Learning transcriptome dynamics for discovery of optimal genetic reporters of novel compounds

Aqib Hasnain*1, Shara Balakrishnan2, Dennis M. Joshy1, Steven B. Haase3, Jen Smith4, and Enoch Yeung1

1Department of Mechanical Engineering, University of California Santa Barbara
2Department of Electrical and Computer Engineering, University of California Santa Barbara
3California Nanosystems Institute, University of California Santa Barbara
4Department of Biology, Duke University

Abstract
Accelerating the design of synthetic biological circuits requires expanding the currently available genetic toolkit. Although whole-cell biosensors have been successfully engineered and deployed, particularly in applications such as environmental and medical diagnostics, novel sensing applications necessitate the discovery and optimization of novel biosensors. Here, we address this issue of the limited repertoire of biosensors by developing a data-driven, transcriptome-wide approach to discover perturbation-inducible genes from time-series RNA sequencing data, guiding the design of synthetic transcriptional reporters. By combining techniques from dynamical systems and control theory, we show that high-dimensional transcriptome dynamics can be efficiently represented and used to rank genes based on their ability to report the perturbation-specific cell state. We extract, construct, and validate 15 functional biosensors for the organophosphate malathion in the underutilized host organism Pseudomonas fluorescens SB25, provide a computational approach to aggregate individual biosensor responses to facilitate enhanced reporting, and exemplify their ability to be useful outside the lab by detecting malathion in the environment. The library of living malathion sensors can be optimized for use in environmental diagnostics while the developed machine learning tool can be applied to discover perturbation-inducible gene expression systems in the compendium of host organisms.

Introduction
The aim of synthetic biology is to design and construct living systems to possess desired functionality; this is done by developing, characterizing, and assembling biological parts in cells, creating living devices [1]. Synthetic biological circuits were first engineered in the year 2000 when Gardner et al. [2] constructed a two-node genetic bistable switch and Elowitz and Leibler [3] constructed a three-node genetic oscillator (known as the repressilator), paving the way for fine-tuned control of gene expression. Since, notable breakthroughs have emerged in post-transcriptional and translational control [4,6], optogenetic control [7], eventually leading to control of metabolic pathways [8,9] and neural-like computing [10]. Although the aforementioned genetic circuits exhibit distinct behavior, their design is implemented with a shared set of biomolecular parts, limiting the range of functionality that can be achieved.

As was the case for the genetic switch and repressilator, much of the engineering workflow for optimizing the design of genetic circuits has relied on iteratively replacing parts to minimize discrepancies between actual and desired behavior [11,13]. By parts, we are referring to DNA sequences which comprise the elementary building blocks of genetic circuits; for example, protein-coding genes, promoters, terminators, and ribosome binding sites to name only a few [12]. The initial pool of parts that were curated for use by synthetic biologists in bottom-up design were largely derived from E. coli and since has expanded into a library containing parts from a diverse set of microorganisms, from bacteriophage [14] to yeast [15].

The expansion of the genetic toolkit for circuit design remains an ongoing challenge as substantial effort is required to mine, design, characterize, and optimize biological parts [16–20]. While a significant amount of attention has been placed on optimizing and characterizing existing biological parts for genetic circuit design, less attention has been placed on mining biological parts. This has resulted in much needed insulation and biological orthogonalization strategies [21] for mitigating inadvertent intra-circuit and inter-circuit-host interactions. Moreover, programmatic tools have been developed to automate the design of genetic circuits that implement logical operations using a set of well-characterized parts in model organisms [22,25]. However, since biological parts and circuits are characterized and optimized within a single model organism and often not evaluated in application relevant organisms, there is no guarantee that the parts can be “taken off the shelf” for use in engineering novel host organisms. An increased focus on mining biological parts from novel host organisms will provide an expansion of the existing genetic toolkit from which synthetic biologists can browse and select from.

Transcriptional genetic sensors are a class of biological components that control the activity of promoters [26] and have been used to construct whole-cell (living) biosensors [27,29]. A large portion of transcriptional sensors rely on transcription factor-promoter pairs [30] and have been used in whole-cell biosensing for detection of heavy metals [31], pesticides and herbicides [32,34], waterborne pathogens [35], disease biomarkers [36,37] and many more applications discussed in [38]. Since microbes are found in virtually all terrestrial environments, one could imagine

*Please address correspondence to aqib@ucsb.edu
that there would be no shortage of transcriptional genetic sensors for novel sensing applications. However, given a novel sensing application for a target compound or perturbation, transcriptional genetic sensors are typically unknown a priori. Moreover, a complete methodology for discovering sensors for the target analyte in novel organisms does not yet exist.

The transcriptional activity of an organism can be measured through RNA sequencing (RNA-seq) to produce a snapshot of the bulk cell state subject to intrinsic and extrinsic perturbations. The typical approach for identifying upregulated and downregulated genes across experimental conditions is to apply differential expression analysis [39][40]. A major pitfall with differential expression analysis is its lack of statistical power when faced with a sparse number of biological replicates. That is to say that the false-positive rate increases drastically when only a small number of biological replicates are available [41] as is often the case due to the costliness of RNA-seq. A related issue arises in that one must sacrifice time points for biological replicates, reducing the fidelity of the dynamical process being studied. As most biological processes are dynamic, time-series profiles are essential for accurate modeling of these processes. Furthermore, differential expression analysis provides no information beyond which genes are upregulated/downregulated [42]. An analysis of expression dynamics provides a potential route to design a sensing scheme for a target analyte for which no single sensor exists.

A typical RNA-seq dataset contains hundreds to tens of thousands of genes; despite that, a subset of genes, which we call encoder genes, are typically sufficient for representing the underlying biological variation in the dataset. This is explained by the fact that variations in many genes are not due to the biological process of interest [43] and that many genes have correlated expression levels [44]. The task of identifying a subset of the state (genes) which recapitulate the entire state (transcriptome/cell state) and explain the variations of interest is well studied in the field of dynamics and controls in the form of optimal filtering and sensor placement [45][46]. In the context of dynamic transcriptional networks, sensor placement is concerned with inferring the underlying cell state based on minimal measurements; this introduces the concept of observability of a dynamical system [47].

The transcriptome is observable if it can be reconstructed from the subset of genes that have been measured. In other words, these genes encode the required information to predict the dynamics of the entire transcriptome. Hence the name, encoder genes. To the best of our knowledge, measures of observability have not been applied to genetic networks to identify genetic sensors, biomarkers, or other key genes.

Overall, a systematic approach for identifying genetic sensors from RNA-seq datasets is still an open and challenging issue. In this work, we develop a machine learning methodology to extract numerous endogenous biological sensors for analytes of interest from time-series gene expression data (Figure 1). Our approach consists of three key steps, each of which is depicted in the middle panel of Figure 1. Briefly, the first step adapts dynamic mode decomposition (DMD) [48][50] to learn the transcriptome dynamics from time-series RNA-seq data. Beyond the scope of sensor discovery, we show how the dynamic modes can be utilized to cluster genes by their temporal response. The second step involves assigning sampling weights to each gene that quantify the contribution to maximizing observability of the cell state [47][51][52]. The sampling weights provide a machine learned ranking of the genes based on their contribution to observability of the system, and using this ranking, encoder genes may be selected. To ensure the ranking is identifying genes which can recapitulate the cell state, the final step is to measure how well a chosen subset of genes can reconstruct the cell state. To validate our proposed methodology, we use our method to generate a library of 15 synthetic genetic reporters for the pesticide malathion [53][55], an organophosphate commonly used for insect control, in the bacterium Pseudomonas fluorescens SBW25. The library is composed of encoder genes identified by our proposed machine learning methodology. The transcriptional sensors play distinct biological roles in their host and exhibit unique malathion response curves. Our method uses no prior knowledge of genes involved in malathion sensing or metabolism. Moreover, we use no data source beyond RNA-seq, thereby providing a cost and computationally efficient approach for transcriptional sensor identification.

Results

Induction of malathion elicits fast host response. To start, we will first introduce the time-series RNA-seq dataset that we will use throughout this work. The transcriptional activation and repression of the soil microbe Pseudomonas fluorescens SBW25 was induced by malathion at a molar concentration of 1.29 µM (425 ng/µL) for the following two reasons: i) it is a moderate amount that can typically be found in streams and ground water after recent pesticide use based on studies done in the United States, Malaysia, China, Japan, and India [56][57], and ii) the characteristic concentration of a metabolite in bacteria is on the order of 0.1 – 10 µM [58]. Malathion is an organophosphorous synthetic insecticide used mainly in agricultural settings [59] while SBW25 is a strain of bacteria that colonizes soil, water, and plant surface environments [60]. This makes the soil-dwelling strain a prime candidate for identification of transcriptional genetic sensors for the detection of malathion.

To enable rapid harvesting and instantaneous freezing of cell cultures, we made use of a custom-built vacuum manifold, enabling fast arrest of transcriptional dynamics (Supplementary Figure 6 and Methods). Following malathion induction, cells were harvested at 10 minute intervals for 80 minutes, obtaining a total of 9 time points across two biological replicates that were sequenced. As the focus of our study is on identifying trends and correlations across time, we heavily favored time points in the trade-off between time points and biological replicates. To identify candidate sensor genes for malathion induction and subsequently build synthetic transcriptional reporters, we also collected samples from a cell culture that was not induced with malathion. See the Methods section for further details on cell culturing and harvesting.

RNA sequencing (RNA-seq) provides a snapshot of the entire transcriptome i.e. the presence and quantity of RNA in a sample at a given moment in time. In this work, we examine the fold change response given by first normalizing the raw counts to obtain transcripts per million (TPM) [61] followed by calculating the fold change of the malathion condition with respect to the negative control. The implication is that the fold change is the cell state, z(k) for some time point k, we are concerned with for discovery of genetic sensors. Of the nearly 6000 known genes in the SBW25 genome, a large fraction of them were not expressed at a level above the accuracy threshold [62][63]. To proceed, we used the fold change response to filter out non-dynamical genes. This yielded a total of 550 candidate genes, which we then subjected to a dynamic mode decomposition (DMD) analysis (Supplementary Figure 6). The DMD approach is a dimensionality reduction technique that identifies the underlying dynamic modes of a system from time-series data [64].

Given our goal of extracting salient biosensors from time-series gene expression data, we first model the dynamical process that is driven by the input of malathion on the SBW25 transcriptome. We consider malathion as a step input to the cell culture and as an impulse to the cells. This is motivated by the fact that biomolecular systems often respond to the derivative of the input and not the input itself (e.g. the absolute concentration...
Figure 1: **Transcriptional genetic sensors underlying the response from environmental perturbations can be extracted using data-driven sensor placement.** Bulk RNA sequencing (RNA-seq) measures transcript abundance over time following transcriptome perturbations. Our method starts by applying dynamic mode decomposition (DMD) to the fold change response to discover dynamic modes which govern the evolution of the cell state. The dynamic modes are used to design a state observer (gene sampling weights) that maximize the observability of the transcriptome dynamics. Measurements from a subset of genes (encoder genes) informed by the gene sampling weights are then used to reconstruct the cell state. Our method returns: 1) a dynamics matrix (or equivalently, a set of dynamic modes) describing how expression of gene \( i \) at time \( t \) is impacted by gene \( j \) and time \( t - 1 \). and 2) gene sampling weights. The outcome, demonstrated in this work, is a library of synthetic sensor promoters (genetic reporters) that are used to detect an analyte of interest. Since each genetic reporter has a unique response to the same perturbation, the library can be artificially fused to produce a purely virtual sensor for enhanced reporting.

Dynamic mode decomposition unearths modes of host cell response.

Dynamic mode decomposition (DMD) is a time-series dimensionality reduction algorithm that was developed in the fluid dynamics community to extract coherent structures and reconstruct dynamical systems from high-dimensional data [48]. Recently, several works have adapted and applied DMD to biological systems in various contexts [64-66], choosing DMD for its ability to i) reproduce dynamic data over traditionally static methods such as principal component analysis [67] and ii) represent the dynamics of high-dimensional processes (e.g. gene interaction networks) using only a relatively small number of modes.

