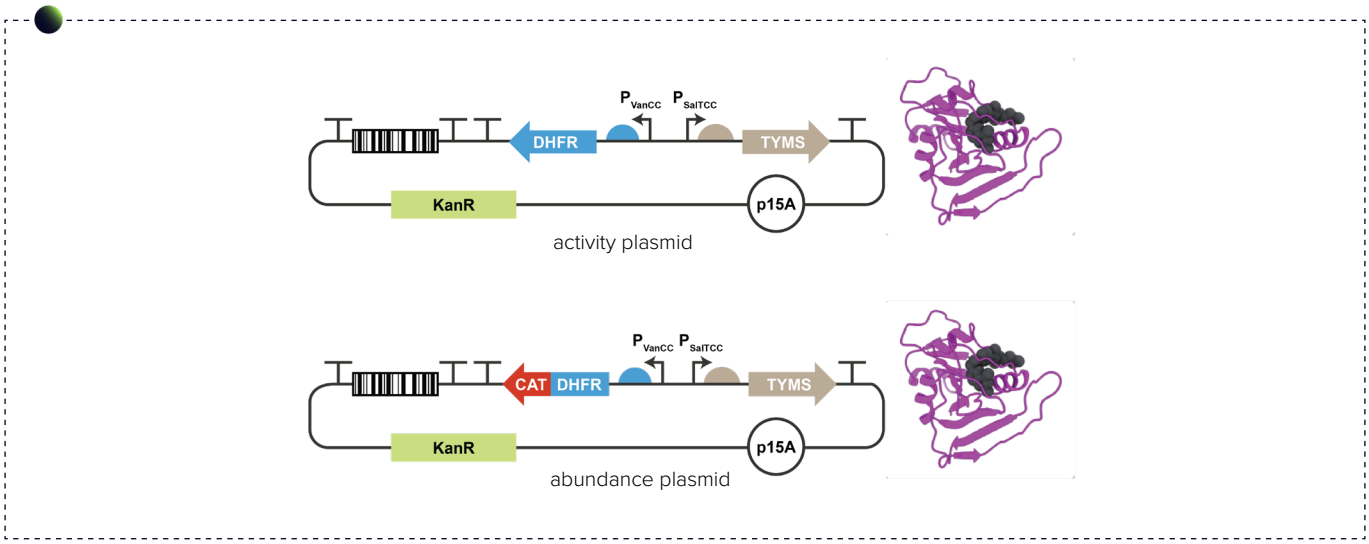
Reviewed

GROQ-seq Platform Expansion: Design of growth-coupled measurements of Dihydrofolate Reductase *in vivo* Biochemistry



A proposal for onboarding dihydrofolate reductase (DHFR) to the growth-based quantitative sequencing (GROQ-seq) platform.

- Links DHFR activity and abundance to growth using gene circuits.
- Measures k_{cat} and K_m in vivo from growth rate data.
- Uses calibration variants to enable quantitative inference.
- Captures single mutants, key double mutants, and random multimutants across several sublibraries of bacterial DHFR variants.

This work was supported by The Align Foundation, which receives philanthropic funding in part from Griffin Catalyst. The Align Foundation is a non-profit research organization operating under open science principles with the goal of improving science research with programmable experiments. The Align Foundation is working to accelerate community-driven science with the use of automated labs to pioneer robust data collection methods and curated, high-fidelity, public biological datasets amenable to machine learning.

DOI: 10.5281/zenodo.15277418

Contributors

Project Lead:

Kimberly A. Reynolds – The University of Texas Southwestern Medical Center

Additional Proposal Author:

Karolina Filipowska – The University of Texas Southwestern Medical Center

The Align Foundation:

Dana Cortade – Technical Project Manager

Peter Kelly – Co-Founder, Head of Science

Reviewers:

Benjamin Lehner – Wellcome Sanger Institute

Craig Markin – University of Manchester

David Ross – National Institute of Standards and Technology (NIST)

Justin B. Kinney – Cold Spring Harbor Laboratory

Neel H. Shah – Columbia University

Timothy Whitehead – University of Colorado Boulder

Willow Coyote-Maestas – University of California, San Francisco

Additional Acknowledgments:

We acknowledge the following people for their work on science communication:

Olesia Bushkova

Naomi Hagelund

Rachel Sevey

Overview

Our ability to predict enzyme sequence–activity relationships is constrained by limited biochemical data. To address this, we will establish a generalizable assay that permits inference of quantitative catalytic parameters (k_{cat}, K_m) from growth data in high throughput. We will use this assay to map sequence–activity encodings for the model metabolic enzyme dihydrofolate reductase (DHFR). Then, we will extend the assay to additional enzymes with key roles in metabolism and industrial biosynthesis. We anticipate our approach can be used to understand catalytic variation across homologs from different species, characterize variants associated with disease, and design enzymes of industrial and medicinal utility.

DOI: 10.5281/zenodo.15277418

Significance and Impact

Motivation and key challenges

Enzymes are DNA-encoded catalysts capable of accelerating chemical reaction rates up to 17 orders of magnitude[1]. Enzyme activity underpins all of cellular metabolism, and enzyme mutations play a central role in disease and evolutionary adaptation. A quantitative understanding of the relationship between enzyme sequence variation and catalytic function would improve our ability to (1) interpret disease-associated mutations, (2) engineer enzymes with modified function, and (3) understand the molecular origins of catalysis. However, existing computational tools that predict the biochemical effects of mutation show only limited quantitative power. We critically need large gold-standard experimental datasets that comprehensively map the effect of mutations on catalytic parameters (enzyme abundance [E], k_{cat}, and K_m) to develop predictive computational models and better understand the sequence encoding of catalysis.

Standard approaches for determining steady-state Michaelis Menten biochemical parameters require protein expression, purification, and *in vitro* characterization of enzyme velocity over numerous substrate concentrations. This relatively laborious process requires days to weeks of experiment time per mutant. Thus, many *in vitro* studies focus on a few carefully chosen mutations and cannot broadly explore the connection between sequence and catalysis. Recently, tools have become available for high-throughput *in vitro* biochemical measurements on microfluidic chips[2]. While very promising, these methods are currently limited to enzymes that can be coupled to a fluorescent reporter of activity and are not yet readily available to non-specialist labs. Moreover, because these assays are performed *in vitro*, the measured enzymatic parameters may differ from those in the cellular milieu.

A different strategy for rapid assessment of enzymatic function lies in growth-coupled assays. In these experiments, cell growth rate — under assay conditions that select for enzyme function— is used as a proxy for enzyme activity[3,4]. Because growth rate can readily be measured for thousands to millions of variants with next-generation sequencing, these Multiplexed Assays of Variant Effects (MAVEs) can be extraordinarily high throughput[5]. At present, limited availability of biochemical data means that many computational models for sequence–activity relationships are trained to predict fitness scores from these sorts of growth-based functional assays rather than true biochemical measurements[6–8]. However, the fitness score does not easily disambiguate between growth rate defects caused by variation in enzyme k_{cat}, K_m, or intracellular abundance. For example, an enzyme mutation with a deleterious fitness score may disrupt catalytic activity, unfold the protein, or some combination thereof. Moreover, the reported fitness score depends on specific conditions of the assay: selection strength, promoter strength, and strain genetic background can all impact the effect of a mutation on growth rate[9–11]. In our own prior work, we identified mutations of DHFR that switched from beneficial to deleterious for growth depending on *E. coli* strain background[4]. Given this, it is not obvious what computational models trained on these types of fitness data learn. Are the resulting models predicting activity, stability, or something else? It is also not obvious that models trained on fitness scores from one experimental dataset should generalize to predict mutational effects in a different strain or assay design.

To address the acute need for more direct characterization of enzyme catalytic function, we will develop a high-throughput *in vivo* assay that reports quantitative biochemical parameters–k_{cat}, K_m, and enzyme abundance [E]–rather than a single fitness-based measure. Recent work on two peptide binding proteins (PDZ and SH3) showed that measuring the growth rate effect of mutations in different assay conditions and genetic backgrounds can constrain inference of quantitative biophysical parameters (K_d, $\Delta\Delta G_{fold}$) for thousands of mutants[12]. Key elements of this approach included: (1) the use of two distinct assays to separately measure intracellular abundance and peptide binding and (2) the measurement of mutational effects in different genetic backgrounds to reveal effects that would otherwise be masked. Recent work has extended this strategy to disentangle the impact of mutations on stability and enzymatic function, but has yet to resolve quantitative biochemical parameters like k_{cat} and K_m[13]. Here, we will follow a new strategy to learn these parameters from growth rate data.

