Title: System Identification Using Compressed Sensing Reveals Signaling-Decoding System by Gene Expression

Authors:

Takaho Tsuchiya[1], Masashi Fujii[1,2], Naoki Matsuda[1], Katsuyuki Kunida[1], Shinsuke Uda[3],

Hiroyuki Kubota[3], Katsumi Konishi[4] and Shinya Kuroda[1,5,*]

Affiliations:

[1] Department of Biological Sciences, Graduate School of Science, University of Tokyo,

Bunkyo-ku, Tokyo 113-0033, Japan

[2] Molecular Genetics Research Laboratory, Graduate School of Science, University of Tokyo,

Bunkyo-ku, Tokyo 113-0033, Japan

[3] Division of Integrated Omics, Research Center for Transomics Medicine, Medical Institute of

Bioregulation, Kyushu University, 3-1-1 Maidashi, Higashi-ku, Fukuoka, Fukuoka 812-8582,

Japan

[4] Department of Computer Science, Faculty of Informatics, Kogakuin University, 1-24-2

Nishi-shinjuku, Shinjuku-ku, Tokyo 163-8677, Japan

[5] CREST, Japan Science and Technology Corporation, Bunkyo-ku, Tokyo 113-0033, Japan

Corresponding Author:

*Correspondence to: skuroda@bs.s.u-tokyo.ac.jp

1

2 **Lead Contact:**

3 Shinya Kuroda, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033, Japan

4 Email: skuroda@bs.s.u-tokyo.ac.jp

5 Tel.: +81-3-5841-4697

6 Fax: +81-3-5841-4698

7

1   **SUMMARY**

2   Cells decode information of signaling activation at a scale of tens of minutes by downstream

3   gene expression with a scale of hours to days, leading to cell fate decisions such as cell

4   differentiation. However, no system identification method with such different time scales exists.

5   Here we used compressed sensing technology and developed a system identification method

6   using data of different time scales by recovering signals of missing time points. We measured

7   phosphorylation of ERK and CREB, immediate early gene expression products, and mRNAs of

8   decoder genes for neurite elongation in PC12 cell differentiation and performed system

9   identification, revealing the input–output relationships between signaling and gene expression

10   with sensitivity such as graded or switch-like response and with time delay and gain,

11   representing signal transfer efficiency. We predicted and validated the identified system using

12   pharmacological perturbation. Thus, we provide a versatile method for system identification

13   using data with different time scales.

14

15   **Keywords:**

16   cell differentiation, ERK, gene expression, system identification, compressed sensing, signal

17   recovery, different time-scale data

18

19

1   **Highlights:**

2   • We developed a system identification method using compressed sensing.

3   • This method allowed us to find a pathway using data of different time scales.

4   • We identified a selective signaling-decoding system by gene expression.

5   • We validated the identified system by pharmacological perturbation.

6

7   **eTOC Blurb:**

8   We describe a system identification method of molecular networks with different time-scale data

9   using a signal recovery technique in compressed sensing.

10

4

INTRODUCTION

In intracellular signaling systems, information of an extracellular stimulus is encoded into combinations of distinct temporal patterns of phosphorylation of intracellular signaling molecules that are selectively decoded by downstream gene expression, leading to cell fate decisions such as cell differentiation, proliferation, and death (Behar and Hoffmann, 2010; Purvis and Lahav, 2013). For instance, in rat adrenal pheochromocytoma PC12 cells, nerve growth factor (NGF) induces cell differentiation mainly through sustained phosphorylation of ERK (Gotoh et al., 1990; Marshall, 1995; Qiu and Green, 1992; Traverse et al., 1992), whereas pituitary adenylate cyclase-activating polypeptide (PACAP) induces cell differentiation mainly through cAMP-dependent CREB phosphorylation (Akimoto et al., 2013; Gerdin and Eiden, 2007; Saito et al., 2013; Vaudry et al., 2002; Watanabe et al., 2012). We showed that cell differentiation in PC12 cells can be divided into two processes: a latent processes (0–12 h after the stimulation) in preparation for neurite extension and a subsequent neurite extension process (12–24 h) (Chung et al., 2010). We identified the three genes essential for cell differentiation, *Metrnl*, *Dclk1*, and *Serpinb1a*, which are induced during the latent process and required for subsequent neurite extension, and named LP (latent process) genes (Watanabe et al., 2012). Although NGF and PACAP selectively induce the different combinations and temporal patterns of signaling molecules, both growth factors commonly induce the LP genes (Watanabe et al., 2012). The expression levels of LP genes, but not the phosphorylation level of ERK, correlate with neurite length regardless of growth factors (Watanabe et al., 2012), indicating that the LP genes are the decoders of neurite length. Thus, how the distinct patterns of signaling molecules

