioRad #7321010)

Sequencing, Sanger:
- Performed by Microsynth

Sequencing, NGS:
- MiSeq PE300 and HiSeqV4 SR100 runs performed by VBCF
- NovaSeq SR100 runs performed by VBCF or OIST sequencing facilities
- NovaSeq PE250 run performed by Microsynth

Synthetic DNA:
- Synthetic Gene Fragments (Twist Bioscience)
- gBlocks Gene Fragments (Integrated DNA Technologies)

Tubes, various:
- 0.2 ml 8-Strip PCR Tube, Individually Attached Flat Caps (StarLab #A1402-3700)
- 1.5 ml ClearView Snap-Cap microtubes (Sigma #T4816-250EA)
- 1.5 ml Graduated Tube Natural and EasyGrip Cap (StarLab #E1415-2230 and #E1480-0399)
- TubeSpin Bioreactor 50 (TPP #87050)
- Various generic 0.5 ml, 1.5 ml, 2 ml, 15 ml, and 50 ml tubes

4.12.2. List of oligos

As a general rule, oligos longer than 50 nt were ordered PAGE-purified.

238 *Fw: backbone for expression vector*
CGCGGGGAAGACGTACTAACTGACCTACTAGTAGCGG

239 *Rv: backbone for expression vector*
GGTGCAGAAGACATCTCCTGCAACCCACTAGTCTCT

292 *Rv: barcoded primer for mutagenic PCR*
ACTTCGGAAGACCAACCTNNNNNNNNNNNNNNNNNNNTGTTCCTAGGCGCCGCTCATCATCA

293 *Fw: primer for mutagenic PCR*
GATTCATTAATCACTCTGGAAGACCAAATG

321 *Fw: before GFP in mKate2 expression construct*
GCCCGGCGTCTACTATGTG

327 *Fw: E. coli genome, before 5' homology arm*
TCGCCATCTTGTCGAGGAAC

329 *Rv: E. coli genome, after 3' homology arm*
TGGCGGAACAGGCGTATATC

331 *Fw: homemade general-use destination vector*
AATCGGTCACTGTGGCACGT

332 *Rv: homemade general-use destination vector*
AAGAGTGCATCCTGCCGCAC

350 *Oligo filler for library circularization*
ATAAAGGTCTCAAGGTCGCCCTGAGCCGCTACTACCAATGAGAGACCAATAT

351 *Oligo filler for library circularization*
ATATTGGTCTCTCATTGGTAGTAGCGGCTCAGGGCGACCTTGAGACCTTTAT

356 *Rv: plasmid backbone of mKate2-GFP vector*
TAGACGTCAGGTGGCACTTTT

359 *Fw: plasmid backbone of mKate2-GFP vector*
GTGAGCAAAAGGCCAGCAAA

366 *TruSeq Universal Read 1 Adapter*
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

367-9 *Fw: amacGFP C-terminus + partial TruSeq Read 1 adapter*
CCCTACACGACGCTCTTCCGATCT[2-4N]GTGAAGTTCGAGGGCGACACACTG

370-2 *Rv: GFP (any) C-terminus + partial TruSeq Read 2 adapter*
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC[2-4N]ACCACAGAGTACTTCGTGGTCTCA

373-5 *Fw: GFP (any) N-terminus + partial TruSeq Read 1 adapter*
CCCTACACGACGCTCTTCCGATCT[2-4N]GATGATGAGCGGCGCCTAGGAACA

376-8 *Rv: amacGFP N-terminus + partial TruSeq Read 2 adapter*
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC[2-4N]GTCCATGCCCTTCAGCTCGATGCG

379-86 *Rv: TruSeq Indexed Read 2 Adapters*
CAAGCAGAAGACGGCATACGAGAT[6N]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

387 *Rv: post-barcodes region + partial TruSeq Read 2 adapter*
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCNNAACCGCCGAGGTCAGTTTCGCC

388 *Oligo filler for library circularization*
ATAAAGAAGACAAAGGTTGAGACCACGAAGTACTCTGTGGTCTCAAATGAAGTCTTCAATAT

389 *Oligo filler for library circularization*
ATATTGAAGACTTCATTTGAGACCACAGAGTACTTCGTGGTCTCAACCTTTGTCTTCTTTAT

421-3 *Rv: cgreGFP N-terminus + partial TruSeq Read 2 adapter*
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC[2-4N]TCCCCAGGATGTTGCCGTTTGACT

424-6 *Fw: cgreGFP C-terminus + partial TruSeq Read 1 adapter*
CCCTACACGACGCTCTTCCGATCT[2-4N]GTTCGACAATGACGGCCAGTACGA

427-30 *Rv: more TruSeq Indexed Read 2 Adapters (see 379-86)*

431-3 *Rv: ppluGFP2 N-terminus + partial TruSeq Read 2 adapter*
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCNNAGCCTGTGCCCACTACCTTGAAGTC

434-6 *Fw: ppluGFP C-terminus + partial TruSeq Read 1 adapter*
CCCTACACGACGCTCTTCCGATCTNNGCATTGAAAAGTACGAGGACGGCG

DI5 *Fw: i5 Illumina primer for dual-indexing*
AATGATACGGCGACCACCGAGATCTACAC[8N]ACACGACGCTCTTCCGATCT

DI7 *Rv: i7 Illumina primer for dual-indexing*
CAAGCAGAAGACGGCATACGAGAT[8N]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

5. References

Adan A, Alizada G, Kiraz Y, Baran Y, Nalbant A. **Flow cytometry: basic principles and applications.** *Crit Rev Biotechnol.* 2017 Mar;37(2):163-176. doi: 10.3109/07388551.2015.1128876. Epub 2016 Jan 14. PMID: 26767547.