Assay concept

Our *in vivo* biochemistry workflow will consist of two assays to resolve mutational effects on intracellular abundance ([E]) and catalysis (k_{cat}, K_m) (**Fig. 1)**. First, to examine mutational effects on intracellular abundance, we will fuse our gene of interest (GOI) to an antibiotic resistance marker (e.g., chloramphenicol acyltransferase, CAT), and quantify growth rates under varied antibiotic concentrations (Fig. 1, gray panel). We refer to the GOI–CAT fusion construct as the abundance selection plasmid. This idea follows from a simple approach for selecting on protein solubility using CAT fusions, originally proposed by Maxwell *et al.*[14]. The basic concept is that well-expressed enzyme-CAT fusions should display higher chloramphenicol resistance than variants that misfold or show limited solubility. In our assay, the growth media conditions will be supplemented so that they do not select for enzymatic function, but instead only report on expression of the antibiotic resistance marker. Moreover, we will inoculate the cells at low density and limit selection in antibiotic to minimize the potential of shared resistance (e.g., mediated by some cells robustly expressing CAT and reducing the concentration of chloramphenicol in the culture). The resulting growth rate data as a function of antibiotic concentration will provide us with a measure of abundance ([E]).

Then, to quantify the effect of mutation on catalytic parameters, we will follow a strategy akin to *in vitro* Michaelis–Menten steady-state kinetics. The idea is to quantify enzyme velocity as a function of varied substrate concentration. We will titrate intracellular substrate abundance by varying the expression (and intracellular abundance) of the enzyme immediately upstream of our gene of interest (**Fig. 2A**). Practically, this will be implemented by including the gene encoding the upstream enzyme (UE) on the selection plasmid (alongside the GOI) and varying

DOI: 10.5281/zenodo.15277418

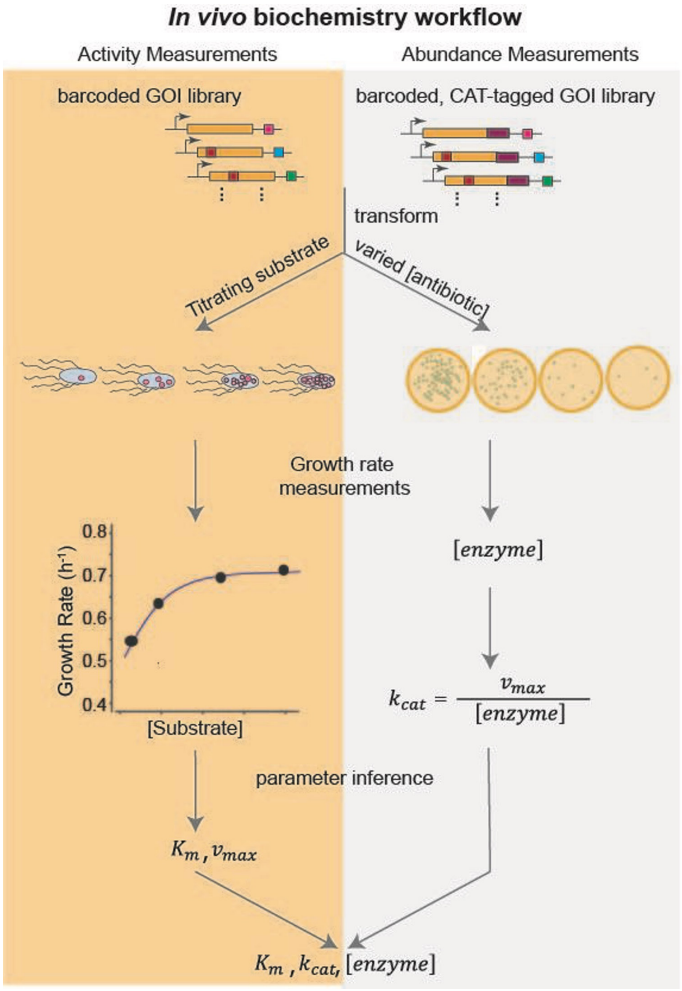


Figure 1: Schematic of the in vivo biochemistry assay.

DOI: 10.5281/zenodo.15277418

abundance with an inducible promoter (**Fig. 2B**). In the case of DHFR (our primary model system), the upstream enzyme is thymidylate synthase (TYMS, encoded by the gene *thyA*). We will make use of the Marionette system of high-dynamic-range inducible promoters to separately control the expression of DHFR and TYMS15. After consultation with Dr. David Ross, we propose to use the vanillic acid (P_{VanCC}) and sodium salicylate (P_{SalTTC}) promoters to control expression of DHFR (encoded by the gene *folA*) and TYMS (encoded by the gene *thyA*), respectively. These two promoters show minimal cross-talk upon induction. We refer to this dual-promoter, dual-enzyme construct as the function selection plasmid. We will measure the growth rate effect of mutations in the gene of interest (DHFR) while titrating the induction (and thus intracellular availability) of TYMS. Taking growth rate as a proxy for enzyme velocity, we will infer quantitative biochemical parameters (V_{max}, K_m) from the function selection data. Finally, we will use the abundance selection data to resolve V_{max} into the catalytic turnover (k_{cat}) and concentration ([E]). It is important to note that this approach will yield data on the K_m for DHF, but does not consider the K_m for the NADPH cofactor. Because NADPH is a critical cofactor for many reactions, we expect that any efforts to titrate intracellular abundance would have pleiotropic (and difficult to interpret) impacts on growth rate.

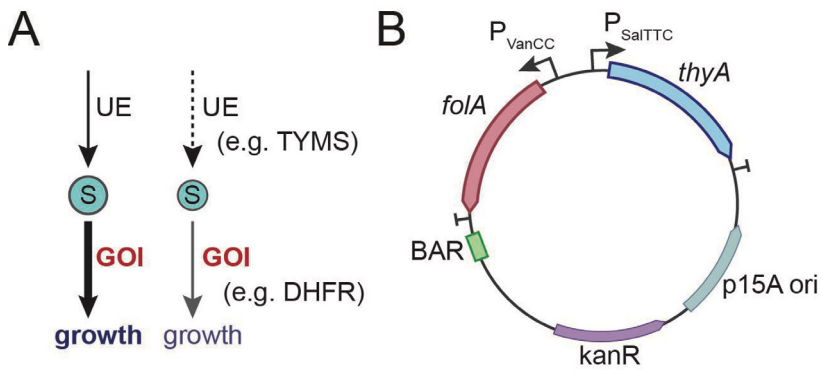


Figure 2: Varying upstream enzyme (UE) abundance to control intracellular substrate (S).

A) A schematic of the approach. Reducing UE abundance decreases intracellular substrate pools and accordingly impacts growth.

B) For the DHFR enzyme (encoded by *folA*), the UE is thymidylate synthase (encoded by *thyA*). Both *folA* and *thyA* are expressed from a single plasmid in our selection system. We will vary *thyA* abundance via induction of a sodium salicylate–responsive promoter. DHFR variants will be associated with unique DNA barcodes following the enzyme stop codon (see also Gene of Interest Region).

Model system

To develop our assay, we begin with DHFR, a well-studied, essential metabolic enzyme important to human health and disease[16]. We have selected this enzyme as the starting point for the biochemistry *in vivo* assay for three reasons.

First, a rich history of biochemistry and structural biology in this system provides excellent context for validating and interpreting our results. For example, we have curated a set of 44 mutations with *in vitro* measurements of k_{cat} and K_m from the literature, which we will use to test and refine our assay[9].

Second, DHFR is a frequent target of antibiotics (e.g., trimethoprim) and chemotherapeutics (e.g., methotrexate)[16]. A more complete understanding of sequence–activity relationships for this enzyme is important to understanding the evolution of drug resistance, and could aid in the design of more specific allosteric inhibitors.

Finally, DHFR is a model system for understanding the connection between long-range conformational change and catalysis[17]. *E. coli* DHFR catalysis proceeds through five chemical intermediates; this cycle is associated with micro-to millisecond conformational dynamics distributed throughout the enzyme[18]. Consistent with this dynamic view of catalysis, work from our group has shown that mutations throughout the enzyme can modify function[4,9], and that surfaces distal to the active site can be used to engineer new regulation at latent allosteric "hot spots"[19,20].

Yet despite this extensive prior work, it is unclear what pattern of mutations in DHFR tune specific biochemical properties such as k_{cat}, K_m, and stability. The number of positions, structural locations, and overlap among mutations that control these properties is unknown. The work proposed here will provide a quantitative map of how mutations throughout the enzyme tune specific catalytic properties.

DOI: 10.5281/zenodo.15277418

DHFR catalyzes the stereospecific reduction of 7,8-dihydrofolate (DHF) to 5,6,7,8-tetrahydrofolate (THF) using NADPH as a cofactor[18]. THF then serves as a carrier for activated one-carbon units in downstream metabolic processes, including the biosynthesis of purines, thymidine, methionine, and glycine. Consequently, DHFR activity is strongly linked to cell growth[16]. Under appropriate selective conditions, growth rate can serve as a proxy for enzyme velocity–the product of activity and abundance. The starting point for our new *in vivo* biochemistry method is a high-throughput growth-based assay previously established in our lab (**Fig. 3**)[20]. Our current assay displays a strong correlation between DHFR velocity and *E. coli* growth rate over four orders of magnitude, but like other MAVES, cannot discriminate between changes in k_{cat}, K_m, or abundance (**Fig. 3B**). We have used this assay to quantify the influence of genetic background on DHFR mutational tolerance; in total, we have quantified the growth rate effect of nearly all 3,021 *E. coli* DHFR single mutations (19 substitutions at 159 positions) under two different DHFR expression conditions (altered ribosome binding sites) and three altered genetic backgrounds (mutations in TYMS)—over 15,000 measurements total[4,9]. This assay and its accompanying data provide a foundation for now developing the *in vitro* biochemistry assay proposed here.