5

1   are decoded by LP gene expression is critical for understanding the unknown mechanism

2   underlying cell differentiation in PC12 cells. Decoding the combinations and temporal patterns

3   of signaling molecules by downstream gene expression is a general mechanism underlying

4   various cellular functions (Behar and Hoffmann, 2010; Purvis and Lahav, 2013; Sumit et al.,

5   2017).

6   Mathematical modeling is useful for the analysis of decoding mechanisms (Janes and

7   Lauffenburger, 2013). If the signaling pathways are well characterized, kinetic modeling based

8   on biochemical reactions reported in the literature is often used (Janes and Lauffenburger, 2006;

9   Kholodenko et al., 2012; Price and Shmulevich, 2007). For example, growth factor–dependent

10   ERK activation in PC12 cells has been modelled by the kinetic model based on prior knowledge

11   of pathway information (Brightman and Fell, 2000; Filippi et al., 2016; Nakakuki et al., 2010;

12   Ryu et al., 2015; Santos et al., 2007; Sasagawa et al., 2005). In general, however, decoding by

13   downstream genes involves more complex processes such as transcription and translation and

14   information on the precise pathway is not available.

15   To identify decoding mechanisms by gene expression, the system identification method (also

16   referred to as data-driven modeling) was developed for identifying quantitative input–output

17   relationships from time series data without detailed knowledge of signaling pathways (Janes and

18   Lauffenburger, 2006; Janes and Yaffe, 2006; Kholodenko et al., 2012; Ljung, 2010; Price and

19   Shmulevich, 2007; Zechner et al., 2016). We previously developed a system identification

20   method based on time series data of signaling molecules and gene expression, denoted as the

21   nonlinear autoregressive exogenous (NARX) model, and applied it to the signaling-dependent

1   immediate early gene (IEG) expression during cell differentiation in PC12 cells (Saito et al.,

2   2013). The NARX model involves the determination of lag-order numbers and use of the Hill

3   equation and the linear autoregressive exogenous (ARX) model (Saito et al., 2013).

4   Determination of lag-order numbers infers the selection of input molecules (*Input*) for an output

5   molecule (*Output)*, which is referred to as the *Input-Output* (*I-O*). The Hill equation characterizes

6   sensitivity with a nonlinear dose-response curve (Hill, 1910). The linear ARX model

7   characterizes temporal changes with time constant and gain, the latter of which is an *I-O*

8   amplitude ratio, and indicates signal transfer efficiency (Ljung, 1998). The NARX model

9   requires equally spaced dense time series data. If the time scale between upstream and

10   downstream are similar, such as signaling molecules (scale of tens of minutes) and IEG

11   expression (a few hours) in PC12 cells, it is not difficult to acquire a sufficient number of equally

12   spaced dense time series data (Saito et al., 2013). However, if the time scale of upstream and

13   downstream molecules is largely different, such as signaling molecules (tens of minutes) and LP

14   gene expression (a day) (Doupé and Perrimon, 2014), it is technically difficult to obtain

15   sufficient equally spaced dense time series data because of experimental and budget limitations.

16   Measuring gene expression often requires a longer time scale than measuring protein

17   phosphorylation. Obtaining equally spaced dense time series data with a longer time scale is

18   labor and cost intensive, because, unlike live-cell imaging experiments, snapshot experiments

19   such as western blotting, RT-PCR, and quantitative image cytometry (QIC) (Ozaki et al., 2010)

20   require the same number of experiments as the number of time points. In addition, experimental

21   noise and variation increases as the number of experiments increases because differences in

1 experimental conditions such as plates, gels, reagents, and cell culture conditions increase as well.