Alieva NO, Konzen KA, Field SF, Meleshkevitch EA, Hunt ME, Beltran-Ramirez V, Miller DJ, Wiedenmann J, Salih A, Matz MV. **Diversity and evolution of coral fluorescent proteins.** *PLoS One.* 2008 Jul 16;3(7):e2680. doi: 10.1371/journal.pone.0002680. PMID: 18648549; PMCID: PMC2481297.

Andrews BT, Schoenfish AR, Roy M, Waldo G, Jennings PA. **The rough energy landscape of superfolder GFP is linked to the chromophore.** *J Mol Biol.* 2007 Oct 19;373(2):476-90. doi: 10.1016/j.jmb.2007.07.071. Epub 2007 Aug 15. PMID: 17822714; PMCID: PMC2695656.

Armougom F, Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, Keduas V, Notredame C. **Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee.** *Nucleic Acids Res.* 2006 Jul 1;34(Web Server issue):W604-8. doi: 10.1093/nar/gkl092. PMID: 16845081; PMCID: PMC1538866.

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. **Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection.** *Mol Syst Biol.* 2006;2:2006.0008. doi: 10.1038/msb4100050. Epub 2006 Feb 21. PMID: 16738554; PMCID: PMC1681482.

Bailey SF, Alonso Morales LA, Kassen R. **Effects of Synonymous Mutations beyond Codon Bias: The Evidence for Adaptive Synonymous Substitutions from Microbial Evolution Experiments.** *Genome Biol Evol.* 2021 Sep 1;13(9):evab141. doi: 10.1093/gbe/evab141. PMID: 34132772; PMCID: PMC8410137.

Banerjee S, Schenkelberg CD, Jordan TB, Reimertz JM, Crone EE, Crone DE, Bystroff C. **Mispacking and the Fitness Landscape of the Green Fluorescent Protein Chromophore Milieu.** *Biochemistry.* 2017 Feb 7;56(5):736-747. doi: 10.1021/acs.biochem.6b00800. Epub 2017 Jan 24. Erratum in: Biochemistry. 2023 Oct 3;62(19):2894. PMID: 28074648; PMCID: PMC6193456.

Bassalo MC, Garst AD, Halweg-Edwards AL, Grau WC, Domaille DW, Mutalik VK, Arkin AP, Gill RT. **Rapid and Efficient One-Step Metabolic Pathway Integration in E. coli.** *ACS Synth Biol.* 2016 Jul 15;5(7):561-8. doi: 10.1021/acssynbio.5b00187. Epub 2016 Apr 22. PMID: 27072506.

Bateson W, Saunders ER, Punnett RC, Hurst CC. **Experimental Studies in the Physiology of Heredity.** 1905 *Reports to the Evolution Committee of the Royal Society, Report II. London, UK: Harrison and Sons.*

Baumann D, Cook M, Ma L, Mushegian A, Sanders E, Schwartz J, Yu CR. **A family of GFP-like proteins with different spectral properties in lancelet Branchiostoma floridae.** *Biol Direct.* 2008 Jul 3;3:28. doi: 10.1186/1745-6150-3-28. PMID: 18598356; PMCID: PMC2467403.

Bazykin GA. **Changing preferences: deformation of single position amino acid fitness landscapes and evolution of proteins.** *Biol Lett.* 2015 Oct;11(10):20150315. doi: 10.1098/rsbl.2015.0315. PMID: 26445980; PMCID: PMC4650171.

Bershtein S, Segal M, Bekerman R, Tokuriki N, Tawfik DS. **Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein.** *Nature.* 2006 Dec 14;444(7121):929-32. doi: 10.1038/nature05385. Epub 2006 Nov 19. PMID: 17122770.

Bigman LS, Levy Y. **Stability Effects of Protein Mutations: The Role of Long-Range Contacts.** *J Phys Chem B.* 2018 Dec 13;122(49):11450-11459. doi: 10.1021/acs.jpcb.8b07379. Epub 2018 Sep 25. PMID: 30198717.

Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH. **Thermodynamic prediction of protein neutrality.** *Proc Natl Acad Sci U S A.* 2005 Jan 18;102(3):606-11. doi: 10.1073/pnas.0406744102. Epub 2005 Jan 11. PMID: 15644440; PMCID: PMC545518.

Bolton KL, Ptashkin RN, Gao T, Braunstein L, Devlin SM, Kelly D, Patel M, Berthon A, Syed A, Yabe M, Coombs CC, Caltabellotta NM, Walsh M, Offit K, Stadler Z, Mandelker D, Schulman J, Patel A, Philip J, Bernard E, Gundem G, Ossa JEA, Levine M, Martinez JSM, Farnoud N, Glodzik D, Li S, Robson ME, Lee C, Pharoah PDP, Stopsack KH, Spitzer B, Mantha S, Fagin J, Boucai L, Gibson CJ, Ebert BL, Young AL, Druley T, Takahashi K, Gillis N, Ball M, Padron E, Hyman DM, Baselga J, Norton L, Gardos S, Klimek VM, Scher H, Bajorin D, Paraiso E, Benayed R, Arcila ME, Ladanyi M, Solit DB, Berger MF, Tallman M, Garcia-Closas M, Chatterjee N, Diaz LA Jr, Levine RL, Morton LM, Zehir A, Papaemmanuil E. **Cancer therapy shapes the fitness landscape of clonal hematopoiesis.** *Nat Genet.* 2020 Nov;52(11):1219-1226. doi: 10.1038/s41588-020-00710-0. Epub 2020 Oct 26. PMID: 33106634; PMCID: PMC7891089.

