bioRxiv preprint doi: https://doi.org/10.1101/2022.11.29.518354; this version posted November 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A general model-based causal inference overcomes the curse of synchrony and indirect effect

Se Ho Park
Department of Mathematics, University of Wisconsin-Madison, WI 53706, United States and Biomedical Mathematics Group, Institute for Basic Science, Daejeon 34126, Republic of Korea

Seokmin Ha and Jae Kyoung Kim*
Department of Mathematical Sciences, KAIST, Daejeon 34141, Republic of Korea and Biomedical Mathematics Group, Institute for Basic Science, Daejeon 34126, Republic of Korea

To identify causation, model-free inference methods, such as Granger Causality, have been widely used due to their flexibility. However, they have difficulty distinguishing synchrony and indirect effects from direct causation, leading to false predictions. To overcome this, model-based inference methods were developed that test the reproducibility of data with a specific mechanistic model to infer causality. However, they can only be applied to systems described by a specific model, greatly limiting their applicability. Here, we address this limitation by deriving an easily-testable condition for a general ODE model to reproduce time-series data. We built a user-friendly computational package, GOBI (General ODE-Based Inference), which is applicable to nearly any system described by ODE. GOBI successfully inferred positive and negative regulations in various networks at both molecular and population levels, unlike existing model-free methods. Thus, this accurate and broadly-applicable inference method is a powerful tool for understanding complex dynamical systems.

I. INTRODUCTION

Identifying a causal interaction is crucial to understand the underlying mechanism of systems in nature. A recent surge in time-series data collection with advanced technology offers opportunities to computationally uncover causation [1]. Various model-free methods, such as Granger Causality (GC) [2] and Convergent Cross Mapping (CCM) [3], have been widely used to infer causation from time-series data. Although they are easy to implement and broadly applicable [4–10], they usually struggle to differentiate synchrony (i.e., similar periods among components) versus causality [11–15] and distinguish between direct and indirect causation [16–20]. For instance, when oscillatory time-series data is given, nearly all-to-all connected networks are inferred [12]. To prevent such false positive predictions, model-free methods have been improved (e.g., Partial Cross Mapping (PCM) [20]), but further investigation is needed to show their universal validity.

Alternatively, model-based methods infer causality by testing the reproducibility of time-series data with mechanistic models. Although testing the reproducibility is computationally expensive, as long as the underlying model is accurate, the model-based inference method is accurate even in the presence of synchrony in time series and indirect effect [21–27]. However, the inference results strongly depend on the choice of model, and inaccurate model imposition can result in false positive predictions, limiting their applicability. To overcome this limit, inference methods using flexible models were developed [28–34]. In particular, the most recent method, ION [12], infers causation from X to Y described by the general ODE model between two components, i.e., $\frac{dY}{dt} = f(X, Y)$. However, ION is applicable only when every component is affected by at most one another component.

Here, we develop a model-based method that infers interactions among multiple components described by the general ODE model:

$$\frac{dY}{dt} = f(\boldsymbol{X}) = f(X_1, X_2, \cdots, X_N), \tag{1}$$

where f can be any smooth and monotonic increasing or decreasing functions of X_i and X_N is Y in the presence of self-regulation. Thus, our approach completely resolves the fundamental limit of model-based inference: strong dependence on a chosen model. Furthermore, we derive the simple condition for the reproducibility of time-series with Eq. (1), which does not require computationally expensive fitting, unlike previous model-based approaches. To facilitate our approach, we develop a user-friendly computational package, GOBI (General ODE-Based Inference). GOBI successfully infers causal relationships in gene regulatory networks, ecological system, and cardiovascular disease caused by air pollution from synchronous time-series data, with which popular model-free methods fail at inference. Furthermore, GOBI can also distinguish between direct and indirect causation even from noisy time-series data. Because GOBI is both accurate and broadly applicable, which have not been achieved by previous model-free or model-based inference methods, it can be a powerful tool in understanding complex dynamical systems.

* jaekkim@kaist.ac.kr

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.29.518354; this version posted November 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

II. RESULTS

A. Inferring regulation types from time-series

We first illustrate the common properties of time series generated by either positive or negative causation with simple examples. When the input signal X positively regulates Y ($X \to Y$) (Fig. 1a), $\dot{Y}$ increases whenever X increases. Thus, for any pair of time point t and t^* with which $X^d(t,t^*) := X(t) - X(t^*) > 0$, then $\dot{Y}^d(t,t^*) := \dot{Y}(t) - \dot{Y}(t^*) > 0$. Similarly, when X negatively regulates Y ($X \dashv Y$) (Fig. 1c left), if $X^d(t,t^*) < 0$, then $\dot{Y}^d(t,t^*) > 0$. Thus, in the presence of either positive ($\sigma = +$) or negative ($\sigma = -$) regulation, the following *regulation-detection function* is always positive (Fig. 1b and c):

$$I^Y_{X^\sigma}(t,t^*) := \sigma X^d(t,t^*) \cdot \dot{Y}^d(t,t^*) \tag{2}$$

defined on (t,t^*) such that $\sigma X^d(t,t^*) > 0$.

This idea can be extended to a case with multiple causes. For instance, when X_1 and X_2 positively regulate Y together (Fig. 1d), if both $X_1^d > 0$ and $X_2^d > 0$, then $\dot{Y}^d > 0$. This leads to the positivity of the regulation-detection function for $\begin{smallmatrix} X_1 \to \\ X_2 \to \end{smallmatrix} Y$, $I^Y_{X_1^+ X_2^+}(t,t^*) := X_1^d(t,t^*) \cdot X_2^d(t,t^*) \cdot \dot{Y}^d(t,t^*)$, defined for (t,t^*) such that $X_1^d(t,t^*) > 0$ and $X_2^d(t,t^*) > 0$ (Fig. 1e). Similarly, if X_1 and X_2 positively and negatively regulate Y, respectively (Fig. 1g), the regulation-detection function for $\begin{smallmatrix} X_1 \to \\ X_2 \dashv \end{smallmatrix} Y$, $I^Y_{X_1^+ X_2^-}(t,t^*) := X_1^d(t,t^*) \cdot (-X_2^d(t,t^*)) \cdot \dot{Y}^d(t,t^*)$, is positive for (t,t^*) such that $X_1^d(t,t^*) > 0$ and $X_2^d(t,t^*) < 0$ (Fig. 1i). Note that for $\begin{smallmatrix} X_1 \to \\ X_2 \to \end{smallmatrix} Y$ ($\begin{smallmatrix} X_1 \to \\ X_2 \dashv \end{smallmatrix} Y$), $I^Y_{X_1^+ X_2^-}$ ($I^Y_{X_1^+ X_2^+}$) is not always positive (Fig. 1f, h). Thus, the non-positivity of the regulation-detection function can be used to infer the absence of the regulation. The same positive relationships can be seen in other types of 2D regulations (Supplementary Fig. 1).

The positivity and negativity of the regulation-detection function $I^Y_{\boldsymbol{X}^\sigma}$ reflect the presence and absence of regulation, respectively. The sign of the $I^Y_{\boldsymbol{X}^\sigma}$ can be quantified with its normalized integral, *regulation-detection score* $S^Y_{\boldsymbol{X}^\sigma}$ (Eq. (4)). Thus, $S^Y_{\boldsymbol{X}^\sigma} = 1$ in the presence of regulation type σ since the regulation-detection function is positive (see Supplementary Information for details). However, even in the absence of regulation type σ, $S^Y_{\boldsymbol{X}^\sigma}$ can often be one. For instance, when X_1 positively regulates Y and X_2 does not regulate Y (Fig. 1j), $\dot{Y}$ increases whenever X_1 increases regardless of X_2. Thus, both $I^Y_{X_1^+ X_2^+}$ and $I^Y_{X_1^+ X_2^-}$ are positive (Fig. 1k and l). Here, $S^Y_{X_1^+ X_2^+} = S^Y_{X_1^+ X_2^-} = 1$ reflects that X_2 does not affect the regulation $X_1 \to Y$. Thus, to quantify the effect of a new component (e.g., X_2) on an existing regulation (e.g., $X_1 \to Y$), we develop a *regulation-delta function* Δ:

$$\Delta^Y_{X_1^+}(X_2) := S^Y_{X_1^+ X_2^+} - S^Y_{X_1^+ X_2^-}. \tag{3}$$

If $\Delta^Y_{X_1^+}(X_2) = 0$, $S^Y_{X_1^+ X_2^+} = 1$ ($S^Y_{X_1^+ X_2^-} = 1$) does not indicate the presence of $\begin{smallmatrix} X_1 \to \\ X_2 \to \end{smallmatrix} Y$ ($\begin{smallmatrix} X_1 \to \\ X_2 \dashv \end{smallmatrix} Y$).

B. Inferring regulatory network structure

$S^Y_{\boldsymbol{X}^\sigma} = 1$ together with $\Delta \neq 0$ can be used as an indicator of regulation type σ from $\boldsymbol{X}$ to Y. Based on this, we construct a framework for inferring a regulatory network from time-series data (Fig. 2a). To illustrate this, we obtain multiple time-series data simulated with random input signal A and different initial conditions of B and C randomly selected from $[-1, 1]$.