To uncover the diverse modes of the host cell response to malathion induction, we performed (exact) DMD [50] on the transcriptomic dataset (see Methods for the details). Specifically, we perform DMD on the standardized fold change, defined as \( \bar{x}_g = \frac{x_g - \mu_{\text{control}}}{\sigma_g} \), where \( x_g \) is the expression (in TPM) of gene \( g \) and the overbar represents a variable which is transformed to have zero mean and unit variance. DMD allows the learning of low-dimensional linear models from high-dimensional time-series data. Briefly, this implies that quantitative features of a nonlinear model are not captured in our model, e.g. multiple equilibria, and chaos. If these nonlinear features are relevant to the system being studied, one can extend DMD to capture arbitrary nonlinearities, at the cost of needing a larger number of samples [71]. In this section we will describe how modeling the fold change response with DMD enables the identification of biologically relevant temporal patterns that are driven by the malathion perturbation. In the following sections we will show that the modes of the fold-change response will allow us to identify genes which act as reporters for the malathion specific response.

DMD captures transcriptome dynamics by decomposing a gene expression matrix (genes \( \times \) time points) into dynamic modes — each mode characterizes damped, forced, and unforced sinusoidal behavior. Namely, each dynamic mode is associated with a rate of exponential growth or decay and oscillation frequency.
Figure 2: Dynamic mode decomposition provides an interpretable and predictive model of gene expression dynamics. (a) DMD reconstruction of the fold change dynamics subject to an impulse input of malathion. Since the modes are complex-valued, their magnitude and phase are separately visualized. The vector \( \mathbf{z}_k \) is the reconstruction of the fold change at time \( k \), given by the depicted spectral decomposition where \( \mathbf{V} \) are the DMD modes, \( \Lambda \) are the DMD eigenvalues, and \( \mathbf{b} \) are the mode amplitudes (see Methods for detailed description of DMD). (b) The DMD spectrum reveals the growth, decay, and oscillation of each of the 10 dynamic modes that comprise the transcriptomic dataset. Each marker is an eigenvalue, and its diameter is proportional to the magnitude of the corresponding dynamic mode. Eigenvalues inside the unit circle correspond to decaying dynamics, eigenvalues with nonzero imaginary part correspond to oscillatory dynamics, and eigenvalues outside the unit circle correspond to growing dynamics. (c) The eigenvalue scaled amplitudes, \( \lambda_i^k \mathbf{b}_i \), of modes 1, 2, and 6 are visualized (upper) along with the 10 genes whose dynamics are most impacted by each of the modes (lower). The marker used for each mode indicates which eigenvalue it corresponds with in (b). (d) The eight-step prediction is visualized for five randomly selected genes in the transcriptomic dataset. The error bars represent the sample standard deviation across two biological replicates (blue solid curve) and across predictions (orange dashed curve). Magenta squares overlapping each gene’s initial condition are indicating the data that is provided to make predictions. The coefficient of determination, \( R^2 \), for the eight-step prediction across all genes is computed to be 0.92.

phase of each DMD mode, eigenvalue, and amplitude is visualized separately. The magnitude of each DMD mode represents gene-wise coherent activation while the phase represents the relative shift of this activation for the damped (or forced) modes. Here 10 modes are chosen as it is a minimal set of modes that can accurately capture the dynamics while also limiting the presence of instabilities in the model (Supplementary Figure 1). With fewer modes the instabilities disappear, however the model accuracy decreases. With more modes, the accuracy asymptotically approaches 100%, however the number of instabilities increases.

Our DMD analysis of RNA-seq data uncovers three distinct modal responses, namely stable, oscillatory, and unstable, and the response of each modes is characterized by the corresponding DMD eigenvalue, \( \lambda = a + bi \) (here \( i = \sqrt{-1} \)). The real part, \( a \), and the imaginary part, \( b \), are what determine the growth (unstable)/decay (stable) rate and the frequency of oscillation, respectively. We have plotted the 10 DMD eigenvalues relative to the unit circle in Figure 2, and labeled the eigenvalues according to their type. Note that in our model a single eigenvalue is either both stable and oscillatory, unstable and oscillatory, or only stable. Also, since our data are real-valued, any complex eigenvalue must be associated with a complex conjugate pair, explaining the symmetry across the real axis in Figure 2.

The first type of mode that we recover is stable and are characterized by eigenvalues which are inside the unit circle. The magnitude of eigenvalues inside the unit circle are strictly less than one and such a set of stable modes indicate relative decay, that is to say that many genes have a temporal response which only transiently deviate from a neutral fold change (fold change equal to zero for non-standardized trajectories and fold change equal to one for non-standardized trajectories). Stable modes that have eigenvalues nearer to the unit circle are capturing majorly uninhibited genes, while stable modes that are nearer to the origin are capturing genes which converge to neutral fold change exponentially, i.e. they exhibit strong relative decay in their fold change.

The second type of dynamic mode we uncover is oscillatory and are characterized by eigenvalues with nonzero imaginary part. Since gene expression data is always real-valued, oscillatory modes will always come in complex conjugate pairs. Each pair of complex-valued modes then describes a fixed frequency of oscillation, and each gene’s dynamics can be reconstructed from one.
or more of these frequencies. The work of Sirovich found that
the oscillatory modes obtained from DMD represent the genes
underlying the yeast cell cycle, and the frequencies of oscillation
were shown to provide an estimate of the cell cycle period that
agrees with the literature [60].

The third and final type of mode we recover is an unstable
response characterized by eigenvalues whose magnitude is larger
than one. Driven by the impulse input of malathion, many genes
show temporal response that were either upregulated or down-
regulated. If the upregulation and downregulation is persistent
throughout the gene’s temporal profile or occurs at later times,
there must be at least a single mode with eigenvalue outside the
unit circle to be able to capture the underlying unstable response.

This is because DMD is essentially learning a linear state-space
representation of the fold change response and a linear system
can only exhibit three types of limiting behaviors, i) convergence
to the origin (stable), ii) periodic orbits, and iii) divergence to
infinity (unstable). Therefore, for the reconstruction accuracy
to be maximized, DMD eigenvalues with magnitude larger than
one may be necessary. Such eigenvalues are marked with rela-
tive growth in Figure 2b. Though the eigenvalues are outside the
unit circle, they are only marginally so, implying that unstable
trajectories make up only a small portion of the transcriptomic
response to malathion.

Despite the fact that most genes require a superposition of
all of the dynamic modes for accurate reconstruction, we show
that the modes can successfully group genes into interpretable
clusters. Figure 2d (upper) shows the evolution of three dynamic
modes representative of the transcriptomic dataset: modes 1, 2,
and 6, corresponding to stable (modes 1 and 6) and unstable
(modes 2) directions in gene space. The genes which are most
influenced by each of these modes are obtained from the columns
of the DMD modes V and are plotted in the lower part of Figure
2c.

The genes which are most influenced by mode 1 are those which
diverge, in a stable manner, from a neutral fold change while the
genes most influenced by mode 2 are those which diverge away
from neutral fold change, capturing unstable trajectories. This
is consistent with the eigenvalues of mode 1 and mode 6, which
are stable and unstable, respectively. Finally, the genes most
influenced by mode 6 are those with no clear trend present in
their dynamics. In the next section, we will characterize those
genes which contribute to cell state reconstruction and act as
reporters for the malathion specific response. Relatedly, of the
20 genes that are most impacted by mode 1, seven of these genes
contribute highly to cell state reconstruction (they are within the
top 20 genes that contribute to the observability of the system).

The model of the gene expression response to malathion that
we have learned using DMD has been shown to be interpretable,
clustering genes with distinct temporal responses. To instill confi-
dence in the model, we measure the accuracy of reconstruction
using the coefficient of determination, $R^2$, as the metric. The
$R^2$ is computed by feeding an initial condition (the gene expres-
sion at time $t = 0$) to the model and then predicting all subse-
quent time points; for the nine time points in the dataset, this
amounts to two-eight-step predictions across the biological re-
plicates. Specifically, the reconstruction is computed precisely as
depicted in Fig 2a, where $V$, $A$, and $b$ are held constant and
only the time $k$ is updated to obtain the DMD estimate of the
bulk cell state at time $k$. We emphasize that this is distinct from
measuring model accuracy by computing a one-step prediction
for each time point, which gives very little information about
the dynamic process that has been captured. We obtain an $R^2$
of 0.92, showcasing that the low-dimensional model learned via

DMD has accurately captured the dynamics of the fold change
response. To provide a foundation for understanding when linear
models can accurately represent fold change dynamics, we have
shown, in the Supplementary Information, that the fold change
response of two linear systems, under stated assumptions, can be
represented as the solution of a linear system.

The results of this section demonstrate that the set of 10 re-
covered DMD modes, eigenvalues, and amplitudes are indeed bi-
ologically relevant to the dynamics of the malathion response
in the window of time that we have sampled the transcriptome.
The DMD model predictions for five randomly selected genes in
the SBW25 transcriptome are depicted in Figure 2f. These five
genes each exhibit a distinct response, and each are well cap-
tured by our DMD model. Though only five genes are presented,
the result is representative of the whole transcriptome prediction.

A key point then is that gene expression dynamics sampled at
the resolution of minutes can be well approximated by a linear
dynamical system, i.e. by a set of exponentially shrinking and
growing modes. In what follows, we develop a sensor placement
framework, relying on the learned linear dynamical system, to
generate a ranked list of encoder genes, i.e. subsets of genes
which show variation to malathion induction and that can reca-
pitulate the cell state.

Sensor placement for cell state inference and extrac-
tion of genetic sensors. Gene interaction networks are com-
plex systems that induce systematic interdependencies between
genes. That is to say that the expression of most genes, if not
all, depends on the expression of at least one more genes in the
network. These interdependencies make it possible to measure
only a subset of genes to infer the behavior of all other genes [72].
In this section, we will show that time-series measurements of a
subset of genes, called encoder genes, are sufficient to capture
the entire cell state, making the system observable. The system
we are referring to is the transcriptome or fold change dynamics
that we now have a DMD representation for and it is observable
when the complete initial cell state, $z_0$, can be uniquely inferred
from output measurements $y_k$, for times $k = 1, 2, \ldots, T$, where
the measurements are linear combinations of the expression of
all genes (see Methods).

The approach taken in this work for evaluating whether a gene
is an encoder of complete cell state information is to quantify how
much each gene contributes to observability. To do this, we op-
timize a scalar measure of the observability gramian, a matrix
which determines the amount of information that a set of sen-
sors can encode about a system. In the context of transcriptome
dynamics, given the DMD representation of the dynamics and
a chosen gene sensor placement, the gramian quantitatively de-
scribes i) to what degree cell states are observable and ii) which
cell states cannot be observed at all. Increasing i) while decreas-
ing ii) is the aim of many sensor placement techniques; further-
more, many scalar measures of the gramian have been proposed
to determine the sensor placement which maximizes the observ-
ability of the underlying dynamical system [73, 75].

To provide a method which is capable of handling high-
dimensional networks, we optimize the signal energy \( \sum_{t=0}^{T} \| y_t \|^2 \)
of the underlying system as it does not require explicit compu-
tation of the observability gramian. Computing gramians from
unstable and/or high-dimensional systems is computationally ex-
pensive and hence we choose to use the measure which can scale
for a wide array of biological datasets collected from diverse host
organisms. To further emphasize this point, we note that we are
implicitly optimizing over $5.5 \times 10^{29}$ sensor placement combina-
tions, if we choose to select 15 genes from the full set of 624 (624
choose 15). The strategy we employ is to assign gene sampling
weights, \( w_g \), to each gene \( g \) through optimizing sensor placement, i.e. maximizing the signal energy. The significance of the magnitude of each weight is to rank each gene by their contribution to observability, i.e. higher magnitude denotes higher contribution. The Methods section provides quantitative details on the relationship between observability, the observability gramian, and signal energy for sensor placement.

