Precision engineering of biological function with large-scale measurements and machine learning

Authors: Drew S. Tack[1], Peter D. Tonner[1], Abe Pressman[1], Nathanael D. Olson[1], Sasha F. Levy[2,3], Eugenia F. Romantseva[1], Nina Alperovich[1], Olga Vasilyeva[1], David Ross[1*]

Affiliations

1. National Institute of Standards and Technology, Gaithersburg, MD, 20899, USA
2. SLAC National Accelerator Laboratory, Menlo Park, CA, 94025, USA
3. Joint Initiative for Metrology in Biology, Stanford, CA, 94305, USA

*Correspondence to: david.ross@nist.gov.

Abstract

As synthetic biology expands and accelerates into real-world applications, methods for quantitatively and precisely engineering biological function become increasingly relevant. This is particularly true for applications that require programmed sensing to dynamically regulate gene expression in response to stimuli. However, few methods have been described that can engineer biological sensing with any level of quantitative precision. Here, we present two complementary methods for precision engineering of genetic sensors: *in silico* selection and machine-learning-enabled forward engineering. Both methods use a large-scale genotype-phenotype dataset to identify DNA sequences that encode sensors with quantitatively specified dose response. First, we show that *in silico* selection can be used to engineer sensors with a wide range of dose-response curves. To demonstrate *in silico* selection for precise, multi-objective engineering, we simultaneously tune a genetic sensor's sensitivity (EC_{50}) and saturating output to meet quantitative specifications. In addition, we engineer sensors with inverted dose-response and specified EC_{50}. Second, we demonstrate a machine-learning-enabled approach to predictively engineer genetic sensors with mutation combinations that are not present in the large-scale dataset. We show that the interpretable machine learning results can be combined with a biophysical model to engineer sensors with improved inverted dose-response curves.

Introduction

As the field of synthetic biology transitions from a qualitative, trial-and-error endeavour into a mature engineering discipline, methods that enable the engineering of biological function with quantitative precision are required, i.e., to produce an outcome that meets a quantitative specification. This need is particularly acute for genetic sensors, which form the basis for synthetic gene circuits and related approaches for programming cells to regulate gene expression dynamically in response to environmental stimuli.

Most efforts to engineer genetic sensors have been qualitative in nature, e.g., demonstrations of new sensor architectures or sensors that respond to new inputs [1-6]. Those qualitative demonstrations are the necessary first steps in developing a toolkit of sensors for synthetic biology and for demonstrating the variety of cellular control circuits enabled by genetic sensors. However, for many applications, genetic sensors will also need to be engineered with a quantitatively specified dose-response curve matched to each application [2, 4, 7-10]. That dose-response curve is typically described using a version of the Hill equation:

$$G(L) = G_0 + \frac{G_\infty - G_0}{1 + \left(\frac{EC_{50}}{L}\right)^n},$$

where L is the input signal level (e.g., concentration of ligand); $G(L)$ is the regulated gene expression output from the sensor as a function of the input signal; G_0 is the basal output level; G_∞ is the saturating output level; EC_{50} is the input level required to give an output midway between G_0 and G_∞; and n is the Hill coefficient, which quantifies the steepness of the dose response.

Although the importance of tuning the dose response of genetic sensors has been recognized for applications such as engineered living therapeutics, dynamic pathway control, and enzyme engineering [2, 4, 7-9, 11, 12], very few methods have been described that can accomplish the required tuning with any level of quantitative precision or accuracy. With RNA-based genetic sensors (e.g., riboswitches), the relatively predictable biophysics of base-pair interactions has enabled methods to engineer new sensors with quantitatively predictable G_0 and G_∞ [13, 14]. For protein-based genetic sensors, general guidelines have been given for tuning dose-response curves [7, 10, 15, 16], and several methods have been demonstrated to improve sensor performance by reducing EC_{50} or increasing the dynamic range (G_∞/G_0) [17-26]. But no methods have yet been described that can engineer protein-based sensors with specific quantitative values for the parameters of the Hill equation.

Here, we leverage a large-scale, genotype-phenotype dataset to demonstrate two methods for quantitatively precise engineering of protein-based genetic sensors: *in silico* selection, and forward engineering enabled by machine-learning (ML). With *in silico* selection, we mine the large-scale dataset to find DNA sequences that encode genetic sensors that meet quantitative specifications. We show that *in silico* selection can be used to engineer genetic sensors with EC_{50} values spanning a wide range (from 3 µmol/L to over 1000 µmol/L) and with quantitative accuracy (within about 1.3-fold). In addition, we demonstrate *in silico* selection for precise, multi-objective engineering: first, by engineering genetic sensors with both EC_{50} and G_∞ within about 1.2-fold of specified values; and second, by engineering sensors with inverted dose-response and EC_{50} within about 2-fold of specified values. With ML-enabled forward engineering, we use the large-scale dataset to train an interpretable ML model, and we show that the model can predict both EC_{50} and G_∞ for novel combinations of mutations, also with high accuracy (within 1.9-fold and 1.2-fold for EC_{50} and G_∞, respectively). Finally, we use results from the interpretable ML model in combination with guidance from a biophysical model, to engineer new inverted LacI variants with improved EC_{50} and G_∞.

Results

Many previous publications have described the effects of protein mutations on genetic sensor dose-response curves. However, we are not aware of any previous work that has demonstrated the use of protein mutations to tune a genetic sensor dose-response curve to meet quantitative specifications. So, the objectives of this manuscript are to demonstrate methods whereby protein mutations can be used for quantitative tuning of dose-response curves and to test the accuracy and precision of those methods. To that end, the primary statistic we will use to assess different methods is the fold-accuracy: $\exp\left[\mathrm{RMSE}\left(\ln(x)\right)\right]$, where x is the parameter to be tuned (e.g., EC_{50}, G_∞ from the Hill equation), and

$\text{RMSE}\big(\ln(x)\big)$ is the root-mean-square difference between the logarithm of the actual value of *x* and the logarithm of the targeted or predicted value of *x*. We use the logarithmic scale to assess accuracy because the parameters of a genetic sensor dose-response curve can span multiple orders of magnitude and because the resulting fold-accuracy is the most suitable metric for applications of engineered genetic regulatory networks [27].

The methods we demonstrate here both require a large-scale genotype-phenotype dataset as a starting point (e.g., deep mutational scanning). For that, we used a recently published dataset that contains dose-response curves for over 60,000 variants of a protein-based genetic sensor, the *lac* repressor, LacI [28]. Briefly, to create the large-scale genotype-phenotype dataset, error-prone PCR was used to generate a library of LacI variants with an average of 7.0 DNA mutations and 4.4 missense mutations (i.e., amino acid substitutions) per coding sequence. The library was barcoded and a growth-based barcode counting assay was used to measure the dose-response curve, $G(L)$, for every variant in the library. Each dose-response curve was fit to the Hill equation to provide estimates for the Hill equation parameters and their associated uncertainties. In addition, long-read sequencing was used to measure the full-length protein coding sequence for each barcoded variant.

Precision engineering via *in silico* selection

The concept of *in silico* selection is fairly simple: use the large-scale dataset as a lookup table to identify variants with desired phenotypes along with their matching genotypes. That information can then be used to synthesize DNA sequences that will result in the required protein phenotype (i.e., dose-response curve). The keys to successful precision engineering with *in silico* selection are the number of measured variants and the diversity of phenotypes spanned by the large-scale dataset. The dataset must include sufficient diversity to cover the range of functional outcomes needed for the engineering objectives. For example, the LacI dataset includes variants with EC_{50} values from less than 1 µmol/L to over 1000 µmol/L (Fig 1). So, with that dataset, it should be possible to engineer LacI variants with a wide range of EC_{50} values. As a first test of the *in silico* selection approach, we used the genotype-phenotype dataset to identify a set of LacI variants with EC_{50} ranging from about 3 µmol/L to over 1000 µmol/L (and with G_0 and G_∞ near the wild-type values). For each of those variants, we then synthesized the LacI coding sequence, integrated it into a plasmid where it regulated the expression of a fluorescent protein, and measured the resulting *in vivo* dose-response curves using flow cytometry (Fig 2A). The results indicate a fold-accuracy of 1.67 for engineering LacI variants with different EC_{50} values (Fig 2B; where we calculate the fold-accuracy as described above, using EC_{50} reported in the large-scale dataset as the predicted values and EC_{50} determined by flow cytometry as the actual values). However, there is a systematic error between the cytometry measurements and the large-scale dataset: at low EC_{50}, the cytometry result tends to be higher than the large-scale result, while at high EC_{50}, the cytometry result tends to be lower (Fig 2C). After correcting for this systematic error (using a linear fit to the $\ln(EC_{50})$ data shown in Fig 2B for the predicted values), we calculate a best-case fold-accuracy of 1.31 for *in silico* selection of EC_{50}.

In addition to providing quantitative accuracy and precision for a single phenotypic parameter, *in silico* selection is particularly well suited to multi-objective optimization of protein function. With *in silico* selection, one can simply search the large-scale dataset for sequence variants that satisfy multiple criteria simultaneously. This avoids the need for complicated multi-objective Darwinian selection schemes that are necessary for directed evolution. Both EC_{50} and G_∞ need to be quantitatively tuned for optimal dynamic control of a metabolic pathway using a genetic sensor [9]. So, to demonstrate multi-objective optimization with *in silico* selection, we first defined a set of quantitative specifications for EC_{50} and G_∞. For those specifications, we chose a grid of EC_{50} and G_∞ values with EC_{50} equal to 10 µmol/L, 30 µmol/L, or 100 µmol/L, and with G_∞ equal to 16 kMEF or 25 kMEF (the units, MEF, are molecules of equivalent fluorescein from the calibration of cytometry data with fluorescent beads, see Materials and Methods). Next, we used the large-scale dataset to identify the DNA sequences most likely to encode LacI variants with both EC_{50} and G_∞ close to the specified values (after correcting for the systematic error in EC_{50}). In most cases, we chose the top three sequences for each specification (ranked by the probability of EC_{50} within 1.2-fold and G_∞ within 1.1-fold of the target, based on the large-scale measurement uncertainty). For EC_{50} = 100 µmol/L, G_∞ = 16 kMEF, the top two sequences were very similar (encoding for the missense mutation V95M, plus mutations to the disordered loops near the LacI tetramer helix), so for this specification, we also chose the fourth-ranked sequence. The specification EC_{50} = 100 µmol/L, G_∞ = 25 kMEF is very close to the wild-type LacI phenotype, so we did not choose any sequences for that specification. We then synthesized each sequence, integrated it into a plasmid where it regulated the expression of a fluorescent protein, and measured the resulting *in vivo* dose-response curves using flow cytometry (Fig 3A). Comparing the cytometry results with the corresponding multi-objective specifications, the *in silico* selection approach showed good performance, with 1.22-fold and 1.14-fold accuracy for EC_{50} and G_∞, respectively. However, there was some systematic deviation from the targeted G_∞ for specifications with G_∞ = 25 kMEF (Fig 3B).