From each time series, regulation-detection score $S^Y_{\boldsymbol{X}^\sigma}$ is calculated for every 1D regulation type σ (Step 1). Here, for each regulation, $\boldsymbol{X}$ are causes and Y is a target among A, B and C. Because only $A \dashv B$ satisfies the criteria $S^Y_{\boldsymbol{X}^\sigma} = 1$ for every time series, only $A \dashv B$ is inferred as 1D regulation. Note that even for the other regulations, $S^Y_{\boldsymbol{X}^\sigma} = 1$ can occur for a few time series, leading to a false positive prediction. This can be prevented by using multiple time series. Next, $S^Y_{\boldsymbol{X}^\sigma}$ is calculated for every 2D regulation type σ (Step 2). Three types of regulation ($\begin{smallmatrix} A \dashv \\ C \to \end{smallmatrix} B$, $\begin{smallmatrix} A \dashv \\ C \dashv \end{smallmatrix} B$ and $\begin{smallmatrix} A \to \\ B \to \end{smallmatrix} C$) satisfy the criteria $S^Y_{\boldsymbol{X}^\sigma} = 1$ for every time series. Among these, we can identify false positive regulations by using a regulation-delta function (Step 3). $\Delta^B_{A^-}(C)$ is equal to zero for every time series, indicating that $\begin{smallmatrix} A \dashv \\ C \to \end{smallmatrix} B$ and $\begin{smallmatrix} A \dashv \\ C \dashv \end{smallmatrix} B$ are false positive regulations. Thus, $\begin{smallmatrix} A \to \\ B \to \end{smallmatrix} C$ is the only inferred 2D regulation as it satisfies the criteria for the regulation-delta function ($\Delta^C_{A^+}(B) \neq 0$ and $\Delta^C_{B^+}(A) \neq 0$). By merging the inferred 1D and 2D regulations, the regulatory network is successfully inferred. Since there are three components in this system, we inferred up to 2D regulations. If there are N components in the system, we go up to $(N-1)$D regulations (Supplementary Fig. 2).

We have applied the framework to infer regulatory networks from simulated time-series data of various biological models. In most biological systems, the degradation rates of molecules increase as their own concentrations increase; thus we assume that self-regulation is negative for every component in the system. Thus, to detect ND regulation, the $(N+1)$D regulation-detection function and score, including negative self-regulation, is used. For example, to infer 1D positive regulation from X to Y, the criteria $S^Y_{X^+Y^-} = 1$ is used.

From the time series simulated with the Kim-Forger model (Fig. 2b left), describing the negative feedback loop of the mammalian circadian clock [35], using the criteria $S^Y_{\boldsymbol{X}^\sigma Y^-} = 1$, two positive 1D regulations ($M \to P_C$ and $P_C \to P$) and one negative 1D regulation ($P \dashv M$) are inferred (Fig. 2b middle). Among the six different types of 2D regulations ($\begin{smallmatrix} M \to \\ P \to \end{smallmatrix} P_C$, $\begin{smallmatrix} M \to \\ P \dashv \end{smallmatrix} P_C$, $\begin{smallmatrix} P_C \to \\ M \to \end{smallmatrix} P$, $\begin{smallmatrix} P_C \to \\ M \dashv \end{smallmatrix} P$, $\begin{smallmatrix} P \dashv \\ P_C \to \end{smallmatrix} M$ and $\begin{smallmatrix} P \dashv \\ P_C \dashv \end{smallmatrix} M$) satisfying the criteria $S^Y_{\boldsymbol{X}^\sigma Y^-} = 1$ for all the time series, none of them pass the Δ test (i.e., $\Delta^{P_C}_{M^+}(P) = \Delta^P_{P_C^+}(M) = \Delta^M_{P^-}(P_C) = 0$) (Fig. 2b middle). Thus, no 2D regulation is inferred. By merging the

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.29.518354; this version posted November 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

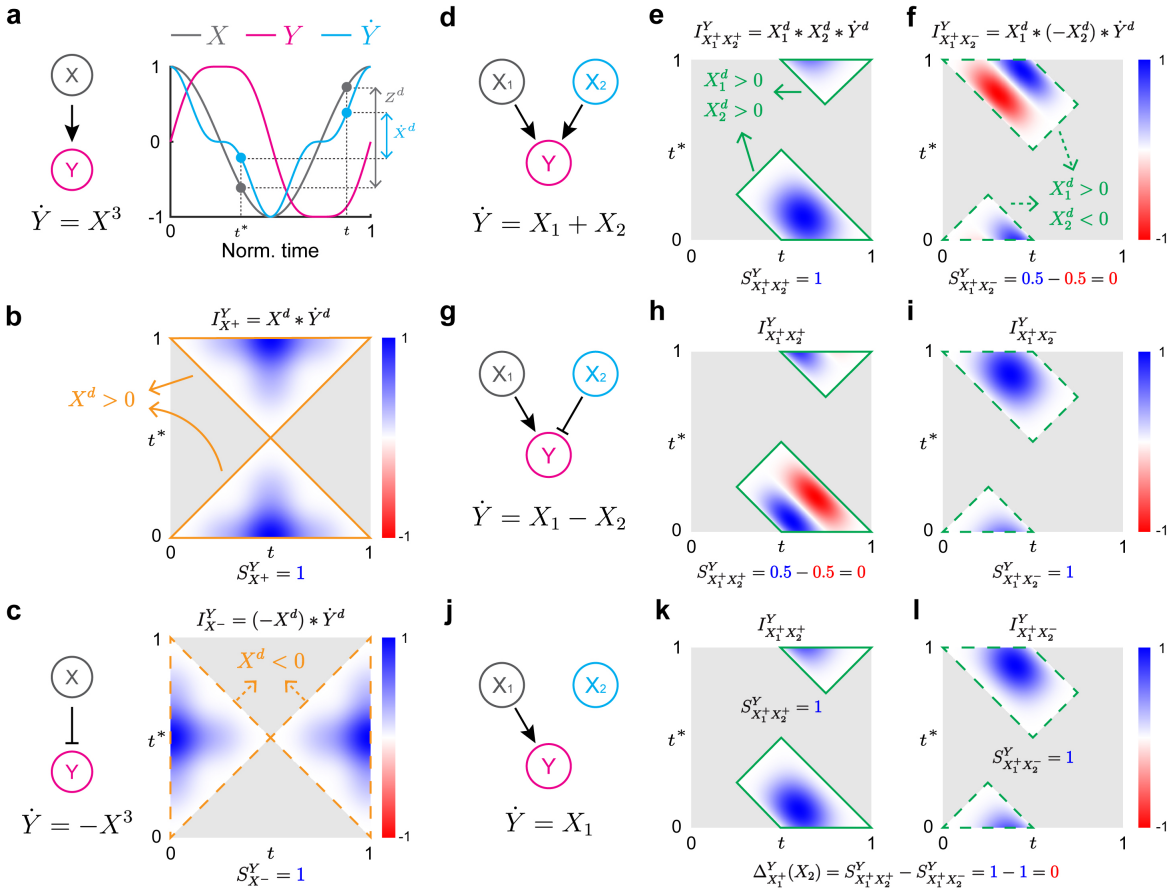

FIG. 1. Inferring regulation types using regulation-detection functions and scores. **a** Because X positively regulates Y, as X increases, $\dot{Y}$ increases. Thus, whenever $X^d(t,t^*) = X(t) - X(t^*) > 0$, $\dot{Y}^d(t,t^*) = \dot{Y}(t) - \dot{Y}(t^*) > 0$. **b** Therefore, when $X^d(t,t^*) > 0$, regulation-detection function $I^Y_{X+}(t,t^*) := X^d(t,t^*) \cdot \dot{Y}^d(t,t^*)$ is always positive. Here, I is in the range $[-1,1]$ since all the time series were normalized. **c** If X negatively regulates Y, $I^Y_{X-} := (-X^d) \cdot \dot{Y}^d$ is always positive when $X^d(t,t^*) < 0$. **d-i** When X_1 and X_2 positively regulate Y, as X_1 and X_2 increase ($X_1^d > 0$, $X_2^d > 0$), $\dot{Y}$ increases ($\dot{Y}^d > 0$) (**d**). Thus, when $X_1^d(t,t^*) > 0$ and $X_2^d(t,t^*) > 0$, $I^Y_{X_1^+X_2^+} := X_1^d \cdot X_2^d \cdot \dot{Y}^d$ is positive (**e**). When X_1 and X_2 positively and negatively regulate Y, respectively (**g**), $I^Y_{X_1^+X_2^-} := X_1^d \cdot (-X_2^d) \cdot \dot{Y}^d$ is always positive when $X_1^d(t,t^*) > 0$ and $X_2^d(t,t^*) < 0$ (**i**). Such positivity disappears for the regulation-detection functions, which do not match with the actual regulation type (**f** and **h**). **j-l** When X_1 positively regulates Y and X_2 does not regulate Y (**j**), both $I^Y_{X_1^+X_2^+} := X_1^d \cdot X_2^d \cdot \dot{Y}^d$ (**k**) and $I^Y_{X_1^+X_2^-} := X_1^d \cdot (-X_2^d) \cdot \dot{Y}^d$ (**l**) are positive because the regulation type of X_2 does not matter. Here, we use $X_1(t) = \cos(2\pi t)$ and $X_2(t) = \sin(2\pi t)$ as the input signal and $Y(0) = 0$ for simulation on $[0,1]$.

three inferred 1D regulations, the negative feedback loop structure is recovered (Fig. 2b right). Our method also successfully inferred the negative feedback loop structure of *Frzilator* [36] (Fig. 2c) and the 4-state Goodwin oscillator [37] (Fig. 2d). Furthermore, our framework correctly inferred the systems having 2D regulations: the Goldbeter model describing the *Drosophila* circadian clock [38] (Fig. 2e) and the regulatory network of the cAMP oscillator of *Dictyostelium* [39] (Fig. 2f) (see Supplementary Information for the equations and parameters of the models and Supplementary Data 1 for detailed inference results).