While the pilot-scale collection will focus on the well-studied enzyme DHFR—a system for which we benefit from prior biochemical data and structural information—we expect that the proposed assay can apply to any growth-linked enzyme. To estimate the potential diversity of reactions that our assay might access, we conducted an analysis of growth-linked enzymatic reactions in *E. coli*. There are 352 genes that are essential for growth in MOPS minimal media[21]. These growth-linked genes present a conservative starting point for our analysis (additional *E. coli* genes are also growth-linked but are non-lethal upon gene deletion). These 352 genes cover 261 unique Enzyme Commission (EC) numbers and 20% of the enzyme-catalyzed reactions in *E. coli*[22]. Thus, our assay has the potential to extend to hundreds of diverse reaction types across both primary and secondary metabolism. In the large-scale collection phase, we will generalize the assay to at least two other growth-linked enzymes in amino acid metabolism. As initial targets, we propose chorismate mutase (CM, encoded by the gene *pheA*), and 4-dihydroxy-tetrahydrodipicolinate synthase (DapA, *dapA*). We will additionally conduct pilot experiments to assess the feasibility of titrating intracellular substrate abundance for approximately 10 additional growth-linked enzymes from varied metabolic processes.

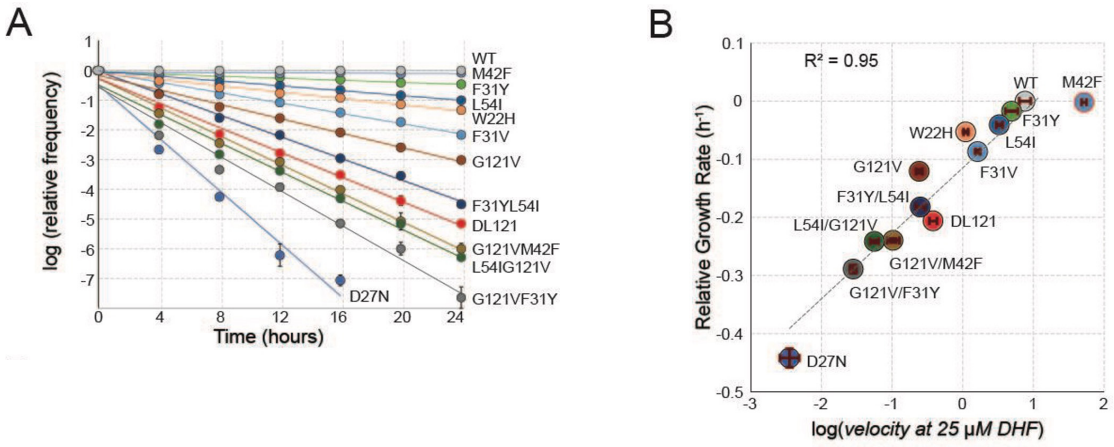


Figure 3: A growth-based assay for DHFR activity.

A) Logarithm of the relative allele frequency over time for 12 DHFR variants, as determined by next-generation sequencing. These mutants were selected to span a range of catalytic activities. Error bars indicate standard error across four replicates. The slope of each line of best fit provides the growth rate difference of a given mutant relative to wild-type (WT).

B) The correlation between relative growth rate as measured in (A) and log(velocity) as measured *in vitro*. Error bars in y indicate standard error over four replicates; error bars in x indicate standard deviation across triplicate *in vitro* measurements.

DOI: 10.5281/zenodo.15277418

A mathematical model linking growth rate to enzyme catalysis

We created a mathematical model to relate high-throughput growth rate measurements to enzyme catalytic parameters[9] (**Fig. 4**). The model consists of two equations: (1) a function relating DHFR and TYMS catalytic parameters to intracellular THF, and (2) a function relating intracellular THF to growth rate. By combining these two equations, one can compute growth rates for arbitrary variation in DHFR and TYMS k_{cat}, K_m, and abundance. Here, we briefly describe the model logic and mathematics.

To create a simplified, analytically solvable model, we reduced the larger folate pathway to a cycle in which DHFR and TYMS catalyze opposing oxidation and reduction reactions (Fig. 4A). In cells, DHFR produces tetrahydrofolate, and a number of other folate metabolic enzymes add and remove varied one-carbon modifications to this THF substrate. TYMS catalyzes the oxidation of 5,10-methylenetetrahydrofolate back to DHF during deoxythymidine synthesis and is the sole enzyme responsible for recycling the reduced folate pool back to the oxidized form. Abstracting the folate pathway to a two-enzyme cycle assumes that DHFR and TYMS dominate turnover of the DHF and THF pools, and reduced folates (the set of one-carbon-modified THF adducts) are considered as a single THF pool. This simplification allows us to solve for the intracellular concentration of THF as a function of DHFR and TYMS steady-state kinetics parameters ($k_{cat\text{-}DHFR}$,$K_{m\text{-}DHFR}$, $k_{cat\text{-}TYMS}$, $K_{m\text{-}TYMS}$), DHFR abundance, and TYMS abundance. The solution takes the form of the Goldbeter–Koshland equation, which was historically used to describe opposing kinase and phosphatase reaction cycles[23,24]. This relatively simplified model showed good correspondence to metabolomics measurements of intracellular THF (R^2 = 0.96, **Fig. 4B**). The next step is then to relate intracellular THF abundance to *E. coli* growth rate. Under the conditions of our experiment, we previously observed a hyperbolic dependence of growth rate on reduced folate abundance[25]. Thus, following a similar approach as Rodrigues *et al.*[26], we fit a single four-parameter sigmoidal function relating growth rate to intracellular THF concentration (**Fig. 4C**). The complete model well-predicts growth rates for a training set of five DHFR mutants in two TYMS backgrounds (R^2= 0.93, **Fig. 4D**) and a test set of seven DHFR single mutants in four different TYMS backgrounds9 (R^2 = 0.72)[9]. In this project, we propose to now run the mathematical model "in reverse": rather than predicting growth rates given *in vitro* biochemical measurements, we will infer *in vitro* biochemical measurements from experimental growth rates. While this model was created with DHFR in mind, we imagine using a simplified but conceptually similar framework for other growth linked enzyme families.

Taken together, we have now: (1) established a high-throughput growth-based assay for DHFR velocity[4,20]; (2) used variation in TYMS activity to alter intracellular [DHF] over at least two orders of magnitude[25]; and (3) created a well-tested mathematical model for relating catalytic parameters to growth rate[9]. The work proposed here will now assemble these components to perform high-throughput biochemistry *in vivo*.

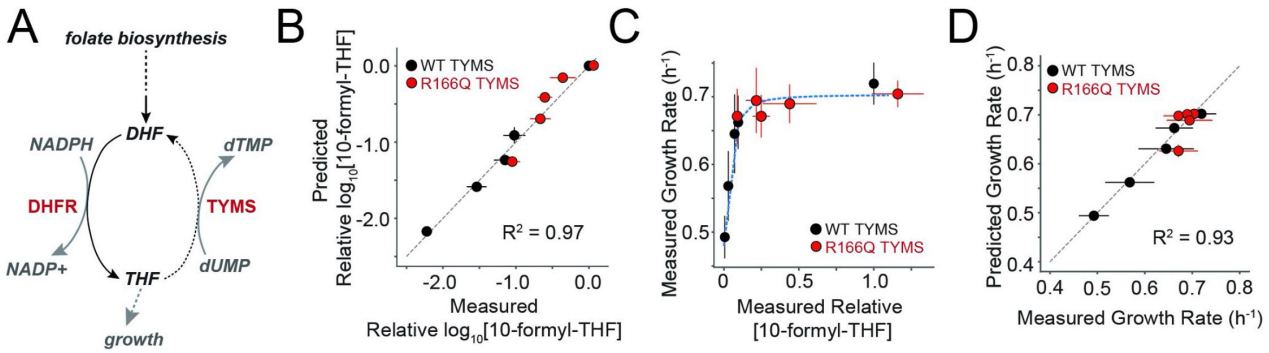


Figure 4: A mathematical model relating biochemical parameters of DHFR and TYMS to growth rate phenotype.

A) A simplified, abstracted version of the DHFR and TYMS system. Dotted lines indicate multiple intermediate reactions that are summarized with a single line.

B) Correlation between experimentally measured log10 [10-formyl-THF] relative abundance and the model prediction. Red points indicate five DHFR variants in the background of TYMS R166Q (a near-catalytically inactive variant) and black indicates the same DHFR variants in the context of WT TYMS. Error bars are the standard deviation across triplicates for both growth rate (y-axis) and 10-formyl-THF abundance (x-axis). The gray dotted line indicates x = y.