As a final test of the *in silico* selection approach, we used it to engineer LacI variants with inverted dose-response ($G_\infty < G_0$) and with specified EC_{50}. To identify sequences from the large-scale dataset, we used criteria similar to those described above to choose the sequences most likely to encode inverted LacI variants with EC_{50} equal to 10 µmol/L, 30 µmol/L, or 100 µmol/L. The dataset contains a much lower density of inverted variants (Fig 1C, $G_\infty/G_0 < 1$). So, for each target specification, there was only a single sequence with a greater than 10% probability of having an EC_{50} within 1.5-fold of the targeted value (based on the uncertainty of the large-scale results). The sparsity of inverted variants is at least partially due to the FACS pre-screening that was applied before the large-scale measurement to reduce the fraction of variants with high G_0 [28], which would have removed all inverted variants from the measured library had it been perfectly efficient.

As before, we synthesized the sequences identified by *in silico* selection, and we measured the *in vivo* dose-response curves for the resulting LacI variants with flow cytometry (Fig 4A). All three variants had inverted dose-response curves with G_0 and G_∞ satisfying the targeted specification (G_0 within 1.3-fold of 25 kMEF and G_∞ < 12.5 kMEF, Fig 4B). However, for each of the sequences, the resulting EC_{50} was higher than the targeted values (by 1.9-fold, 2.6-fold, and 1.6-fold for targeted EC_{50} of 10 µmol/L, 30 µmol/L, and 100 µmol/L, respectively).

To determine whether the deviations from the targeted EC_{50} were due to systematic errors in the large-scale measurement, we synthesized and measured the dose-response for eight additional sequences, selected only based on the inverted phenotype (without a specified EC_{50}). The cytometry results confirm that all eight variants have inverted dose-response curves (Fig 5). Furthermore, the results indicate an accuracy of 2.8-fold for EC_{50} of the inverted variants, with no systematic bias (Fig 6). The lower accuracy for the inverted variants (compared with the results in Fig 2B) is consistent with the estimated uncertainty of the large-scale measurements, and is due to the FACS pre-screening, which reduced the number of barcode reads associated with each inverted variant.

ML-enabled forward engineering

For some applications, it can be important to predict the phenotype resulting from combinations of mutations that are not present in the large-scale dataset (e.g., to apply sequence constraints that could not be easily applied during construction of the large-scale library). In those situations, the large-scale data can be used to train a machine-learning (ML) models that can then be used to predict the phenotype resulting from novel combinations of mutations. To demonstrate this approach, we used the large-scale LacI dataset to train an ML model using LANTERN, a recently described approach that learns interpretable models of genotype-phenotype landscapes and that also provides good predictive accuracy (e.g., as good or better than neural network models) [29]. We used the resulting model to predict EC_{50} and G_{∞} for 33 variants with mutation combinations that are not found in the large-scale dataset – and using only a restricted set of 16 missense mutations. We chose the 16 mutations to give a range of different effects on the dose-response, and we used mutations distributed across the LacI core domain (Fig 7, Table S1) but avoided mutations to the DNA binding domain that might disrupt interactions between LacI and its cognate DNA operator [21]. We then synthesized the LacI sequences for the 33 variants, measured their dose-response with cytometry, and compared the results with the predictions from the LANTERN model. Overall, the prediction accuracy of the LANTERN model was nearly as good as the accuracy of the underlying measurements, with 1.93-fold and 1.19-fold accuracy for EC_{50} and G_{∞}, respectively (Fig 8).

Surprisingly, five of the 33 variants had inverted dose-response curves, and all five had the same missense mutation: V136E. In addition, two double mutants with the V136E mutation had non-monotonic dose-response: the double mutant V136E/G200C had a band-stop dose-response curve (referred to as the "reversed" phenotype in earlier literature [30-36]); and the double mutant V136E/S279T had a more complicated non-monotonic dose-response (high-low-high-low). We did not include the data for V136E/G200C or V136E/S279T in the quantitative comparison (Fig 8), because it did not match the form of the Hill equation. The single mutation V136E, applied to the wild-type background, gives a dose-response with reduced G_{∞} but G_0 and EC_{50} similar to the wild type (Fig 7). Previous work has shown that single mutations that reduce G_{∞} relative to the wild-type can be intermediates toward the evolution of the inverted phenotype [37-39], though V136E is located more on the periphery of the protein structure than the intermediate mutations in those previous studies. The prediction accuracy for the five inverted variants was generally poor, particularly for EC_{50}. This discrepancy was not surprising: the large-scale dataset used to train the model contained few examples of inverted variants, and so the model could not learn to predict them. If we consider only the 28 non-inverted variants tested, the prediction accuracy of the LANTERN model improves significantly for EC_{50} (1.31-fold) but only slightly for G_{∞} (1.17-fold).

In addition to accurately predicting phenotype from genotype, LANTERN learns interpretable models [29]. Part of this interpretability comes from the way LANTERN learns to represent the effect of each mutation. LANTERN represents each mutational effect as a vector in a low dimensional latent space (three dimensions for the LacI dataset), and the combined effect of multiple mutations is simply represented as the sum of the corresponding vectors. The different components of the latent vector space learned by a LANTERN model often resemble a set of latent biophysical parameters (e.g., free energies) that control the protein phenotype. However, the latent parameters learned by a LANTERN model are unlabeled, meaning that while a connection between the parameters learned by LANTERN and biophysical parameters may exist, the model does not identify this connection. But, when an explicit biophysical model is available, it can potentially be linked to the parameters learned by LANTERN. This has been demonstrated qualitatively for a biophysical model of LacI function [40-43] and the LANTERN model trained on the large-scale LacI dataset [29]. More specifically, the first (most significant) latent parameter learned by the LANTERN model seems to correspond to changes to any one of three parameters in the biophysical model (the binding free energy for LacI to its DNA operator, $\Delta\varepsilon_{RA}$; the logarithm of the LacI allosteric constant, $\Delta\varepsilon_{AI}$; or the ligand binding constant for the inactive state of LacI, K_I; using the notation of [40, 42]). The second latent parameter, however, seems to correspond to changes to a single parameter in the biophysical model (the ligand binding constant for the active state of LacI, K_A)

To see if this potential link between LANTERN and biophysics could be used in forward engineering, we attempted to use the LANTERN model results together with insight from the biophysical model to engineer improved inverted LacI variants. Most inverted LacI variants in the large-scale dataset have relatively high EC_{50}, and they are also somewhat leaky (G_∞ > 1000 MEF, compared with G_0 = 158 MEF for wild-type LacI). Based on the biophysical model, both EC_{50} and G_∞ of inverted variants can be reduced by decreasing the ligand binding constant for the active state, K_A, which tentatively corresponds to an increase in the second latent parameter of the LANTERN model. So, we chose three mutations with a significant predicted increase in that second latent parameter (S70R, V80L, and V136E). We synthesized and tested LacI variants composed of those mutations added onto the background sequences for two genetically distinct inverted variants. In both inverted backgrounds, the mutation V80L reduced EC_{50} by a factor of 5 or 6, and reduced G_∞ by a factor of about 1.3 (Fig 9, blue). The other two mutations, however, did not have the intended effect: S70R increased EC_{50} in both inverted backgrounds (Fig 9, orange), and V136E resulted in constitutively high output (Fig 9, green). Although imperfect, this initial test of linking an interpretable, data-driven ML model to a biophysical model to engineer genetic sensors shows promise for engineering difficult-to-access phenotypes that differ significantly from the wild type.

Discussion

We have demonstrated two approaches for precision engineering of genetic sensors and quantitatively evaluated their accuracy and the range of engineered phenotypes they can access. With *in silico* selection, we engineered sensors with EC_{50} values spanning nearly three orders of magnitude with high precision (1.3-fold). In addition, we demonstrated that *in silico* selection can be used for facile, multi-objective engineering to give genetic sensors with specified values for both EC_{50} and G_∞, and with high accuracy relative to pre-defined specifications (1.22-fold and 1.14-fold for EC_{50} and G_∞, respectively). We also showed that *in silico* selection can be used for multi-objective engineering of more difficult and rare phenotypes: inverted sensors with specified EC_{50}, though with lower accuracy due to the relative sparsity of inverted variants in the large-scale dataset (1.6-fold to 2.6-fold for EC_{50}). With ML-enabled forward engineering we demonstrated that an ML model can be trained with a large-scale genotype-phenotype landscape dataset, and that model can then be used to predict the dose-response of new mutation combinations, again with good accuracy (1.3-fold to 1.9-fold for EC_{50} and ~1.2-fold for G_∞). We further demonstrated that an interpretable ML model can be used together with insight from a more explicit biophysical model to engineer inverted genetic sensors with improved EC_{50} and G_∞. To get a baseline for comparison of the performance of the precision engineering approaches, we measured multiple replicate dose-response curves for wild-type LacI (two biological replicates, with a total of 15 technical replicates measured on six different days). Across those wild-type replicates, the geometric standard deviation was 1.16-fold, 1.22-fold, and 1.11-fold, for EC_{50}, G_0, and G_∞, respectively.

For both approaches to precision engineering, it is important that the large-scale dataset contains sequence variants with multiple mutations, i.e., not just data for variants with single amino acid substitutions. Similarly, the dataset must contain results specifically related to each variant in the measured library rather than just an enrichment score associated with each mutation. With *in silico* selection, if we restrict the dataset to only single-mutant variants, the expected probability for success (i.e., engineering a dose-response satisfying the specification) drops significantly (Supplementary Information). Also, there are no single-mutant variants in the dataset expected to satisfy the specifications farthest from the wild-type (inverted dose response; or G_∞ = 16 kMEF and EC_{50} = 10 µmol/L or 30 µmol/L; Table S2). So, with only single mutations, the range of phenotypes that can be engineered becomes more limited. Multi-mutant variants are also important for training the ML model, since multi-mutant data are required to make predictions for new mutation combinations without strong assumptions about the additivity and linearity of mutational effects [44].