C. Inference with noisy time series

In the presence of noise in the time-series data, the regulation-detection score ($S^Y_{\boldsymbol{X}^\sigma}$) may not be one even if there is a regulation type σ from $\boldsymbol{X}$ to Y. For example, in the case of an Incoherent Feed-forward Loop (IFL) which contains $A \dashv B$ (Fig. 3a), S^B_{A-} is not one in the presence of noise (Fig. 3b blue). Thus, for noisy data, we need to relax the criteria $S^Y_{\boldsymbol{X}^\sigma} = 1$ to $S^Y_{\boldsymbol{X}^\sigma} > S^{\text{thres}}$ where $S^{\text{thres}} < 1$ is a threshold. Because S^B_{A-} gets farther away from one as the noise level increases, S^{thres} also needs to be decreased. We choose S^{thres} as $0.9 - 0.005 \times$ (noise level) with which true and false regulations can be distinguished in the majority of cases for

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.29.518354; this version posted November 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

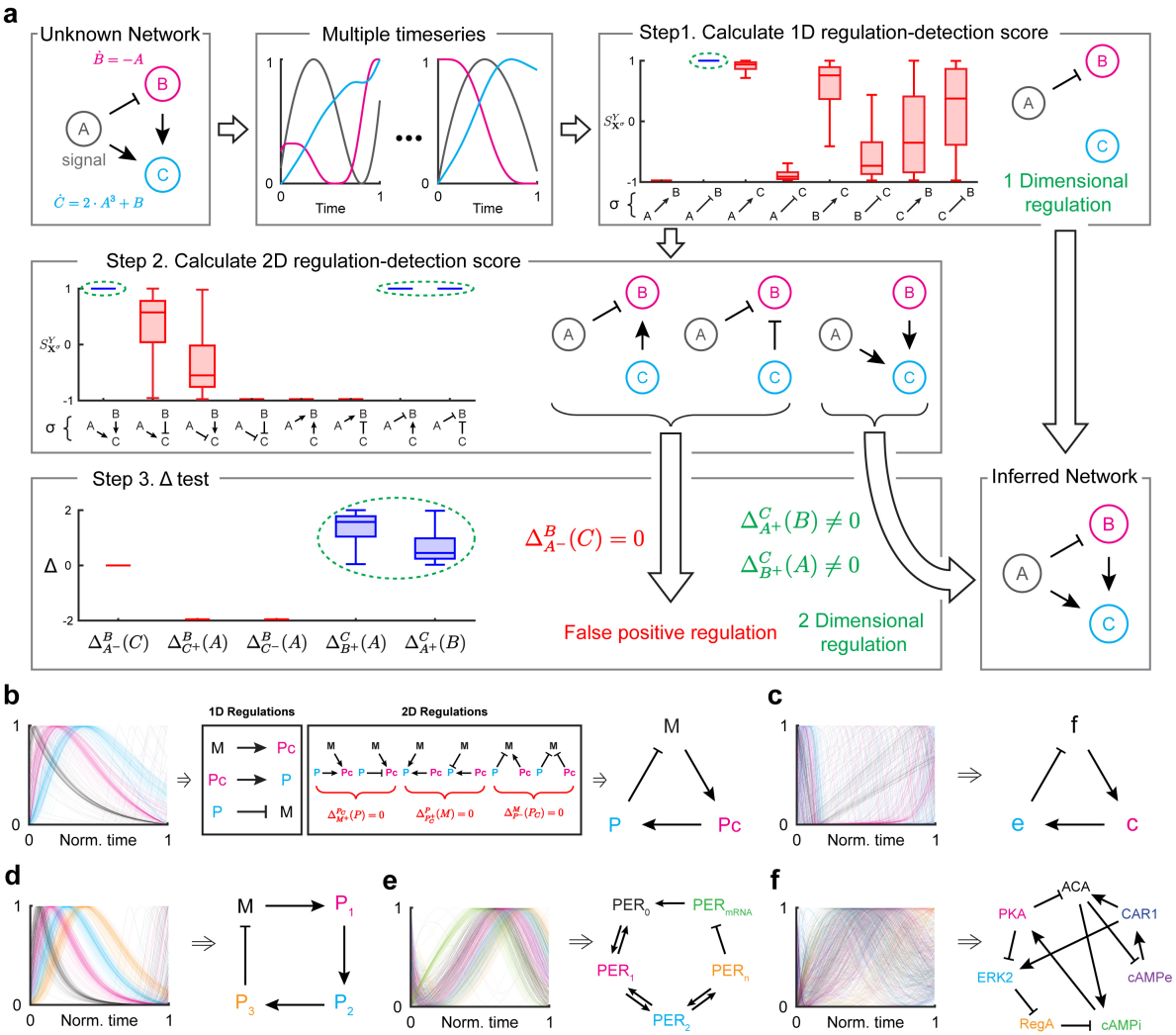

FIG. 2. Framework for inferring regulatory networks. **a** With ODE describing the network (left), various time series are simulated with different initial conditions (middle). Then, from each time series, regulation-detection score $S^Y_{\mathbf{X}\sigma}$ is calculated for every 1D regulation type σ (Step 1). The criteria $S^Y_{\mathbf{X}\sigma} = 1$ infers $A \dashv B$. Next, $S^Y_{\mathbf{X}\sigma}$ is calculated for every 2D regulation type σ (Step 2). Among the three types of regulations with $S^Y_{\mathbf{X}\sigma} = 1$, only one passed the Δ test (Step 3). By merging the inferred 2D regulation with the 1D regulation from Step 1, the regulatory network is successfully inferred. **b-f** This framework successfully infers the network structures of the Kim-Forger model (**b**), Frzilator (**c**), the 4-state Goodwin oscillator (**d**), the Goldbeter model for the *Drosophila* circadian clock (**e**), and the cAMP oscillator of *Dictyostelium* (**f**). For each model, 100 time-series data were simulated from randomly selected initial conditions, which lie in the range of the original limit cycle.

our examples (Fig. 3b and Supplementary Fig. 3e). For instance, S^{thres} (green dashed line, Fig. 3b) overall separates true regulation (Fig. 3b blue) and false regulation (Fig. 3b red). However, $S^Y_{\mathbf{X}\sigma} > S^{\text{thres}}$ ($S^Y_{\mathbf{X}\sigma} < S^{\text{thres}}$) is not always satisfied for true (false) regulation type σ from $\boldsymbol{X}$ to Y (Fig. 3b). Thus, we further use a *Total Regulation Score (TRS)*, the fraction of time-series data satisfying $S^Y_{\mathbf{X}\sigma} > S^{\text{thres}}$ (Fig. 3c left). Then, we use the criteria $\text{TRS}^Y_{\mathbf{X}\sigma} > \text{TRS}^{\text{thres}}$ to infer the regulation. Similar to S^{thres}, $\text{TRS}^{\text{thres}}$ also decreases as the noise level increases. Thus, we use $\text{TRS}^{\text{thres}} = 0.9 - 0.01 \times (\text{noise level})$, which successfully distinguishes between the true and false regulation of IFL (Fig. 3c right) and the other systems (Supplementary Fig. 3f). Note that $\text{TRS}^Y_{\mathbf{X}\sigma}$ is the measure which integrates the weight given on the regulation-detection score reflecting the size of the domain of the regulation-detection function (see Supplementary Information for details). See Method for how to quantify the noise level.