By examining the learned gene sampling weights, we found that nearly all 624 modeled genes contribute, some insignificantly, to the observability of the system. Displayed in Figure 3a (left) are the magnitude of gene sampling weights, \( w \), whose elements have been scaled to be in the range 0 to 1, that maximize the observability of the cell state. We note that the relative magnitude of the weights are what is important, therefore any linear scaling will preserve the information that are contained in the weights. Weights that are negative-valued (not shown here) correspond to downregulated genes and weights that are positive-valued correspond to genes that are upregulated. The higher the magnitude of the gene sampling weight, the more important the gene is likely to be for cell state reconstruction. To test this notion, the sampling weights are artificially grouped into three categories, distinguishing genes which correspond to the top (green), middle (orange), and lower (blue) third for magnitude of sampling weights. Each category contains 208 genes, and next we show the gain in information that can be achieved when sampling from one category over another.
To examine the contribution to observability provided by genes in each of the categories, we perform Monte Carlo simulations to estimate the expected predictability of the initial cell state. From output measurements, \( y_i (i = 1, 2, ..., T) \), that are generated by randomly sampling 50 genes from a specified category (low, mid, high), the cell state, \( z_0 \), is estimated and the coefficient of determination \( R^2 \) between the actual and estimated cell state is computed as a measure of reconstruction accuracy. The simulation is repeated 1000 times for each category and the resulting distributions over the random gene sets are plotted in Figure 3a (right). In the top panel, we can see that when \( T = 2 \) (2 time points are used for reconstruction), predictability of the cell state is low in all cases, and it is highest for the genes in the high category. Specifically, the reconstruction accuracy is three and two times larger in the high category than in the low and mid categories, respectively. Similarly, when the number of time points, \( T \), is increased to eight, exhausting the time points we are modeling before extrapolation, the genes in the high category best reconstruct the cell state. We found that the low and mid category genes are also capable of significant reconstruction of the cell state, exemplifying that there is a rich amount of information encoded in the dynamics. This further highlights the importance of carefully designing experiments that are sufficiently rich in conditions and time points.

Measuring fewer genes for many time points leads to higher cell state reconstruction accuracy than if many genes are measured for fewer time points. This result is demonstrated in Figure 3b which shows how the cell state reconstruction accuracy is affected by two parameters, the number of sampled genes and the number of time points, \( T \), that the genes are measured for. The reconstruction accuracy is again the coefficient of determination, \( R^2 \), between the reconstructed initial condition, \( z_0 \), and the actual initial condition \( z_0 \). For each \( T \), the first data point is generated by sampling only the five genes with the highest sampling weights for \( T \) time points. The complete cell-state is then inferred from these measurements alone and the coefficient of determination between the estimated and actual cell state can be computed (see Methods for a detailed description of the cell state inference algorithm). To compute subsequent data points, the next five genes with maximum sampling weights are simultaneously measured along with previously measured genes, and the cell state is reconstructed again. For the response of SBW25 to malathion, we find that even if only the top five genes are measured but for \( T = 10 \) time points, the cell state reconstruction is still more accurate than if all genes with nonzero sampling weights are measured with \( T \leq 8 \) time points. This signifies that the ability to study the dynamics of a few genes with fine temporal resolution can greatly increase the knowledge of the entire system.

Failure to reconstruct the initial cell state is a result of two mechanisms. The first is that we only have access to the DMD representation of the dynamics, not the true dynamics. Therefore, any output measurements generated using the DMD model will certainly incur an error with respect to the actual dynamics. As error accumulates each time-step, it is possible for the reconstruction accuracy to decrease with increasing time points. In addition to this, if a gene is added to the set of sensors, yet its dynamics are poorly predicted by the model, then it can drag down the cell state reconstruction accuracy. This can be observed in two curves in Figure 3a, namely for \( T = 10 \) and \( T = 9 \). The second hindrance for full cell state reconstruction is when many genes contain redundant information. If two genes have nearly identical gene expression profiles, adding the second gene to the set of measurements provides no useful information for the cell state inference. This may explain the asymptotic behavior of the curves in Figure 3a. There are only relatively few distinct dynamic profiles present in the transcriptomic dataset, and once all distinct profiles have been sampled, no further improvement in reconstruction can occur. This explanation is consistent with the fact that many genes co-express and this fact has even been used to reconstruct dynamic gene regulatory networks.

The gene sampling weights, \( w \), provide a machine learned ranking for discovering genetic sensors. Recall that the fold change was taken to be the state of the system when performing DMD. In so doing, we show that the encoder gene ranking can also predict genes that respond to malathion in a condition specific manner. Specifically, genes which contribute highly to the observability of the system are genes which show prolonged dysregulation in the presence of malathion. This is visualized in Figure 3c where in the top panel the 20 genes which have the largest sampling weights are plotted. Each of the 20 genes show dysregulation from the neutral fold change (0) that is persistent over the course of the time-series. Conversely, the 20 genes with lowest sampling weights show no clear trend or signal of dysregulation.

To show that encoder genes can act as genetic reporters for malathion, we selected a set of 15 genes with which to construct transcriptional reporters from. The 15 time-series profiles generated via RNA-seq (malathion TPM - control TPM) are visualized in Figure 3c. To select this set of 15, the genes were first ranked (out of 624 genes) based on their gene sampling weights with 0 being the highest. Then a randomly chosen subset of 15 genes from the top half of the ranking were used to reconstruct the cell state. The subset of 15 which produced the highest cell state reconstruction accuracy, i.e. which maximize the observability of the cell state, were chosen as the encoder genes with which to design genetic reporters from. Specifically, the observability maximizing set of 15 genes shown in Figure 3c achieve a cell state reconstruction accuracy of 76% when outputs are generated using \( T = 8 \) time points. Of the 15 selected encoder genes, 12 appear to be activated by induction of malathion while the remaining 3 appear to be repressed.

The selected encoder genes are involved in disparate biological processes. Table 1 lists the molecular functions of each of the selected genes based on their Gene Ontology (GO) annotations. Where gene names are not available, we have used protein annotations to denote those genes. It is shown that the set of molecular functions are diverse, indicating that malathion drives the activation and repression the disparate biological processes. This is precisely the goal of our sensor placement framework, to select genes which not only show variation to the biological process of interest and recapitulate the cell state, but also to select genes which are involved in distinct dynamical processes. When synthesized into genetic reporters, as we will show next, these encoder genes exhibit distinct dynamic range, sensitivity, and time-scales in response to malathion.

**Design and characterization of fluorescent malathion sensors.**

To validate the transcriptome-wide analysis for identification of biosensors, the putative promoters of the candidate sensor genes were cloned into a reporter plasmid containing a reporter gene encoding sfGFP (superfolder green fluorescent protein) and transformed into the host SBW25 (Figure 3b). The reporter strains are cloned in an unpoole format, allowing for malathion response curves to be generated at the reporter level as opposed to a pooled study which would incur additional sequencing costs for individual strain isolation.

Malathion reporters are characterized in the laboratory in an...
Figure 4: Our machine learning approach successfully extracted 15 sensors, each with distinct malathion response curves. (a) A map of the plasmid, pBHVK, used to construct the library. The plasmid contains a kanamycin resistance gene as well as a fast-folding sfGFP gene. (b) Hierarchical clustering performed on correlations between each pair of reporter strain response at 1.87 µM malathion. (c) Average per cell sfGFP signal at 0.37 µM (left) and 1.83 µM (right) malathion normalized by signal at 0.0 µM malathion is shown for all 15 engineered strains. (d) Transfer curves (or response curves) for each strain is depicted with markers and their fit to Hill equation kinetics are given by solid lines. The Hill equation parameters are given in Table 1. The promoter sequences corresponding to each reporter and time points for each transfer curve are given in Supplementary Tables 2 and 3, respectively. The error bars represent the standard deviation from the mean across three biological replicates.
to the mass spectrum of analytical standard malathion. Comparing the two mass spectra, we found that they are nearly identical (Supplementary Figs. 2-19). See the Methods section for more details about the use of Spectracide as a source for malathion and Supplementary Figure 3 for the effect of Spectracide on the growth of the reporter strains.

To examine the transcriptional activity of sfGFP, controlled by the encoder gene promoters, cells are grown in rich medium and fluorescence output was measured every three minutes over 24 hours of growth. This resulted in 400 time points per reporter strain, a nearly 45 fold increase over the number of time points obtained via RNA-seq. Prior to starting the experiment and collecting fluorescence measurements, reporter strains were induced with Spectracide to drive the reporter response. Since sfGFP is a stable protein with a long half-life and fast maturation time [78], the result is that each strain serves as a reporter for the rate of transcription initiation. This is distinctly different from the transcript abundance that is measured via RNA-seq due to the instability of mRNA molecules.

Inducing the reporter strains with malathion results in correlated transcriptional activity. To correlate the reporter strains’ activity, first the sfGFP fluorescence is normalized by the OD to give average per cell fluorescence. The Pearson correlation between the average per cell fluorescence of all pairs of reporters is given in Figure 1. From the heatmap, three distinct positively correlated clusters are apparent. The strains cspA2, atpB, fabA, acrA, and petA form the first cluster. The second positively correlated cluster contains uncharacterized protein II, cspA2, and putative ABC transport system. Lastly, glA, putative outer membrane porin A, and lipC form the third cluster. Moreover, we see that the first cluster negatively correlates with the second and that the second cluster negatively correlates with the third. The present correlations thus suggest that the genes within a cluster may have functional dependence in the presence of malathion or they share a transcriptional regulator. This also highlights the role of redundancy in gene expression and has been studied widely in the form of gene co-expression networks or regulons [14].

Examining the transcription initiation driven by malathion at distinct concentrations reveals detailed gene expression dynamics, dependencies of expression on malathion concentration, as well as the correlations. Firstly, the fold change (with respect to 0.0 µM malathion and referred to as the background) reveals oscillatory signals in several strains; the reporters atpB, petA, cspA2, and acrA each contain oscillations that are near in phase at 0.38 µM malathion (Figure 4). As the concentration of malathion is increased, only atpB and petA appear to remain in phase while the signals of the other strains strongly increase. We also see that anti-sigma 28 factor and rpoA oscillate with lower frequency and that anti-sigma 28 factor peaks at a peak around 10 hours after induction while rpoA hits an anti-peak around 10 hours after induction. For the lower malathion concentration, sucC has a large lag time until transcriptional activation occurs, however there is a sharp decrease in the lag time at the higher concentration. The strains acrA, glA, putative outer membrane porin A, putative ABC transport system, and lipC consistently respond within minutes of malathion induction with lipC being the reporter with highest signal over background and acrA the reporter with highest overall signal energy (area under the curve) in early times. Though cspA2 was shown by the RNA-seq data to be repressed by malathion, we find that cspA2 strain is consistently activated in the presence of malathion. Of the remaining repressed promoters, uncharacterized protein II is far more repressed in the presence of malathion across all concentrations tested.

The response curves of the reporter strains to malathion strongly resemble Michaelis-Menten enzyme-substrate kinetics. Such kinetics are characterized by exactly two parameters and mathematically described by Hill functions [63] (Methods). The first parameter is the Hill coefficient or cooperativity, n, which is a measure of how steep the response curve is. This is also denoted as a measure of ultrasensitivity. The second parameter, K_M, is the Michaelis constant and it is equal to the malathion concentration at which the response is half of its minimum value subtracted from its maximum value. Figure 4A shows the malathion response curves of each reporter strain at the time point with maximum fold change with respect to the 0 µM malathion condition. The solid line depicts the fit of a Hill function to the experimentally generated response curves and the parameters of each Hill function are given in Table 1. The response shown is the average fluorescence per cell obtained by normalizing the sfGFP signal by the optical density. See Supplementary Table 4 for the precise time points used here for each strain and see Methods for further details on parameter fitting.