Boyd D, Weiss DS, Chen JC, Beckwith J. **Towards single-copy gene expression systems making gene cloning physiologically relevant: lambda InCh, a simple Escherichia coli plasmid-chromosome shuttle system.** *J Bacteriol.* 2000 Feb;182(3):842-7. doi: 10.1128/JB.182.3.842-847.2000. PMID: 10633125; PMCID: PMC94354.

Brockwell DJ, Radford SE. **Intermediates: ubiquitous species on folding energy landscapes?** *Curr Opin Struct Biol.* 2007 Feb;17(1):30-7. doi: 10.1016/j.sbi.2007.01.003. Epub 2007 Jan 18. PMID: 17239580; PMCID: PMC2706323.

Canale AS, Cote-Hammarlof PA, Flynn JM, Bolon DN. **Evolutionary mechanisms studied through protein fitness landscapes.** *Curr Opin Struct Biol.* 2018 Feb;48:141-148. doi: 10.1016/j.sbi.2018.01.001. Epub 2018 Jan 30. PMID: 29351890.

Chan YH, Venev SV, Zeldovich KB, Matthews CR. **Correlation of fitness landscapes from three orthologous TIM barrels originates from sequence and structure constraints.** *Nat Commun.* 2017 Mar 6;8:14614. doi: 10.1038/ncomms14614. PMID: 28262665; PMCID: PMC5343507.

Chayot R, Montagne B, Mazel D, Ricchetti M. **An end-joining repair mechanism in Escherichia coli.** *Proc Natl Acad Sci U S A.* 2010 Feb 2;107(5):2141-6. doi: 10.1073/pnas.0906355107. Epub 2010 Jan 19. PMID: 20133858; PMCID: PMC2836643.

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. **Biopython: freely available Python tools for computational molecular biology and bioinformatics**. *Bioinformatics*. 2009 June;25(11):1422-1423. doi: 10.1093/bioinformatics/btp163.

Crowther GJ, Napuli AJ, Thomas AP, Chung DJ, Kovzun KV, Leibly DJ, Castaneda LJ, Bhandari J, Damman CJ, Hui R, Hol WG, Buckner FS, Verlinde CL, Zhang Z, Fan E, van Voorhis WC. **Buffer optimization of thermal melt assays of Plasmodium proteins for detection of small-molecule ligands.** *J Biomol Screen.* 2009 Jul;14(6):700-7. doi: 10.1177/1087057109335749. Epub 2009 May 21. PMID: 19470714; PMCID: PMC2819745.

Davidson EA, Windram OP, Bayer TS. **Building synthetic systems to learn nature's design principles.** *Adv Exp Med Biol.* 2012;751:411-29. doi: 10.1007/978-1-4614-3567-9_19. PMID: 22821469.

DeHaan LR, Van Tassel DL. **Useful insights from evolutionary biology for developing perennial grain crops.** *Am J Bot.* 2014 Oct;101(10):1801-19. doi: 10.3732/ajb.1400084. Epub 2014 Oct 8. PMID: 25326622.

de Visser JA, Hermisson J, Wagner GP, Ancel Meyers L, Bagheri-Chaichian H, Blanchard JL, Chao L, Cheverud JM, Elena SF, Fontana W, Gibson G, Hansen TF, Krakauer D, Lewontin RC, Ofria C, Rice SH, von Dassow G, Wagner A, Whitlock MC. **Perspective: Evolution and detection of genetic robustness.** *Evolution*. 2003 Sep;57(9):1959-72. doi: 10.1111/j.0014-3820.2003.tb00377.x. PMID: 14575319.

de Visser JA, Cooper TF, Elena SF. **The causes of epistasis.** Proc Biol Sci. 2011 Dec 22;278(1725):3617-24. doi: 10.1098/rspb.2011.1537. Epub 2011 Oct 5. PMID: 21976687; PMCID: PMC3203509.

Dou J, Vorobieva AA, Sheffler W, Doyle LA, Park H, Bick MJ, Mao B, Foight GW, Lee MY, Gagnon LA, Carter L, Sankaran B, Ovchinnikov S, Marcos E, Huang PS, Vaughan JC, Stoddard BL, Baker D. **De novo design of a fluorescence-activating β-barrel.** *Nature*. 2018 Sep;561(7724):485-491. doi: 10.1038/s41586-018-0509-0. Epub 2018 Sep 12. PMID: 30209393; PMCID: PMC6275156.

Edgar RC. **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res.* 2004 Mar 19;32(5):1792-7. doi: 10.1093/nar/gkh340. PMID: 15034147; PMCID: PMC390337.

Eyre-Walker A, Keightley PD. **The distribution of fitness effects of new mutations.** *Nat Rev Genet.* 2007 Aug;8(8):610-8. doi: 10.1038/nrg2146. PMID: 17637733.

Evdokimov AG, Pokross ME, Egorov NS, Zaraisky AG, Yampolsky IV, Merzlyak EM, Shkoporov AN, Sander I, Lukyanov KA, Chudakov DM. **Structural basis for the fast maturation of Arthropoda green fluorescent protein.** *EMBO Rep.* 2006 Oct;7(10):1006-12. doi: 10.1038/sj.embor.7400787. Epub 2006 Aug 25. PMID: 16936637; PMCID: PMC1618374.

Flynn J, Samant N, Schneider-Nachum G, Tenzin T, Bolon DNA. **Mutational fitness landscape and drug resistance.** *Curr Opin Struct Biol.* 2023 Feb;78:102525. doi: 10.1016/j.sbi.2022.102525. Epub 2023 Jan 6. PMID: 36621152; PMCID: PMC10243218.