Next, we investigate whether the Δ test can distinguish direct and indirect regulations using examples of Coherent Feed-forward Loop (CFL, Fig. 3d) and Single Feed-forward Loop (SFL, Fig. 3e). In CFL, direct regulation of $A \dashv C$ exists. On the other hand, in SFL, only indirect negative regulation from A to C, induced from a regulatory chain $A \dashv B \rightarrow C$, exists.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.29.518354; this version posted November 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

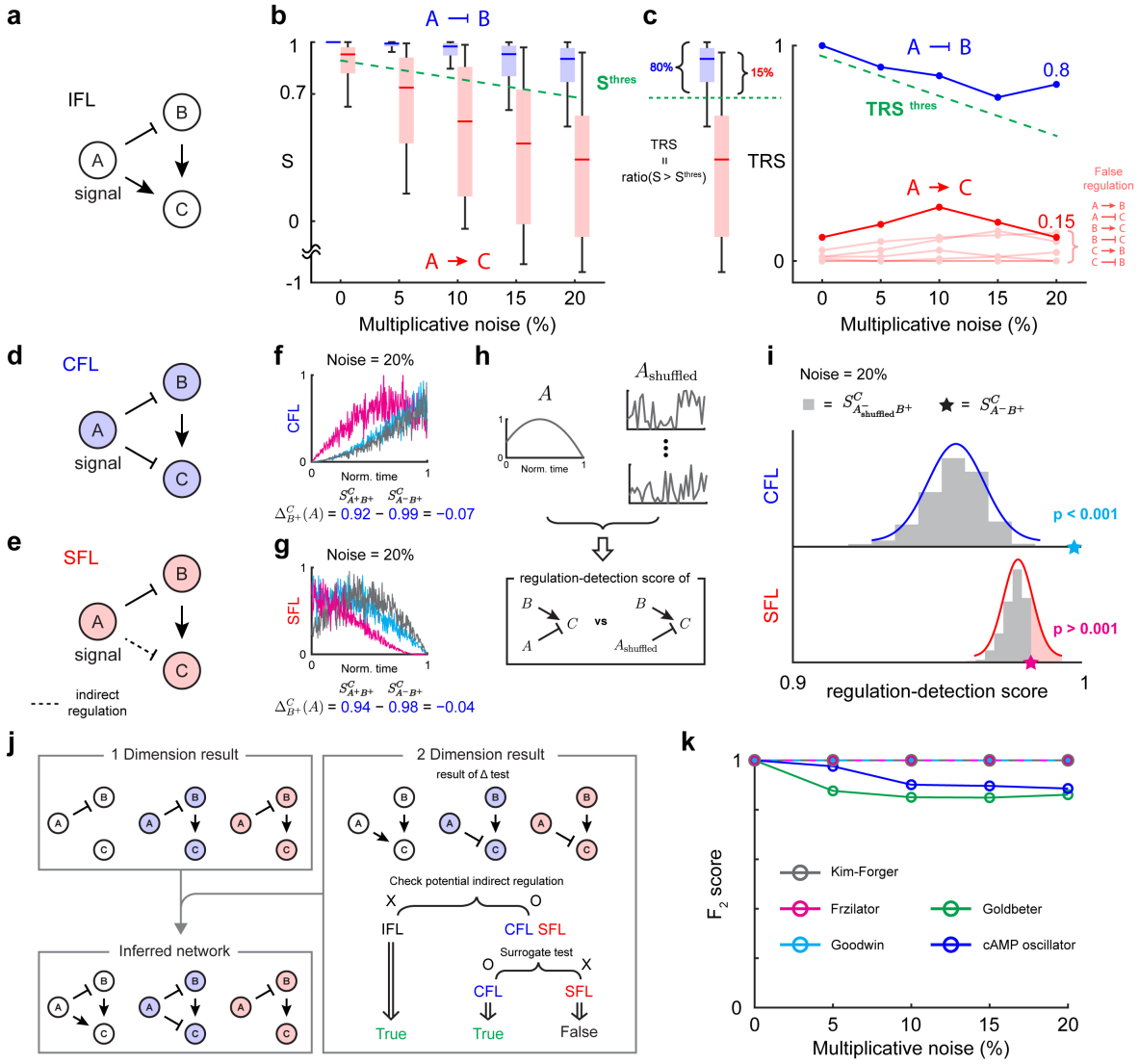

FIG. 3. Extended framework for inferring regulatory network from noisy data. **a** A regulatory network with 1D regulation from A to B and 2D regulation from A and B to C. **b** The threshold for regulation-detection score ($S^{\text{thres}} = 0.9 - 0.005 \times$(noise level), green dashed line) distinguishes true ($A \dashv B$) and false regulation ($A \to C$). **c** The fraction of data satisfying $S^Y_{\boldsymbol{X}^\sigma} > S^{\text{thres}}$, total regulation score ($\text{TRS}^Y_{\boldsymbol{X}^\sigma}$), is used to infer the network. Specifically, $\text{TRS}^Y_{\boldsymbol{X}^\sigma} > \text{TRS}^{\text{thres}}$ is used where $\text{TRS}^{\text{thres}} = 0.9 - 0.01 \times$ (noise level) (green dashed line). **d** In CFL, direct negative regulation exists from A to C. **e** On the other hand, in SFL, the regulatory chain $A \dashv B \to C$ induces an indirect negative regulation from A to C. **f, g** $\Delta^C_{B+}(A)$ cannot distinguish between the direct and indirect regulations in the presence of noise because $S^C_{A-B+} > S^C_{A+B+}$ for both CFL and SFL, indicating the presence of regulation $\overset{A\dashv}{B\to}C$. **h-i** $S^C_{A^-_{\text{shuffled}}B+}$ with the surrogate time series of A can be used to distinguish between the indirect and direct regulations. To disrupt the information of A, the time series of A is shuffled (**h**). In the presence of direct regulation (CFL), but not indirect regulation (SFL), $S^C_{A^-_{\text{shuffled}}B+}$ is significantly smaller than the original S^C_{A-B+} (p-value < 0.001). **j** By including the surrogate test, our extended framework can successfully infer IFL, CFL and SFL even from noisy time series. **k** F_2 score of our inference method when the level of noise increases from 0 to 20%. Here, the mean of the F_2 score for 10 data-sets is calculated. Each data-set consists of 100 time series which are simulated with different initial conditions.

In the presence of noise, the regulation-delta function often fails to distinguish these direct and indirect regulations from A to C in CFL and SFL. Specifically, for both of CFL and SFL with 20% multiplicative noise, S^C_{A-B+} is larger than S^{thres} and $\Delta^C_{B+}(A)$ is strictly negative (Fig. 3f and g) for the most of cases. Here, the sign of Δ is quantified by using a one-tailed Wilcoxon signed rank test (Supplementary Fig. 4a). Thus, the regulation $\overset{A\dashv}{B\to}C$ is inferred from not only CFL but also SFL. This indicates that in the presence of noise, the regulation-delta function can be skewed to the specific type of regulation even for indirect regulation. To prevent such false positive prediction, we developed another criteria. Specifically, we

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.29.518354; this version posted November 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

use a surrogate time series A (A_{shuffled}, Fig. 3h) to destroy the dependence of C on A in the presence of direct regulation ($A \dashv C$). As a result, the regulation-detection score $S^C_{A^-_{\text{shuffled}}B^+}$ is significantly reduced compared to $S^C_{A^-B^+}$ (Fig. 3i top). On the other hand, if A does not directly regulate C, then regulation-detection score $S^C_{A^-_{\text{shuffled}}B^+}$ does not decrease much (Fig. 3i bottom), and $S^C_{A^-B^+}$ is not significantly larger than $S^C_{A^-_{\text{shuffled}}B^+}$. When multiple time series are given, we calculate the p-values for each data and integrate them using Fisher's method. The criteria (the combined p-value $<$ combining $p = 0.001$ for every data) successfully distinguishes between direct and indirect regulation even when the noise varies (Supplementary Fig. 4b).

From the noisy time series, using the criteria $\text{TRS}^Y_{\boldsymbol{X}^\sigma} > \text{TRS}^{\text{thres}}$, all potential 1D (Fig. 3h upper-left) and 2D (Fig. 3h upper-right) regulations are inferred. Then, among the inferred regulations, we need to identify indirect regulations. Unlike IFL, CFL and SFL have a potential indirect regulation. That is, $A \dashv C$ has the potential to be indirect since there is a regulatory chain $A \dashv B \rightarrow C$. In this case, we use a surrogate time series of a potential source of indirect regulation (A) to test whether $S^C_{A^-B^+}$ is significantly larger than $S^C_{A^-_{\text{shuffled}}B^+}$. This reveals that $A \dashv C$ is direct regulation for CFL, but not SFL. Then, merging 1D and 2D results successfully recovers the network structure of IFL, CFL, and SFL even from noisy time series.

Based on TRS and post-filtering tests (Δ test, surrogate test), we develop a user-friendly computational package, General ODE-Based Inference (GOBI), which can be used to infer regulations for systems described by Eq. (1). GOBI successfully infers regulatory networks from simulated time series using ODE models (Fig. 2b-g) in the presence of noise. Here, the F_2 score, the weighted harmonic mean of precision and recall, is nearly one, indicating the nearly perfect recovery of all regulations (Fig. 3k).