We find that there is significant variation across the Hill coefficient, dynamic range, and Michaelis constant in the library of reporters. The Hill coefficient ranges from 1.1 to 21.6, and recalling that this parameter is a measure of sensitivity, the extremes depicted by a small slope in strain fabA and large slope in strain sucC, respectively. The dynamic range, measured as the difference between the maximum signal and the minimum signal, ranges from 80 to 1401 and is obtained by sucC and the repressed uncharacterized protein II, respectively. The Michaelis constant ranges from 0.2 to 1.5, depicted by the shift in malathion concentration at which half of the maximum signal is achieved from fabA and cspA2.

Overall, we find that each synthetic reporter, selected via our data-driven sensor placement framework, is capable of detecting malathion with distinct dynamic ranges and sensitivity. Moreover, we note that two of the selected reporters, ABC transporter and acrA, are membrane transporters and are not expected to be specific to malathion. The above two points motivate combining features from individual reporters to generate a single (virtual) reporter that enhances sensing capabilities. In what follows we demonstrate one approach to achieve such a task.

Superimposing the response of multiple sensors creates an enhanced virtual sensor.

The genetic reporters characterized in the previous section respond to malathion with distinct timescales, amplitudes, and frequencies, each acting as a unique report of the environmental context. However, as explained previously, not every reporter is expected to uniquely respond to malathion. Therefore, when testing for malathion in an environmental scenario, the conclusion given by individual reporters are expected to have a higher false positive rate than if the measurements were aggregated to form a single, combined sensor.

Recognizing the need to construct a multi-component sensor from the reporters in our synthetic promoter library, in this subsection we explore an approach for incorporating each unique temporal response to produce a desired output that provides more information than a single reporter alone. This application of the library views the synthetic reporters as genetic basis functions with fixed expressivity, comprising a single-input-single-output genetic network. Here the single input is malathion and the single output is a virtual sensor. As opposed to a biological sensor, a virtual sensor solely processes data originally gathered by the distinct biological sensors [79]. In our case, the 15 genetic reporters described in the previous section comprise the biological sensors and we aggregate the response measurements from
The usefulness of a virtual sensor in the setting of detection of a novel small compound is two-fold. i) aggregating contrasting responses can only reduce the false-positive rate of a detection event and ii) combining individual sensors in a software-based manner reduces the need for implementation of complex synthetic genetic networks and reduces metabolic burden on the host organism. Taking advantage of the benefits of virtual sensing, we develop an approach for enhancing malathion reporting by aggregating the response of the reporters in our library. Specifically, the weighted superposition of malathion responses are used to produce a desired output signal.

We show that transpositional virtual sensing is capable of detecting environmentally relevant events. Consider a scenario where malathion is discarded in a prohibited site such as a body of water or soil. Such an event might trigger a reference (desired) response that resembles the response of a linear, second-order system to a step input \[ 47 \]. Specifically, the reference response is characterized by a rapid response to malathion followed by lower magnitude, sustained response (Figure 5b). Treating the reporter library as genetic basis functions, we learn the sparse set of coefficients that approximate the reference trajectory (see Methods for details). We find that with only three sensors, the desired response is accurately captured. The strains \( \text{acrA, anti-sigma 28 factor, and cspA2} \) each possess peaks shortly after malathion induction, capturing the peak in the reference. At later times, the superposition of the three strains are able to recapitulate the sustained response.

We now consider a second scenario where we aim to detect malathion from a more subtle source where in pulses of malathion are introduced to the system periodically. Figure 5c depicts the pulse inputs and reference trajectory which is comprised of a linear combination of radial basis functions. We find that for this more complex scenario, superposition of the response of nine reporters is required to approximate the reference trajectory.

In both scenarios, a single genetic reporter would not be sufficient to inform of the type of event that occurred. Furthermore, we have shown how virtual sensing can prove useful for aggregation of measurements from individual sensors without having to clone synthetic multi-component reporters, a difficult task due to the tremendously large size of the design space and the emergent effects seen when composing genetic parts.

**Detecting malathion in environmental samples.**

In the previous section, we discussed how we can virtually enhance the sensing ability of the malathion reporter library in environmentally relevant scenarios. However, the library has only been examined in an ideal laboratory scenario with either pure or processed malathion that has been analyzed with mass spectrometry; it is not yet known if the reporters will be able to sense malathion when induced with actual environmental water samples that have been treated with the insecticide. Confounding factors may be present in the environmental sample such as other small compounds that may make it difficult to deconvolve malathion response from the response due to the confounder. Therefore, in this section we describe an experiment to assess whether or not the malathion concentration can be deduced from our reporters treated with environmental insecticide samples.

In order to test if the genetic reporters can sense malathion from environmental samples, irrigation water was collected from three crops after being sprayed with a mixture of Spectracide (50% malathion) and water (Figure 6). The concentration of the mixture sprayed was either 0, 1, or 8 times the maximum recommended working concentration of Spectracide – 1 fluid ounce from all sensors, increasing the confidence in the conclusion of the event that the sensors have been exposed to.
per gallon of water. To rid the solution of unwanted microbes and particles, the irrigation water was strained and filtered prior to to induction of the genetic reporters (see Methods). The growth and induction protocols all remain the same as for the samples treated with Spectracide in Figure 4c,d.

We found that a total 9 out of the 15 of the reporters were activated by induction of the irrigation water containing malathion. Fig 6a shows the average per cell fluorescence 24 hours after induction of the nine strains subjected to 0, 1, or 8 times the working concentration of Spectracide. The reporters atpB, petA, sucC, rpoA, fabA, and gltA all show a response to malathion at 1x working concentration, while the remaining three did not show significant differences from the negative control in this range. Among the strains in Figure 6a, the strain sucC was activated the most, showing an 80% increase from the 0x to 8x condition after the 24 hour time period. This shows that many of the selected genetic reporters, 60%, are able to detect malathion in environmentally relevant scenarios, and, furthermore, we can use this data to infer the concentration of malathion present in the samples collected from the environment.

The response curves characterized previously in Figure 4d for each of the genetic reporters can be used to make an inference about the amount of malathion present in each environmental sample. Note that we are making the assumption that the response curves characterized for each of the nine reporters can be applied to this new setting of treatment with irrigation water. With this assumption we can then use the fitted Hill equations from Figure 4d and numerically estimate the malathion concentration that reproduces the signal at 1 or 8 times the working concentration of Spectracide. The results obtained are shown in Figure 6b for each of the nine strains. Through this approach, the reporters provide a range of inferred malathion concentrations; at the working concentration of Spectracide, we can infer that the concentration of malathion is in the range 0.48 – 0.97 µM and at 8 times the working concentration of Spectracide, we can infer the concentration of malathion to be in the range 0.82 – 2 µM. It is important to note that for most, if not all, of the characterized reporter strains, 2 µM was the maximum discernable concentration before the signal saturates. Therefore, it is possible the concentration of malathion is higher than 2 µM, however that range cannot be detected by our reporter library.

Discussion

It is often the case that biologists seek to identify key genes which show variation for the biological process of interest. Many tools have been developed or adapted to meet this need e.g. differential expression and principal component analysis to name only a few. However, when using the current tools, there is potential to measure system variables that are redundant which can lead to wasted time and resources. Therefore, we developed an efficient method that identifies the variables that allow for the inference of the complete system. The method combines dynamic mode decomposition (DMD) and observability of dynamical systems to provide a systematic approach for the discovery of perturbation-inducible genes. To extract optimal biosensors from our model, we showed that if the fold change was taken as the state of the system, the encoder genes inform the design of transcriptional reporters that showcase condition specific sensing.

We introduced DMD as a novel tool for analysis of transcriptome dynamics. In this case, we studied bulk transcriptome dynamics at the minutes resolution and showed that the low-dimensional DMD representation accurately predicts the dynamics and clusters genes based on temporal behavior. Our results suggest that DMD is a capable tool for analysis of transcriptomic data and warrants further exploration in single-cell RNA-seq and other ‘omics technologies that aim to infer cell trajectories, pseudotime, and single-cell regulatory networks.

The identification of transcriptional genetic sensors was posed as a design challenge, where a subset of genes are selected to maximize the observability of the cell state. It was shown that a large fraction of genes contribute insignificantly to the cell state observability when only few time points are measured, further validating the common knowledge that genetic networks possess redundancies and are noisy. We also showed that it is significantly more beneficial to measure a sparse set of genes for more time points than to measure more genes for fewer time points.
Our results suggest future joint experimental and computational approaches which limit the amount of resources required to get a full description of the system dynamics. A natural extension of our work is to determine how well measurements from a small library of reporters recapitulate the bulk cell state under unseen conditions. Such studies will inform how RNA-seq data should be collected in the future in order to maximize the reconstruction accuracy and minimize labor and experimental costs.

The machine learning-driven selection of genetic reporters was shown to produce 15 functional biosensors with a variety of malathion response curves. We demonstrated how to aggregate information from each reporter to create a virtual sensor that can be used to infer events of interest. Moreover, we showed that the genetic reporters can be used to detect malathion in environmental settings. More generally, our results and methodology offer an innovative approach that can be used to test perturbation-inducible gene expression systems. We emphasize that our approach takes advantage of the largely untapped resources present in native host genomes and we anticipate that techniques like the one developed here will produce a plethora of parts for synthetic biologists to build useful devices from.

Lastly, our developed approach makes no assumptions on the nature of the underlying system. In that sense, the framework we have developed is general and can be applied to data generated from other omics techniques and from any organism. In the case that a linear response model is insufficient for capturing the transcriptome dynamics, it can be extended to a variety of nonlinear models to capture nonlinear modes of response [71]. An interesting extension of observability to transcriptome dynamics would be to construct state-estimators (also known as observers) of the dynamics for real-time monitoring of gene interaction networks [80]. Such approaches could find potential use in designing and implementing better diagnostic tools for synthetic biologists.

Finally, further refinement of the list of encoder genes could be obtained by fusing ChIP-seq (chromatin immunoprecipitation followed by sequencing) with RNA-seq measurements to discover transcription factors, however such an experimental assay can be prohibitively expensive. The DNA binding sites measured by ChIP-seq alone are not sufficient to infer regulation of transcription. However, together with RNA-seq, the set of encoder genes which causally drive the condition specific response can be uncovered.

### Methods

#### Rapid culture sampling.

For each biological replicate, *Pseudomonas fluorescens* SBW25 glycerol stock was scraped and inoculated in 5 mL of fresh LB broth (Teknova Catalog no. L8022) and was incubated and shaken at 30°C and 200 r.p.m. for 15 hours. The OD$_{600}$ of the 5 mL culture was measured and the entire culture was transferred to 50 mL of fresh LB broth, which was then proceeded by incubation and shaking. Once the OD$_{600}$ of the 50 mL culture reached 0.5, the culture was again passaged into 300 mL of fresh LB broth. The 300 mL culture was grown until OD$_{600}$ of 0.5. Then the culture was split into two 150 mL cultures (one for malathion induction and one for negative control). The two cultures were sampled at evenly spaced intervals in time (see Supplementary Table 1 for sampling volumes and times) and after the 0 minute sample, malathion (Millipore Sigma Catalog no. 36143) was introduced to the positive condition at 1.83 mM. To separate the media from the cells, a vacuum manifold with 3D printed filter holders was constructed and utilized (Supplementary Figure 6). 0.45 µm PVDF membrane filters (Durapore Catalog no. HVLP04700) were placed on the filter holders, a vacuum pump was turned on, and the culture sample was dispensed onto the center of the filter, quickly separating the media from the cells. The filter with the cells was then placed into a 50 mL conical centrifuge tube (Fisher Scientific 149590A) using sterile tweezers. The tube with the filter was then submerged into a liquid nitrogen bath for 10 seconds to flash freeze the sample. The sample were then stored -80°C.