To compare the accuracy demonstrated here with previous work, we are only able to find four examples of quantitative evaluation of predicted vs. measured genetic sensor dose-response. Two of those were for RNA-based sensors, and the other two were focused on engineering the dose-response of protein-based genetic sensors by varying the sequence of the cognate DNA operator (while using the wild-type protein sequences). Those previous publications included quantitative results for G_0 and G_∞ (or the ratio G_∞/G_0), and one included results for $G(L)$, but none of them included quantitative results for EC_{50}. Borujeni et al. developed a biophysical modeling approach to engineer RNA-based genetic sensors [13]. They tested the accuracy of the model by measuring the response of 67 riboswitches and showed that their model could predict the activation ratio, G_∞/G_0, with approximately 2.5-fold accuracy (i.e., within 2-fold of the correct value for 55 % of the tested riboswitches). However, their model was less accurate for calculating the values of G_0 and G_∞ rather than their ratio (~8-fold and ~6-fold accuracy respectively). Angenent-Mari et al. trained several deep neural network models using a large-scale genotype-phenotype dataset for RNA toehold switches [14]. Their best model was able to predict G_0 and G_∞ with about 3-fold accuracy. Yu et al. developed a biophysical model to predict how changes in promoter architecture and sequence affect G_0 and G_∞ [45]. Their model was able to predict G_0 and G_∞ with 1.6-fold accuracy across a set of 8269 designed *lac* operators (i.e., predictions within 2-fold of the true value 87% of the time). Zhou et al. used dose-response measurements for protein-based genetic sensors with 2632 combinatorically designed operator sequences to train regression models for $G(L)$ at each ligand concentration (L). Their best model had a predictive accuracy of about 1.2-fold [46]. By comparison, in our demonstration of the *in silico* selection method, all 16 of the engineered sensors with data shown in Fig 3 had both EC_{50} and G_∞ within 2-fold of the specified target values, and two of the three inverted sensors (Fig 4) had EC_{50} within 2-fold or the target value. Also, our data-driven ML model was able to correctly predict EC_{50} and G_∞ within 2-fold for 76 % and 97 % of the tested LacI variants, respectively.

If we broaden our comparisons to include predictive models for constitutive gene expression, the best-known examples are probably the various models for predicting the translation initiation rate from ribosomal binding site (RBS) sequences [47-52]. In a recent evaluation of several of those models using data for nearly 10,000 RBS sequences, the models' predictive accuracy ranged from approximately 1.85-fold to 11-fold (between 23 % and 74 % predicted within 2-fold of the measured value), with the most recent iteration of the RBS calculator giving the best performance [53]. A biophysics-based model was also demonstrated for terminator strength in *E. coli*, with approximately 3.9-fold accuracy across a set of 582 natural and synthetic designed terminators [54]. More recently, LaFleur et al. developed a biophysical model for the strength of promoters in *E. coli* [55]. That model was able to correctly predict *in vitro* transcription rates with 1.6-fold accuracy across a set of 5388 designed promoters (i.e., within 2-fold of the correct value 92 % of the time), though it was less accurate for *in vivo* systems (approximately 2-fold accuracy). Similar predictive models of promoter function have been developed for eukaryotic cells [56-59]. However, those reports only evaluated model performance using the correlation coefficient, and the data comparing predicted and measured results are not available as part of the reports' data supplements. So, it is not possible to estimate the predictive fold-accuracy of those models with the available information.

In summary, the precision engineering approaches described here have very good accuracy compared with previous quantitative results. The question of how accurate an engineering method would need to be will depend on specific applications. Beal et al. have estimated that a target accuracy of 1.5-fold would be sufficient for most applications requiring engineered genetic regulatory networks [27].

The use of interpretable ML modeling in conjunction with a biophysical model also has the potential to become a useful engineering approach, as demonstrated here for the engineering of improved inverted LacI variants. But more rigorous methods would be needed to link the latent parameters of the ML model to the biophysical parameters before that approach could be used for engineering with quantitative precision. An alternative would be to fit the large-scale dataset directly with a biophysical model, if an appropriate model is available. One outstanding problem is that estimation of biophysical parameters from phenotype measurements can be ambiguous [60, 61]. A large-scale measurement approach, with measurements of many different multi-mutation combinations could help to overcome ambiguity, since it provides information on mutational effects across many different genetic backgrounds that can help resolve those ambiguities [62]. However, that kind of approach will probably prove much more challenging for protein-based genetic sensors, where the same change to the dose-response curve can be explained by changes to several different biophysical parameters as shown by Razo-Mejia et al. [42] and demonstrated in our experience fitting the large-scale LacI dataset with a LANTERN model as discussed above.

For most applications, there will be some shift in context between the large-scale measurement and the application (e.g., a change in strain, growth conditions, and/or the genes that are regulated by the sensor). Ultimately, successful use of the methods described here will depend on the ability to predict how a genetic sensor's dose-response curve will change in response to those types of context shifts. The types of biophysical models discussed above, whether used in conjunction with interpretable ML or fit directly to data, provide a promising solution to the challenge of predicting function across different contexts. For example, Razo-Mejia et al. developed a biophysical model for allosteric regulation with LacI, and showed that it could accurately predict changes to the dose-response curve due to changes in LacI copy number or the interaction strength between LacI and its cognate operator [42]. Chure, Kaczmarek, and Phillips then demonstrated that the same model could accurately predict changes in the basal output level, G_0, due to cell growth at different temperatures and with different carbon sources [63]. Notably, Chure, Razo-Mejia, et al. showed that the model could also be used to predict changes in dose-response resulting from combinations of mutations (using single-mutant data) [40]. Although they did not include a quantitative evaluation of the accuracy of those predictions, it appears to be quite good (e.g., six of six predicted EC_{50} within 2-fold of the correct value, based on a visual inspection of Fig. 5A in [40]). Sochor showed that a similar biophysical model could be used to predict the *in vivo* dose-response curve of LacI using data from *in vitro* transcription measurements [64]. Finally, the model developed by LaFleur et al. [55] can predict changes in gene expression due to changes in sequence context upstream and downstream of a promoter site. So, although quantitative prediction of the effects of different biological contexts remains one of the outstanding challenges in the field [65], for genetic sensors at least, promising solutions exist. Admittedly, if biophysical models (or other means) are needed to correct for shifts in context between the large-scale measurement and the application, that will add an additional layer of uncertainty in the use of the methods described here. But that just highlights the need for the best possible quantitative accuracy of the underlying large-scale measurements.

Currently, we are aware of only one large-scale dataset with quantitative results for the dose-response curves of a protein-based genetic sensor: the LacI dataset used here [28]. So, it is not yet possible to fully assess the generalizability of the methods presented here to other proteins. As an indication of the possible generalizability, though, we can compare the basic requirements of our methods with the requirements for directed evolution: both rely on the ability to generate phenotypic diversity via protein mutations. Directed evolution and related methods have been used to qualitatively improve a large variety of protein-based genetic sensors [17-26], in some cases with a single round of mutagenesis and a library diversity comparable to number of variants in the LacI dataset (10^4 to 10^5 variants) [19-21, 26]. Furthermore, in an approach similar to the *in silico* selection method described here, Ogawa et al. used deep mutational scanning data for a library of single-mutant XylS variants to identify mutations that alter the ligand specific of that protein-based genetic sensor [66]. So, as large-scale genotype-phenotype measurements become more accessible, we expect that the type of precision engineering approaches described here could be readily generalized to engineer different types of genetic sensors or other complex biological functions.

Compared with our approach, directed evolution has the advantage that it can be implemented with very large libraries of sensor protein variants: as many as 10^8, compared with ~10^5 for the LacI dataset used here. So, we think that directed evolution methods will remain important for engineering new, hard-to-access protein functions, such as sensitivity to new ligands [6, 10, 67]. However, it would be very difficult to implement a directed evolution method for precision sensor engineering, for example to give a quantitatively specified EC_{50}. Similarly, promising new methods have been demonstrated for *de novo* computational design of genetic sensors [68], but those methods are unlikely to provide quantitative precision on their own. Therefore, we expect that methods like those described here will ultimate be used in conjunction with directed evolution or computational design, to provide quantitative precision when that is needed for real-world applications.

Materials and Methods

Large-scale dataset

The large-scale dataset for LacI dose-response curves is described in ref[28]. It includes the estimated Hill equation parameters, EC_{50}, G_0 and G_∞, for over 60,000 variants of the LacI genetic sensor, measured in *E. coli*. Those Hill equation parameter estimates, and their associated uncertainties, were obtained by fitting the measured dose-response curve of every variant to the Hill equation. That dataset is available via the NIST Science Data Portal, with the identifier ark:/88434/mds2-2259 (https://data.nist.gov/od/id/mds2-2259 or https://doi.org/10.18434/M32259). Here, we used the Hill equation parameter estimates and uncertainties as they are reported in that dataset.

In silico selection

For the *in silico* selection results shown in Fig. 3, LacI variants were chosen from the large-scale dataset based on the following criteria:

1. EC_{50} within 1.2-fold of the target value (after correcting for systematic errors, see Fig. 2C)
2. G_∞ within 1.1-fold of the target value
3. $G_0 < 2$ kMEF

Those criteria were first applied using the median values reported in the dataset for G_0, G_∞, and EC_{50}. That resulted in multiple LacI variants for each specification (between 18 and 1513). To identify the best variants to synthesize and test, the uncertainty information reported in the dataset was then used to estimate the probability for success of each variant: more specifically, the posterior samples reported in the dataset (from Bayesian estimation of the Hill equation parameters) were used to calculate the probability that each variant would meet the listed criteria. The variants were then ranked based on their probability of success; and the highest ranking three variants were selected for testing.

For the *in silico* selection results shown in Fig. 4, a similar procedure was used to choose LacI variants, with the following criteria:

1. EC_{50} within 1.5-fold of the target value
2. $G_\infty < 12.5$ kMEF
3. 19.2 kMEF $< G_0 <$ 32.5 kMEF

When applied to the median values for G_0, G_∞, and EC_{50}, those criteria were only met by one or two LacI variants for each specification. Also, the calculated probability to meet the listed criteria was greater than 10% for only one variant per specification. So, only a single variant was selected for each specification.

Strains, plasmids, and culture conditions

All reported measurements were completed using *E. coli* strain MG1655Δ*lac* [69], in which the lactose operon of *E. coli* strain MG1655 (ATCC #47076) was replaced with the bleomycin resistance gene from *Streptoalloteichus hindustanus* (*Shble*).

Dose-response curves were measured with flow cytometry using *E. coli* MG1655Δ*lac* transformed with variants of the pVER plasmid, described previously [28]. The plasmid contained different variants of the *lacI* coding DNA sequence (CDS), as described in the text, and an expression cassette with enhanced yellow fluorescent protein (eYFP) under the control of the lactose operator (*lacO*). The *lacI* CDS was verified with Sanger sequencing for each variant.