2 Therefore, in reality, for a longer time scale experiment, unequally spaced sparse time series data

3 rather than equally spaced dense time series data are desired. For example, under conditions in

4 which stimulation by cell growth factors triggers rapid and transient phosphorylation and slow

5 and sustained gene expression, time series data should be obtained with dense time points during

6 the transient phase and eventually with sparse time points. The timing and dynamic

7 characteristics of temporal changes may differ between upstream and downstream molecules,

8 such that time points and intervals for measuring upstream and downstream molecules may be

9 different. Thus, a system identification method using unequally spaced sparse time series data

10 with different time scale needs to be developed.

11 To solve this problem, here we used the signal recovery technique based on a low-rank

12 approach proposed in the field of compressed sensing to generate a sufficient number of time

13 points for equally spaced dense time series data from unequally spaced sparse time series data

14 with different time points and intervals. We applied this nonlinear system identification method

15 to the signaling-dependent gene expression underlying cell differentiation in PC12 cells and

16 identified the signaling-decoding system by gene expression.

17 Unequally spaced sparse time series data can be regarded as equally spaced dense time series

18 data with missing time points, and therefore we can generate equally spaced dense time series

19 data by applying a signal recovery technique, which has been studied in the field of compressed

20 sensing (Candès and Wakin, 2008; Donoho, 2006). Compressed sensing is a signal processing

21 method for efficient data acquisition by recovering missing signals/images from a small number

1   of randomly sampled signals including unequally spaced sparse data based on sparseness of a

2   vector (Candès et al., 2008) or low rankness of a matrix (Fazel, 2002). Both the sparse approach

3   and the low-rank approach have been applied to various fields, such as sampling and

4   reconstructing magnetic resonance images (Lustig et al., 2008; Ongie and Jacob, 2016),

5   super-resolution imaging (Candès and Fernandez-Granda, 2014; Yang et al., 2010), image

6   inpainting (Takahashi et al., 2012; Takahashi et al., 2016), and collaborative filtering (Candès

7   and Recht, 2009). In this study, we applied a matrix rank minimization algorithm (Konishi et al.,

8   2014) to recover missing time points from unequally spaced time series data, and we generated

9   equally spaced time series data with the same time points from signaling and gene expression

10   data with different time scales. We previously developed a system idenfitication method from

11   equally spaced dense time series data of signaling and gene expression using the NARX model

12   (Saito, 2013). We developed a new system identification method from unequally spaced sparse

13   time series data with different time scales by integrating this signal recovery method using the

14   matrix rank minimization algorithm (Konishi et al., 2014) and the NARX model (Saito et al.,

15   2013). We applied the method to system identification of signaling-dependent gene expression in

16   cell differentiationin PC12 cells, revealing a selective signaling-decoding mechansim by gene

17   expression.

18

19   **RESULTS**

20   **Signal Recovery Using Compressed Sensing from Unequally Spaced Data**

21   In this study, we regarded unequally spaced sparse time series data as equally spaced dense time

1   series data with missing time points, and equally spaced time series data were generated by

2   restoring missing time points using a low-rank approach (Konishi et al., 2014). In the low-rank

3   approach for image recovery, we assumed that the value of each pixel is represented by a linear

4   combination of its neighbor pixels, which is mathematically represented by an autoregressive

5   (AR) model. Then a Hankel-like matrix composed of pixel values has a low rank because each

6   column is represented by the linear combination of the other columns (Figures 1 and S1A). This

7   means that the Hankel-like matrix is a low-rank matrix whose rank is determined by the system

8   order. Missing data can be recovered by estimating missing elements of the matrix so that the

9   rank of this matrix [Y] is r. When system order r is unknown, based on the idea that the system

10   can be described with as few parameters as possible, missing elements of this Hankel-like matrix

11   are recovered so as to minimize the rank of the matrix [Y]. Based on the low rankness of the

12   Hankel matrix, the signal recovery problem of the missing pixles can be formulated as a matrix

13   rank minimization problem, and we can restore an image by solving this problem (Takahashi et

14   al., 2012; Takahashi et al., 2016) (Figure 1).