C) The relationship between the experimentally measured relative abundance of [10-formyl-THF] and growth rate. The blue dotted line is the model best fit. Color coding is identical to (B).

D) Growth rate predictions from biochemical parameters given the two-equation model. Error bars in x indicate the standard deviation across triplicates. Color coding is identical to (B).

DOI: 10.5281/zenodo.15277418

Key outcomes

Completion of this work is anticipated to yield the following outcomes:

1. A comprehensive genotype-to-phenotype map for DHFR that quantifies the effect of all 3,021 possible single mutations on intracellular enzyme abundance [E], catalytic turnover (k_{cat}), and the substrate Michaelis constant (K_m)

   a. These data will reveal the extent to which [E], k_{cat}, and K_m can be separately tuned by mutations, or if these properties are encoded by overlapping positions.

   b. We will use the data to identify allosteric surface sites that can manipulate enzyme k_{cat} and K_m. These surfaces represent potential targets for the design of small-molecule allosteric inhibitors.

   c. We will additionally compare the pattern of mutational tolerance in our experiments to natural sequence variation. This will reveal how residues tuning k_{cat}, K_m, and [E] are conserved–or vary–across species with different metabolic lifestyles (commensal, free-living, etc).

   d. We anticipate using these data to train a machine learning model (e.g., Augmented Potts model) that can predict mutational effects on specific biochemical parameters.

2. Quantification of biochemical variation in DHFR activity across enzyme orthologs sampled from diverse species.

   a. This will provide a quantitative picture of how catalysis varies across species.

   b. These data will sample a wider breadth of sequence space, representing a rich training set for machine learning and deep learning algorithms.

3. A general strategy for high-throughput measurement of biochemical parameters using growth-based data.

   a. We will extend our approach to two additional enzymes in amino acid metabolism. We expect that this methodology should generalize to any enzyme where a growth-based selection is possible.

Anticipated Impact.

We envision a future where bioengineers can reliably design enzymes that are not just active, but satisfy specified constraints on k_{cat}, K_m, and enzyme abundance [E]. Accomplishing this level of precision requires new computational models–and centrally, abundant biochemical data for training and testing these models.

While AlphaFold was trained using 10^5 structures sampling a diversity of folds (from the Protein Data Bank or PDB), the largest comparable database of biochemical measurements (BRENDA) contains about an order of magnitude less data (~38,000 measurements of catalytic power)[27]. Moreover, these measurements are spread across 8,424 enzyme classes, meaning that we lack dense data relating sequence to quantitative biochemical parameters for most enzymes.

Our work has the potential to transform the scale of biochemical data collection: our proposed method will enable measurements of catalytic parameters (k_{cat}, K_m) at a scale that exceeds the entire BRENDA database of literature-curated biochemical constants in a single experiment.

DOI: 10.5281/zenodo.15277418

Gene of Interest (GOI) Region

For the singleplex and pooled fitness assays, we will make use of two related constructs: an abundance selection plasmid and a function selection plasmid. The first construct is necessary to disentangle the contributions of [E] and k_{cat} to growth. The second series of constructs is necessary to resolve k_{cat} and K_m.

The function selection plasmid includes both the GOI (*folA*) and the gene encoding the UE (*thyA*). Each gene will be separately controlled by an inducible promoter (the vanillic acid-inducible P_{VanCC} for *folA*, and the sodium salicylate-inducible P_{SalTTC} for *thyA*) and followed by a double transcriptional terminator. To insulate expression of each gene, we have included (1) a self-cleaving ribozyme and (2) bicistronic leader peptide at the 5' end of each gene[28,29]. The ribozyme helps ensure consistent transcription (and a consistent 5' end of each transcript), while the bicistronic leader peptide promotes consistent translation. Together, these elements will help minimize mutant-to-mutant expression changes emerging from variation in transcription and translation.

The assembled plasmid construct has a kanamycin resistance marker and a low-copy p15A origin suitable for amplicon-based sequencing experiments. By varying the induction of *thyA* relative to *folA*, we can titrate the intracellular abundance of dihydrofolate substrate during our selection experiments. In prior work, we established that variation in the enzymatic velocity of TYMS can alter intracellular substrate pools for DHFR over two orders of magnitude[25]. For context, DHFR has a K_m of 1.1 µM for DHF, and native *E. coli* have an intracellular concentration of 25 µM DHF.

For this work, our goal is to identify a range of induction conditions that sample intracellular substrate concentrations ranging from ~0.1 – 200 µM. This should yield a range of differential growth rate effects from near-dead to native-like growth for the WT DHFR enzyme. We will confirm intracellular concentrations of substrate by mass spectrometry for all promoter/RBS combinations spanning the range of expected growth phenotypes, in collaboration with Dr. Jun Park at UCLA.

We will include a barcode region following the terminator of the GOI. This barcode follows the standard design from AlignToInnovate, using two half barcodes to increase the likelihood of unique barcode assignments during plasmid construction. Importantly, because our construct uses non-native promoters, we remove the potential for endogenous transcriptional regulation (e.g., feedback control) that could complicate the interpretation of our experimental results.

The abundance selection plasmid is nearly identical to the function selection plasmid, but contains a fusion of the GOI (DHFR) with the resistance marker chloramphenicol acyltransferase (CAT). Under media conditions that contain chloramphenicol but are not selective for DHFR, we can use the growth rate of strains transformed with this construct as a reporter of intracellular DHFR abundance. Because growth rates diverge exponentially over time, this approach has the potential to be far more sensitive to small variations in intracellular abundance than tagging the GOI with a fluorescent reporter.

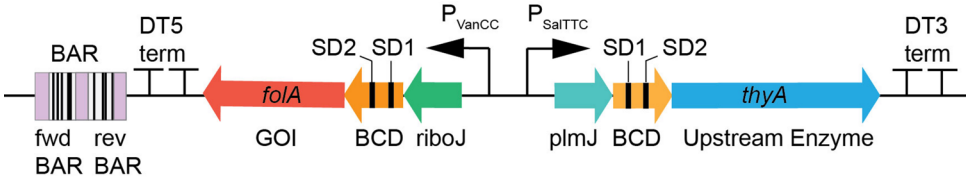


Figure 5: The function selection cassette. A complete plasmid map of our proposed function selection construct can be found here.

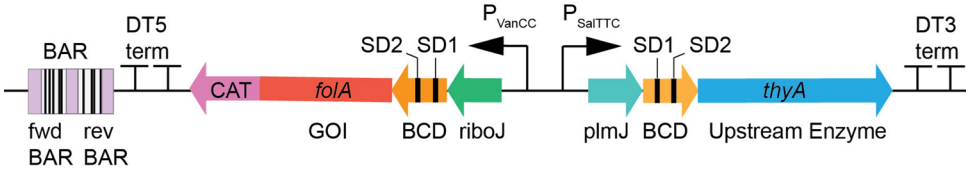


Figure 6: The abundance selection cassette. A complete plasmid map of our proposed abundance selection construct can be found here.

DOI: 10.5281/zenodo.15277418

Proposed Hosts and Plasmids

Host strains: Our host strain for DHFR assay development (pilot-scale and large-scale data collection) will be *E. coli* ER2566 *ΔfolA ΔthyA*[19]. This *E. coli* expression strain is a derivative of BL21(DE3) and contains double knockout for the genes encoding DHFR (*folA*) and the UE TYMS (*thyA*). We have verified the genotype of this strain by sequencing and have used it in the collection of several published deep mutational scanning datasets[9,17,4]. We will edit this strain to genomically integrate the vanillic acid and sodium salicylate sensor cassettes (VanR and NahR, respectively). We will make these genomic integrations using ORBIT, a new approach for high-efficiency *E. coli* genome editing that we have previously used in our lab[30]. Continuing the DHFR selections in the same strain background (but with the Marionette sensor machinery) will allow us to compare the new mutational data to our earlier work.

To extend our assay to CM and DapA (during the large-scale collection phase) we will construct auxotrophic knockouts of the *E. coli* K-12 MG1655 strain. The DapA measurement strains will be deleted for *dapA* and the upstream enzyme *asd*, while the CM measurement strains will be deleted for *pheA*, *tyrA* (which also contains a CM domain), and the upstream enzyme *aroC*. Again, we will make these deletions with ORBIT. We will verify both auxotrophic strains by sequencing the deleted loci and phenotyping in selective media.

Plasmids: Our plasmid constructs are derived from a Marionette-system plasmid designed by Dr. David Ross at NIST. The plasmid includes the GOI and UE under the control of independently inducible promoters, a kanamycin resistance cassette, and a p15A origin of replication. Expression of both genes will be insulated using a self-cleaving ribozyme and a bicistronic leader peptide. Complete maps for the proposed function and abundance selection plasmids are linked above.