All cultures were grown in a rich M9 media (3 g/L KH_2PO_4, 6.78 g/L Na_2HPO_4, 0.5 g/L NaCl, 1 g/L NH_4Cl, 0.1 mmol/L $CaCl_2$, 2 mmol/L $MgSO_4$, 4 % glycerol, and 20 g/L casamino acids) supplemented with 50 µg/mL kanamycin.

For flow cytometry measurements, *E. coli* cultures were grown in a laboratory automation system that included an automated liquid handler (Hamilton, STAR), an automated plate sealer (4titude, a4S), an automated de-sealer (Brooks, XPeel), and two multi-mode plate readers (BioTek, Neo2SM).

Cultures were grown in clear polystyrene 96-well plates with 1.1 mL square wells (4titude, 4ti-0255). The culture volume per well was 0.5 mL. Before incubation, each 96-well growth plate was sealed by the automated plate sealer with a gas permeable membrane (4titude, 4ti-0598). Growth plates were incubated in one of the multi-mode plate readers at 37 °C with a 1 °C gradient applied from the bottom to the top of the incubation chamber to minimize condensation on the inside of the membrane. The plate readers were set for double-orbital shaking at 807 cycles per minute. Optical density at 600 nm (OD600) was measured every 5 minutes during incubation, with continuous shaking applied between measurements (optical density at 700 nm and YFP fluorescence were also measured every 5 minutes). After incubation, the automated de-sealer was used to remove the gas-permeable membrane from each 96-well plate to enable automated passaging of cultures and sample preparation for flow cytometry measurements.

For each measurement, starter cultures were prepared from glycerol freezer stock in 5 mL of rich M9 media in a 14 mL snap-cap culture tubes. Starter cultures were incubated at 37 °C with orbital shaking at 300 rpm for between 4 h and 24 h prior to loading the automation system. The automation system then prepared 96-well growth plates, sealed and de-sealed the growth plates, incubated the growth plates, and prepared flow cytometry sample plates. The automated culture protocol consisted of the following steps:

1. Prepare first growth plate, with 450 µL rich M9 media in each well.
2. Pipette 50 µL of starter culture into each well in rows B-G of the plate (leaving rows A and H blank).
    a. Use a *E. coli* containing a different lacI variant for each row.
3. Seal first growth plate with gas permeable membrane.
4. Incubate plate in plate reader for 12 h to 14 h.
    a. Grow to stationary to provide a reproducible starting point for each measurement.
5. Prepare second growth plate with 490 µL in each well.
    a. Dilution series of isopropyl-β-D-thiogalactopyranoside (IPTG): 11 columns of a 2-fold serial dilution gradient and one column with zero IPTG.
6. Ten minutes before the end of the incubation cycle for the first growth plate, move the second growth plate to a heated station set to 47 °C.
    a. Ten minutes at 47 °C will pre-warm the media in the plate to 37 °C.
7. De-seal the first growth plate (after completion of the stationary-phase incubation cycle).
8. Pipette 10 µL from each well in the first growth plate to the corresponding well in the second growth plate.
    a. 50-fold dilution; using a 96-channel pipetting head.
9. Seal second growth plate with gas permeable membrane.
10. Incubate second growth plate in plate reader for 160 minutes.
    a. Sufficient for approximately 10-fold increase in cell density or 3.3 doublings.
11. Prepare third growth plate with 450 µL in each well.
    a. Same dilution series as in second growth plate.
12. Ten minutes before the end of the incubation cycle for the second growth plate, move the third growth plate to a heated station set to 47 °C.
13. De-seal the second growth plate (after completion of the 160 minute incubation cycle).

14. Pipette 50 µL from each well in the second growth plate to the corresponding well in the third growth plate.
    a. 10-fold dilution; using a 96-channel pipetting head.
15. Seal third growth plate with gas permeable membrane.
16. Incubate third growth plate in plate reader for 160 minutes.
17. Prepare flow cytometry sample plate (round-bottom 96-well plate, Falcon, 351177).
    a. Each well in rows B-G: 195 µL 1x PBS with 170 µg/mL chloramphenicol (Fisher BioReagents, cat. #BP904-100).
    b. Rows A and H: PBS blanks, focusing fluid blanks, and space for calibration bead sample
18. At the end of the incubation cycle for the third growth plate, pipette 5 µL from each well to the corresponding well in the flow cytometry sample plate.

At the end of the automated culture protocol, the flow cytometry sample plate was transferred to the flow cytometry autosampler for measurement.

Flow cytometry

Flow cytometry samples were measured with an Attune NxT flow cytometer equipped with a 96-well plate autosampler using a 488 nm excitation laser and a 530 nm ± 15 nm bandpass emission filter. Blank samples were measured with each batch of cell measurements, and an automated gating algorithm was used to discriminate cell events from non-cell events [70]. Fluorescence calibration beads (Spherotech, part no. RCP-30-20A) were also measured with each batch of samples to facilitate calibration of flow cytometry data to molecules of equivalent fluorescein (MEF) [71-73].

For each LacI variant, the dose-response curve was taken to be the geometric mean fluorescence from flow cytometry as a function of the IPTG concentration in the media of the third growth plate. For many variants, data from multiple measurements were used, e.g., from biological or technical replicates, or data across multiple, overlapping IPTG dilution series to extend the range of inducer concentrations. For some biological and/or technical replicates, the cytometry results differed significantly from the consensus results from other replicates (i.e., G_∞ more than 1.25-fold different from the consensus value). Data for those outlier replicates were not used. The Hill equation parameters and their associated uncertainties were determined by fitting all of the non-outlier cytometry data for each variant to the Hill equation using Bayesian parameter estimation by Markov Chain Monte Carlo (MCMC) sampling with PyStan [74].

LANTERN ML modeling

LANTERN was fit to the LacI dataset with methods described in Ref[29]. In this model, LANTERN learns to predict observed phenotypes $y \in R^D$ given a one-hot encoded form of the genotype $x \in \{0,1\}^p$ in two key steps. First, the genotype is projected to a low dimensional space $z = Wx$, where $W \in R^{K \times p}$ and $K \ll p$. Second, LANTERN learns a smooth non-linear surface connecting this low dimensional space to observed phenotypes: $y = f(z)$. Both the matrix W and function $f(z)$ are unknown parameters and are learned by LANTERN in the form of an approximate variational posterior [75].

To quantify the predictive uncertainty of the LANTERN model for individual variants, we approximated the posterior predictive distribution for each variant under the learned model. This was done by taking Monte Carlo draws from learned approximate posterior (fifty draws were taken for each variant). Then, the mean and standard deviation of these draws were used to summarize the posterior predictive interval, as shown in Fig 8.

Acknowledgments

We would like to thank Elizabeth Strychalski, Samuel Schaffter, and Edward Eisenstein for thoughtful comments on the manuscript. S.F.L. is supported by NIH grants R01 HG011676 and R01 AI164530.

Author Contributions

D.S.T., P.D.T., and D.R. designed the experiment.

D.S.T, S.L., and D.R. developed the experimental workflow.

E.F.R., and D.R. programmed automated protocols.

D.S.T., N.A., O.V., and D.R. performed flow cytometry experiments.

P.D.T. performed the machine learning analysis and predictions.

D.S.T. and D.R. wrote the manuscript.

All authors contributed to the manuscript.

Conflict of Interest

The authors declare that they have no conflict of interest. Certain commercial equipment, instruments, or materials are identified to adequately specify experimental procedures. Such identification neither implies recommendation nor endorsement by the National Institute of Standards and Technology nor that the equipment, instruments, or materials identified are necessarily the best for the purpose.

Figures

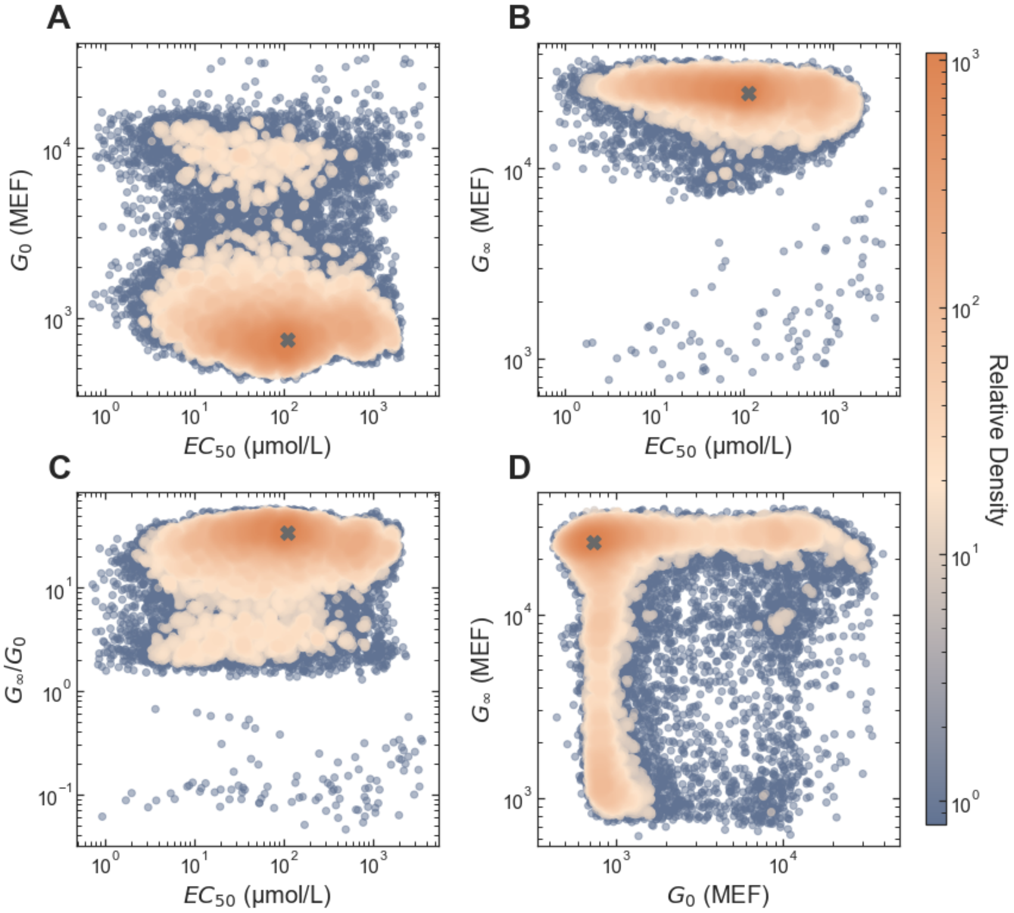

Figure 1. Diversity of dose-response phenotypes in the large-scale dataset. The colored points are the values as reported in the genotype-phenotype dataset, with colors indicating the relative density of similar phenotypes. The gray ‘X’ in each plot shows the parameter values for the wild-type LacI dose-response curve.