Fourrage C, Swann K, Gonzalez Garcia JR, Campbell AK, Houliston E. **An endogenous green fluorescent protein-photoprotein pair in Clytia hemisphaerica eggs shows co-targeting to mitochondria and efficient bioluminescence energy transfer.** *Open Biol.* 2014 Apr 9;4(4):130206. doi: 10.1098/rsob.130206. PMID: 24718596; PMCID: PMC4043110.

Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, Fields S. **High-resolution mapping of protein sequence-function relationships.** *Nat Methods.* 2010 Sep;7(9):741-6. doi: 10.1038/nmeth.1492. Epub 2010 Aug 15. PMID: 20711194; PMCID: PMC2938879.

Fragata I, Blanckaert A, Dias Louro MA, Liberles DA, Bank C. **Evolution in the light of fitness landscape theory.** *Trends Ecol Evol.* 2019 Jan;34(1):69-82. doi: 10.1016/j.tree.2018.10.009. Epub 2018 Dec 21. PMID: 30583805.

Gong LI, Suchard MA, Bloom JD. **Stability-mediated epistasis constrains the evolution of an influenza protein.** *Elife*. 2013 May 14;2:e00631. doi: 10.7554/eLife.00631. PMID: 23682315; PMCID: PMC3654441.

Gonzalez Somermeyer L, Fleiss A, Mishin AS, Bozhanova NG, Igolkina AA, Meiler J, Alaball Pujol ME, Putintseva EV, Sarkisyan KS, Kondrashov FA. **Heterogeneity of the GFP fitness landscape and data-driven protein design.** *Elife*. 2022 May 5;11:e75842. doi: 10.7554/eLife.75842. PMID: 35510622; PMCID: PMC9119679.

Greenbury SF, Schaper S, Ahnert SE, Louis AA. **Genetic Correlations Greatly Increase Mutational Robustness and Can Both Reduce and Enhance Evolvability.** *PLoS Comput Biol.* 2016 Mar 3;12(3):e1004773. doi: 10.1371/journal.pcbi.1004773. PMID: 26937652; PMCID: PMC4777517.

Gupta A, Zaman L, Strobel HM, Gallie J, Burmeister AR, Kerr B, Tamar ES, Kishony R, Meyer JR. **Host-parasite coevolution promotes**

innovation through deformations in fitness landscapes. *Elife*. 2022 Jul 6;11:e76162. doi: 10.7554/eLife.76162. PMID: 35793223; PMCID: PMC9259030.

Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E, Methé B, DeSantis TZ; Human Microbiome Consortium; Petrosino JF, Knight R, Birren BW. **Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons.** *Genome Res.* 2011 Mar;21(3):494-504. doi: 10.1101/gr.112730.110. Epub 2011 Jan 6. PMID: 21212162; PMCID: PMC3044863.

Hartl FU, Hayer-Hartl M. **Converging concepts of protein folding in vitro and in vivo.** *Nat Struct Mol Biol.* 2009 Jun;16(6):574-81. doi: 10.1038/nsmb.1591. PMID: 19491934.

Hartman EC, Tullman-Ercek D. **Learning from protein fitness landscapes: a review of mutability, epistasis, and evolution.** *Curr. Opin. Sys. Biol.* 2019(14):25-31. ISSN 2452-3100. doi: 10.1016/j.coisb.2019.02.006.

Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, Verkuil R, Tran VQ, Deaton J, Wiggert M, Badkundri R, Shafkat I, Gong J, Derry A, Molina RS, Thomas N, Khan Y, Mishra C, Kim C, Bartie LJ, Nemeth M, Hsu PD, Sercu T, Candido S, Rives A. **Simulating 500 million years of evolution with a language model.** *bioRxiv* 2024.07.01.600583; doi: https://doi.org/10.1101/2024.07.01.600583

Hietpas RT, Jensen JD, Bolon DN. **Experimental illumination of a fitness landscape.** *Proc Natl Acad Sci U S A.* 2011 May 10;108(19):7896-901. doi: 10.1073/pnas.1016024108. Epub 2011 Apr 4. PMID: 21464309; PMCID: PMC3093508.

Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CP, Springer M, Sander C, Marks DS. **Mutation effects predicted from sequence co-variation.** *Nat Biotechnol.* 2017 Feb;35(2):128-135. doi: 10.1038/nbt.3769. Epub 2017 Jan 16. PMID: 28092658; PMCID: PMC5383098.

Huang S. **Genetic and non-genetic instability in tumor progression: link between the fitness landscape and the epigenetic landscape of cancer cells.** *Cancer Metastasis Rev.* 2013 Dec;32(3-4):423-48. doi: 10.1007/s10555-013-9435-7. PMID: 23640024.

Ignatova Z, Gierasch LM. **Monitoring protein stability and aggregation in vivo by real-time fluorescent labeling.** *Proc Natl Acad Sci U S A.* 2004 Jan 13;101(2):523-8. doi: 10.1073/pnas.0304533101. Epub 2003 Dec 30. PMID: 14701904; PMCID: PMC327180.

Ishida Y, David VA, Eizirik E, Schäffer AA, Neelam BA, Roelke ME, Hannah SS, O'Brien SJ, Menotti-Raymond M. **A homozygous single-base deletion in MLPH causes the dilute coat color phenotype in the domestic cat.** *Genomics.* 2006 Dec;88(6):698-705. doi: 10.1016/j.ygeno.2006.06.006. Epub 2006 Jul 24. PMID: 16860533.