Proposed Protein Targets

For the pilot-scale dataset and a portion of the large-scale dataset, we will focus on the model metabolic enzyme DHFR. As described above, we have chosen DHFR as our starting point because:

1. Extensive experimental data and knowledge of this system will accelerate assay development and allow robust validation.
2. DHFR plays a key role in human health and disease. More complete knowledge of sequence–activity relationships will assist in the interpretation of drug resistance mutations and permit identification of potential druggable allosteric sites.
3. DHFR is a model system for understanding how conformational change governs catalysis. These data will provide a comprehensive map of how mutations throughout the structure govern distinct catalytic properties.

During the pilot-scale stage, we will collect growth rate measurements for an established DHFR saturation mutagenesis library[4] (3,021 mutations in total) and a more extensive library of double mutants (~30,000 mutations in total). Prior work indicates that measuring the impact of a given single mutation in approximately 10 different double-mutant backgrounds assists in disambiguating mutational effects on different underlying biochemical parameters[12]. With this in mind, we will design our double-mutant library to contain all possible single mutations in the background of ten mutations selected to have moderate effects on catalysis and stability.

In total, we will obtain approximately a dozen growth rates for each single and double mutant: six in different chloramphenicol concentrations, and six to ten under different levels of UE induction. These data will be used to infer k_{cat}, K_m and [E] for each single mutant (see Current Developmental Stage for a description of our inference strategy and preliminary results). We propose additional sub-pilots to characterize: (1) a library generated by error-prone PCR that additionally captures indels, and (2) deep mutational scanning of two additional DHFR orthologs (from *H. sapiens* and *C. elegans*).

During the large-scale stage, we will additionally measure k_{cat}, K_m and [E] for a series of 100 orthologous DHFR sequences sampled across diverse species. These sequences will be selected to capture increasing amounts of sequence variation: we will start from *E. coli* strain-level variation wherein there are only a few mutations per DHFR, and move to phylum-level variation, where sequences of DHFR are only 20% identical to each other.

Together, these data will provide (1) extensive mutational data in one sequence background and (2) shallow sequence data across many sequence backgrounds. This combination will be important for training and testing computational models. For example, one can directly test how well a model trained on deep mutational scanning data from one enzyme can predict the catalytic properties of increasingly sequence-diverged orthologs.

In the large-scale stage we will also extend the assay to at least two additional enzymes with key roles in metabolism and industrial biosynthesis: (1) chorismate mutase (tyrosine/phenylalanine biosynthesis) and (2) 4-hydroxy-tetrahydrodipicolinate synthase (lysine biosynthesis). For these enzymes, we will again first consider single-mutant variation, and then expand our dataset to sample diversity across orthologs. This will help to establish the generality of our biochemistry *in vivo* approach.

DOI: 10.5281/zenodo.15277418

Proposed Controls for Assay Development

Stage 0 – De-risking

1. Comparison of turbidostat and plate-based culture conditions

Our current sequencing-based growth rate measurements are performed under continuous culture conditions in a turbidostat. This approach maintains the cell population in exponential-phase growth and allows for excellent control of environmental conditions. However, turbidostats are not readily available to most labs and can be complex to assemble and maintain. With this in mind, our goal is to transition the sequencing-based growth rate measurements from continuous culture to a 96-well plate format. During the de-risking phase, we will conduct a head-to-head comparison of sequencing-based growth rate measurements under turbidostat culture conditions (using our current protocols) and in 96-well plates with serial passaging (following protocols adapted from Dr. David Ross).

For these experiments, we will use 12 barcoded DHFR variants that span a range of catalytic activities and growth rate effects (see mutant list in Table 2). These variants are already cloned into a pTet-Duet vector under control of a T7 promoter, as described in published work. We will transform these constructs into our existing ER2566 *ΔfolA ΔthyA* selection strain.

The mixed population will be cultured either (1) in the turbidostat or (2) in 96-well plates. We will use 96-well plates identical to those at NIST and follow an analogous protocol of five serial passages every three hours (15-hour total outgrowth times). Both experiments will be run in quadruplicate.

We will collect samples every three hours from both the turbidostat and plate-based assay; these samples will be prepared for NGS. We will then correlate the growth rates measured in the turbidostat and plate-based selections, with the goal of establishing a normalization function that scales data from one platform to align with that of the other. We will also compare the variation in the measurements between the two culture platforms.

2. Pilot measurements for the protein abundance assay

To establish the feasibility of the abundance assay, we will characterize the growth rate impact of fusing a CAT tag to DHFR. For these experiments, we will use an existing set of three constructs wherein DHFR is under the translational control of different RBS sequences: one with high translational efficiency (RBSI, 'AAGGAG'), one with medium translation efficiency (RBSII, 'AAGGAA'), and one with low translational efficiency (RBSIII, 'AATGAG').

We have prior Western blot data quantifying the abundance of all three RBS variants in ER2566 *ΔfolA ΔthyA* grown in selective conditions. We will create an in-frame C-terminal fusion of the CAT tag to DHFR in all three RBS backgrounds.

First, we will assess the impact of the CAT tag on DHFR function. To do this, we will measure the growth rate of strains carrying DHFR with and without the CAT tag (for all three RBS backgrounds) under conditions selective for DHFR activity. These experiments will be performed in the absence of chloramphenicol. The difference in growth rate between the tagged and untagged variants will then reflect the impact of the CAT tag on function.

If the CAT tag has a large effect on growth rate rescue, we will switch to a different strategy for measuring abundance (an N-terminal CAT tag or a fluorescent tag). If the CAT tag has a minimal effect, we will then measure growth rate for all three RBS backgrounds of DHFR over a range of 12 chloramphenicol concentrations. These experiments will be performed under conditions that are non-selective for DHFR activity.

We will then compare the IC_{50} for chloramphenicol to the DHFR intracellular abundance for each of the three RBS variants. This will allow us to test whether strains with higher intracellular abundance show a higher IC_{50} for chloramphenicol.

Stage 1 - Development and Tuning

During the Development and Tuning stage we will optimize the dynamic range of both the function selection assay and abundance selection assay. In both cases, we will make use of a Singleplex Assay that relies on growth curve fitness measurements in 96-well plates. In our current setup, we monitor optical density at 600 nm over the course of approximately 24 hours in a BioTek plate reader with stacker, incubated at 37 °C. Cells are periodically mixed by shaking, and the plates are sealed with AeraSeal. Growth rates are then fit by linear regression to the $\log(OD_{600})$ vs time relationship.

Tuning for the function selection assay

We will first identify six to ten induction conditions that vary abundance of the UE (TYMS for DHFR) and achieve intracellular substrate concentrations ranging from ~0.1 to 200 µM. This series of conditions should yield a range of growth rate phenotypes from near-dead to native-like growth for both WT DHFR and a variant with highly reduced activity (DHFR M42F/G121V).

In general, selection on DHFR should be the most stringent under conditions of high TYMS expression, and the least under conditions of reduced TYMS expression. Indeed, prior work shows that inhibition of DHFR can be rescued by strong loss-of-function mutations in TYMS[25,31], and that increases in TYMS activity restrict mutational variation in DHFR9. This is because TYMS not only produces the substrate of DHFR but also siphons off a substantial portion of the reduced folate pool (the product of DHFR, THF, by converting it back to DHF, (Fig. 4A).

Prior mathematical modeling in our group provides a quantitative prediction of growth rate dependency on TYMS expression for both WT DHFR and DHFR M42F/G121V (**Fig. 7A, B**). This modeling will help guide our choice of induction conditions and sets expectations for what we might see experimentally.

We will first clone WT DHFR and DHFR M42F/G121V (two of our calibration controls) into the function selection plasmid using Gibson assembly. We will transform this plasmid into the host strain *E. coli* ER2566 *ΔfolA ΔthyA*. Then, we will use the Singleplex growth rate assay to measure the effects of TYMS expression variation.

We envision a grid-based layout of induction conditions on

DOI: 10.5281/zenodo.15277418

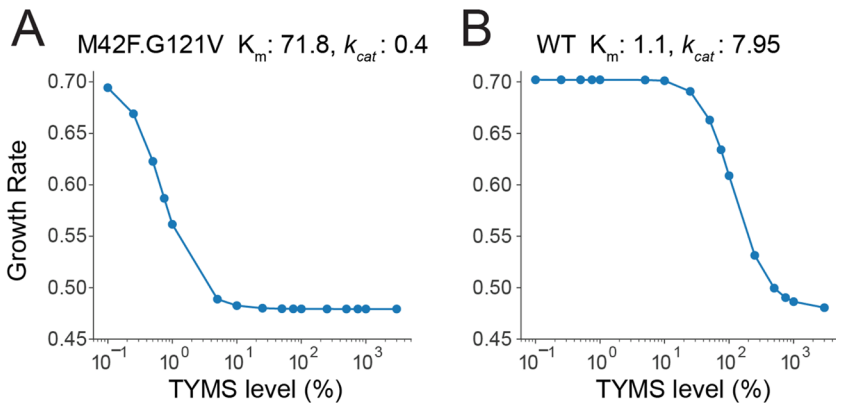


Figure 7: The relationship between TYMS abundance and growth rate. **A)** The dependence of growth rate on TYMS abundance for DHFR M42F/G121V; **B)** The dependence of growth rate on TYMS abundance for WT DHFR.

our 96-well plate that combinatorially samples variation in GOI and UE induction. We will attempt to choose a final set of approximately six to ten induction conditions that span a range of growth rate effects for both WT DHFR and DHFR M42F/G121V. Given that fast enzymes will require higher TYMS expression to achieve high concentrations of intracellular DHF, it is possible that we may need to consider two separate sets of induction conditions that can resolve catalytic parameters for "fast" and "slow" enzymes.