#### RNA extraction.

To extract the RNA, first the filter-harvested cells were resuspended in 2 mL RNAprotect Bacterial Reagent (Qiagen Catalog no. 76506), then pelleted in a centrifuge. To lyse the cells, the pellet was then resuspended in 200 µL of TE Buffer containing 1 mg/mL lysozyme. The RNA was then extracted from the lysed cells using RNeasy Mini Kit (Catalog no. 74104), and the samples were DNase treated and concentrated using Zymo RNA Clean and Concentrator (Catalog no. R1019).

#### RNA library preparation and sequencing.

Bacterial RNA was depleted using NEBNext Bacterial RNA Depletion Kit (Catalog no. E7850X). The indexed cDNA library was generated using NEBNext Ultra II Directional RNA Library Prep (Catalog no. E7765L) and NEBNext Multiplex oligos for Illumina (Catalog no. E6609S). In total, 40 samples (two biological replicates, 10 time points, two conditions) were prepped and sequenced. The library was sequenced at the Genetics Core in the Biological Nanostructures Laboratory at the University of California, Santa Barbara on an Illumina NextSeq with High Output, 150 Cycle, paired end settings.

#### Pre-processing of sequencing data.

The raw reads were trimmed for adapters and quality using Trimmomatic [81]. The reads were then pseudoaligned with Kallisto [82] to the *Pseudomonas fluorescens* SBW25 transcriptome generated using GFFRead [83] and GenBank genome AM181176.4. The normalized gene expression of transcripts per million (TPM), which takes into account sequencing depth and gene length, are used for modeling and analysis. Genes with an average TPM less than 100 in all experimental conditions were discarded for further analysis.

#### Malathion reporter library cloning.

For the reporter plasmid cassette design, first, the closest intergenic region to the gene target larger than 100 base pairs (bp) was identified based on the open reading frame of the sequenced genome of *Pseudomonas fluorescens* SBW25 (GenBank genome AM181176.4). Primers were designed to include the entire intergenic region in order to capture any transcription-regulator binding sites surrounding the promoter (Figure 4). The identified intergenic regions were amplified using the primers and this is what we refer to as ‘promoter regions’ following the terminology of [84]. The promoter regions were cloned into a cassette on the plasmid backbone pBHVK (Supplementary Figure 3) containing a bicistronic ribosome binding site and superfolder GFP (sfGFP) as the reporter gene. Lastly, a cloning site was placed in the cassette so that the cloned promoter controls transcriptional activity of sfGFP. The promoters were assembled onto the plasmid backbone pBHVK (see Supplementary Fig. 4 via Golden Gate Assembly [85] using NEB Golden Gate Assembly Kit (Catalog no. E1601S). Because of the potential of arcing during electrotransformation of *Pseudomonas fluorescens* SBW25 with Golden Gate reaction buffers, the plasmids are first subcloned into *E. coli* Mach1 (Thermo Fisher Scientific Catalog no. C862003) following the manufacturer’s protocol for chemical transformation. Between three and six colonies are selected for each strain and the reporter cassette was sent for sequencing at Eurofins Genomics. Then the plasmid DNA was prepared from cultures of transformed Mach1 cells using Qiagen Spin Miniprep Kit (Catalog no. 27106) followed by chemical transformation into SBW25. SBW25 was made chemically competent by washing a culture at OD$_{600}$ of 0.3 with a solution of 10% glycerol two times, then resuspending in 500 µL of 10% glycerol. The plasmid DNA is added to 80 µL of the cell suspension and kept at 4°C for 30 minutes, then the cells were electroporated with 1600 V, 200 Ω, and 25 µF. The cells were immediately resuspended in 300 µL of SOC Broth (Fischer Scientific Catalog no. MT46000CR), recovered for 2 hours at 30°C in a shaking incubator, and plated onto 1.5% LB Agar plates with 50 µg/mL Kanamycin.

Again, three to six colonies of each strain have their reporter cassette sequenced at Eurofins Genomics and simultaneously glycerol stocks of each colony is prepared for long term storage.

#### Photobleaching of Spectracide.

Spectracide malathion insect spray (Spectracide Catalog no. 071131209006) was utilized as the environmentally relevant source of malathion for the reporter library testing and contains 50% malathion. Spectracide is an opaque liquid. We found that we can remove the opaque substances by photobleaching a 5% Spectracide solution (in LB) in a Synergy H1
plate reader (Biotek), at 30°C and 800 r.p.m. OD<sub>600</sub> and fluorescence (excitation 485nm, emission 528nm) were measured every 3 minutes for 8 hours. To ensure malathion remained in solution after photobleaching, the mass spectrum was analyzed at the University of California, Santa Barbara Mass Spectroscopy Facility. From this we determined that malathion is stable for the course of the photobleaching (Supplementary Figures S6 to S9).

Plate reader assays to measure response curves and doubling times. Scrape off culture from glycerol stocks of each strain and inoculate 3 mL of LB (Kanamycin 50 μg/mL) in 10 mL 24 deep-well plate sealed with a breathable film (Spectrum Chemical Catalog no. 630-11763) and grown at 30°C overnight in a shaker incubator. The overnight cultures were diluted to an OD<sub>600</sub> of 0.1 in 2 mL of LB and the cultures were grown for an additional 2 hours. 250 μL of this culture was then transferred to a 96 well optically-transparent microtiter plate. Photobleached spectracide (50% malathion) solution at a concentration of 1 fluid ounces. The strained samples were then centrifuged to separate dense, through from each plant were first strained using a 40 µm cell strainer. The strained samples were then centrifuged to separate dense, through from each plant were first strained using a 40 µm cell strainer.

Collection and cleanup of irrigation water treated with Spectracide. Three cabbage plants were each potted in 5 gallon buckets with fresh soil (Harvest Supreme) and a water catchment tray was placed under the plants to catch flow through. The first plant was sprayed with water containing no malathion and the flow through was collected in a 1 L pyrex bottle. The second plant was sprayed with a Spectracide (50% malathion) solution at a concentration of 1 fluid ounces. The strained samples were then centrifuged to separate dense, soil particles from the Spectracide solution. Finally, the supernatant was vacuum filtered through a 0.22 µm membrane before induction of the reporters. The protocol for induction of the reporters with the irrigation water is the same as above.

Computing the dynamic mode decomposition. We now discuss the details of applying dynamic mode decomposition (DMD) to time-series data obtained from sequencing. As mentioned previously, many algorithms have been developed to compute the DMD modes, eigenvalues, and amplitudes, and a key requirement of almost all of the techniques is that the time points are spaced uniformly in time. In our work we begin by collecting the data for a single experimental condition into a time-ordered matrix, X, which contains a total of m x r data snapshots for a data set with m time points and r replicates. For response to malathion, each X<sup>(i)</sup> corresponds to the gene expression vector at time i in replicate j and is in the (i + m x j)th column of the data matrix X where i Є {0, 1, m – 1} and j Є {1, 2, r}. For gene expression data obtained from RNA-seq, each data snapshot typically contains thousands of rows denoted by n. The n x m matrix data for the response to malathion is then given by

\[
X_{\text{malathion}} = \begin{bmatrix}
    x^{(1)}_1 & x^{(1)}_2 & \cdots & x^{(1)}_{m-1} & x^{(2)}_1 & x^{(2)}_2 & \cdots & x^{(2)}_{m-1} & \cdots
\end{bmatrix}
\]

where each x Є R<sup>n</sup> represents the gene expression given in transcripts per million (TPM) from the malathion condition. Similarly, the data matrix for the control condition is constructed. The fold change data matrix, Z, is subsequently computed as Z = X<sub>malathion</sub> / X<sub>control</sub> and where H<sub>malathion</sub> is the Hadamard (elementwise division) operator. Next we compute the mean-subtracted and standard deviation-normalized data matrix \( \bar{\mathbf{Z}} \)

\[
\bar{\mathbf{Z}} = \begin{bmatrix}
    \overline{z}_0 - \mu_{0:m-1} & \overline{z}_1 - \mu_{0:m-1} & \cdots & \overline{z}_{m-1} - \mu_{0:m-1}
\end{bmatrix}
\]

(2)

where \( \mu_{0:m-1} \) is the vector of time-averages of each gene and \( \sigma^2_{0:m-1} \) is the vector of time-standard deviations of each gene. The divisions in Eq. (1) are performed element-wise. We see that \( \bar{\mathbf{Z}} \) is obtained by removing the time-averages from each gene and standardizing the time-variances of each gene. The mean subtraction operation is motivated by the fact that the mean of the data corresponds to the eigenvalue \( \lambda = 1 \), which is always an eigenvalue of the Koopman operator, the operator that DMD ultimately aims to approximate, and not one that we are particularly interested in. The normalization by the standard deviation is performed so that the magnitude of the fold change has no implication on the connectivity of the learned dynamical system.

The algorithm we make use of to compute the dynamic mode decomposition (and the approximation of the Koopman operator) is exact DMD [90], which aims to identify the best-fit linear relationship between the following time-shifted data matrices

\[
\mathbf{Z}_p = [\overline{z}_0 \quad \overline{z}_1 \quad \cdots \quad \overline{z}_{m-2}] \\
\mathbf{Z}_f = [\overline{z}_1 \quad \overline{z}_2 \quad \cdots \quad \overline{z}_{m-1}]
\]

such that

\[
\bar{\mathbf{Z}}_f = \mathbf{K}\bar{\mathbf{Z}}_p + \mathbf{r}
\]

(3)

Rearranging the above equation, it is shown that \( \mathbf{K} \) is related to \( \mathbf{r} \) through a similarity transformation as shown in Eq. (4)

\[
\mathbf{K} = \mathbf{U}^\dagger\mathbf{Z}_f\mathbf{W}^{2}\mathbf{W}^\dagger
\]

(5)

meaning that the eigenvalues of \( \mathbf{K} \), \( \lambda \), are equivalent to the k leading eigenvalues of \( \mathbf{K} \) while the eigenvectors of \( \mathbf{K} \) and \( \mathbf{r} \) are related to the k leading eigenvectors of \( \mathbf{K} \). v, by \( \mathbf{v} = \mathbf{U}^\dagger\mathbf{r} \). This eigendecomposition then allows the fold change response to be written as the following spectral decomposition

\[
\mathbf{z}_k = \sum_{j=1}^{k-1} v_j^\dagger b_j + \mathbf{V}^\dagger \mathbf{b}
\]

(6)

where \( \mathbf{V} \) is a matrix whose columns are the eigenvectors (DMD modes) \( \mathbf{y}_m \) and \( \mathbf{b} \) is a vector of amplitudes corresponding to the gene expression at the initial time point as \( \mathbf{b} = \mathbf{V}^\dagger\mathbf{z}_0 \). Here \( \dagger \) represents the Moore-Penrose pseudoinverse of a matrix.

Using the above spectral decomposition, the modes can then be evolved in time for m – 1 time steps to reconstruct the data from knowledge of the initial condition. Evolving past the mth time point allows for forecasting of the fold change response. To measure the accuracy of reconstruction we use the coefficient of determination

\[
R^2 = 1 - \frac{\sum_{k=m}^{\infty} \mathbf{z}_k^\dagger (\mathbf{z}_k - \hat{\mathbf{z}}_k)}{\sum_{k=0}^{\infty} \mathbf{z}_k^\dagger (\mathbf{z}_k - \bar{\mathbf{z}})}
\]

(7)

where \( \mathbf{z} \) is the vector of each gene’s mean expression, formally \( \mathbf{z}^{(i)} = \sum_{k=0}^{\infty} \mathbf{z}_k^\dagger (\mathbf{z}_k - \bar{\mathbf{z}}) \). This is the prediction of \( \mathbf{z}_k \) given by the model starting from the initial condition.

Computing the gene sampling weights. Here we describe our methodology for ranking genes based on their contribution to the observability of the dynamical system learned via dynamic mode decomposition.