D. Successful network inferences from experimentally measured time series

We use GOBI to infer regulatory networks from experimentally measured time series. From the population data of two unicellular ciliates *Paramecium aurelia* (P) and *Didinium nasutum* (D) [3, 40] (Fig. 4a left), the network between the prey (P) and predator (D) is successfully inferred (Fig. 4a and Supplementary Fig. 6a).

Next, we apply GOBI to the time series of the synthetic genetic oscillator, which consists of *Tetracycline repressor* ($TetR$) and RNA polymerase sigma factor (σ^{28}) [41] (Fig. 4b left). While the time series are measured under different conditions after adding purified $TetR$ or inactivating intrinsic $TetR$, our method consistently infers the negative feedback loop based on two direct regulations $\sigma^{28} \rightarrow TetR$ and $TetR \dashv \sigma^{28}$ for all cases (Fig. 4b middle and Supplementary Fig. 6b). This indicates that our method can infer regulations even when the data are achieved from different conditions since we do not specify the specific equations with parameters in Eq. (1). Here, since depletion of a component typically increases as its own concentration increases, self-regulation is assumed to be negative (Fig. 4b right, dashed arrow).

We next investigate the time-series data from a slightly more complex synthetic oscillator, the three-gene repressilator [42] (Fig. 4c left). Assuming negative self-regulation, the criteria $\text{TRS}^Y_{\boldsymbol{X}^\sigma Y^-} > \text{TRS}^{\text{thres}}$ infers three negative 1D regulations and three 2D regulations (Fig. 4c middle). Among the 2D regulations, positive regulations are inferred as indirect as they do not pass the surrogate test (Fig. 4c middle, dashed arrow). Thus, among the inferred 2D regulations, only the negative regulations, consistent with the inferred 1D regulations, are inferred as direct regulations. Gathering these results, GOBI successfully infers the network structure of the repressilator (Fig. 4c right and Supplementary Fig. 6c). Note that although our method infers the regulations among proteins as direct, in fact, $mRNA$ exists as an intermediate step between the negative regulations among the proteins. This happens due to the short translation time in *E.coli* [44] and thus the similar shape and phase of the $mRNA$ and protein profiles. This indicates that our method infers indirect regulations with a short intermediate step as direct regulations.

From the time series measuring the amount of four cofactors present at the estrogen-sensitive $pS2$ promoter after treatment with estradiol [43, 45](Fig. 4d left), four 1D regulations ($HDAC \dashv hER$, $TRIP1 \dashv hER$, $HDAC \dashv POLII$ and $hER \rightarrow POLII$) satisfy the criteria $\text{TRS}^Y_{\boldsymbol{X}^\sigma Y^-} > \text{TRS}^{\text{thres}}$. However, we exclude them because hER and $POLII$ have two causes, forming 2D regulations, although the 1D criteria assumes a single cause (Fig. 4d middle, dashed box). If both regulations are effective, they will be identified as 2D regulations. Indeed, among the 10 candidates for 2D regulations, most of them include the four inferred 1D regulations. Via Δ test and surrogate test, two 1D regulations ($hER \rightarrow POLII$ and $HDAC \dashv hER$) and one 2D regulation ($^{POLII\rightarrow}_{TRIP1\rightarrow}HDAC$) are inferred (Supplementary Fig. 6d). While we are not able to further infer 3D regulations due to the limited amount of data, the inferred regulations are supported by the experiments. That is, estradiol triggers the binding of hER to the $pS2$ promoter to recruit $PolII$ [43], supporting $hER \rightarrow POLII$. Also, inhibition of $POLII$ phosphorylation blocks the recruitment of $HDAC$ but does not affect the $APIS$ engagement at the $pS2$ promoter [43], supporting $POLII \rightarrow HDAC$ and no regulation from $POLII$ to $TRIP1$, which is a surrogate measure of $APIS$. Without inhibition of $POLII$, $HDAC$ is recruited after the $APIS$ engagement, and when the $HDAC$ has maximum occupation, then the $pS2$ promoter becomes refractory to hER [43], supporting $TRIP1 \rightarrow HDAC \dashv hER$. Interestingly, the in-

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.29.518354; this version posted November 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

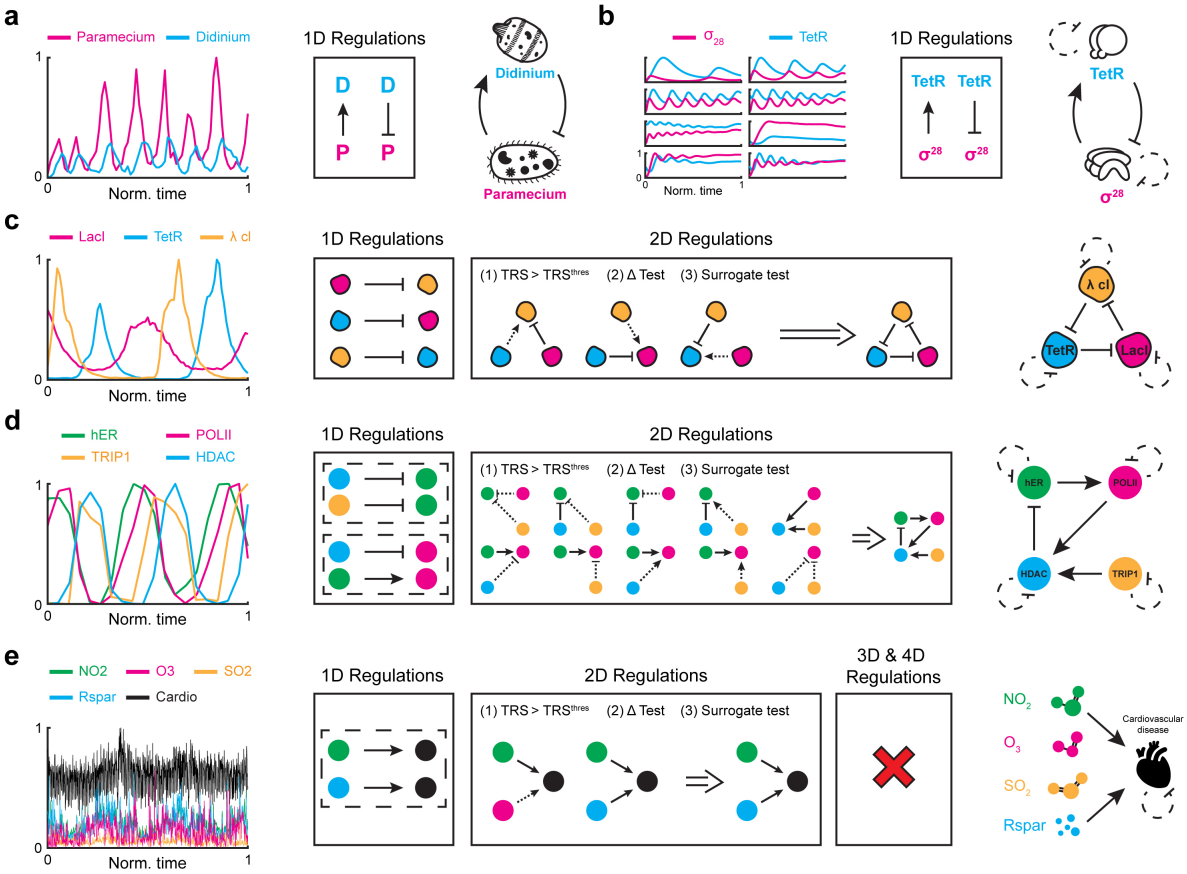

FIG. 4. Inferring regulatory networks from experimental data. **a** GOBI successfully infers predatory interaction from 30-day abundance time-series data of two unicellular ciliates *Paramecium aurelia* and *Didinium nasutum* (data taken from [3, 40]). **b** GOBI successfully infers the negative feedback loop of the synthetic genetic oscillator with a repressor $TetR$ and activator σ^{28} (data is taken from [41]). **c** From time-series data of a three-gene repressilator (data taken from [42]), GOBI successfully infers the underlying network. Three direct negative 1D regulations are inferred. Among the three 2D regulations having high TRS, only negative regulations pass the Δ test and surrogate test. **d** From time series measuring the amount of cofactors present at the estrogen-sensitive $pS2$ promoter after treatment with estradiol (data taken from [43]), four 1D regulations have high TRS. However, they are not inferred because they share a common target (dashed box). Among 10 regulations having high TRS, one 2D regulation and two 1D regulations are inferred, passing the Δ test and surrogate test. **e** From 1000-day time-series data of daily air pollutants and cardiovascular disease occurrence in the city of Hong Kong (data taken from [20]), GOBI finds direct positive causal links from NO_2 and $Rspar$ to the disease.

ferred network contains a negative feedback loop which is required to generate sustained oscillations [46].