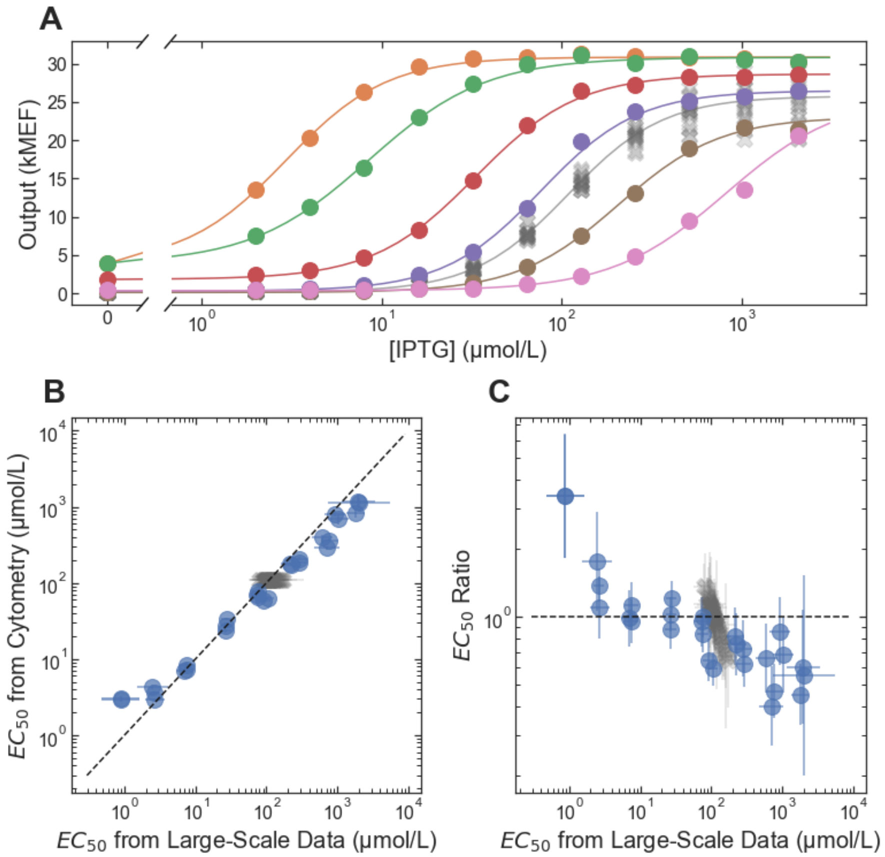

Figure 2. Accuracy and precision of *EC_{50}* from *in silico* selection. (A) Example dose-response curves for LacI variants selected to span a wide range of *EC_{50}* values. Each variant is plotted with a different color, with lines showing the fits to the dose-response using the Hill equation. The wild-type dose response is plotted with the gray 'X' markers. (B) *EC_{50}* from the flow cytometry measurements plotted vs. *EC_{50}* from the large-scale dataset. The dashed line indicates equality between the cytometry and large-scale results. (C) The ratio: (*EC_{50}* from flow cytometry) ÷ (*EC_{50}* from the large-scale dataset) plotted vs. *EC_{50}* from the large-scale dataset. In both B and C, results for non-wild-type LacI variants are plotted with blue circles, and results for wild-type LacI are plotted with gray X's (there were multiple copies of the wild-type in the large-scale dataset, each plotted separately). Error bars indicate ± one standard deviation.

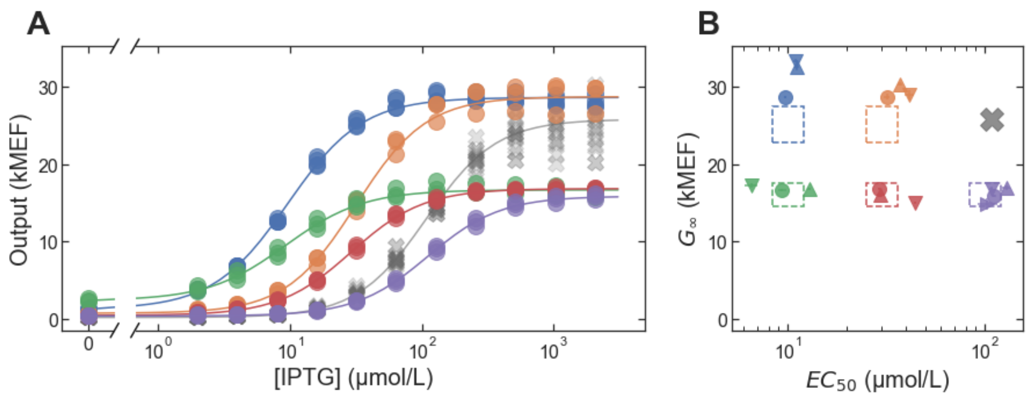

Figure 3. Multi-objective *in silico* selection of LacI variants with different EC_{50} and G_∞ values. (A) Example dose-response curves for LacI variants selected to satisfy multi-objective specifications for EC_{50} and G_∞. One variant is plotted for each target specification, each with a different color and with lines showing the fits to the dose-response using the Hill equation. The wild-type dose response is plotted with the gray 'X' markers. (B) Evaluation of multi-objective selection performance. The dashed rectangles show the target specifications in a 2D plot of G_∞ vs. EC_{50}, with a different color for each specification. For each specification, three or four distinct LacI variants were selected, and the resulting G_∞ and EC_{50} values (from cytometry) for those variants are plotted with different markers (with marker color indicating the targeted specification). Error bars indicate ± one standard deviation and are typically smaller than the markers.

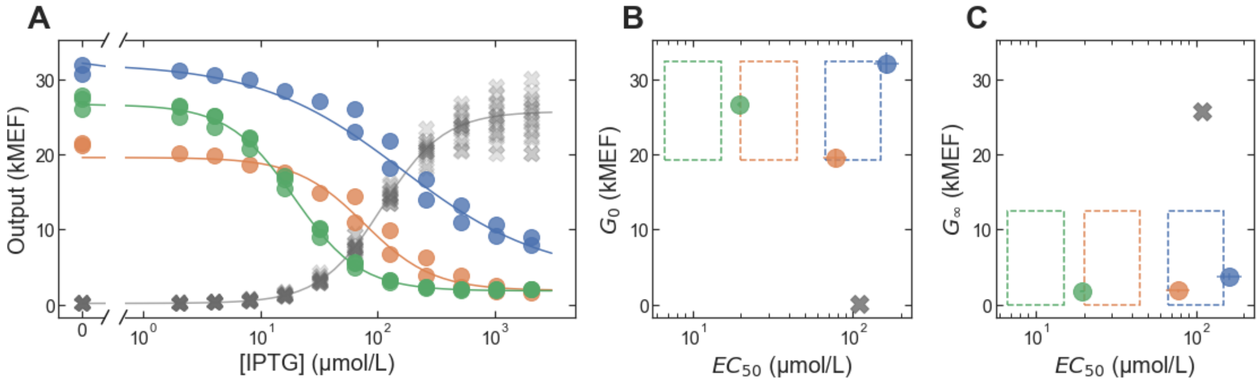

Figure 4. Multi-objective *in silico* selection of inverted LacI variants. (A) Dose-response curves for LacI variants selected to have inverted dose-response curves with specified EC_{50}. One variant is plotted for each target specification, each with a different color and with lines showing the fits to the dose-response using the Hill equation. The wild-type dose response is plotted with the gray 'X' markers. (B-C) Evaluation of multi-objective selection performance. The dashed rectangles show the target specifications in a 2D plot of G_0 (B) and G_∞ (C) vs. EC_{50}, each with a different color. For each specification, one LacI variant was selected, and the resulting G_0, G_∞ and EC_{50} values (from cytometry) for those variants are plotted (with marker color indicating the targeted specification). For comparison, the wild-type G_0, G_∞ and EC_{50} are plotted with gray 'X' markers. Error bars indicate ± one standard deviation and are typically smaller than the markers.

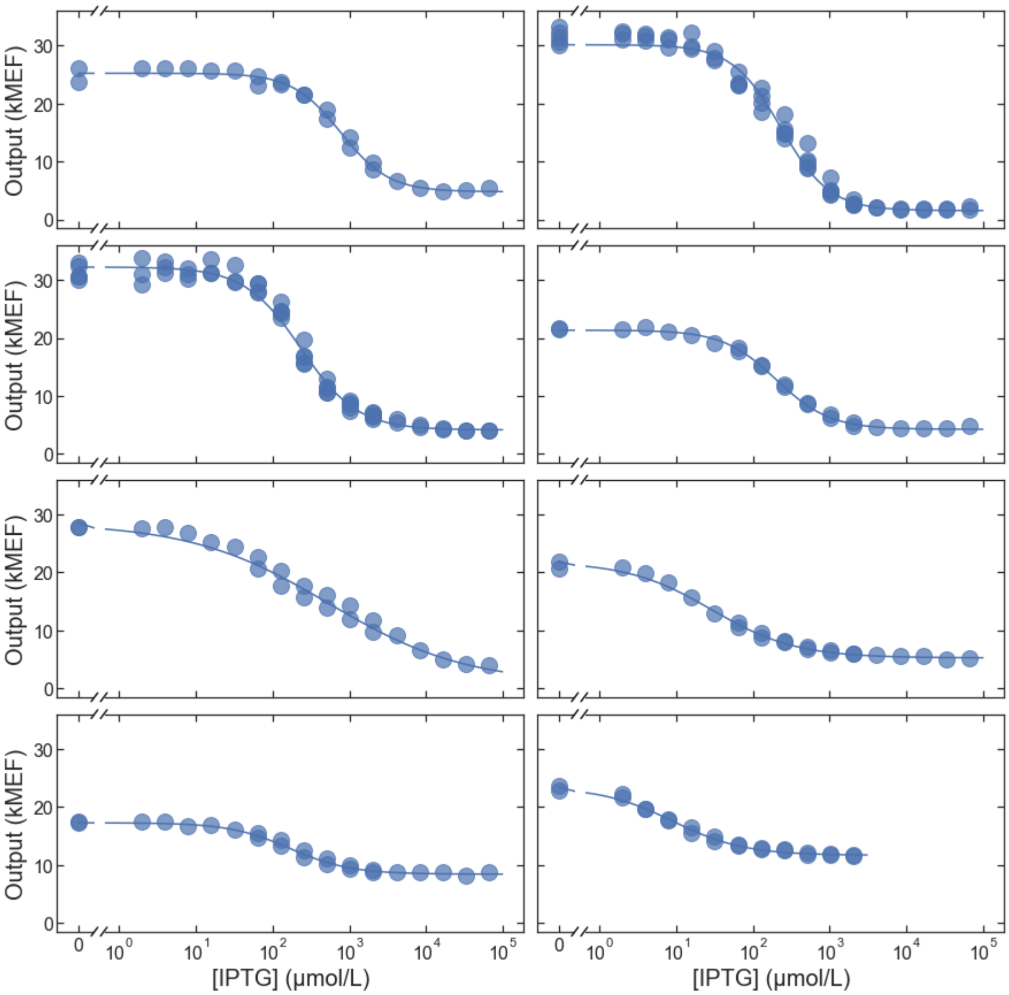

Figure 5. Additional inverted variants. Dose-response curves for eight additional inverted LacI variants selected to test the accuracy of the large-scale measurements.