15   We performed system identification from unequally spaced time series data of input

16   molecules (*Inputs*) and output molecules (*Outputs*). Although an AR model is used for image

17   recovery, we used an ARX model where the value at a time point is represented by a linear

18   combination of two kinds of signals, *Inputs* and *Outputs*. Therefore, we modified the

19   rank-minimization-based signal recovery method of the AR model to the ARX model and

20   performed system identification (Figures 1 and S1B). Several methods for system identification

21   using a linear ARX model with signal recovery of missing points of input and output based on

10

1   matrix rank minimization have been proposed (Liu et al., 2013; Markovsky, 2013). They can

2   recover missing time series input–output data even when missing time points of input are not

3   equal to those of output.

4   However, we cannot directly apply the method because we used the NARX model rather than

5   the ARX model due to the nonlinearity of signaling-dependent gene expression (Kudo et al.,

6   2016; Saito et al., 2013). Therefore, by combining the nonlinear ARX system identification

7   method (Saito et al., 2013) and the signal recovery method based on the matrix rank

8   minimization problem (Konishi et al., 2014), we derived the signal recovery algorithm applicable

9   to the nonlinear ARX system and performed system identification using recovered equally

10   spaced time series input–output data (see "NARX Model and Data Representation" and

11   "Extension ARX System Identification from Unequally Spaced Time Series Data to the NARX

12   System" sections in the STAR METHODS).

13

14   **System Identification by Integrating Signal Recovery and the NARX Model**

15   In the NARX model used in our previous work, time series data of *Inputs* are nonlinearly

16   transformed using the Hill equation, which are then used as inputs for the ARX model (Saito et

17   al., 2013) (Figure 2A). The Hill equation, which is nonlinear transformation function $f(x)$

18   widely used in biochemistry (Hill, 1910), can represent sensitivity with a graded or switch-like

19   response by the values of *n* and *K* (Figure 2A). The ARX model in the NARX model can

20   represent how the *Output* efficiently responds to the temporal change of the nonlinearly

21   transformed *Inputs* by the time constant and gain (Figure 2A). Thus, from the estimated

1   parameters of the Hill equation and ARX model, the sensitivity with graded or switch-like

2   response and the time constant and gain are obtained, respectively. In this study, the parameters

3   of this NARX model were estimated using a signal recovery scheme based on a low-rank

4   approach (Konishi et al., 2014), as follows (Figure 2B, see details in "Procedure for System

5   Identification by Integrating Signal Recovery and the NARX Model" section in the STAR

6   METHODS).

7   　　To estimate the *I-O* relationship, we selected a combination of *Inputs* for each *Output* and

8   prepared a data set of all combinations of *Inputs* for each *Output*. The each data set was divided

9   into test dataset for one stimulation condition and training data set for the rest of two stimulation

10   conditions, leave-one-out (LOO) cross-validation was performed. We estimated the parameters

11   of the NARX model (the NARX parameters) for the training data set in each *Input–Output*

12   combination using the following method. First, the initial values of n and K are given by

13    $n = 1$ and a random number, respectively, and nonlinear transformation of input unequally

14   spaced time series data by using the Hill equation was performed (Figure 2B, step i). A

15   Hankel-like matrix was constructed from unequally spaced time series *Output* data and from

16   unequally spaced time series *Inputs* nonlinearly transformed by the Hill equation. Next, signal

17   recovery was performed with an iterative partial matrix shrinkage (IPMS) algorithm to minimize

18   the rank of the Hankel-like matrix composed of *Output* and *Inputs* transformed by the Hill

19   equation (Konishi et al., 2014) (Figure 2B, step ii). The rank of the recovered Hankel-like matrix

20   corresponds to the lag order; from the recovered Hankel-like matrix, the parameters of the ARX

21   model $\boldsymbol{a}$ and $\boldsymbol{b}$ were uniquely obtained (Figure 2B, step iii; Figure S1B). Further estimation of

1   n and K was performed by using recovered data, ARX parameters obtained until step iii, and

2   other combination of n and K given random numbers (Figure 2B, step iv). By using the

3   inverse function of the Hill equation, we recovered the missing time points data of input before

4   transformation by the Hill equation. For the other 200 combinations of n and K given by

5   random numbers, we performed simulation of the NARX model using the recovered data, ARX

6   parameters obtained until step iii, and the given combination of n and K.