Finally, we investigate five time series of air pollutants and cardiovascular disease occurrence in Hong Kong from 1994 to 1997 [47] (Fig. 4e left). Since our goal is to identify which pollutants cause cardiovascular disease, we fix the disease as a target. Also, we assume the negative self-regulation of disease reflecting death. While two positive causal links from NO_2 and respirable suspended particulates ($Rspar$) to the disease are identified as 1D regulation (Fig. 4e middle), we exclude them because they share the same target. (Fig. 4e middle, dashed box). Among two inferred 2D regulations, one of them passes the Δ test and surrogate test (Fig. 4d middle). Furthermore, no 3D and 4D regulation is inferred (Supplementary Fig. 6e). The inferred network indicates that both NO_2 and $Rspar$ are major causes of cardiovascular diseases (Fig. 4e right). Indeed, it was reported that NO_2 and $Rspar$ are associated with hospital admissions and mortality due to cardiovascular disease, respectively [48].

E. Comparison between our framework and other model-free inference methods

Here, we compare our framework with popular model-free methods, i.e., GC, CCM and PCM, by using the experimental time-series data in the previous section (Fig. 4a-e). Unlike our method, the model-free methods can only infer the presence of regulation and not its type (i.e. positive and negative). Thus, the arrows represent inferred regulations, which could be either positive or negative.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.29.518354; this version posted November 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

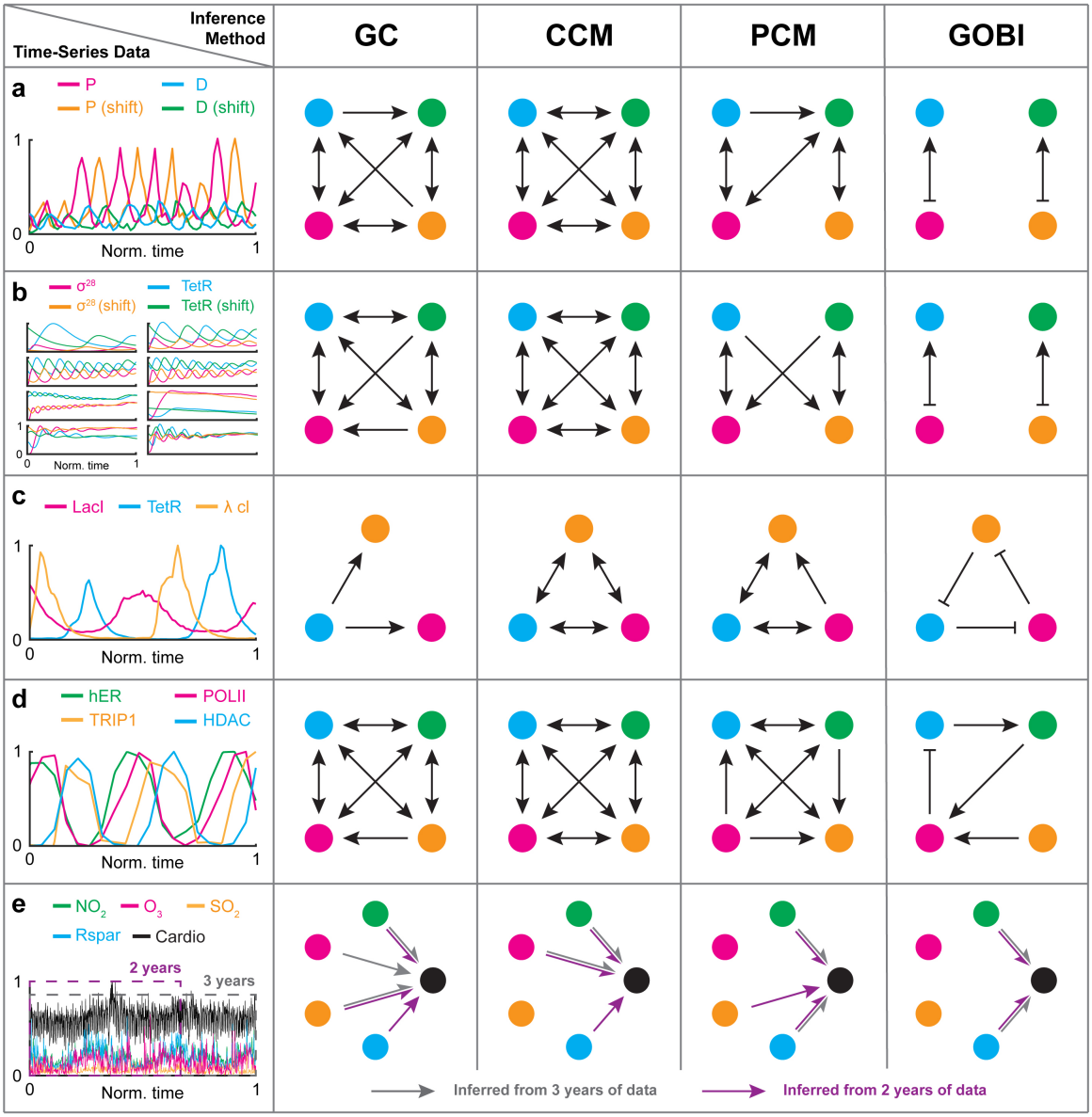

FIG. 5. Model free-methods, but not our method, make a false prediction due to the presence of synchrony and indirect effect. **a-e** We apply our method and popular model-free methods (i.e., GC, CCM, PCM) to various experimental time-series data obtained from the prey-predator system (**a**); genetic oscillator (**b**); repressilator (**c**); cofactors at the $pS2$ promoter (**d**); and air pollutants and cardiovascular disease (**e**). For the prey-predator system and genetic oscillator, each time series is duplicated and the phase is shifted by about half of the period. For the air pollutants and cardiovascular disease data, we test the methods on three years of data (**e** grey) and on two years of data (**e** purple).

For the prey-predator system and genetic oscillator (Fig. 4a,b), we changed them to more challenging cases: each time series is duplicated and shifted about half of its period to increase the number of components. While our method successfully detects two independent negative feedback loops (Fig. 5a,b), model-free methods infer false positive predictions (e.g., P to D^{shift} in Fig. 5a) because they usually misidentify synchrony as causality.

For a similar reason, synchrony obscures the inference of the model-free methods for the repressilator (Fig. 5c). Moreover, the model-free methods fail to distinguish between direct and indirect regulations. For example, they infer the indirect causation $TetR \rightarrow \lambda cI$ induced by the regulatory chain $TetR \dashv LacI \dashv \lambda cI$ unlike our method. Similarly, due to synchrony and indirect effect, for the system of cofactors at the $pS2$ promoter, model-free methods infer an almost fully connected causal network unlike our method (Fig. 5d).

When we use three years of data (full-length data) of air pollutants and cardiovascular disease, PCM infers the same structure as GOBI infers, i.e., only NO_2 and $Rspar$ cause the disease (Fig. 5e grey) [20]. On the other hand,

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.29.518354; this version posted November 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

when only part of the data (i.e. two years of data) is used, only GOBI infers the same structure (Fig. 5e purple). This indicates that GOBI is more reliable and accurate than the model-free methods.

III. DISCUSSION

We develop an inference method that does not suffer from the weakness of model-free and model-based inference methods. We derive the conditions for interactions satisfying the general ODE (Eq. (1)). As this allows us to easily check the reproducibility of given time-series data with the general ODE (i.e., the existence of ODE satisfying given time-series data) without fitting, the computational cost is dramatically reduced compared to the previous model-based approaches. Importantly, as our method can be applied to any system described by general ODE (Eq. (1)), it does not suffer from the fundamental limit of the model-based approach (i.e., requirement of a priori model accurately describing the system). In addition, our method also does not run the serious risk of misidentifying synchrony as causality, unlike the previous model-free approaches. Furthermore, our method successfully distinguishes direct from indirect causal relations by adopting the surrogate test (Fig. 3). In this way, our framework dramatically reduces the false positive predictions which are the inherent flaw of the model-free inference method (Fig. 5). Taken together, we developed an accurate and broadly applicable inference method that can uncover unknown functional relationships underlying the system from their output time-series data (Fig. 4).

In our approach, we assumed that when X causes Y, X causes Y either positively or negatively. Thus, our approach cannot capture the causation when X causes Y both positively and negatively or when the type of causation changes over time. It would be an interesting future work to derive the condition of reproducibility without assuming a fixed causation type (i.e. the monotonicity of f in Eq. (1)). Because our method tests the reproducibility of time-series data using necessary conditions, false positive causations can be predicted. To resolve this, we used multiple time-series data and performed post-filtering tests (i.e., Δ test and surrogate test). Thus, to infer high-dimensional regulations, a large amount of data is required. Lastly, while we considered the general form of ODE, an interesting future direction would be to extend our work to models that describe interactions including time delays.

IV. METHODS

A. Computational package for inferring regulatory network

Here, we describe the key steps of our computational package, GOBI (Github link will be provided upon acceptance). For the experimental time-series data $\boldsymbol{X}(t) = (X_1(t), X_2(t), \cdots, X_N(t))$, $\boldsymbol{X}(t)$ can be interpolated with either the 'spline', or 'fourier' method, chosen by the user. Also, the derivative of $\boldsymbol{X}(t)$ is computed using the MATLAB function 'gradient'.