Finally, we will quantify intracellular folate concentrations ([DHF], [THF]) for strains carrying the function selection plasmid under each induction condition by mass spectrometry.

Tuning for the abundance selection assay

We will identify a range of chloramphenicol concentrations that optimize the dynamic range of the abundance selection assay. To do this, we will consider: WT DHFR, a variant with increased intracellular abundance (DHFR M42F), and a highly destabilized, low-abundance variant (DHFR I41A)[4]. We will clone these three variants as in-frame fusions with CAT (see also abundance selection plasmid GOI cassette). All plasmids will be transformed into the host strain *E. coli* ER2566 *ΔfolA ΔthyA*.

We will then use the Singleplex growth rate assay to measure growth curve fitness of each construct under a dozen chloramphenicol concentrations in nutrient-replete media. From these data, we will fit a chloramphenicol IC_{50} for each DHFR-CAT fusion. We will select a final set of six to eight chloramphenicol concentrations that maximize our ability to distinguish between the low-abundance construct DHFR I41A and the high abundance construct DHFR M42F.

DHFR Variant	Intracellular Abundance	Expected Result
WT DHFR	52 molecules/cell	Intermediate IC_{50} for chloramphenicol
DHFR M42F	1,119.1 molecules/cell	High IC_{50} for chloramphenicol
DHFR I41A	65.6 molecules/cell	Low IC_{50} for chloramphenicol

Table 1: DHFR variants for abundance selection assay tuning, and expected growth phenotypes in chloramphenicol.

Stage 2 – Normalization and Calibration Controls

During Stage 2, we will consider additional calibration controls for both the function selection and abundance assays. Our goal during this stage is to optimize assay resolution.

For the function selection assay, we will consider 12 DHFR variants spanning a range of catalytic activities and inspect our ability to resolve V_{max} (the product of k_{cat} and [E]), and K_m.

For the abundance selection assay, we will consider 12 DHFR variants spanning a range of intracellular abundance (as previously determined by lysate-based assays[4]). Here we will examine the relationship between chloramphenicol IC_{50} and intracellular abundance ([E]).

In both cases we will measure relative growth rates using small scale barcode-based fitness measurements, as these will be most analogous to our downstream experiments. All measurements will be made in quadruplicate in a single run (day). We

DOI: 10.5281/zenodo.15277418

have selected 12 DHFR calibration controls for each assay with a 96-well plate format in mind; these 12 controls can be arrayed against 8 induction conditions (for the function singleplex assay) or 8 chloramphenicol concentrations (for the abundance singleplex assay). Thus we can run all calibration controls in one plate.

For the abundance assay, we will also conduct orthogonal measurements of abundance using a fluorescent protein tag (sfGFP) for four DHFR variants spanning a range of abundance values . Comparing these measurements to those with the CAT tag will help establish our confidence in the measurements.

DHFR Variant	k_{cat} (s^{-1})	K_m (μM)	Std. dev k_{cat}	Std. dev K_m
M42F	79.20	13.00	1.72	1.00
WT	7.95	1.10	0.38	0.20
F31Y	20.61	80.00	2.12	14.00
T113V	32.90	21.40	0.50	1.10
L54I	7.88	35.00	0.28	3.40
W22H	1.89	18.00	0.06	1.20
F31V	8.65	108.00	0.29	6.80
G121V	0.30	6.10	0.01	0.60
F31Y/L54I	1.94	168.30	0.16	21.40
M42F/G121V	0.40	71.80	0.04	13.20
F31Y/G121V	0.13	90.60	0.01	7.40
D27N	0.05	330.00	0.01	78.90

Table 2: Calibration controls for the function selection assay.

We will also carry out the activity assay for four DHFR variants spanning a range of catalytic efficiencies in both the presence and absence of the CAT tag. This will allow us to again quantify the impact (if any) of the CAT tag on our estimations of activity.

Finally, it may become necessary to consider product inhibition of DHFR as a separate factor that could impact growth rate under conditions of high intracellular [THF]. If we observe unexpectedly large growth rate defects for DHFR variants under low TYMS expression conditions (which would align with high intracellular [THF]) we can perform additional *in vitro* measurements of product inhibition for the mutants of interest.

DHFR Variant	Intracellular Abundance (molecules/cell)	Standard Deviation
E154V	37	5.1
L156Y	47.3	3.1
WT	52	9.2
I41A	65.6	5.9
L24V	83.6	6.8
H45S	101.9	10.1
R98Y	166.1	16.2
Q102L	239.5	27
M42Y	360.9	2.6
T113V	418.3	19.1
W47L	558.3	94.5
M42F	1119.1	17.6

Table 3: Calibration controls for the abundance selection assay.

Pilot-Scale Collection

We have an established DHFR DMS library that covers all possible single mutants across the 159 amino acid coding sequence[4]. We will barcode this library and clone it in the context of (1) the function selection plasmid and (2) the abundance selection plasmid (with CAT fusion) using Gibson assembly.

We will additionally construct three other sources of variation:

1. Targeted double mutants (~30K). These will be constructed to include all possible single mutants in the background of ten mutations with moderate effects on stability and activity. This library will be important for disambiguating the impact of mutations on stability and biochemical parameters in our analysis.
2. A library of insertions and deletions (indels) and an error-prone PCR as a source of random variation.
3. Saturation mutagenesis across two additional DHFR orthologs (*H. sapiens* DHFR and *C. elegans* DHFR). These data will allow us to compare the pattern of residues that encode biochemical function across species. Prior structural characterization by NMR indicates that these two DHFR variants show altered dynamics associated with catalysis relative to the *E. coli* enzyme.

For all libraries, we will measure the effect of mutation on growth rate using barcode-based fitness measurements. All measurements for a given library will be conducted in a single run (day) in quadruplicate, starting from four independently transformed replicates. Together, the growth rate data from these libraries will be used to infer k_{cat}, K_m, and [E] across the entire DHFR sequence (see also Current Developmental stage).

DOI: 10.5281/zenodo.15277418

Large-Scale Collection

During the large scale collection phase, we propose expanding the dataset in three ways. First, we will leverage the established DHFR assay to measure catalytic activities for 100 orthologs sampled across species. Each DHFR sequence can be assembled from approximately five ~200 nucleotide oligos. We will order these as an oligo pool and include orthogonal priming sites on the oligos encoding each DHFR variant. Then, we will use PCR with orthogonal primers to amplify out the oligonucleotides corresponding to each DHFR variant and assemble using overlap-extension PCR. As described above, these orthologs will be selected to span a range of sequence variation relative to *E. coli* DHFR. To our knowledge, this will be the first deep survey of catalytic variation across an entire protein family. We intend to examine the association between catalytic proficiency, phylogeny, and bacterial lifestyle. Because these data will capture a broader swath of sequence variation, they will also be important to training and testing computational models of DHFR enzymatic function.

Second, we will extend the *in vivo* biochemistry approach to at least two additional enzymes in amino acid metabolism: Chorismate Mutase (CM, encoded by the gene *pheA*), and 4-dhydroxy-tetrahydrodipicolinate synthase (DapA, *dapA*) (Fig. 8). Amino acid metabolism enzymes are interesting targets for assay development because they are essential to growth in minimal media, and selection can be tuned by the addition of exogenous amino acids. *E. coli* knockouts of CM are auxotrophic for tyrosine and phenylalanine[32,33], while knockouts of DapA are auxotrophic for lysine[34]. Moreover, there is a large industrial market for downstream pathway products. CM is the branch point for the shikimate pathway, which is used to produce several aromatic compounds with pharmaceutical value[35]. Lysine is also used in pharmaceuticals, animal feedstocks, and cosmetics[36]. A better understanding of sequence–activity relationships for enzymes in these pathways would facilitate biosynthetic engineering and yield optimization.

The CM domain of *pheA* is ~100 amino acids long and forms a domain swapped dimer. This complex catalyzes the conversion of chorismate to prephenate by a Claisen condensation; prephenate is then used to produce tyrosine and phenylalanine (Fig. 8A). It has become a model system for computational enzyme design due to its small size, ease of selection, and the apparent lack of covalent interactions between substrate and enzyme[3,37]. Similar to DHFR, the availability of an established growth based selection and extensive high quality *in vitro* biochemical data will facilitate our efforts to establish and validate an *in vivo* biochemistry assay for CM activity.