We start by introducing the energy of a signal in discrete-time as

\[
E_g = \sum_{i=0}^{\infty} \|y_i\|^2
\]

(8)

which is closely related to the idea of energy in the physical sense and where \( \mathbf{y} = \mathbf{W}\mathbf{z} \) are measurements of the system state and \( \mathbf{W} \in \mathbb{R}^{p \times n} \).
Rewriting the signal energy using the recursion for $y$ given as $y_t = W^Tz_0$, we can reveal the connection between energy and observability

$$E_y = \sum_{i=0}^{\infty} \alpha_i^2 (\sum_{i=0}^{\infty} K_i^T W^T W K_i z_0)$$

$$= \sum_{i=0}^{\infty} \alpha_i^2 \langle \sum_{i=0}^{\infty} K_i^T W^T W K_i \rangle z_0$$

$$= \sum_{i=0}^{\infty} \alpha_i^2 X_i z_0$$

where $X_i$ is the infinite-horizon observability gramian, a symmetric matrix that is unique if the magnitude of $K$ all have magnitude less than 1. The observability gramian describes how much gain will be obtained in $y$, given an initial condition $z_0$. It simultaneously gives a measure of how well the initial condition $z_0$ can be estimated given only measurements of the system state $y$.

We use the observability gramian along with the measure of energy it provides to optimize for the gene sampling weights in the rows of $W$ that maximize the signal energy $E_y$. Formally, the objective function is given as

$$\max_{W \in \mathbb{R}^{p \times n}} \alpha_i^2 \langle \sum_{i=0}^{\infty} K_i^T W^T W K_i \rangle z_0$$

subject to $W W^T = I_p \times p$. (10)

where we seek the matrix $W$ that maximizes the observability of the cell state $z_0$. The constraint above enforces the following three points:

i) the length of each row vector in $W$ is not important, we are only concerned with the direction and the constraint sets the length of each row vector to be equal to 1, ii) the maximization problem is well-posed, i.e. the objective cannot blow up to infinity with the length constraint, and iii) the rows of $W$ form $p$ vectors of an orthonormal basis for $\mathbb{R}^p$, i.e. $W W^T = I_p \times p$. Each row vector in $W$ can then be viewed as a set of weights, each orthogonal to one another, that rank genes based on their contribution to the observability of the system. The optimization problem (10) represents a quadratic program with linear constraints, and the rows of $W$ which maximize the objective are the $p$ eigenvectors corresponding to the $p$ eigenvalues with highest magnitude of the Gram matrix

$$G = \sum_{i=0}^{\infty} K_i^T W^T W K_i$$

Since $G \in \mathbb{R}^{n \times n}$ is a sum of quadratic forms, the result is that $G$ has non-negative, real-valued eigenvalues. If the eigendecomposition is $G = QDQ^{-1}$, then the solution to the optimization problem Eq. (10) is

$$W = \begin{bmatrix} q_1^T \\ \vdots \\ q_p^T \end{bmatrix}$$

(12)

where $q_1$ through $q_p$ are the top eigenvectors of the Gram matrix $G$.

The proof of the solution to the optimization problem is provided in the Supplementary Information. The single set of gene sampling weights that maximize the observability are precisely $q_1$, and from here on out we can refer to these weights $w$.

Since transcriptomic data sets typically have few initial conditions, i.e. biological and technical replicates, before solving for $w$ we enrich our data set with $N$ synthetic initial conditions that are randomly sampled as $\mathcal{U}(\min(z_0^{(i)}), \max(z_0^{(i)}))$ where $j$ is in $\{1, 2, \ldots, r\}$ and $r$ is the number of replicates. The motivation for the artificial data sampling is given in [27], where it is shown that artificially generated data points improve the estimate of the DMD model when the data set is affected by noise. $N$ is chosen to be equal to the number of genes to ensure the matrix of initial conditions has full rank. Another issue that we have addressed are the instabilities present in the DMD eigenvalues.

Consequently, the observability gramian is not unique and the sum in Eq. (11) diverges to infinity. To mend this issue, we compute the finite-horizon Gram matrix, where the sum in Eq. (9) and Eq. (11) is from 0 to $m$. This allows for the computation of the finite-horizon signal energy from Eq. (9) where the bounds on the sum are now from $i = 0$ to $i = m$.

Once $w$ is obtained by solving Eq. (10), then measurements $y_t$, for $t$ in $\{0, 1, \ldots, T\}$, are generated from $y_t = W^T K^T z_0$ while keeping only the $q$ elements of $w$ with largest magnitude as nonzero. All other elements of $w$ are set to zero to simulate the sampling of only selected genes. To reconstruct $z_0$ using only the measurements, we form the following observability matrix from the known sampling weights, $w$ and the dynamics matrix $K$.

$$\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_T \end{bmatrix} = \begin{bmatrix} W^T \\ \vdots \\ W^T K^T \end{bmatrix} z_0 = O_T z_0$$

and using the Moore-Penrose pseudoinverse we can obtain an estimate of the initial condition as follows

Increasing $q$ while keeping $T$ constant results in increasing reconstruction accuracy until a critical value of $q$ such that the reconstruction accuracy plateaus; a similar scenario holds for keeping $q$ constant and increasing $T$. When both $T$ and $q$ surpass the critical values, perfect reconstruction may be achieved.

When the computation of the Gram matrix, $G$, is not computationally feasible, as can be the case when the dimensionality of the data are relatively high compared to that of bacterial transcription networks that we are dealing with here, the reduced order dynamics given by DMD can be used to compute an approximation to the leading eigenvalues and eigenvectors. The reduced order $G$ is then given by

$$\hat{G} = \sum_{i=0}^{\infty} \bar{K}_i^T \bar{z}_i \bar{z}_i^T U \bar{K}_i^T$$

where $\bar{K}$ and $U$ are given in Eq. (4). Supplementary Figure 2 shows the approximation of the leading eigenvalues and eigenvectors of $G$ by $\hat{G}$.

Fitting the response curves to Hill kinetics. The malathion response curves for each sensor were fit to Hill functions of the form

$$y = y_{\min} + \left( y_{\max} - y_{\min} \right) \frac{u^n}{K_M + u^n} = H_{act}(u)$$

for activated sensors and

$$y = y_{\max} - \left( y_{\max} - y_{\min} \right) \frac{u^n}{K_M + u^n} = H_{rep}(u)$$

for repressed sensors. The parameter $n$ is a measure of ultrasensitivity [85] how steep the response curve is and is known as the Hill coefficient. The Michaelis constant, $K_M$, is equivalent to the malathion concentration at which the sensor response, $y$ (measured in OD normalized arbitrary fluorescence units), is half of $(y_{\max} - y_{\min})$. The input $u$ represents the malathion concentration in millimolar.

The objective function used to determine the parameters of the Hill equations is shown below

$$\min_{c,R_M} \sum_{i=1}^{n_c} (y_i - H(u_i))^2$$

where $H$ is the Hill function of the activator or repressor and $n_c$ is the number of data points and is equivalent to the number of malathion concentrations times the number of replicates. The Levenberg-Marquardt algorithm is used to solve a nonlinear least squares problem to obtain a solution to optimization problem [86].

Approximating reference curves with genetic basis functions. Here we describe the treatment of the transcriptional sensors as genetic basis functions and how to use them to approximate reference curves. For this task, we work with the mean fold change of malathion response at 2.24 mM with respect to the zero malathion condition. The mean is taken across biological replicates for each of the $n_c$ reporters. OD normalized arbitrary fluorescence units (which can alternatively be viewed as average per cell fluorescence). We start by collecting the
mean change response of each sensor at a particular instant in time, \( \vec{y}_t \), into a \( n_s \times M \) data matrix, \( Y \)

\[
Y = \begin{bmatrix} y_0 & y_1 & \ldots & y_{M-1} \end{bmatrix}
\]

(19)

where \( M \) denotes the number of time points. Then a desired response vector, \( s \), is generated corresponding to the desired reference trajectory.

For example, the first reference trajectory (Figure 5) used in this work is generated by

\[
s_1(t) = 1.5 + 2.43e^{-0.44t}\sin 0.33t
\]

(20)

which corresponds to a second-order underdamped system subject to a step input of 1.5 (arbitrary units). The second reference trajectory is generated by the superposition of radial basis functions

\[
s_2(t) = e^{-\frac{(t-0)^2}{2}} + 0.2e^{-\frac{(t-9)^2}{2}} + e^{-\frac{(t-15)^2}{2}} + e^{-\frac{(t-21)^2}{2}} + 1.
\]

(21)

The two functions were sampled at the time points corresponding to the sensor response measurements to obtain the vector \( s \).

Attending to realistic constraints surrounding genetic circuit design, data acquisition, and cost, we seek to identify the fewest combination of transcriptional sensors that can be used to recapitulate the desired response \( s \). This can be described mathematically using the following cost function

\[
\min_{\beta \in \mathbb{R}^{n_s}} ||s - \beta^T Y||_2^2 + \gamma ||\beta||_1
\]

(22)

where \( ||s||_2 \) is the Euclidean norm, quantifying the distance of a vector from the origin. The term \( ||\beta||_1 \) is the 1-norm and adding this quantity to the cost function has been shown to promote sparsity in the minimizer \[89\]. As \( \gamma \) increases, the number of sensors to recapitulate the desired response decreases. However if \( \gamma \) is too large, the sparse set of coefficients may be unable to accurately describe \( s \). This optimization problem represents a linear program with linear constraints and the minimizer is obtained using the splitting conic solver \[89\].

Data availability

The data generated from RNA sequencing are available at GEO Accession GSE200822: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE200822

The DNA sequencing data for the reporter strains and the kinetic data generated from the spectrophotometer are available at: https://github.com/AqibHasnain/transcriptome-dynamics-dmd-observability

Code availability

All codes used in this study are available at: https://github.com/AqibHasnain/transcriptome-dynamics-dmd-observability or available from the author’s upon request.

Acknowledgments

This work was supported by DARPA, AFRL under contract numbers FA8750-17-C-0229, HR001117C0092, HR001117C0094, DEAC0576RL01830. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Project Agency, the Department of Defense, or the United States government. This work was also funded, in part, by the Department of Energy’s Biological and Environmental Research office, under the DOE Scientific Focus Area: Secure Biosystems Design project, via funding from Pacific Northwest National Laboratory subcontract numbers 545157 and 490521. This work received partially funding from the Army Young Investigator Award W911NF-20-1-0165 and the Army Research Office Grants W911NF-19-D-001, W911-NF-19-F-037, and W911-NF-19-0026. We acknowledge the use of the Biological Nanostructures Laboratory within the California NanoSystems Institute, supported by the University of California, Santa Barbara and the University of California, Office of the President.

Authors contributions

A.H. and E.Y. designed research and experiments. A.H. performed experiments, performed formal analysis, analyzed data, and wrote the manuscript. S.B. assisted with RNA-seq sample collection and virtual sensor analysis. D.M.J. assisted with cloning of reporter strains; J.S. performed the mRNA library prep and sequencing; S.B.H. assisted in conceptualization and designing the time-series RNA-seq experiment; E.Y. supervised research and secured funding. A.H. revised the manuscript with inputs from all authors.

Competing interests

The authors declare no competing interests.