1. *Regulation-detection region*

For the ND regulation (Eq. (1)) with regulation type σ, the regulation-detection region ($R_{\boldsymbol{X}^\sigma}$) is defined as the set of (t, t^*) on the domain of time series $[0, \tau)^2$ satisfying $\sigma(i)X_i^d(t, t^*) > 0$ for all i. For example, with the positive 1D regulation $X \to Y$ ($\sigma = +$), R_{X^+} is the set of (t, t^*) where $X^d > 0$. For the 2D regulation $\begin{smallmatrix} X_1 \to Y \\ X_2 \dashv \end{smallmatrix}$ ($\sigma = (+, -)$), $R_{X_1^+ X_2^-}$ is the set of (t, t^*) satisfying both $X_1^d > 0$ and $X_2^d < 0$. The size of the regulation-detection region (size($R_{\boldsymbol{X}^\sigma}$)) is the fraction of $R_{\boldsymbol{X}^\sigma}$ over the domain $[0, \tau)^2$. In the presence of noise, we only consider a region which is not small (i.e., size($R_{\boldsymbol{X}^\sigma}$) $> R^{\text{thres}}$) to avoid an error from the noise. The value of R^{thres} can be chosen from 0 to 0.1, and the choice of R^{thres} does not significantly affect the results (Supplementary Fig. 3a). However, a small value of R^{thres} is recommended for inferring high dimensional regulations since the average of size($R_{\boldsymbol{X}^\sigma}$) decreases exponentially as dimension increases (see Supplementary Information for details).

2. *Regulation-detection function and score*

When the regulation type σ from $\boldsymbol{X} = (X_1, X_2, \cdots, X_N)$ to Y exists, the following regulation-detection function ($I^Y_{\boldsymbol{X}^\sigma}$) defined on regulation-detection region R_{X^σ} is always positive.

$$I^Y_{\boldsymbol{X}^\sigma} := \dot{Y}^d \cdot \prod_{i=1}^{N} \sigma(i) X_i^d.$$

Thus, the following regulation-detection score ($S^Y_{\boldsymbol{X}^\sigma}$) is one:

$$S^Y_{\boldsymbol{X}^\sigma} := \frac{\iint_{R_{\boldsymbol{X}^\sigma}} I^Y_{\boldsymbol{X}^\sigma}(t, t^*) dt dt^*}{\iint_{R_{\boldsymbol{X}^\sigma}} |I^Y_{\boldsymbol{X}^\sigma}(t, t^*)| dt dt^*} \tag{4}$$

(see Supplementary Information for details). However, this is not true anymore in the presence of noise. Thus, we relax the criteria from $S^Y_{\boldsymbol{X}^\sigma} = 1$ to $S^Y_{\boldsymbol{X}^\sigma} > S^{\text{thres}}$. Among the data which has nonempty $R_{\boldsymbol{X}^\sigma}$ (i.e., $R_{\boldsymbol{X}^\sigma} > R^{\text{thres}}$), the fraction of data satisfying the criteria $S^Y_{\boldsymbol{X}^\sigma} > S^{\text{thres}}$ is called Total Regulation Score ($\text{TRS}^Y_{\boldsymbol{X}^\sigma}$). Finally, we infer the regulation from noisy time-series data using the criteria $\text{TRS}^Y_{\boldsymbol{X}^\sigma} > \text{TRS}^{\text{thres}}$ for noisy time-series data. $S^{\text{thres}} = 0.9 - 0.005 \times$ (noise level) and $\text{TRS}^{\text{thres}} = 0.9 - 0.01 \times$(noise level) are used (Fig. 3a-c and Supplementary Fig. 3). The noise level of the time series is approximated using the mean square of the residual between the noisy and fitted time series (Supplementary Fig. 5).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.29.518354; this version posted November 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

3. Δ test

When we add any regulation on an existing true regulation, the regulation-detection score is always one (Fig. 1j-l). Thus, to test whether the additional regulation is effective, we consider $\Delta^Y_{\boldsymbol{X}^\sigma}(X_{\text{new}}) = S^Y_{\boldsymbol{X}^\sigma X^+_{\text{new}}} - S^Y_{\boldsymbol{X}^\sigma X^-_{\text{new}}}$, where $S^Y_{\boldsymbol{X}^\sigma X^+_{\text{new}}}$ ($S^Y_{\boldsymbol{X}^\sigma X^-_{\text{new}}}$) is the regulation-detection score when the new component (X_{new}) is positively (negatively) added to the existing regulation type σ. Because $\Delta^Y_{\boldsymbol{X}^\sigma}(X_{\text{new}}) = 0$ reflects that the new component (X_{new}) does not have any regulatory role, the newly added regulation is inferred only when $\Delta^Y_{\boldsymbol{X}^\sigma}(X_{\text{new}}) \neq 0$ for some data. In particular, $\Delta > 0$ ($\Delta < 0$) represents that the new component adds positive (negative) regulation. In the presence of noise, the positive (negative) regulation is inferred if $\Delta \geq 0$ ($\Delta \leq 0$) consistently for all time series. If the number of time series is greater than 25, the sign of Δ is quantified by a one-tailed Wilcoxon signed rank test. We set the critical value of significance as 0.01, but it can be chosen by the user.

4. Surrogate test

Indirect regulation is induced by the chain of direct regulations. For example, in SFL (Fig. 3e), regulatory chain $A \dashv B \rightarrow C$ induces the indirect negative regulation $A \dashv C$. In the presence of noise, the Δ test sometimes fails to distinguish between direct and indirect regulations (Fig. 3d-g). Thus, after the Δ test, if the inferred regulation has the potential to be indirect, we additionally perform the surrogate test to determine whether the inferred regulation is direct or indirect. Specifically, for each candidate of indirect regulation, we shuffle the time series of cause using the MATLAB function 'perm' and then calculate the regulation-detection scores. Then, we test whether the original regulation-detection score is significantly larger than the shuffled ones by using a one-tailed Z test. In the presence of the k number of time-series data, we can get the k number of p-values (p_i, $i = 1, 2, \cdots, k$). Thus, we combined them into one test statistic (χ^2) using Fisher's method, $\chi^2_{2k} \sim -2\sum_{i=1}^{k} \log(p_i)$. We set the critical value of the significance of Fisher's method by combining $p_i = 0.001$ for all the data, but it can also be chosen by the user.

5. Model-free methods

For GC, we rejected the null hypothesis that Y does not Granger cause X, and thereby inferred direct regulations by using the F statistic with a significance level of 95% [2]. For Convergent cross mapping (CCM) [3] and Partial cross mapping (PCM) [20], we choose an appropriate embedding dimension using the false nearest neighbor algorithm. Also, we select a time lag producing the first minimum of delayed mutual information. Specifically, we used embedding dimension 2 for the prey-predator, genetic oscillator and estradiol data-sets; and 3 for the repressilator and air pollutants and cardiovascular disease data-sets. Also, we used time lag 2 for prey-predator; 3 $\sim$ 10 for genetic oscillator (there are eight different time-series data-sets); 10 for therepressilator; 15 for the estradiol data-set; and 3 for the air pollutants and cardiovascular disease data-set.

B. *in silico* time-series data

With the ODE describing the system, we simulate the time-series data using the MATLAB function 'ode45'. The sampling rate is 100 points per period for all the examples (Fig. 1, 2, 3). For the multiple time-series data (Fig. 2, 3), we generate 100 different time series with different initial conditions. Then, before applying our method, we normalize each time series by re-scaling to have minimum 0 and maximum 1. For noisy time series, we add multiplicative noise sampled randomly from a normal distribution with mean 0 and standard deviation given by the noise level. For example, for 10% multiplicative noise, we add the noise $X(t_i) \cdot \epsilon$ to $X(t_i)$, where $\epsilon \sim N(0, 0.1^2)$. Before applying our method, all the simulated noisy time series are fitted using the MATLAB function 'fourier4'. However, if the noise level is too high, 'fourier4' tends to overfit and capture the noise. Thus, in the presence of a high level of noise, 'fourier2' is recommended for smoothing.

C. Experimental time-series data

For the experimental data, we first calculate the period of data by using the first peak of auto-correlation. Then, we cut the time series into periods (Fig. 4a,b). Specifically, we cut the prey-predator time series every five days to generate seven different time series (Fig. 4a). When the number of cycles in the data is low (<5), to generate enough multiple time series (Fig. 4c-e), we cut the data using the moving-window technique. That is, we choose the window whose size is the period of the time series. Then, along the time series, we move the window until the next window overlaps with the current window by 90%. Then, the time series in every window is used for our approach. We did this for the repressilator (Fig. 4c); estradiol data-set (Fig. 4d); and air pollution and cardiovascular disease data (Fig. 4e). For instance, we used time-series data of air pollutants and cardiovascular disease with a window size of one year and an overlap of 11 months (i.e., move the window for a month) to generate 23 data-sets. Before this, the time series of disease admissions are smoothed using a simple moving average with a window width of seven days to avoid the effect of days of the week. Each time series is interpolated using the MATLAB function 'spline' (Fig. 4a-d) or 'fourier2' (Fig. 4e) depending on the noise level of the time-series

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.29.518354; this version posted November 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

data.