DapA catalyzes the first biochemical step unique to lysine biosynthesis, condensing pyruvate and L-aspartate 4-semialdehyde to create 4-hydroxy-2,3,4,5-tetrahydrodipicolinate (Fig. 8B). DapA is a larger enzyme in comparison to CM and DHFR: it is a tetramer composed of four 292 amino acid monomers. While perhaps less well-studied than CM or DHFR, ***in vitro*** biochemical data exist for a number of mutant variants[38,39].This provides a starting place for assay optimization and validation. Intriguingly, DapA is subject to allosteric feedback inhibition by lysine. Several known mutations disrupt this feedback inhibition[40]. This opens the possibility of mapping mutational effects on both catalysis AND allostery: we could conduct our assay in parallel for both native DapA and a DapA mutant lacking feedback inhibition. This would reveal the pattern of mutations that impact catalysis in the presence and absence of lysine-mediated feedback.

Finally, we will conduct pilot experiments to establish the feasibility of titrating intracellular substrate concentrations for 5-10 additional growth-linked enzymes. This is necessary to better characterize the generalizability of our *in vivo* biochemistry approach and identify other candidate systems. We will select 10 enzymes from the list of 352 growth coupled enzymes in *E. coli*, prioritizing enzymes that are (1) conditionally essential (selection can be titrated through the addition of an exogenous media supplement), (2) less than 250 amino acids long (to facilitate library construction), (3) sample diverse metabolic processes, and (4) monomeric. For each enzyme we will identify candidate upstream enzymes that could potentially modulate intracellular substrate abundance (typically the enzyme immediately upstream in the metabolic pathway). Then, we will use an approach (developed in my lab) called titratable CRISPRi[41] to create gradated knockdowns of the upstream enzyme. We will measure the growth rate impact of these knockdowns under selective and non-selective (nutrient-supplemented) conditions. Additionally, we will use targeted metabolomics to measure the impact of gene repression on intracellular substrate levels. If the upstream enzyme generates gradated variation in substrate availability upon expression titration, it suggests that the upstream enzyme/target enzyme pair is a suitable candidate for our *in vivo* biochemistry approach. Next steps in development would include creating a double knockout strain for selection, assembling the upstream enzyme/target enzyme pair onto the Marionette vector, and repeating the tuning/calibration experiments.

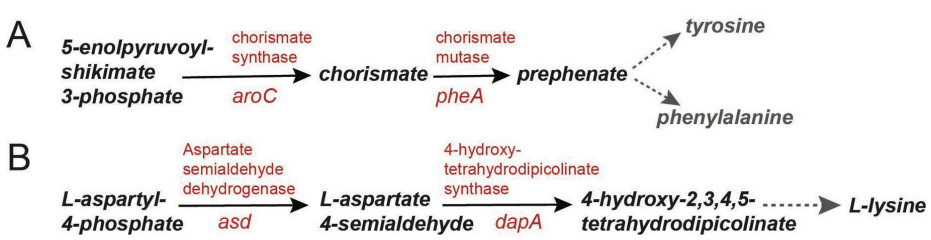


Figure 8: The CM and DapA reactions. Enzyme and gene names are in red, metabolites in black.

DOI: 10.5281/zenodo.15277418

Current Developmental Stage

As an initial test of our ability to infer quantitative biochemical parameters from growth rate data, we considered an existing dataset comprising growth rate measurements for DHFR point mutants in the background of three TYMS variants. These TYMS variants span a range of catalytic velocities, and thus modulate intracellular DHF (and THF) concentrations. We asked if this initial dataset is sufficient to constrain inference of k_{cat} and K_m. More specifically, we used Markov Chain Monte Carlo (MCMC) to sample a wide range of DHFR k_{cat} and K_m values. For each parameter pair, we then computed the predicted growth rate in each of the three TYMS backgrounds using our existing mathematical model (see also A mathematical model linking growth rate to enzyme catalysis). Then, we computed the difference between the experimentally observed growth rates and computational growth rate predictions for a particular k_{cat} and K_m combination (Fig. 9A). A small difference indicates good agreement between the experiment and model, and suggests that a given DHFR k_{cat} and K_m are consistent with the experimental observations. This process allowed us to estimate quantitative biochemical parameters and a confidence interval around them for each mutation. We found that measurements in only three TYMS backgrounds already substantially constrain the space of possible DHFR k_{cat} and K_m values (under assumptions of fixed [E]) for some mutants, however for others we lack sufficient data to resolve variation (Fig. 9B-D). Model simulations suggest that collecting additional growth rate data across ~7 measurement strains (as proposed here) will constrain fits of k_{cat} and K_m for most variants. In our future work, we will additionally assess Bayesian inference as a more robust approach to estimating biochemical parameters from our growth rate data.

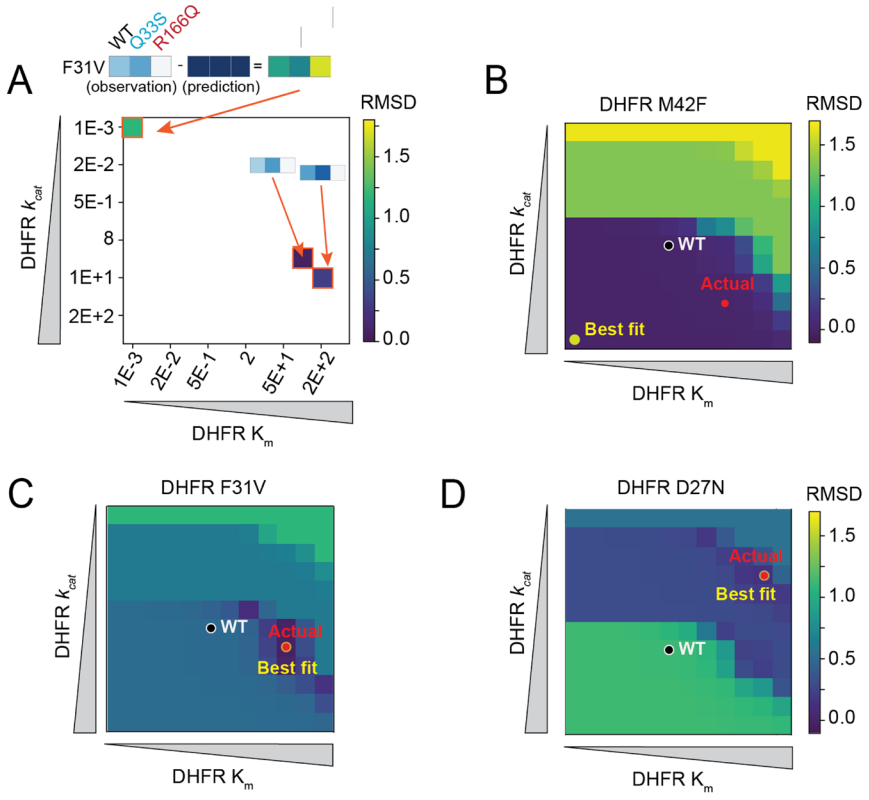


Figure 9: Inference of biochemical parameters from growth data. **A)** We computed the RMSD between the experimentally measured and computationally predicted growth rates over a range of DHFR k_{cat} and K_m values. Low RMSDs indicate that a particular k_{cat} / K_m combination leads to predicted growth rates that are more consistent with the experimental measurements. **B-D)** Agreement between the *in vitro* characterized k_{cat} and K_m values (red dot) and the model best fit (lowest RMSD, yellow dot) for three DHFR mutants. For some mutants (F31V, D27N) the k_{cat} and K_m values are well-constrained by the three available growth rate measurements, in other cases (M42F) more data is needed.

DOI: 10.5281/zenodo.15277418

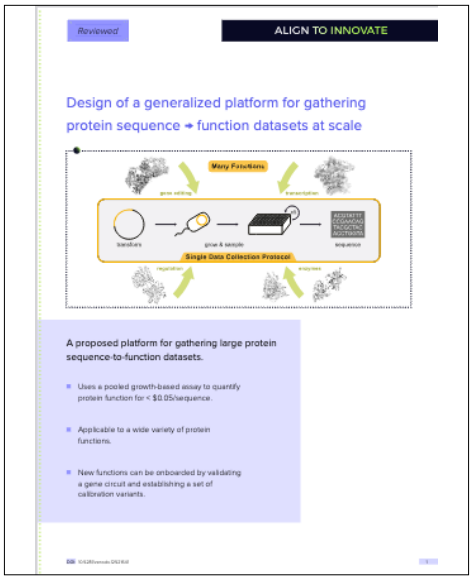
Suggested Reading

Below is a suggested curated reading list to better understand this document and its context:

This proposed dataset platform is part of Align's Open Dataset Initiative, which pioneers new ways to identify, collect, and share large datasets in life science. Read more about the Open Datasets initiative here:

"Design of a generalized platform for gathering protein sequence → function datasets at scale"