Iverson SV, Haddock TL, Beal J, Densmore DM. **CIDAR MoClo: Improved MoClo Assembly Standard and New E. coli Part Library Enable Rapid Combinatorial Design for Synthetic and Traditional Biology.** *ACS Synth Biol.* 2016 Jan 15;5(1):99-103. doi: 10.1021/acssynbio.5b00124. Epub 2015 Nov 4. PMID: 26479688.

Jacquier H, Birgy A, Le Nagard H, Mechulam Y, Schmitt E, Glodt J, Bercot B, Petit E, Poulain J, Barnaud G, Gros PA, Tenaillon O. **Capturing the mutational landscape of the beta-lactamase TEM-1.** *Proc Natl Acad Sci U S A.* 2013 Aug 6;110(32):13067-72. doi: 10.1073/pnas.1215206110. Epub 2013 Jul 22. PMID: 23878237; PMCID: PMC3740883.

Jiang L, Mishra P, Hietpas RT, Zeldovich KB, Bolon DN. **Latent effects of Hsp90 mutants revealed at reduced expression levels.** *PLoS Genet.* 2013 Jun;9(6):e1003600. doi: 10.1371/journal.pgen.1003600. Epub 2013 Jun 27. PMID: 23825969; PMCID: PMC3694843.

Johnson MS, Martsul A, Kryazhimskiy S, Desai MM. **Higher-fitness yeast genotypes are less robust to deleterious mutations.** *Science*. 2019 Oct 25;366(6464):490-493. doi: 10.1126/science.aay4199. PMID: 31649199; PMCID: PMC7204892.

Johnston KE, Almhjell PJ, Watkins-Dulaney EJ, Liu G, Porter NJ, Yang J, Arnold FH. **A combinatorially complete epistatic fitness landscape in an enzyme active site.** *Proc Natl Acad Sci U S A.* 2024 Aug 6;121(32):e2400439121. doi: 10.1073/pnas.2400439121. Epub 2024 Jul 29. PMID: 39074291; PMCID: PMC11317637.

Kaplan J. **The end of the adaptive landscape metaphor?** *Biology and Philosophy*, 2008 Vol. 23(5), pp. 625-38. DOI: 10.1007/s10539-008-9116-z.

Kashimoto R, Hisata K, Shinzato C, Satoh N, Shoguchi E. **Expansion and Diversification of Fluorescent Protein Genes in Fifteen *Acropora* Species during the Evolution of Acroporid Corals.** *Genes (Basel).* 2021 Mar 11;12(3):397. doi: 10.3390/genes12030397. PMID: 33799612; PMCID: PMC8001845.

Kellogg EH, Leaver-Fay A, Baker D. **Role of conformational sampling in computing mutation-induced changes in protein structure and stability.** *Proteins.* 2011 Mar;79(3):830-8. doi: 10.1002/prot.22921. Epub 2010 Dec 3. PMID: 21287615; PMCID: PMC3760476.

Kondrashov AS, Sunyaev S, Kondrashov FA. **Dobzhansky-Muller incompatibilities in protein evolution.** *Proc Natl Acad Sci U S A.* 2002 Nov 12;99(23):14878-83. doi: 10.1073/pnas.232565499. Epub 2002 Oct 28. PMID: 12403824; PMCID: PMC137512.

Kondrashov DA, Kondrashov FA. **Topological features of rugged fitness landscapes in sequence space.** *Trends Genet.* 2015 Jan;31(1):24-33. doi: 10.1016/j.tig.2014.09.009. Epub 2014 Oct 15. PMID: 25438718.

Kulathinal RJ, Bettencourt BR, Hartl DL. **Compensated deleterious mutations in insect genomes.** *Science*. 2004 Nov 26;306(5701):1553-4. doi: 10.1126/science.1100522. Epub 2004 Oct 21. PMID: 15498973.

Kumar A, Natarajan C, Moriyama H, Witt CC, Weber RE, Fago A, Storz JF. **Stability-Mediated Epistasis Restricts Accessible Mutational Pathways in the Functional Evolution of Avian Hemoglobin.** *Mol Biol Evol.* 2017 May 1;34(5):1240-1251. doi: 10.1093/molbev/msx085. PMID: 28201714; PMCID: PMC5400398.

Labas YA, Gurskaya NG, Yanushevich YG, Fradkov AF, Lukyanov KA, Lukyanov SA, Matz MV. **Diversity and evolution of the green fluorescent protein family.** *Proc Natl Acad Sci U S A.* 2002 Apr 2;99(7):4256-61. doi: 10.1073/pnas.062552299. PMID: 11929996; PMCID: PMC123635.

Lambert GG, Depernet H, Gotthard G, Schultz DT, Navizet I, Lambert T, Adams SR, Torreblanca-Zanca A, Chu M, Bindels DS, Levesque V, Nero Moffatt J, Salih A, Royant A, Shaner NC. **Aequorea's secrets revealed: New fluorescent proteins with unique properties for bioimaging and biosensing.** *PLoS Biol.* 2020 Nov 2;18(11):e3000936. doi: 10.1371/journal.pbio.3000936. PMID: 33137097; PMCID: PMC7660908.

Leclère L, Horin C, Chevalier S, Lapébie P, Dru P, Peron S, Jager M, Condamine T, Pottin K, Romano S, Steger J, Sinigaglia C, Barreau C, Quiroga Artigas G, Ruggiero A, Fourrage C, Kraus JEM, Poulain J, Aury JM, Wincker P, Quéinnec E, Technau U, Manuel M, Momose T, Houliston E, Copley RR. **The genome of the jellyfish Clytia hemisphaerica and the evolution of the cnidarian life-cycle.** *Nat Ecol Evol.* 2019 May;3(5):801-810. doi: 10.1038/s41559-019-0833-2. Epub 2019 Mar 11. PMID: 30858591.