7   We calculated the Akaike information criterion (AIC) from the residual sum of squares

8   between the experiment and simulation, number of parameters, and number of data to determine

9   the parameters n and K. AIC is a measure of the relative quality of statistical models based on the

10 trade-off between the goodness-of-fit of the model and the complexity of the model (Akaike,

11 1974). In step iv, we selected the combination of n and K with the minimum AIC and carried

12 out signal recovery again using these n and K. We repeated steps i–iv 500 times, and selected

13 the n and K and ARX parameters that minimize AIC for the training data set $AIC_{training}$ in

14 total (Figure 2B, step v). Let parameters with minimum $AIC_{training}$ be parameters obtained

15 from the training data set (Figure 2B, step v). Once these parameters were obtained, test data

16 (still unequally spaced time series data) was added to the recovered Hankel-like matrix and

17 signal recovery of the test data was performed (Figure 2B, step vi). With the parameters of the

18 NARX model estimated from the training data set, we simulated the NARX model for test data

19 and calculated RSS_{LOO}^s, the residual sum of squares between experiment and simulation for test

20 data set by stimulation condition s (NGF, PACAP, or PMA) (Figure 2B, step vii).

21   Because RSS_{LOO}^s was obtained for each combination of training and test data set s, we took

13

1 the sum of RSS_{LOO}^s for test data set s as RSS_{LOO} (Figure 2B, step viii). We obtained the

2 combination of *Inputs* as the identified *I-O* relationship that minimizes RSS_{LOO} for all

3 combination of *Inputs* (Figure 2B, step ix). The *I-O* relationship indicates that a set of *Inputs* are

4 selected as upstream molecules for each *Output*. In the final step, using this combination of input

5 molecules, the parameters of the final NARX model were estimated by the procedure from step i

6 to step v using all stimulation conditions as training data sets (Figure 2B, step x). Note that we

7 used two different criterions $AIC_{training}$ and RSS_{LOO}; $AIC_{trainig}$ to determine *n*, *k*, and ARX

8 parameters, and RSS_{LOO} to select *Inputs* in order to save computational cost.

9 These estimated NARX parameters were used for further study (Figs. 3–6). The sensitivity

10 with graded or switch-like response was obtained from the parameters of the Hill equation, and

11 the gain and time constant were obtained from the parameters of the ARX model (Figure 2B, see

12 "Calculation of Gain and Time Constant from the Linear ARX Model" section in the STAR

13 METHODS). An example of the transformation of *Inputs* by the Hill equation and signal

14 recovery following the ARX model and simulated *Output* is shown in Figure S2. We applied this

15 method to identify the signaling-decoding system by gene expression underlying cell

16 differentiation in PC12 cells using unequally spaced time series data with different time scales.

17

18 **System Identification of Signaling-Dependent Gene Expression**

19 We stimulated PC12 cells by NGF, PACAP, and PMA and measured the amount of

20 phosphorylated ERK1 and ERK2 (pERK) and CREB (pCREB) and protein abundance of

21 products of the IEGs, such as c-Jun, c-Fos, Egr1, FosB, and JunB by using QIC (Ozaki et al.,

14

1   2010) (Figure 4). We chose these growth factors because they use different signaling pathways:

2   NGF, PACAP, and PMA use Ras-, cAMP-, and PKC-dependent signaling pathways, respectively

3   (Farah and Sossin, 2012; Gerdin and Eiden, 2007; Ravni et al., 2006; Vaudry et al., 2002). We

4   also measured mRNA expression of LP genes such as *Metrnl*, *Dclk1*, and *Serpinb1a* using

5   qRT-PCR (Figure 4). We measured the signaling molecules and gene expression with different

6   sets of the time points because of the different time scales of temporal changes in signaling

7   molecules and gene expression (Figure 4). Using the unequally spaced time series data with the

8   different sets of the time points, we performed the system identification using integration of

9   signal recovery and the NARX model (Figure 5A–C).

10   Using these time series data sets, we selected three sets of *Inputs–Outputs* combinations from

11   upstream to downstream and performed system identification for each set (Figure 5A). The

12   system identification consists estimating the *I-O* relationship, dose-response by the Hill equation,

13   and gain and time constant by the linear ARX model (Figure 3, see "Calculation of Gain and

14   Time Constant from the Linear ARX Model" section in the STAR METHODS).

15   We selected pERK and pCREB as *Input* candidates for each *Output*, c-Jun, c-Fos, and Egr1,

16   based on previous studies (Akimoto et al., 2013; Saito et al., 2013; Watanabe et al., 2012) (Figure

17   5A). We selected pERK, pCREB, c-Jun, c-Fos, and Egr1 as *Input* candidates for each *Output*,

18   FosB and JunB (Akimoto et al., 2013; Saito et al., 2013; Watanabe et al., 2012) (Figure 5A). We

19   selected pERK, pCREB, c-Jun, c-Fos, Egr1, FosB, and JunB as *Input* candidates for each *Output*,

20   *Metrnl*, *Serpinb1a*, and *Dclk1* (Watanabe et al., 2012) (Figure 5A).