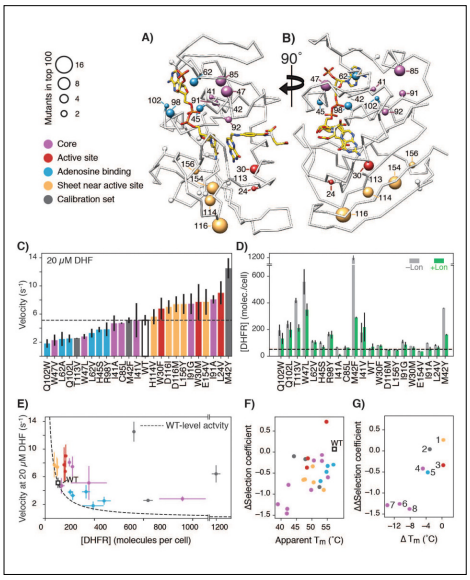
February 2024, Zenodo



Read more about the Reynold's labs' prior work identifying mutations of DHFR that switched from beneficial to deleterious for growth depending on *E. coli* strain background here:

"Altered expression of a quality control protease in E. coli reshapes the in vivo mutational landscape of a model enzyme"

July 2020, eLife



DOI: 10.5281/zenodo.15277418

References

1. Wolfenden, R. Benchmark reaction rates, the stability of biological molecules in water, and the evolution of catalytic power in enzymes. *Annu Rev Biochem* **80**, 645–667 (2011).
2. Markin, C. J. *et al*. Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics. *Science* **373**, eabf8761 (2021).
3. Russ, W. P. *et al*. An evolution-based model for designing chorismate mutase enzymes. *Science* **369**, 440–445 (2020).
4. Thompson, S., Zhang, Y., Ingle, C., Reynolds, K. A. & Kortemme, T. Altered expression of a quality control protease in *E. coli* reshapes the *in vivo* mutational landscape of a model enzyme. *eLife* **9**, e53476 (2020).
5. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat Methods* **11**, 801–807 (2014).
6. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat Methods* **15**, 816–822 (2018).
7. Hsu, C., Nisonoff, H., Fannjiang, C. & Listgarten, J. Learning protein fitness models from evolutionary and assay-labeled data. *Nat Biotechnol* **40**, 1114–1122 (2022).
8. Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. Low-N protein engineering with data-efficient deep learning. *bioRxiv* 2020.01.23.917682 (2020) doi:10.1101/2020.01.23.917682.
9. Nguyen, T. N., Ingle, C., Thompson, S. & Reynolds, K. A. The genetic landscape of a metabolic interaction. *Nat Commun* **15**, 3351 (2024).
10. Stiffler, M. A., Hekstra, D. R. & Ranganathan, R. Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. *Cell* **160**, 882–892 (2015).
11. Jiang, L., Mishra, P., Hietpas, R. T., Zeldovich, K. B. & Bolon, D. N. A. Latent Effects of Hsp90 Mutants Revealed at Reduced Expression Levels. *PLOS Genetics* **9**, e1003600 (2013).
12. Faure, A. J. *et al*. Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175–183 (2022).
13. Weng, C., Faure, A. J., Escobedo, A. & Lehner, B. The energetic and allosteric landscape for KRAS inhibition. *Nature* **626**, 643–652 (2024).
14. Maxwell, K. L., Mittermaier, A. K., Forman-Kay, J. D. & Davidson, A. R. A simple *in vivo* assay for increased protein solubility. *Protein Science* **8**, 1908–1911 (1999).
15. Meyer, A. J., Segall-Shapiro, T. H., Glassey, E., Zhang, J. & Voigt, C. A. Escherichia coli 'Marionette' strains with 12 highly optimized small-molecule sensors. *Nat. Chem. Biol.* **15**, 196–204 (2019).
16. Ducker, G. S. & Rabinowitz, J. D. One-Carbon Metabolism in Health and Disease. *Cell Metabolism* **25**, 27–42 (2017).
17. Schnell, J. R., Dyson, H. J. & Wright, P. E. Structure, Dynamics, and Catalytic Function of Dihydrofolate Reductase. *Annual Review of Biophysics and Biomolecular Structure* **33**, 119–140 (2004).
18. Fierke, C. A., Johnson, K. A. & Benkovic, S. J. Construction and evaluation of the kinetic scheme associated with dihydrofolate reductase from Escherichia coli. *Biochemistry* **26**, 4085–4092 (1987).
19. Reynolds, K. A., McLaughlin, R. N. & Ranganathan, R. Hot Spots for Allosteric Regulation on Protein Surfaces. *Cell* **147**, 1564–1575 (2011).
20. McCormick, J. W., Russo, M. A., Thompson, S., Blevins, A. & Reynolds, K. A. Structurally distributed surface sites tune allosteric regulation. *Elife* **10**, e68346 (2021).
21. Baba, T. *et al*. Construction of Escherichia coli K-2 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology* **2**, 2006.0008 (2006).
22. Karp, P. D. *et al*. The EcoCyc Database (2023). *EcoSal Plus* 11, eesp00022023 (2023).
23. Goldbeter, A. & Koshland, D. E. An amplified sensitivity arising from covalent modification in biological systems. *Proceedings of the National Academy of Sciences* **78**, 6840–6844 (1981).
24. Ferrell, J. E. & Ha, S. H. Ultrasensitivity part I: Michaelian responses and zero-order ultrasensitivity. *Trends in Biochemical Sciences* **39**, 496–503 (2014).
25. Schober, A. F. *et al*. A two-enzyme adaptive unit within bacterial folate metabolism. *Cell Reports* **27**, 3359-3370. e7 (2019).
26. Rodrigues, J. V. *et al*. Biophysical principles predict fitness landscapes of drug resistance. *PNAS* **113**, E1470–E1478 (2016).
27. Chang, A. *et al*. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Research* **49**, D498–D508 (2021).
28. Lou, C., Stanton, B., Chen, Y.-J., Munsky, B. & Voigt, C. A. Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nat Biotechnol* **30**, 1137–1142 (2012).
29. Jansen, Z. *et al*. Interrogating the Function of Bicistronic Translational Control Elements to Improve Consistency of Gene Expression. *ACS Synth. Biol.* **12**, 1608–1615 (2023).
30. Saunders, S. H. & Ahmed, A. M. ORBIT for *E. coli*: kilobase-scale oligonucleotide recombineering at high throughput and high efficiency. *Nucleic Acids Research* **52**, e43 (2024).

DOI: 10.5281/zenodo.15277418

31. King, C. H., Shlaes, D. M. & Dul, M. J. Infection caused by thymidine-requiring, trimethoprim-resistant bacteria. *Journal of Clinical Microbiology* **18**, 79–83 (1983).

32. Neuenschwander, M., Butz, M., Heintz, C., Kast, P. & Hilvert, D. A simple selection strategy for evolving highly efficient enzymes. *Nat Biotechnol* **25**, 1145–1147 (2007).

33. Kast, P., Asif-Ullah, M., Jiang, N. & Hilvert, D. Exploring the active site of chorismate mutase by combinatorial mutagenesis and selection: the importance of electrostatic catalysis. *Proceedings of the National Academy of Sciences* **93**, 5043–5048 (1996).

34. Bukhari, A. I. & Taylor, A. L. Genetic analysis of diaminopimelic acid- and lysine-requiring mutants of Escherichia coli. *J Bacteriol* **105**, 844–854 (1971).

35. Jiang, M. & Zhang, H. Engineering the shikimate pathway for biosynthesis of molecules with pharmaceutical activities in *E. coli*. *Current Opinion in Biotechnology* **42**, 1–6 (2016).

36. Bassalo, M. C. *et al*. Deep scanning lysine metabolism in Escherichia coli. *Molecular Systems Biology* **14**, e8371 (2018).

37. Lassila, J. K., Keeffe, J. R., Oelschlaeger, P. & Mayo, S. L. Computationally designed variants of Escherichia coli chorismate mutase show altered catalytic activity. *Protein Eng Des Sel* **18**, 161–163 (2005).

38. Soares da Costa, T. P. *et al*. How essential is the 'essential' active-site lysine in dihydrodipicolinate synthase? *Biochimie* **92**, 837–845 (2010).

39. Griffin, M. D. W., Dobson, R. C. J., Gerrard, J. A. & Perugini, M. A. Exploring the dihydrodipicolinate synthase tetramer: how resilient is the dimer-dimer interface? A*rch Biochem Biophys* **494**, 58–63 (2010).

40. Geng, F., Chen, Z., Zheng, P., Sun, J. & Zeng, A.-P. Exploring the allosteric mechanism of dihydrodipicolinate synthase by reverse engineering of the allosteric inhibitor binding sites and its application for lysine production. *Appl Microbiol Biotechnol* **97**, 1963–1971 (2013).

41. Mathis, A. D., Otto, R. M. & Reynolds, K. A. A simplified strategy for titrating gene expression reveals new relationships between genotype, environment, and bacterial growth. *Nucleic Acids Res* doi:10.1093/nar/gkaa1073.

DOI: 10.5281/zenodo.15277418