Lee PY, Costumbrado J, Hsu CY, Kim YH. **Agarose gel electrophoresis for the separation of DNA fragments.** *J Vis Exp.* 2012 Apr 20;(62):3923. doi: 10.3791/3923. PMID: 22546956; PMCID: PMC4846332.

Lisanza SL, Gershon JM, Tipps S, Arnoldt L, Hendel S, Sims JN, Li X, Baker D. **Joint Generation of Protein Sequence and Structure with RoseTTAFold Sequence Space Diffusion.** *bioRxiv* 2023.05.08.539766; doi: https://doi.org/10.1101/2023.05.08.539766

Lozovsky ER, Chookajorn T, Brown KM, Imwong M, Shaw PJ, Kamchonwongpaisan S, Neafsey DE, Weinreich DM, Hartl DL. **Stepwise acquisition of pyrimethamine resistance in the malaria parasite.** *Proc Natl Acad Sci U S A.* 2009 Jul 21;106(29):12025-30. doi: 10.1073/pnas.0905922106. Epub 2009 Jul 8. PMID: 19587242; PMCID: PMC2715478.

Luo WX, Cheng T, Guan BQ, Li SW, Miao J, Zhang J, Xia NS. **Variants of green fluorescent protein GFPxm.** *Mar Biotechnol (NY).* 2006 Sep-Oct;8(5):560-6. doi: 10.1007/s10126-006-6006-8. Epub 2006 Jul 3. PMID: 17072681.

Mackay TF. **Epistasis and quantitative traits: using model organisms to study gene-gene interactions.** *Nat Rev Genet.* 2014 Jan;15(1):22-33. doi: 10.1038/nrg3627. Epub 2013 Dec 3. PMID: 24296533; PMCID: PMC3918431.

Malikova NP, Visser NV, van Hoek A, Skakun VV, Vysotski ES, Lee J, Visser AJ. **Green-fluorescent protein from the bioluminescent jellyfish Clytia gregaria is an obligate dimer and does not form a stable complex with the Ca(2+)-discharged photoprotein clytin.** *Biochemistry.* 2011 May 24;50(20):4232-41. doi: 10.1021/bi101671p. Epub 2011 Apr 27. PMID: 21425831.

Markova SV, Burakova LP, Frank LA, Golz S, Korostileva KA, Vysotski ES. **Green-fluorescent protein from the bioluminescent jellyfish Clytia gregaria: cDNA cloning, expression, and characterization of novel recombinant protein.** *Photochem Photobiol Sci.* 2010 Jun;9(6):757-65. doi: 10.1039/c0pp00023j. Epub 2010 May 5. PMID: 20442953.

Melamed D, Young DL, Gamble CE, Miller CR, Fields S. **Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein.** *RNA.* 2013 Nov;19(11):1537-51. doi: 10.1261/rna.040709.113. Epub 2013 Sep 24. PMID: 24064791; PMCID: PMC3851721.

Nowak MA, Boerlijst MC, Cooke J, Smith JM. **Evolution of genetic redundancy.** *Nature.* 1997 Jul 10;388(6638):167-71. doi: 10.1038/40618. PMID: 9217155.

Ogden PJ, Kelsic ED, Sinai S, Church GM. **Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design.** *Science.* 2019 Nov 29;366(6469):1139-1143. doi: 10.1126/science.aaw2900. PMID: 31780559; PMCID: PMC7197022.

Olson CA, Wu NC, Sun R. **A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain.** *Curr Biol.* 2014 Nov 17;24(22):2643-51. doi: 10.1016/j.cub.2014.09.072. Epub 2014 Oct 16. PMID: 25455030; PMCID: PMC4254498.

O'Maille PE, Malone A, Dellas N, Andes Hess B Jr, Smentek L, Sheehan I, Greenhagen BT, Chappell J, Manning G, Noel JP. **Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases.** *Nat Chem Biol.* 2008 Oct;4(10):617-23. doi: 10.1038/nchembio.113. Epub 2008 Sep 7. PMID: 18776889; PMCID: PMC2664519.

Orr HA. **The population genetics of speciation: the evolution of hybrid incompatibilities.** *Genetics.* 1995 Apr;139(4):1805-13. doi: 10.1093/genetics/139.4.1805. PMID: 7789779; PMCID: PMC1206504.

Orr HA, Turelli M. **The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities.** *Evolution.* 2001 Jun;55(6):1085-94. doi: 10.1111/j.0014-3820.2001.tb00628.x. PMID: 11475044.

Orr HA. **Fitness and its role in evolutionary genetics.** *Nat Rev Genet.* 2009 Aug;10(8):531-9. doi: 10.1038/nrg2603. PMID: 19546856; PMCID: PMC2753274.

Palmer AC, Toprak E, Baym M, Kim S, Veres A, Bershtein S, Kishony R. **Delayed commitment to evolutionary fate in antibiotic resistance fitness landscapes.** *Nat Commun.* 2015 Jun 10;6:7385. doi: 10.1038/ncomms8385. PMID: 26060115; PMCID: PMC4548896.

Pérez-Losada M, Arenas M, Galán JC, Palero F, González-Candelas F. **Recombination in viruses: mechanisms, methods of study, and evolutionary consequences.** *Infect Genet Evol.* 2015 Mar;30:296-307. doi: 10.1016/j.meegid.2014.12.022. Epub 2014 Dec 23. PMID: 25541518; PMCID: PMC7106159.

Podgornaia AI, Laub MT. **Pervasive degeneracy and epistasis in a protein-protein interface.** *Science*. 2015 Feb 6;347(6222):673-7. doi: 10.1126/science.1257360. PMID: 25657251.