21   For c-Jun and Egr1, pERK was selected as an *Input*, and for c-Fos, pERK and pCREB were

15

1   selected as *Inputs* (Figure 5). For FosB, c-Jun and c-Fos were selected as *Inputs*, and for JunB,

2   pCREB was selected as *Input* (Figure 5B). For *Metrnl*, FosB, c-Fos and JunB were selected as an

3   *Inputs*; however, contributions of c-Fos and JunB were negligible (Figure S2), indicating that

4   FosB is a main *Input* for *Metrnl*. For *Serpinb1a* and *Dclk1*, JunB was selected as an *Input* (Figure

5   5B). It is noteworthy that FosB and JunB, but not signaling molecules and other IEGs, were

6   mainly selected as *Inputs* of the LP genes and the inputs for *Metrnl* and *Dclk1* were different

7   despite their similar temporal patterns.

8   We characterized the dose-response by the Hill equation and gain and time constant by the

9   linear ARX model (Figure 5C, Table 1). The dose-responses from c-Jun and c-Fos to FosB

10   showed typical switch-like responses, whereas others showed graded or weaker switch-like

11   responses. Note that the gain from the converted c-Jun to FosB was much smaller than that from

12   the converted c-Fos (Table 1), indicating that FosB is mainly regulated by c-Fos but not c-Jun.

13   The time constants for c-Jun, c-Fos, Egr1, *Metrnl*, and *Dclk1* were less than 1 h, whereas those

14   for FosB, JunB, and *Serpinb1a* were more than 100 min (Table 1), indicating that induction of

15   FosB and JunB temporally limit the overall induction of the LP genes from signaling molecules.

16   The transformation of *Inputs* by the Hill equation followed by the ARX model is shown in Figure

17   S2. In addition, when we integrated these three sets of the NARX model and simulated the

18   response using only pERK and pCREB as *Inputs*, we obtained a similar result (Figure S3).

19

20   **Prediction and Validation of the Identified System by Pharmacological Perturbation**

21   We validated the identified system by pharmacological perturbation. One of the key issues in

16

1 PC12 cell differentiation is whether ERK or CREB phosphorylation mediates expression of the

2 downstream genes (Ravni et al., 2006; Vaudry et al., 2002; Watanabe et al., 2012). Therefore, we

3 selectively inhibited ERK phosphorylation by a specific MEK inhibitor, trametinib (Gilmartin et

4 al., 2011; Watanabe et al., 2013; Yamaguchi et al., 2007) in PACAP-stimulated PC12 cells. We

5 found that PACAP-induced ERK phosphorylation, but not CREB phosphorylation, was

6 specifically inhibited by trametinib (Figure 6A, black dots).

7 For c-Jun, c-Fos, Egr1, and JunB, we recovered signals of the unequally spaced time series

8 data of *Inputs* and *Output*. For c-Jun, c-Fos, and Egr1, we simulated *Outputs* responses using

9 these recovered data and the identified NARX model (Figure 6A black lines, see also

10 "Simulation of the Integrated NARX Model" section in the STAR METHODS). For other

11 downstream molecules, FosB, JunB, *Metrnl*, *Serpinb1a*, and *Dclk1*, we used the recovered data

12 of pCREB and the simulated time series data of c-Jun, c-Fos, and Egr1 as *Inputs* for the

13 identified NARX model (Figure 6A, black lines). The simulated time courses of *Outputs* were

14 similar with those in experiments, except those of FosB and *Metrnl* (Figure 6A, black lines and

15 black dots). In the simulation, FosB and *Metrnl* did not respond to PACAP in the presence of

16 trametinib, whereas in the experiment both molecules did so, suggesting the possibility of failure