Ryan Chambers, Trevor Marks, and Kirk Fields for construction of the vacuum manifold. We thank Jamiree Harrison for engaging in insightful discussions on linear systems theory.
Bibliography


<table>
<thead>
<tr>
<th>Malathion reporter</th>
<th>Locus tag</th>
<th>Molecular function</th>
<th>Act./Rep.</th>
<th>$y_{\text{min}}$</th>
<th>$y_{\text{max}}$</th>
<th>$K_M$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>atpB</td>
<td>PFLU_0124</td>
<td>• proton-transporting ATP synthase activity, rotational mechanism</td>
<td>Activated</td>
<td>1467</td>
<td>1783</td>
<td>0.6</td>
<td>4.5</td>
</tr>
<tr>
<td>petA</td>
<td>PFLU_0841</td>
<td>• 2 iron, 2 sulfur cluster binding, • metal ion binding • ubiquinol-cytochrome-c reductase activity</td>
<td>Activated</td>
<td>853</td>
<td>1125</td>
<td>1.4</td>
<td>2.4</td>
</tr>
<tr>
<td>sucC</td>
<td>PFLU_1823</td>
<td>• ATP binding • magnesium ion binding • succinate-CoA ligase activity</td>
<td>Activated</td>
<td>257</td>
<td>337</td>
<td>0.4</td>
<td>21.6</td>
</tr>
<tr>
<td>rpoA</td>
<td>PFLU_2502</td>
<td>• DNA binding • protein dimerization activity • DNA-directed 5'-3' RNA polymerase activity</td>
<td>Activated</td>
<td>1256</td>
<td>1542</td>
<td>0.9</td>
<td>3.0</td>
</tr>
<tr>
<td>fabA</td>
<td>PFLU_1836</td>
<td>• dehydratase activity • isomerase activity</td>
<td>Activated</td>
<td>292</td>
<td>373</td>
<td>0.2</td>
<td>1.1</td>
</tr>
<tr>
<td>anti-sigma 28 factor</td>
<td>PFLU_4736</td>
<td>• Negative regulator of flagellin synthesis</td>
<td>Activated</td>
<td>339</td>
<td>535</td>
<td>0.7</td>
<td>1.5</td>
</tr>
<tr>
<td>Uncharacterized protein I</td>
<td>PFLU_2761</td>
<td></td>
<td>Activated</td>
<td>2465</td>
<td>3110</td>
<td>0.5</td>
<td>2.7</td>
</tr>
<tr>
<td>cspA2</td>
<td>PFLU_4150</td>
<td>• major cold shock protein</td>
<td>Activated</td>
<td>706</td>
<td>1186</td>
<td>1.5</td>
<td>5.3</td>
</tr>
<tr>
<td>Putative ABC transport protein</td>
<td>PFLU_0176</td>
<td>• ligand-gated ion channel activity</td>
<td>Activated</td>
<td>584</td>
<td>1083</td>
<td>1.0</td>
<td>2.0</td>
</tr>
<tr>
<td>gcrA</td>
<td>PFLU_1815</td>
<td>• citrate (Si)-synthase activity</td>
<td>Activated</td>
<td>238</td>
<td>458</td>
<td>0.9</td>
<td>1.9</td>
</tr>
<tr>
<td>lpxC</td>
<td>PFLU_0953</td>
<td>• metal ion binding • deacetylase activity</td>
<td>Activated</td>
<td>1017</td>
<td>2418</td>
<td>0.4</td>
<td>8.7</td>
</tr>
<tr>
<td>Uncharacterized protein II</td>
<td>PFLU_4358</td>
<td></td>
<td>Repressed</td>
<td>1073</td>
<td>3387</td>
<td>0.3</td>
<td>1.9</td>
</tr>
<tr>
<td>capB</td>
<td>PFLU_1302A</td>
<td>• cold shock protein</td>
<td>Repressed</td>
<td>9616</td>
<td>10543</td>
<td>1.0</td>
<td>8.6</td>
</tr>
<tr>
<td>Putative outer membrane porin A protein</td>
<td>PFLU_4612</td>
<td>• porin activity</td>
<td>Activated</td>
<td>642</td>
<td>1172</td>
<td>0.6</td>
<td>1.5</td>
</tr>
<tr>
<td>acrA</td>
<td>PFLU_1380</td>
<td>• transmembrane transporter activity</td>
<td>Activated</td>
<td>354</td>
<td>682</td>
<td>0.9</td>
<td>2.9</td>
</tr>
</tbody>
</table>

Table 1: Encoder library metadata and transfer curve parameters for the fitted Hill equations in Fig. 4d.
Learning transcriptome dynamics for discovery of optimal genetic reporters of novel compounds

Aqib Hasnain et al.

1 Supplementary Text

1.1 Observability maximization for transcriptome dynamics

Here we derive the solution to the observability maximization problem briefly outlined in the Methods section. Recall that we have a state-space representation of the transcriptome dynamics as

\[ x_{t+1} = Kx_t \]
\[ y = Wx_t \]

where \( x \in \mathbb{R}^n \) is the (hidden) cell state, \( K \) is the state transition matrix, \( W \) are the unknown gene sampling weights, and \( y \in \mathbb{R}^p \) are the \( p \) measurements. The objective, \( J \), is formulated by the signal energy (or output energy) of the system

\[ J = \sum_{i=1}^{m} y_i^T y_i = \sum_{i=0}^{m} x_0^T K^T W^T W K^i x_0, \]

and we seek the gene sampling weights \( W \) which maximize the objective

\[ \max_{W \in \mathbb{R}^{p \times n}} J \]

subject to \( WW^T = I_{p \times p} \). The constraint enforces that the rows of \( W \) are orthogonal to each other and that the length of each row be equal to 1. This further avoids the issue of the objective blowing up to infinity. The solution to the above optimization problem is obtained by forming the Lagrangian dual problem and finding the maxima of the dual objective in terms of the dual variable (a \( p \times p \) matrix), \( D \), i.e.

\[ \max_{W \in \mathbb{R}^{p \times n}} J + \mathcal{L} \]

where \( \mathcal{L} = -\text{tr}((WW^T - I_{p\times p})D) \) and \( \text{tr}() \) denotes the trace operator. Differentiating the dual objective with respect to \( WW^T \) and equating to 0, we have

\[ \frac{\partial (J + \mathcal{L})}{\partial WW^T} = \frac{\partial}{\partial WW^T} \left( \sum_{i=0}^{m} x_0^T K^T W^T W K^i x_0 - \text{tr}((WW^T - I_{p\times p})D) \right) \]
\[ = \frac{\partial}{\partial WW^T} \left( \sum_{i=0}^{m} \text{tr}(x_0^T K^T W^T W K^i x_0) - \text{tr}((WW^T - I_{p\times p})D) \right) \]
\[ = \frac{\partial}{\partial WW^T} \left( \sum_{i=0}^{m} \text{tr}(WK^i x_0 x_0^T K^T W^T) - \text{tr}((WW^T - I_{p\times p})D) \right) \]
\[ = \frac{\partial}{\partial WW^T} \left( \sum_{i=0}^{m} \text{tr}(WG^iT W^T) - \text{tr}((WW^T - I_{p\times p})D) \right) \]
\[ = \frac{\partial}{\partial WW^T} \left( \text{tr}(W \sum_{i=0}^{m} G^iT W^T) - \text{tr}((WW^T - I_{p\times p})D) \right) \]
\[ = \frac{\partial}{\partial WW^T} \left( \text{tr}(GW W^T) - \text{tr}((WW^T - I_{p\times p})D) \right) \]
\[ = 2GW^T - 2W^T D = 0 \]
where the second equality comes from the fact that $J$ is a sum of $m$ scalars and so applying the trace operator has no effect on the sum, the third equality uses the cyclic property of the trace of products, and the fifth equality uses the fact that $\text{tr}(A) + \text{tr}(B) = \text{tr}(A + B)$. Finally, the Gram matrix, $G$, is defined to be $G = \sum_i \mathbf{G}^{(i)} = \sum_i \mathbf{K}_i \mathbf{x}_0^i \mathbf{x}_0^i \mathbf{K}_i^T$, a sum of quadratic forms, which is itself a quadratic form and therefore a symmetric matrix with non-negative, real-valued eigenvalues. From the final equality in Eq. (5) we have $GW^T = W^T D$ (6) which says columns of the eigenvectors of $G$ are the rows of gene sampling weights $W$. Moreover, the eigenvector of $G$ corresponding to the eigenvalue with largest magnitude in $D$ is the maximizer when $p = 1$.  

1.2 Fold change dynamics of two linear systems

We have reasoned in the main text that the gene expression dynamics of each experimental condition are well approximated by a linear state-space representation. We then define the dynamics as

$$\begin{align*}
\frac{dx_{on}}{dt} &= ax_{on} + bu \\
\frac{dx_{off}}{dt} &= ax_{off}
\end{align*}$$

(7)

where here $x_{on}$ and $x_{off}$ are scalar variables for ease of analysis. The variables represent the dynamics in the case where the input is present (on) and when the input is absent (off), respectively. The input $u$ represents the scalar input of a small molecule, e.g. malathion, that drives the expression of genes in the on condition through a step input, i.e. $u(t) = 1$ for all $t > 0$. The solution of the linear ordinary differential equations above are given by

$$\begin{align*}
x_{on}(t) &= e^{at}x_0 + \int_0^t e^{a(t-\tau)}bu(\tau)d\tau \\
x_{off}(t) &= e^{at}x_0
\end{align*}$$

(8)

where $x(0) = x_0$ for both $x_{on}$ and $x_{off}$. We want to show that the fold change response is given by the solution of a linear dynamical system. Taking the fold change of $x_{on}$ to $x_{off}$ we have

$$x_{fc}(t) = \frac{x_{on}(t)}{x_{off}(t)} = 1 + \int_0^t e^{-at} \frac{b}{x_0} d\tau
$$

$$= 1 + \frac{b}{ax_0} \left[ e^{-at} - e^{at} \right]$$

$$= 1 + \alpha - \alpha e^{at}.$$

(9)

To show that there exists a linear ordinary differential equation (ODE) that gives rise to the above solution $x_{fc}(t)$, we apply the steps to solve linear ODEs using integrating factors but in reverse order. We know in advance that the integrating factor should take the form $e^{at}$ and we start by dividing both sides of (9) by this integrating factor

$$e^{-at}x_{fc} = e^{-at}(1 + \alpha) - \alpha.$$

(10)

We next differentiate both sides and integrate both sides with respect to $t$

$$\int \frac{d}{dt} \left( e^{-at}x_{fc} \right) dt = \int ae^{-at}dt - \int \alpha ae^{-at}dt,$$

(11)

then once again differentiating both sides gives

$$\frac{d}{dt} \left( e^{-at}x_{fc} \right) = ae^{-at} - \alpha ae^{-at}.$$

(12)

Applying the product rule to the left hand side, we have

$$e^{-at} \frac{dx_{fc}}{dt} - ae^{-at}x_{fc} = ae^{-at} - \alpha ae^{-at}$$

$$= e^{-at}(a - \alphaa).$$

(13)
Finally, multiplying through by the integrating factor, $e^{at}$, and solving for $\frac{dx_{tc}}{dt}$, we obtain

$$\frac{dx_{tc}}{dt} = ax_{tc} + a - \alpha a$$

which is a linear first order ODE, i.e. a linear dynamical system with a step input and $\alpha = \frac{b}{x_{eq}}$. The importance of this result is to to be able to say that if the dynamics of the transcriptome in each experimental condition are well represented by a linear system, than the fold change dynamics, under the stated assumptions, can also be well represented by a linear system.

We briefly remark on the extension to the multivariate case. Under the assumption that the system dynamics, $A$, is diagonalizable, the above analysis holds. One such transformation which diagonalizes the system is given by the set of eigenvectors of $A$. Formally, if we now have system dynamics with state, $x \in \mathbb{R}^n$, such that

$$\frac{dx_{on}}{dt} = Ax_{on} + Bu$$

$$\frac{dx_{off}}{dt} = Ax_{off},$$

applying the transformation $\tilde{x} = T^{-1}x$, where $T \in \mathbb{R}^{n \times n}$ is the matrix of eigenvectors of $A$, results in the transformed systems

$$\frac{d\tilde{x}_{on}}{dt} = D\tilde{x}_{on} + \tilde{B}u$$

$$\frac{d\tilde{x}_{off}}{dt} = D\tilde{x}_{off},$$

where $\tilde{B} = T^{-1}B$. To solve for the fold change dynamics in the multivariate case, we cast the state coordinates into a diagonal matrix, i.e. $\text{diag}(\tilde{x})$, and compute $\text{diag}(\tilde{x}_{on})(\text{diag}(\tilde{x}_{off}))^{-1}$. Since the solution in each coordinate is uncoupled from other coordinates, we then have $n$ solutions, each as in Eq. (9).