Poelwijk FJ, Kiviet DJ, Weinreich DM, Tans SJ. **Empirical fitness landscapes reveal accessible evolutionary paths.** *Nature*. 2007 Jan 25;445(7126):383-6. doi: 10.1038/nature05451. PMID: 17251971.

Poelwijk FJ, Tănase-Nicola S, Kiviet DJ, Tans SJ. **Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes.** *J Theor Biol.* 2011 Mar 7;272(1):141-4. doi: 10.1016/j.jtbi.2010.12.015. Epub 2010 Dec 16. PMID: 21167837.

Poelwijk FJ, Socolich M, Ranganathan R. **Learning the pattern of epistasis linking genotype and phenotype in a protein.** *Nat Commun.* 2019 Sep 16;10(1):4213. doi: 10.1038/s41467-019-12130-8. PMID: 31527666; PMCID: PMC6746860.

Pokusaeva VO, Usmanova DR, Putintseva EV, Espinar L, Sarkisyan KS, Mishin AS, Bogatyreva NS, Ivankov DN, Akopyan AV, Avvakumov SY, Povolotskaya IS, Filion GJ, Carey LB, Kondrashov FA. **An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape.** *PLoS Genet.* 2019 Apr 10;15(4):e1008079. doi: 10.1371/journal.pgen.1008079. PMID: 30969963; PMCID: PMC6476524.

Poole ES, Brown CM, Tate WP. **The identity of the base following the stop codon determines the efficiency of in vivo translational termination in Escherichia coli.** *EMBO J.* 1995 Jan 3;14(1):151-8. doi: 10.1002/j.1460-2075.1995.tb06985.x. PMID: 7828587; PMCID: PMC398062.

Reddy G, Liu Z, Thirumalai D. **Denaturant-dependent folding of GFP.** *Proc Natl Acad Sci U S A*. 2012 Oct 30;109(44):17832-8. doi: 10.1073/pnas.1201808109. Epub 2012 Jul 9. PMID: 22778437; PMCID: PMC3497794.

Sanjuán R, Moya A, Elena SF. **The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus.** *Proc Natl Acad Sci U S A.* 2004 Jun 1;101(22):8396-401. doi: 10.1073/pnas.0400146101. Epub 2004 May 24. PMID: 15159545; PMCID: PMC420405.

Saona R, Kondrashov FA, Khudiakova KA. **Relation Between the Number of Peaks and the Number of Reciprocal Sign Epistatic Interactions.** *Bull Math Biol.* 2022 Jun 17;84(8):74. doi: 10.1007/s11538-022-01029-z. Erratum in: Bull Math Biol. 2023 Jan 21;85(3):17. PMID: 35713756; PMCID: PMC9205815.

Sarkisyan KS, Yampolsky IV, Solntsev KM, Lukyanov SA, Lukyanov KA, Mishin AS. **Tryptophan-based chromophore in fluorescent proteins can be anionic.** *Sci Rep.* 2012;2:608. doi: 10.1038/srep00608. Epub 2012 Aug 29. PMID: 22934131; PMCID: PMC3429880.

Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, Ivankov DN, Bozhanova NG, Baranov MS, Soylemez O, Bogatyreva NS, Vlasov PK, Egorov ES, Logacheva MD, Kondrashov AS, Chudakov DM, Putintseva EV, Mamedov IZ, Tawfik DS, Lukyanov KA, Kondrashov FA. **Local fitness landscape of the green fluorescent protein**. *Nature*. 2016 May 19;533(7603):397-401. doi: 10.1038/nature17995. Epub 2016 May 11. PMID: 27193686; PMCID: PMC4968632.

Shagin DA, Barsova EV, Yanushevich YG, Fradkov AF, Lukyanov KA, Labas YA, Semenova TN, Ugalde JA, Meyers A, Nunez JM, Widder EA, Lukyanov SA, Matz MV. **GFP-like proteins as ubiquitous metazoan superfamily: evolution of functional features and structural complexity.** *Mol Biol Evol.* 2004 May;21(5):841-50. doi: 10.1093/molbev/msh079. Epub 2004 Feb 12. PMID: 14963095.

Sharan SK, Thomason LC, Kuznetsov SG, Court DL. **Recombineering: a homologous recombination-based method of genetic engineering.** *Nat Protoc.* 2009;4(2):206-23. doi: 10.1038/nprot.2008.227. PMID: 19180090; PMCID: PMC2790811.

Shinobu A, Palm GJ, Schierbeek AJ, Agmon N. **Visualizing proton antenna in a high-resolution green fluorescent protein structure.** *J Am Chem Soc.* 2010 Aug 18;132(32):11093-102. doi: 10.1021/ja1010652. PMID: 20698675.

Schmidt-Küntzel A, Nelson G, David VA, Schäffer AA, Eizirik E, Roelke ME, Kehler JS, Hannah SS, O'Brien SJ, Menotti-Raymond M. **A domestic cat X chromosome linkage map and the sex-linked orange locus: mapping of orange, multiple origins and epistasis over nonagouti.** *Genetics*. 2009 Apr;181(4):1415-25. doi: 10.1534/genetics.108.095240. Epub 2009 Feb 2. PMID: 19189955; PMCID: PMC2666509.

Smith JM. **Natural selection and the concept of a protein space.** *Nature.* 1970 Feb 7;225(5232):563-4. doi: 10.1038/225563a0. PMID: 5411867.