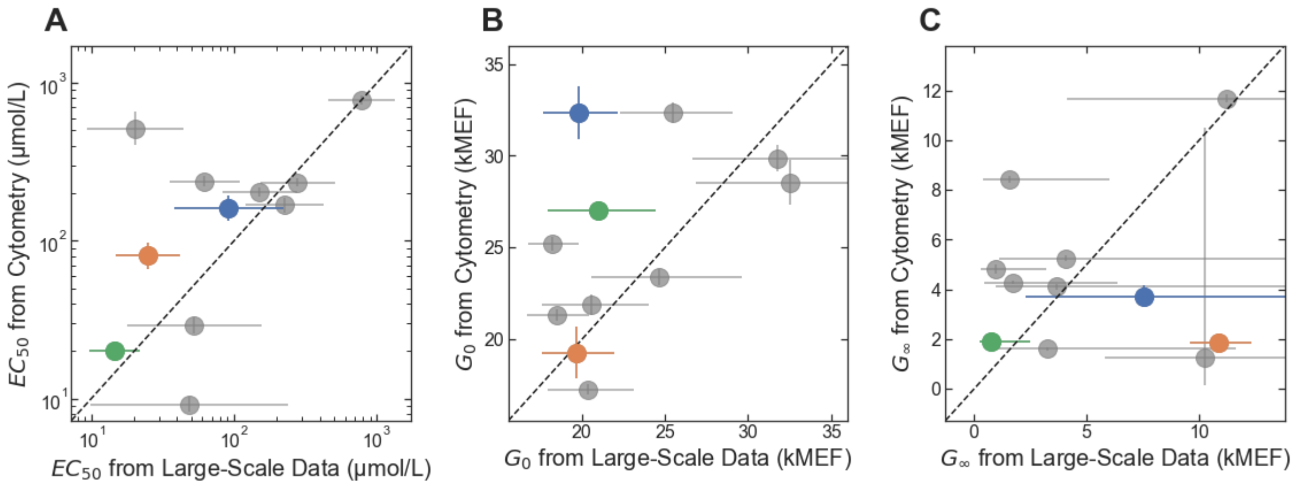

Figure 6. Accuracy of large-scale measurement for inverted variants. (A) EC_{50} from the flow cytometry measurements plotted vs. EC_{50} from the large-scale dataset. (B) G_0 from the flow cytometry measurements plotted vs. G_0 from the large-scale dataset. (C) G_∞ from the flow cytometry measurements plotted vs. G_∞ from the large-scale dataset. In all three plots, results for the inverted variants selected to have specified EC_{50} are plotted with markers colored to match the results in Fig. 4; results for additional inverted variants are plotted with gray markers. Error bars indicate ± one standard deviation.

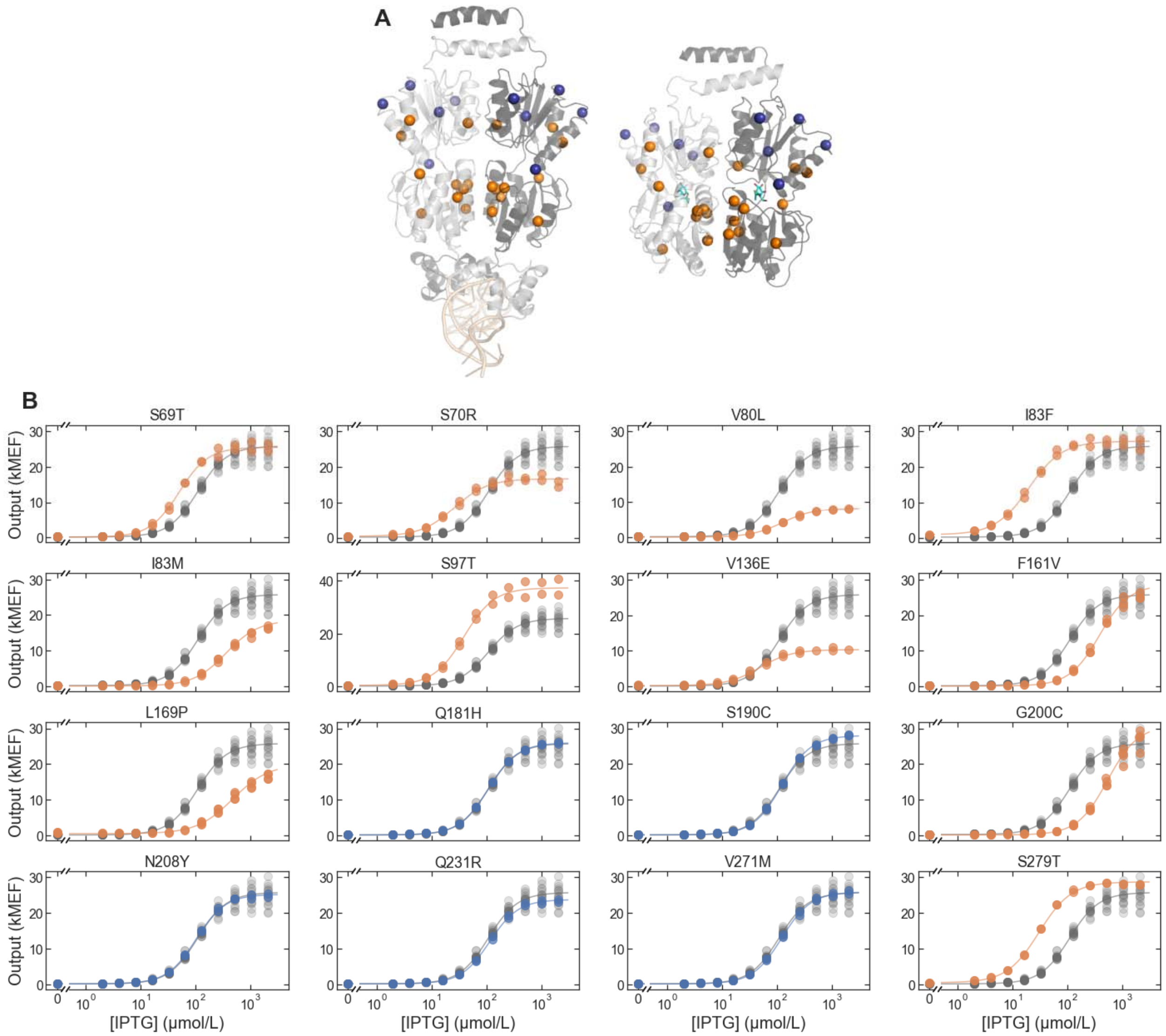

Figure 7. Mutations used for ML-enabled forward engineering. (A) LacI protein structure showing location of mutations. The DNA-binding configuration is shown on the left (DNA at the bottom of the structure in light orange, PDB ID: 1LBG [76]) and the ligand-binding configuration is shown on the right (IPTG in cyan, PDB ID: 1LBH [76]). Both configurations are shown with the view oriented along the protein dimer interface, with one monomer in light gray and the other monomer in dark gray. Colored spheres highlight the positions of mutations used for ML-enabled forward engineering, with silent mutations in blue and non-silent mutations in orange. (B) Dose-response of single-mutant LacI variants with each of the mutations used for ML-enabled forward engineering. In each plot, the single-mutant dose-response is plotted in blue (for silent mutations) or orange (for non-silent mutations), and the wild-type dose response is plotted in gray.

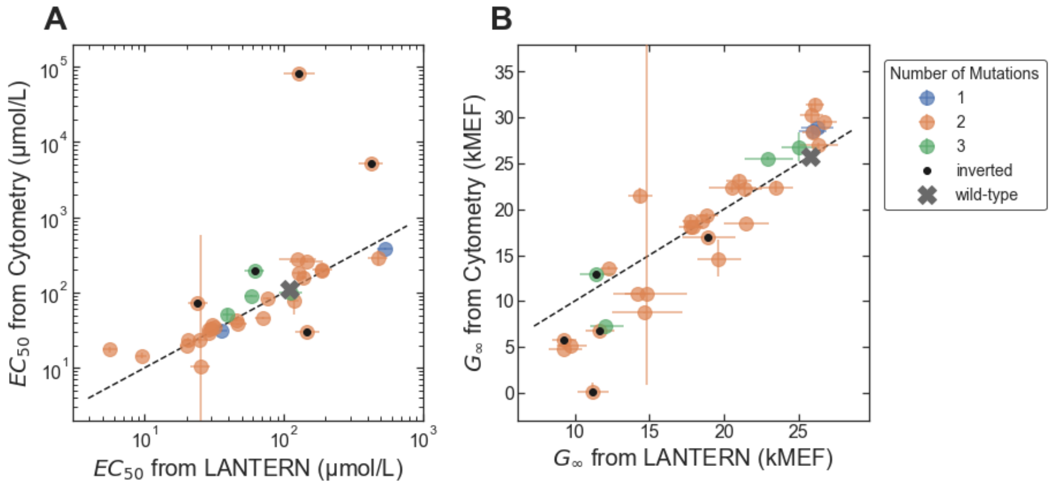

Figure 8. Accuracy of ML-enabled forward engineering. (A) EC_{50} from the flow cytometry measurements plotted vs. EC_{50} predicted by the LANTERN ML model. (B) G_∞ from the flow cytometry measurements plotted vs. G_∞ predicted by the LANTERN ML model. In each plot, results for LacI variants with different numbers of mutations are plotted with different colors. Results for the five unexpectedly inverted variants are marked with black dots. Error bars indicate ± one standard deviation.