The case where the above derivation does not hold when the eigenvalues of $A$ have zero real part, i.e. they are exactly zero or have purely sinusoidal response (corresponding to periodic orbits). In this case, the fold change in the coordinate corresponding to zero eigenvalues will approach infinity or it will not be possible to represent the fold change dynamics as a sum of weighted exponentials, e.g. $\tan(x)$. However, such a case would be improbable in a data-driven application for gene regulatory networks. Moreover, any eigenvalue with magnitude zero does not contribute to the dynamics of the system and should be removed from the model.

2 Supplementary Figures
Figure 1: The eigenvalues of the DMD operator plotted in the complex plane for varying number of modes.

Figure 2: (Left) Approximation of the eigenvalues of the Gram matrix by the reduced order model given by DMD. The full Gram matrix eigenvalues are given in blue circles and the reduced Gram matrix eigenvalues are given in orange squares. (Right) Approximation of the leading eigenvector of the Gram matrix by the reduced order model given by DMD. This eigenvector corresponds to the gene sampling weights in the main text.
Figure 3: The full plasmid map of pBHVK with the reporter cassette. The two BsaI cut sites on either side of the promoter, pJ23150, are used in Golden Gate Assembly to replace the promoter sequence with a promoter used for malathion sensing. A bicistronic design is used for the ribosome binding site, BCD1. A terminator from the set of Voigt lab terminators is used, T14. For fluorescent reporting, super folder GFP (sfGFP) is used. See Table 3 for sequences of the terminator, ribosome binding site, and sfGFP. See figure 2 for sequences of the promoters used in the sensor library.
Figure 4: Growth curves of each malathion reporter subject to malathion induction by means of Spectracide.

Figure 5: Comparison of the reconstruction accuracy if genes were sampled according to observability ranked sampling (solid line) vs. random sampling (dashed line).
Figure 6: Vacuum manifold design for rapid sampling of mRNA dynamics.
Figure 7: Mass spectrum of malathion (Millipore Sigma Catalog no. 36143) given by time-of-flight mass spectrometry. The theoretical mass spectrum is shown in the upper spectrum and the measured mass spectrum is shown in the lower spectrum.
Figure 8: Mass spectrum of Spectracide (replicate 1) (Spectracide Catalog no. 071121309006) given by time-of-flight mass spectrometry. The theoretical mass spectrum of malathion is shown in the upper spectrum and the measured mass spectrum of Spectracide is shown in the lower spectrum.
Figure 9: Mass spectrum of Spectracide (replicate 2) (Spectracide Catalog no. 071121309006) given by time-of-flight mass spectrometry. The theoretical mass spectrum of malathion is shown in the upper spectrum and the measured mass spectrum of Spectracide is shown in the lower spectrum.
Figure 10: Mass spectrum of Spectracide (replicate 3) (Spectracide Catalog no. 071121309006) given by time-of-flight mass spectrometry. The theoretical mass spectrum of malathion is shown in the upper spectrum and the measured mass spectrum of Spectracide is shown in the lower spectrum.
Figure 11: Mass spectrum of Spectracide (replicate 4) (Spectracide Catalog no. 071121309006) given by time-of-flight mass spectrometry. The theoretical mass spectrum of malathion is shown in the upper spectrum and the measured mass spectrum of Spectracide is shown in the lower spectrum.
Figure 12: Mass spectrum of a 5% Spectracide in LB broth (replicate 1) given by time-of-flight mass spectrometry. The theoretical mass spectrum of malathion is shown in the upper spectrum and the measured mass spectrum of the solution is shown in the lower spectrum.
Figure 13: Mass spectrum of a 5% Spectracide in LB broth (replicate 2) given by time-of-flight mass spectrometry. The theoretical mass spectrum of malathion is shown in the upper spectrum and the measured mass spectrum of the solution is shown in the lower spectrum.
Figure 14: Mass spectrum of a 5% Spectracide in LB broth (replicate 3) given by time-of-flight mass spectrometry. The theoretical mass spectrum of malathion is shown in the upper spectrum and the measured mass spectrum of the solution is shown in the lower spectrum.
Figure 15: Mass spectrum of a 5% Spectracide in LB broth (replicate 4) given by time-of-flight mass spectrometry. The theoretical mass spectrum of malathion is shown in the upper spectrum and the measured mass spectrum of the solution is shown in the lower spectrum.
Figure 16: Mass spectrum of a 5% Spectracide in LB broth after photobleaching (replicate 1) given by time-of-flight mass spectrometry. The theoretical mass spectrum of malathion is shown in the upper spectrum and the measured mass spectrum of the photobleached solution is shown in the lower spectrum.
Figure 17: Mass spectrum of a 5% Spectracide in LB broth after photobleaching (replicate 2) given by time-of-flight mass spectrometry. The theoretical mass spectrum of malathion is shown in the upper spectrum and the measured mass spectrum of the photobleached solution is shown in the lower spectrum.
Figure 18: Mass spectrum of a 5% Spectracide in LB broth after photobleaching (replicate 3) given by time-of-flight mass spectrometry. The theoretical mass spectrum of malathion is shown in the upper spectrum and the measured mass spectrum of the photobleached solution is shown in the lower spectrum.
Figure 19: Mass spectrum of a 5% Spectracide in LB broth after photobleaching (replicate 4) given by time-of-flight mass spectrometry. The theoretical mass spectrum of malathion is shown in the upper spectrum and the measured mass spectrum of the photobleached solution is shown in the lower spectrum.
3 Supplementary Tables

<table>
<thead>
<tr>
<th>Time point (minutes)</th>
<th>OD&lt;sub&gt;600&lt;/sub&gt;</th>
<th>Volume harvested (mL)</th>
<th>Malathion induction</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.5</td>
<td>10</td>
<td>X</td>
</tr>
<tr>
<td>10</td>
<td>–</td>
<td>10</td>
<td>–</td>
</tr>
<tr>
<td>20</td>
<td>–</td>
<td>10</td>
<td>–</td>
</tr>
<tr>
<td>30</td>
<td>–</td>
<td>10</td>
<td>–</td>
</tr>
<tr>
<td>40</td>
<td>–</td>
<td>10</td>
<td>–</td>
</tr>
<tr>
<td>50</td>
<td>–</td>
<td>10</td>
<td>–</td>
</tr>
<tr>
<td>60</td>
<td>1.0</td>
<td>10</td>
<td>–</td>
</tr>
<tr>
<td>70</td>
<td>–</td>
<td>10</td>
<td>–</td>
</tr>
<tr>
<td>80</td>
<td>–</td>
<td>10</td>
<td>–</td>
</tr>
<tr>
<td>90</td>
<td>–</td>
<td>10</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 1: Metadata for the time-series RNAseq experiment.

<table>
<thead>
<tr>
<th>Sensor</th>
<th>Strand</th>
<th>Loci</th>
<th>Promoter sequence</th>
</tr>
</thead>
</table>
| PFLU   | Antisense | 6709164 - 6709017 | AACATTTGCTTATGTAGCGCTGATCGGAAATCCTACCCGCGCAATTTGGAATAGGGCAGAACCGCCCTATATAGCTCTC GCGCGCAATTTGTCGCCCAAATAATATGCGCATATGATGAGATGTCCTCCAGCCGACCCAGGAGCAATTCATTCCCGGAGGCAAGAAGGTG
| PFLU   | Sense   | 950865 - 951131 | TTCCGCTTTACCGCTCACAACAAAGCAGCCGCCCTTGTTCTCTGTACATAGCTCCAGGAGGAGCATTGGAATAGTGTCCTCCAGCCGACCCAGGAGCAATTCATTCCCGGAGGCAAGAAGGTG
| PFLU   | Sense   | 5213048 - 5213188 | AGTGCTGGCAGGACGCTGCTGGGTTTTTTCTACATCTGTGCACGCTATTCGGTCGAGAGGAGCGGATATGCTGGAATAGTGTCCTCCAGCCGACCCAGGAGCAATTCATTCCCGGAGGCAAGAAGGTG
| PFLU   | Sense   | 1989934 - 1990137 | GCGAGATAATAAGAAACACGCAGGAGTTGGCCCGTCTCAGCGATTCGGCGAGGAGGAGCGGATATGCTGGAATAGTGTCCTCCAGCCGACCCAGGAGCAATTCATTCCCGGAGGCAAGAAGGTG
| PFLU   | Antisense | 4158693 - 4158135 | CTGTGACACGTTCGCGAAAGCGCGAGTTGGCCCGTCTCAGCGATTCGGCGAGGAGGAGCGGATATGCTGGAATAGTGTCCTCCAGCCGACCCAGGAGCAATTCATTCCCGGAGGCAAGAAGGTG
| PFLU   | Antisense | 6038217-6038089 | GTGTGATCCGCTTGAAGCCCGGCAGCTAGTGCGCTGCCGGGTTGATTATTTGTTATTACAGCGATATTATCTCGCGCCCTATTTCTTGGCTTCCGGGGCGTAGGTAGCTGTCAAAGTGCTTGTTATCTGGTAGGAC

The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.
<table>
<thead>
<tr>
<th>Part</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCD1</td>
<td>GGCCCAAGTTCACTTTAAAAAGGAGATCAAACATGAAAGCAATTTTCGTAATTGAAAACACATCTTAAATGCACAGGACAGACTTTCT</td>
</tr>
<tr>
<td>T14</td>
<td>AACGCATGAGAAAGCCCGGAAGATCACCTTCCGGGGGCTTTTTTATTGCGC</td>
</tr>
<tr>
<td>sfGFP</td>
<td>ATGCCTTAAGGCGAGAGCTTTACTGTTGTCGCCACTTATTTCGTAATTGAAAACACATCTTAAATGCACAGGACAGACTTTCT</td>
</tr>
</tbody>
</table>

Table 3: Sequences for the parts used in the reporter cassette.
<table>
<thead>
<tr>
<th>Malathion reporter</th>
<th>Locus tag</th>
<th>Time point (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>atpB</td>
<td>PFLU_6124</td>
<td>1.0</td>
</tr>
<tr>
<td>petA</td>
<td>PFLU_0841</td>
<td>2.0</td>
</tr>
<tr>
<td>anti-sigma 28 factor</td>
<td>PFLU_4736</td>
<td>3.2</td>
</tr>
<tr>
<td>sucC</td>
<td>PFLU_1823</td>
<td>8.1</td>
</tr>
<tr>
<td>Uncharacterized protein I</td>
<td>PFLU_3761</td>
<td>12.9</td>
</tr>
<tr>
<td>rpoA</td>
<td>PFLU_5502</td>
<td>15.0</td>
</tr>
<tr>
<td>fabA</td>
<td>PFLU_1836</td>
<td>14.0</td>
</tr>
<tr>
<td>Putative ABC transport protein</td>
<td>PFLU_0376</td>
<td>0.9</td>
</tr>
<tr>
<td>gltA</td>
<td>PFLU_1815</td>
<td>3.2</td>
</tr>
<tr>
<td>lpxC</td>
<td>PFLU_0953</td>
<td>0.7</td>
</tr>
<tr>
<td>acrA</td>
<td>PFLU_1380</td>
<td>3.1</td>
</tr>
<tr>
<td>Putative outer membrane porin A protein</td>
<td>PFLU_4612</td>
<td>2.0</td>
</tr>
<tr>
<td>cspA2</td>
<td>PFLU_4150</td>
<td>2.4</td>
</tr>
<tr>
<td>capB</td>
<td>PFLU_1302A</td>
<td>8.5</td>
</tr>
<tr>
<td>Uncharacterized protein II</td>
<td>PFLU_1358</td>
<td>5.6</td>
</tr>
</tbody>
</table>

Table 4: The time points at which the Hill functions are fit to each reporters’ response.