Sorokina I, Mushegian AR, Koonin EV. **Is Protein Folding a Thermodynamically Unfavorable, Active, Energy-Dependent Process?** *Int J Mol Sci*. 2022 Jan 4;23(1):521. doi: 10.3390/ijms23010521. PMID: 35008947; PMCID: PMC8745595.

Starr TN, Thornton JW. **Epistasis in protein evolution.** *Protein Sci.* 2016 Jul;25(7):1204-18. doi: 10.1002/pro.2897. Epub 2016 Feb 28. PMID: 26833806; PMCID: PMC4918427.

Stoler N, Nekrutenko A. **Sequencing error profiles of Illumina sequencing instruments**. *NAR Genom Bioinform.* 2021 Mar 27;3(1):lqab019. doi: 10.1093/nargab/lqab019. PMID: 33817639; PMCID: PMC8002175.

Takahashi-Kariyazono S, Satta Y, Terai Y. **Genetic diversity of fluorescent protein genes generated by gene duplication and**

alternative splicing in reef-building corals. *Zoological Lett.* 2015 Jul 21;1:23. doi: 10.1186/s40851-015-0020-5. PMID: 26605068; PMCID: PMC4657232.

Tsien RY. **The green fluorescent protein.** *Annu Rev Biochem.* 1998;67:509-44. doi: 10.1146/annurev.biochem.67.1.509. PMID: 9759496.

Watson CJ, Papula AL, Poon GYP, Wong WH, Young AL, Druley TE, Fisher DS, Blundell JR. **The evolutionary dynamics and fitness landscape of clonal hematopoiesis.** *Science.* 2020 Mar 27;367(6485):1449-1454. doi: 10.1126/science.aay9333. PMID: 32217721.

Weinreich DM, Lan Y, Wylie CS, Heckendorn RB. **Should evolutionary geneticists worry about higher-order epistasis?** *Curr Opin Genet Dev.* 2013 Dec;23(6):700-7. doi: 10.1016/j.gde.2013.10.007. Epub 2013 Nov 27. PMID: 24290990; PMCID: PMC4313208.

Wilmann PG, Battad J, Petersen J, Wilce MC, Dove S, Devenish RJ, Prescott M, Rossjohn J. **The 2.1A crystal structure of copGFP, a representative member of the copepod clade within the green fluorescent protein superfamily.** *J Mol Biol.* 2006 Jun 16;359(4):890-900. doi: 10.1016/j.jmb.2006.04.002. Epub 2006 Apr 25. PMID: 16697009.

Wloch DM, Szafraniec K, Borts RH, Korona R. **Direct estimate of the mutation rate and the distribution of fitness effects in the yeast *Saccharomyces cerevisiae.*** *Genetics.* 2001 Oct;159(2):441-52. doi: 10.1093/genetics/159.2.441. PMID: 11606524; PMCID: PMC1461830.

Wood TI, Barondeau DP, Hitomi C, Kassmann CJ, Tainer JA, Getzoff ED. **Defining the role of arginine 96 in green fluorescent protein fluorophore biosynthesis.** *Biochemistry.* 2005 Dec 13;44(49):16211-20. doi: 10.1021/bi051388j. PMID: 16331981.

Wrenbeck EE, Faber MS, Whitehead TA. **Deep sequencing methods for protein engineering and design.** *Curr Opin Struct Biol.* 2017 Aug;45:36-44. doi: 10.1016/j.sbi.2016.11.001. Epub 2016 Nov 22. PMID: 27886568; PMCID: PMC5440218.

Wright S. **The roles of mutation, inbreeding, crossbreeding and selection in evolution.** Proceedings of the Sixth International Congress of Genetics, 1932, pp. 356-366.

Wylie CS, Shakhnovich EI. **A biophysical protein folding model accounts for most mutational fitness effects in viruses.** *Proc Natl Acad Sci U S A.* 2011 Jun 14;108(24):9916-21. doi: 10.1073/pnas.1017572108. Epub 2011 May 24. PMID: 21610162; PMCID: PMC3116435.

Xia NS, Luo WX, Zhang J, Xie XY, Yang HJ, Li SW, Chen M, Ng MH. **Bioluminescence of Aequorea macrodactyla, a common jellyfish species in the East China Sea.** *Mar Biotechnol (NY).* 2002 Mar;4(2):155-62. doi: 10.1007/s10126-001-0081-7. PMID: 14961275.

Yue JX, Holland ND, Holland LZ, Deheyn DD. **The evolution of genes encoding for green fluorescent proteins: insights from cephalochordates (amphioxus).** *Sci Rep.* 2016 Jun 17;6:28350. doi: 10.1038/srep28350. PMID: 27311567; PMCID: PMC4911609.

Zeldovich KB, Chen P, Shakhnovich EI. **Protein stability imposes limits on organism complexity and speed of molecular evolution.** *Proc Natl Acad Sci U S A.* 2007 Oct 9;104(41):16152-7. doi: 10.1073/pnas.0705366104. Epub 2007 Oct 3. PMID: 17913881; PMCID: PMC2042177.

Zhang Z, Wang L, Gao Y, Zhang J, Zhenirovskyy M, Alexov E. **Predicting folding free energy changes upon single point mutations.** *Bioinformatics.* 2012 Mar 1;28(5):664-71. doi: 10.1093/bioinformatics/bts005. Epub 2012 Jan 11. PMID: 22238268; PMCID: PMC3289912.

Zhou L, Lei XH, Bochner BR, Wanner BL. **Phenotype microarray analysis of Escherichia coli K-12 mutants with deletions of all two-component systems.** *J Bacteriol.* 2003 Aug;185(16):4956-72. doi: 10.1128/JB.185.16.4956-4972.2003. PMID: 12897016; PMCID: PMC166450.