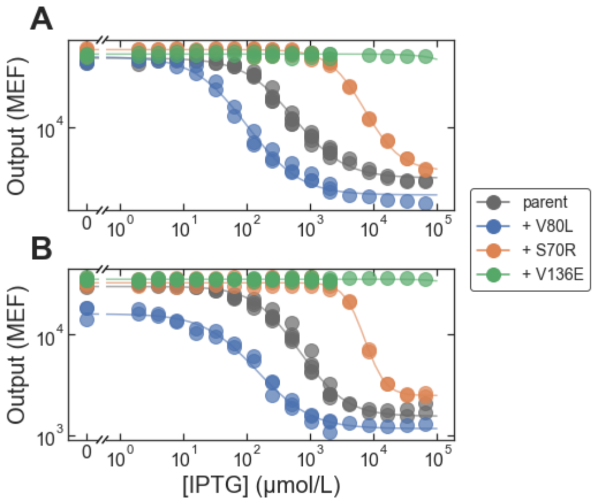

Figure 9. Forward engineering to improve inverted sensors. Each plot shows dose-response curves for a 'parent' inverted LacI variant and for that parent with the addition of mutations chosen to improve the inverted variant (by reducing EC_{50} and G_∞). (A) The parent variant has three missense mutations: A87P, V301M, and E357G. (B) The parent variant has five missense mutations: V96E, T154I, S158R, V238D, M254I, and V264I.

References

1. Shi S, Ang EL, Zhao H. In vivo biosensors: mechanisms, development, and applications. Journal of Industrial Microbiology and Biotechnology. 2018;45(7):491-516. doi: 10.1007/s10295-018-2004-x.

2. De Paepe B, Peters G, Coussement P, Maertens J, De Mey M. Tailor-made transcriptional biosensors for optimizing microbial cell factories. Journal of Industrial Microbiology & Biotechnology. 2017;44(4):623-45. doi: 10.1007/s10295-016-1862-3.

3. Dykstra PB, Kaplan M, Smolke CD. Engineering synthetic RNA devices for cell control. Nature Reviews Genetics. 2022;23(4):215-28. doi: 10.1038/s41576-021-00436-7.

4. Liu D, Evans T, Zhang F. Applications and advances of metabolite biosensors for metabolic engineering. Metabolic Engineering. 2015;31:35-43. doi: https://doi.org/10.1016/j.ymben.2015.06.008.

5. Koch M, Pandi A, Borkowski O, Batista AC, Faulon J-L. Custom-made transcriptional biosensors for metabolic engineering. Current Opinion in Biotechnology. 2019;59:78-84. doi: https://doi.org/10.1016/j.copbio.2019.02.016.

6. Galvão TC, de Lorenzo V. Transcriptional regulators à la carte: engineering new effector specificities in bacterial regulatory proteins. Current Opinion in Biotechnology. 2006;17(1):34-42. doi: https://doi.org/10.1016/j.copbio.2005.12.002.

7. Mannan AA, Liu D, Zhang F, Oyarzún DA. Fundamental Design Principles for Transcription-Factor-Based Metabolite Biosensors. ACS Synthetic Biology. 2017;6(10):1851-9. doi: 10.1021/acssynbio.7b00172.

8. Ang J, Harris E, Hussey BJ, Kil R, McMillen DR. Tuning Response Curves for Synthetic Biology. ACS Synthetic Biology. 2013;2(10):547-67. doi: 10.1021/sb4000564.

9. Verma BK, Mannan AA, Zhang F, Oyarzún DA. Trade-Offs in Biosensor Optimization for Dynamic Pathway Engineering. ACS Synthetic Biology. 2022;11(1):228-40. doi: 10.1021/acssynbio.1c00391.

10. Zhang J, Pang Q, Wang Q, Qi Q, Wang Q. Modular tuning engineering and versatile applications of genetically encoded biosensors. Critical Reviews in Biotechnology. 2021:1-18. doi: 10.1080/07388551.2021.1982858.

11. Ozdemir T, Fedorec AJH, Danino T, Barnes CP. Synthetic Biology and Engineered Live Biotherapeutics: Toward Increasing System Complexity. Cell Systems. 2018;7(1):5-16. doi: https://doi.org/10.1016/j.cels.2018.06.008.

12. Lim HG, Jang S, Jang S, Seo SW, Jung GY. Design and optimization of genetically encoded biosensors for high-throughput screening of chemicals. Current Opinion in Biotechnology. 2018;54:18-25. doi: https://doi.org/10.1016/j.copbio.2018.01.011.

13. Borujeni AE, Mishler DM, Wang J, Huso W, Salis HM. Automated physics-based design of synthetic riboswitches from diverse RNA aptamers. Nucleic Acids Research. 2016;44(1):1-13. doi: 10.1093/nar/gkv1289.

14. Angenent-Mari NM, Garruss AS, Soenksen LR, Church G, Collins JJ. A deep learning approach to programmable RNA switches. Nature Communications. 2020;11(1):5057. doi: 10.1038/s41467-020-18677-1.

15. Brophy JAN, Voigt CA. Principles of genetic circuit design. Nature Methods. 2014;11(5):508-20. doi: 10.1038/nmeth.2926.

16. De Paepe B, Maertens J, Vanholme B, De Mey M. Modularization and Response Curve Engineering of a Naringenin-Responsive Transcriptional Biosensor. ACS Synthetic Biology. 2018;7(5):1303-14. doi: 10.1021/acssynbio.7b00419.

17. Meyer AJ, Segall-Shapiro TH, Glassey E, Zhang J, Voigt CA. Escherichia coli “Marionette” strains with 12 highly optimized small-molecule sensors. Nature Chemical Biology. 2019;15(2):196-204. doi: 10.1038/s41589-018-0168-3.

18. Satya Lakshmi O, Rao NM. Evolving Lac repressor for enhanced inducibility. Protein Engineering, Design and Selection. 2008;22(2):53-8. doi: 10.1093/protein/gzn069.

19. Saeki K, Tominaga M, Kawai-Noma S, Saito K, Umeno D. Rapid Diversification of BetI-Based Transcriptional Switches for the Control of Biosynthetic Pathways and Genetic Circuits. ACS Synthetic Biology. 2016;5(11):1201-10. doi: 10.1021/acssynbio.5b00230.

20. Chong H, Ching CB. Development of Colorimetric-Based Whole-Cell Biosensor for Organophosphorus Compounds by Engineering Transcription Regulator DmpR. ACS Synthetic Biology. 2016;5(11):1290-8. doi: 10.1021/acssynbio.6b00061.

21. Snoek T, Chaberski EK, Ambri F, Kol S, Bjørn SP, Pang B, et al. Evolution-guided engineering of small-molecule biosensors. Nucleic Acids Research. 2020;48(1):e3-e. doi: 10.1093/nar/gkz954.

22. Miller CA, Ho JML, Bennett MR. Strategies for Improving Small-Molecule Biosensors in Bacteria. Biosensors. 2022;12(2):64. PubMed PMID: doi:10.3390/bios12020064.

23. Spisak S, Ostermeier M. Engineered protein switches for exogenous control of gene expression. Biochemical Society Transactions. 2020;48(5):2205-12. doi: 10.1042/bst20200441.

24. Lee Sung K, Chou Howard H, Pfleger Brian F, Newman Jack D, Yoshikuni Y, Keasling Jay D. Directed Evolution of AraC for Improved Compatibility of Arabinose- and Lactose-Inducible Promoters. Appl Environ Microb. 2007;73(18):5711-5. doi: 10.1128/AEM.00791-07.

25. Tashiro Y, Kimura Y, Furubayashi M, Tanaka A, Terakubo K, Saito K, et al. Directed evolution of the autoinducer selectivity of Vibrio fischeri LuxR. The Journal of General and Applied Microbiology. 2016;62(5):240-7. doi: 10.2323/jgam.2016.04.005.

26. Ike K, Arasawa Y, Koizumi S, Mihashi S, Kawai-Noma S, Saito K, et al. Evolutionary Design of Choline-Inducible and -Repressible T7-Based Induction Systems. ACS Synthetic Biology. 2015;4(12):1352-60. doi: 10.1021/acssynbio.5b00107.

27. Beal J, Teague B, Sexton JT, Castillo-Hair S, DeLateur NA, Samineni M, et al. Meeting Measurement Precision Requirements for Effective Engineering of Genetic Regulatory Networks. ACS Synthetic Biology. 2022;11(3):1196-207. doi: 10.1021/acssynbio.1c00488.

28. Tack DS, Tonner PD, Pressman A, Olson ND, Levy SF, Romantseva EF, et al. The genotype-phenotype landscape of an allosteric protein. Molecular Systems Biology. 2021;17(3):e10179. doi: https://doi.org/10.15252/msb.202010179.

29. Tonner Peter D, Pressman A, Ross D. Interpretable modeling of genotype–phenotype landscapes with state-of-the-art predictive power. Proceedings of the National Academy of Sciences. 2022;119(26):e2114021119. doi: 10.1073/pnas.2114021119.

30. Sadler JR, Novick A. PROPERTIES OF REPRESSOR AND KINETICS OF ITS ACTION. Journal of Molecular Biology. 1965;12(2):305-27. doi: 10.1016/s0022-2836(65)80255-8. PubMed PMID: WOS:A19656603600001.

31. Chamness GC, Willson CD. AN UNUSUAL LAC REPRESSOR MUTANT. Journal of Molecular Biology. 1970;53(3):561-5. doi: 10.1016/0022-2836(70)90084-7. PubMed PMID: WOS:A1970H871100019.

32. Jobe A, Bourgeois S. LAC REPRESSOR-OPERATOR INTERACTION VII. REPRESSOR WITH UNIQUE BINDING PROPERTIES - X86 REPRESSOR. Journal of Molecular Biology. 1972;72(1):139-52. doi: 10.1016/0022-2836(72)90075-7. PubMed PMID: WOS:A1972O225800013.

33. Betz JL, Sadler JR. TIGHT-BINDING REPRESSORS OF LACTOSE OPERON. Journal of Molecular Biology. 1976;105(2):293-319. doi: 10.1016/0022-2836(76)90113-3. PubMed PMID: WOS:A1976CA11900008.

34. Schmitz A, Coulondre C, Miller JH. GENETIC STUDIES OF LAC REPRESSOR V. REPRESSORS WHICH BIND OPERATOR MORE TIGHTLY GENERATED BY SUPPRESSION AND REVERSION OF NONSENSE MUTATIONS. Journal of Molecular Biology. 1978;123(3):431-54. doi: 10.1016/0022-2836(78)90089-x. PubMed PMID: WOS:A1978FM59000008.