REFERENCES

[1] Saint-Antoine, M. M. & Singh, A. Network inference in systems biology: recent developments, challenges, and applications. *Current opinion in biotechnology* **63**, 89–98 (2020).

[2] Granger, C. W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society* 424–438 (1969).

[3] Sugihara, G. *et al.* Detecting causality in complex ecosystems. *science* **338**, 496–500 (2012).

[4] Pourzanjani, A., Herzog, E. D. & Petzold, L. R. On the inference of functional circadian networks using granger causality. *PLoS One* **10**, e0137540 (2015).

[5] Runge, J. *et al.* Inferring causation from time series in earth system sciences. *Nature communications* **10**, 1–13 (2019).

[6] Kamiński, M., Ding, M., Truccolo, W. A. & Bressler, S. L. Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological cybernetics* **85**, 145–157 (2001).

[7] Deyle, E. R., Maher, M. C., Hernandez, R. D., Basu, S. & Sugihara, G. Global environmental drivers of influenza. *Proceedings of the National Academy of Sciences* **113**, 13081–13086 (2016).

[8] Ma, H. *et al.* Detection of time delays and directional interactions based on time series from complex dynamical systems. *Physical Review E* **96**, 012221 (2017).

[9] Tsonis, A. A. *et al.* Dynamical evidence for causality between galactic cosmic rays and interannual variation in global temperature. *Proceedings of the National Academy of Sciences* **112**, 3253–3256 (2015).

[10] Ye, H., Deyle, E. R., Gilarranz, L. J. & Sugihara, G. Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific reports* **5**, 1–9 (2015).

[11] Stokes, P. A. & Purdon, P. L. A study of problems encountered in granger causality analysis from a neuroscience perspective. *Proceedings of the national academy of sciences* **114**, E7063–E7072 (2017).

[12] Tyler, J., Forger, D. & Kim, J. K. Inferring causality in biological oscillators. *Bioinformatics* **38**, 196–203 (2022).

[13] Nawrath, J. *et al.* Distinguishing direct from indirect interactions in oscillatory networks with multiple time scales. *Physical review letters* **104**, 038701 (2010).

[14] Schelter, B. *et al.* Direct or indirect? graphical models for neural oscillators. *Journal of Physiology-Paris* **99**, 37–46 (2006).

[15] Cobey, S. & Baskerville, E. B. Limits to causal inference with state-space reconstruction for infectious disease. *PloS one* **11**, e0169050 (2016).

[16] Guo, S., Seth, A. K., Kendrick, K. M., Zhou, C. & Feng, J. Partial granger causality—eliminating exogenous inputs and latent variables. *Journal of neuroscience methods* **172**, 79–93 (2008).

[17] Frenzel, S. & Pompe, B. Partial mutual information for coupling analysis of multivariate time series. *Physical review letters* **99**, 204101 (2007).

[18] Zhao, J., Zhou, Y., Zhang, X. & Chen, L. Part mutual information for quantifying direct associations in networks. *Proceedings of the National Academy of Sciences* **113**, 5130–5135 (2016).

[19] Runge, J., Petoukhov, V. & Kurths, J. Quantifying the strength and delay of climatic interactions: The ambiguities of cross correlation and a novel measure based on graphical models. *Journal of climate* **27**, 720–739 (2014).

[20] Leng, S. *et al.* Partial cross mapping eliminates indirect causal influences. *Nature communications* **11**, 1–9 (2020).

[21] Gotoh, T. *et al.* Model-driven experimental approach reveals the complex regulatory distribution of p53 by the circadian factor period 2. *Proceedings of the National Academy of Sciences* **113**, 13516–13521 (2016).

[22] Lillacci, G. & Khammash, M. Parameter estimation and model selection in computational biology. *PLoS computational biology* **6**, e1000696 (2010).

[23] McBride, D. & Petzold, L. Model-based inference of a directed network of circadian neurons. *Journal of Biological Rhythms* **33**, 515–522 (2018).

[24] Pitt, J. A. & Banga, J. R. Parameter estimation in models of biological oscillators: an automated regularised estimation approach. *BMC bioinformatics* **20**, 1–17 (2019).

[25] Radde, N. & Kaderali, L. Inference of an oscillating model for the yeast cell cycle. *Discrete Applied Mathematics* **157**, 2285–2295 (2009).

[26] Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M. P. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface* **6**, 187–202 (2009).

[27] Trejo Banos, D., Millar, A. J. & Sanguinetti, G. A bayesian approach for structure learning in oscillating regulatory networks. *Bioinformatics* **31**, 3617–3624 (2015).

[28] Brunton, S. L., Proctor, J. L. & Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences* **113**, 3932–3937 (2016).

[29] Kim, J. K. & Forger, D. B. On the existence and uniqueness of biological clock models matching experimental data. *SIAM Journal on Applied Mathematics* **72**, 1842–1855 (2012).

[30] Konopka, T. & Rooman, M. Gene expression model (in) validation by fourier analysis. *BMC systems biology* **4**, 1–12 (2010).

[31] Mangan, N. M., Brunton, S. L., Proctor, J. L. & Kutz, J. N. Inferring biological networks by sparse identification of nonlinear dynamics. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications* **2**, 52–63 (2016).

[32] McGoff, K. A. *et al.* The local edge machine: inference of dynamic models of gene regulation. *Genome biology* **17**, 1–13 (2016).

[33] Pigolotti, S., Krishna, S. & Jensen, M. H. Oscillation patterns in negative feedback loops. *Proceedings of the National Academy of Sciences* **104**, 6533–6537 (2007).

[34] Pigolotti, S., Krishna, S. & Jensen, M. H. Symbolic dynamics of biological feedback networks. *Physical review letters* **102**, 088701 (2009).

[35] Kim, J. K. & Forger, D. B. A mechanism for robust circadian timekeeping via stoichiometric balance. *Molecular systems biology* **8**, 630 (2012).

[36] Igoshin, O. A., Goldbeter, A., Kaiser, D. & Oster, G. A biochemical oscillator explains several aspects of myxo-

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.29.518354; this version posted November 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

coccus xanthus behavior during development. *Proceedings of the National Academy of Sciences* **101**, 15760–15765 (2004).

[37] Goodwin, B. C. Oscillatory behavior in enzymatic control processes. *Advances in enzyme regulation* **3**, 425–437 (1965).

[38] Goldbeter, A. A model for circadian oscillations in the drosophila period protein (per). *Proceedings of the Royal Society of London. Series B: Biological Sciences* **261**, 319–324 (1995).

[39] Maeda, M. *et al.* Periodic signaling controlled by an oscillatory circuit that includes protein kinases erk2 and pka. *Science* **304**, 875–878 (2004).

[40] Veilleux, B. G. The analysis of a predatory interaction between didinium and paramecium (1976).

[41] Aufinger, L., Brenner, J. & Simmel, F. C. Complex dynamics in a synchronized cell-free genetic clock. *Nature communications* **13**, 1–9 (2022).

[42] Potvin-Trottier, L., Lord, N. D., Vinnicombe, G. & Paulsson, J. Synchronous long-term oscillations in a synthetic gene circuit. *Nature* **538**, 514–517 (2016).

[43] Métivier, R. *et al.* Estrogen receptor-α directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter. *Cell* **115**, 751–763 (2003).

[44] Choi, B. *et al.* Bayesian inference of distributed time delay in transcriptional and translational regulation. *Bioinformatics* **36**, 586–593 (2020).

[45] Lemaire, V., Lee, C. F., Lei, J., Métivier, R. & Glass, L. Sequential recruitment and combinatorial assembling of multiprotein complexes in transcriptional activation. *Physical review letters* **96**, 198102 (2006).

[46] Novák, B. & Tyson, J. J. Design principles of biochemical oscillators. *Nature reviews Molecular cell biology* **9**, 981–991 (2008).

[47] Wong, T. W. *et al.* Air pollution and hospital admissions for respiratory and cardiovascular diseases in hong kong. *Occupational and environmental medicine* **56**, 679–683 (1999).

[48] Milojevic, A. *et al.* Short-term effects of air pollution on a range of cardiovascular events in england and wales: case-crossover analysis of the minap database, hospital admissions and mortality. *Heart* **100**, 1093–1098 (2014).

ACKNOWLEDGMENTS

We thank Seokjoo Chae, Hyukpyo Hong and Yun Min Song for valuable comments. This work was supported by Samsung Science and Technology Foundation SSTF-BA1902-01 (to J.K.K.) and Institute for Basic Science IBS-R029-C3 (to J.K.K.).

AUTHOR CONTRIBUTIONS

S.H.P., S.H. and J.K.K. designed the research. S.H.P. and S.H. developed the method. S.H.P. performed computation. S.H.P. analyzed data. J.K.K. supervised the project. All authors wrote the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.