35. Miller JH, Schmeissner U. GENETIC-STUDIES OF THE LAC REPRESSOR X. ANALYSIS OF MISSENSE MUTATIONS IN THE LACI GENE. Journal of Molecular Biology. 1979;131(2):223-48. doi: 10.1016/0022-2836(79)90074-3. PubMed PMID: WOS:A1979HE03000005.

36. Miller JH, Coulondre C, Hofer M, Schmeissner U, Sommer H, Schmitz A, et al. GENETIC-STUDIES OF THE LAC REPRESSOR IX. GENERATION OF ALTERED PROTEINS BY THE SUPPRESSION OF NONSENSE MUTATIONS. Journal of Molecular Biology. 1979;131(2):191-222. doi: 10.1016/0022-2836(79)90073-1. PubMed PMID: WOS:A1979HE03000004.

37. Poelwijk Frank J, de Vos Marjon GJ, Tans Sander J. Tradeoffs and Optimality in the Evolution of Gene Regulation. Cell. 2011;146(3):462-70. doi: https://doi.org/10.1016/j.cell.2011.06.035.

38. Meyer S, Ramot R, Kishore Inampudi K, Luo B, Lin C, Amere S, et al. Engineering alternate cooperative-communications in the lactose repressor protein scaffold. Protein Engineering, Design and Selection. 2013;26(6):433-43. doi: 10.1093/protein/gzt013.

39. Richards DH, Meyer S, Wilson CJ. Fourteen Ways to Reroute Cooperative Communication in the Lactose Repressor: Engineering Regulatory Proteins with Alternate Repressive Functions. ACS Synthetic Biology. 2017;6(1):6-12. doi: 10.1021/acssynbio.6b00048.

40. Chure G, Razo-Mejia M, Belliveau NM, Einav T, Kaczmarek ZA, Barnes SL, et al. Predictive shifts in free energy couple mutations to their phenotypic consequences. Proceedings of the National Academy of Sciences. 2019;116(37):18275-84. doi: doi:10.1073/pnas.1907869116.

41. Marzen S, Garcia HG, Phillips R. Statistical Mechanics of Monod–Wyman–Changeux (MWC) Models. Journal of Molecular Biology. 2013;425(9):1433-60. doi: https://doi.org/10.1016/j.jmb.2013.03.013.

42. Razo-Mejia M, Barnes SL, Belliveau NM, Chure G, Einav T, Lewis M, et al. Tuning Transcriptional Regulation through Signaling: A Predictive Theory of Allosteric Induction. Cell Systems. 2018;6(4):456-69.e10. doi: 10.1016/j.cels.2018.02.004.

43. Weinert FM, Brewster RC, Rydenfelt M, Phillips R, Kegel WK. Scaling of Gene Expression with Transcription-Factor Fugacity. Physical Review Letters. 2014;113(25):258101. doi: 10.1103/PhysRevLett.113.258101.

44. Domingo J, Baeza-Centurion P, Lehner B. The Causes and Consequences of Genetic Interactions (Epistasis). Annual Review of Genomics and Human Genetics. 2019;20(1):433-60. doi: 10.1146/annurev-genom-083118-014857.

45. Yu TC, Liu WL, Brinck MS, Davis JE, Shek J, Bower G, et al. Multiplexed characterization of rationally designed promoter architectures deconstructs combinatorial logic for IPTG-inducible systems. Nature Communications. 2021;12(1):325. doi: 10.1038/s41467-020-20094-3.

46. Zhou Y, Yuan Y, Wu Y, Li L, Jameel A, Xing X-H, et al. Encoding Genetic Circuits with DNA Barcodes Paves the Way for Machine Learning-Assisted Metabolite Biosensor Response Curve Profiling in Yeast. ACS Synthetic Biology. 2022;11(2):977-89. doi: 10.1021/acssynbio.1c00595.

47. Salis HM. The Ribosome Binding Site Calculator. Elsevier; 2011. p. 19-42.

48. Salis HM, Mirsky EA, Voigt CA. Automated design of synthetic ribosome binding sites to control protein expression. Nature Biotechnology. 2009;27(10):946-50. doi: 10.1038/nbt.1568.

49. Na D, Lee S, Lee D. Mathematical modeling of translation initiation for the estimation of its efficiency to computationally design mRNA sequences with desired expression levels in prokaryotes. BMC Systems Biology. 2010;4(1):71. doi: 10.1186/1752-0509-4-71.

50. Seo SW, Yang J-S, Kim I, Yang J, Min BE, Kim S, et al. Predictive design of mRNA translation initiation region to control prokaryotic translation efficiency. Metabolic Engineering. 2013;15:67-74. doi: 10.1016/j.ymben.2012.10.006.

51. Espah Borujeni A, Channarasappa AS, Salis HM. Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. Nucleic Acids Research. 2014;42(4):2646-59. doi: 10.1093/nar/gkt1139.

52. Bonde MT, Pedersen M, Klausen MS, Jensen SI, Wulff T, Harrison S, et al. Predictable tuning of protein expression in bacteria. Nature Methods. 2016;13(3):233-6. doi: 10.1038/nmeth.3727.

53. Reis AC, Salis HM. An Automated Model Test System for Systematic Development and Improvement of Gene Expression Models. ACS Synthetic Biology. 2020;9(11):3145-56. doi: 10.1021/acssynbio.0c00394.

54. Chen Y-J, Liu P, Nielsen AAK, Brophy JAN, Clancy K, Peterson T, et al. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. Nature Methods. 2013;10(7):659-64. doi: 10.1038/nmeth.2515.

55. LaFleur TL, Hossain A, Salis HM. Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria. Nature Communications. 2022;13(1):5159. doi: 10.1038/s41467-022-32829-5.

56. de Boer CG, Vaishnav ED, Sadeh R, Abeyta EL, Friedman N, Regev A. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. Nature Biotechnology. 2020;38(1):56-65. doi: 10.1038/s41587-019-0315-8.

57. Grossman Sharon R, Zhang X, Wang L, Engreitz J, Melnikov A, Rogov P, et al. Systematic dissection of genomic features determining transcription factor binding and enhancer function. Proceedings of the National Academy of Sciences. 2017;114(7):E1291-E300. doi: 10.1073/pnas.1621150114.

58. Mogno I, Kwasnieski JC, Cohen BA. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. Genome Research. 2013;23(11):1908-15. doi: 10.1101/gr.157891.113.

59. van Dijk D, Sharon E, Lotan-Pompan M, Weinberger A, Segal E, Carey LB. Large-scale mapping of gene regulatory logic reveals context-dependent repression by transcriptional activators. Genome Research. 2017;27(1):87-94. doi: 10.1101/gr.212316.116.

60. Li X, Lehner B. Biophysical ambiguities prevent accurate genetic prediction. Nature Communications. 2020;11(1):4923. doi: 10.1038/s41467-020-18694-0.

61. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP. Universally Sloppy Parameter Sensitivities in Systems Biology Models. PLOS Computational Biology. 2007;3(10):e189. doi: 10.1371/journal.pcbi.0030189.

62. Faure AJ, Domingo J, Schmiedel JM, Hidalgo-Carcedo C, Diss G, Lehner B. Mapping the energetic and allosteric landscapes of protein binding domains. Nature. 2022;604(7904):175-83. doi: 10.1038/s41586-022-04586-4.

63. Chure G, Kaczmarek ZA, Phillips R. Physiological Adaptability and Parametric Versatility in a Simple Genetic Circuit. bioRxiv. 2019:2019.12.19.878462. doi: 10.1101/2019.12.19.878462.

64. Sochor MA. In vitro transcription accurately predicts lac repressor phenotype in vivo in Escherichia coli. PeerJ. 2014;2:e498. doi: https://doi.org/10.7717/peerj.498.

65. Ilia K, Del Vecchio D. Squaring a Circle: To What Extent Are Traditional Circuit Analogies Impeding Synthetic Biology? GEN Biotechnology. 2022;1(2):150-5. doi: 10.1089/genbio.2021.0014.

66. Ogawa Y, Katsuyama Y, Ohnishi Y. Engineering of the Ligand Specificity of Transcriptional Regulator XylS by Deep Mutational Scanning. ACS Synthetic Biology. 2022;11(1):473-85. doi: 10.1021/acssynbio.1c00564.

67. Libis V, Delépine B, Faulon J-L. Sensing new chemicals with bacterial transcription factors. Current Opinion in Microbiology. 2016;33:105-12. doi: https://doi.org/10.1016/j.mib.2016.07.006.

68. Glasgow Anum A, Huang Y-M, Mandell Daniel J, Thompson M, Ritterson R, Loshbaugh Amanda L, et al. Computational design of a modular protein sense-response system. Science. 2019;366(6468):1024-8. doi: 10.1126/science.aax8780.

69. Sarkar S, Tack D, Ross D. Sparse estimation of mutual information landscapes quantifies information transmission through cellular biochemical reaction networks. Communications Biology. 2020;3(1):203. doi: 10.1038/s42003-020-0901-9.

70. Ross D. Automated analysis of bacterial flow cytometry data with FlowGateNIST. PLOS ONE. 2021;16(8):e0250753. doi: 10.1371/journal.pone.0250753.

71. Castillo-Hair SM, Sexton JT, Landry BP, Olson EJ, Igoshin OA, Tabor JJ. FlowCal: A User-Friendly, Open Source Software Tool for Automatically Converting Flow Cytometry Data from Arbitrary to Calibrated Units. ACS Synthetic Biology. 2016;5(7):774-80. doi: 10.1021/acssynbio.5b00284.

72. Gaigalas A, Wang L, DeRose PC. Assignment of the Number of Equivalent Reference Fluorophores to Dyed Microspheres. Journal of Research of the National Institute of Standards and Technology. 2016;121:264-81.

73. Schwartz A, Gaigalas AK, Wang L, Marti GE, Vogt RF, Fernandez-Repollet E. Formalization of the MESF unit of fluorescence intensity. Cytometry. 2004;57B(1):1-6. doi: 10.1002/cyto.b.10066.

74. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A Probabilistic Programming Language. Journal of Statistical Software. 2017;76(1):1 - 32. doi: 10.18637/jss.v076.i01.

75. Blei DM, Kucukelbir A, McAuliffe JD. Variational Inference: A Review for Statisticians. Journal of the American Statistical Association. 2017;112(518):859-77. doi: 10.1080/01621459.2017.1285773.

76. Lewis M, Chang G, Horton NC, Kercher MA, Pace HC, Schumacher MA, et al. Crystal Structure of the Lactose Operon Repressor and Its Complexes with DNA and Inducer. Science. 1996;271(5253):1247-54. doi: doi:10.1126/science.271.5253.1247.