17 of the system identification of FosB and/or *Metrnl*. Therefore, we investigated whether FosB and

18 *Metrnl* can be reasonably reproduced when experimental and recovered data of c-Fos and c-Jun

19 and of FosB, respectively, were used rather than the simulated ones (Figure 6B). When

20 experimental and recovered data were used as *Inputs*, *Metrnl*, but not FosB, responded to PACAP

21 in the presence of trametinib both in the simulation and experiment (Figure 6B), indicating that

17

1 the failure of the system identification of FosB and *Metrnl* in Figure 6A arose from the failure of

2 the system identification of FosB. Thus, all *Outputs* except FosB showed similar responses in the

3 experiment and simulation when the experimental and recovered data were used as *Inputs*,

4 indicating that in most cases the identified system is validated by pharmacological perturbation.

5

6 **DISCUSSION**

7 In this study, we identified the system from signaling molecules to gene expression using the

8 unequally spaced time series data for 720 min after the stimulation. Given that expression levels

9 of the LP genes were highly correlated with neurite length regardless of growth factors

10 (Watanabe et al., 2012) and expression continues for 720 min after the initial addition of NGF

11 (Chung et al., 2010), the identified system is the selective growth factor–signaling decoding

12 system for neurite length information, one of the most critical steps for cell differentiation in

13 PC12 cells.

14   We previously identified the systems leading from pERK and pCREB to the IEGs using the

15 equally spaced dense time series data with a uniform 3-min interval during 180 min (Saito et al.,

16 2013). The identified *I-O* relationships in this study are the same, except for the inputs of FosB

17 and JunB. In this study, for FosB c-Jun was selected as an *Input* in addition to c-Fos. However,

18 the gain from the converted c-Jun to FosB was much smaller than that from the converted c-Fos

19 (Table 1), indicating that the effect of c-Jun is negligible. For JunB, c-Fos was not selected as an

20 *Input* in this study, whereas c-Fos was selected in the previous study (Saito et al., 2013). The gain

21 from the converted c-Fos to JunB at lower frequency was much smaller than that from the

1   converted pCREB (Saito et al., 2013), indicating that the effect of c-Fos is negligible in the

2   previous study. Thus, the identified *I-O* relationships in this study are consistent with our

3   previous work.

4   The estimated NARX parameters were also generally consistent with those in our previous

5   study (Saito et al., 2013). The peak of c-Fos by NGF stimulation was approximately 0.9, whereas

6   it was approximately 0.6 in our previous study (Saito et al., 2013). The difference may come

7   from the difference in the algorithm for parameter estimation, because of the procedure of signal

8   recovery is included in this study. Overall, the inferred the *I-O* relationship, the Hill equation,

9   and the linear ARX model in this study are consistent with our earlier observations, indicating

10   that system identification using unequally spaced time series data can give the same performance

11   as using equally spaced time series data and that the system is time invariant during 720 min.

12   Furthermore, the identified system can reasonably reproduce the time series data using

13   extrapolated data with trametinib, except for FosB. Taken together, these results demonstrate the

14   validity of the predicted response of the identified systems. The reason for the failure of system

15   identification of FosB is unclear, but it may reflect the failure of parameter estimation due to the

16   insufficient number of experimental data points, the limitation of the NARX model structure, or

17   the existence of unknown regulatory molecules. Further studies are necessary to address these

18   issues.

19   One key issue is whether ERK and/or CREB mediates cell differentiation through

20   downstream gene expression in PC12 cells (Ravni et al., 2006; Vaudry et al., 2002; Watanabe et

21   al., 2012). We previously found that the LP genes are not induced by NGF in the presence of

19

1   U0126, another MEK inhibitor (Watanabe et al., 2012), and that the MEK inhibitor blocks

2   NGF-induced phosphorylation of both ERK and CREB in PC12 cells (Akimoto et al., 2013; Uda

3   et al., 2013; Watanabe et al., 2012). By contrast, the MEK inhibitor blocked phosphorylation of

4   ERK, but not CREB, in PACAP-stimulated PC12 cells (Akimoto et al., 2013; Uda et al., 2013)

5   (Figure 6), suggesting that PACAP induces phosphorylation of CREB through a

6   cAMP-dependent pathway, rather than the ERK pathway. These results demonstrate that NGF

7   selectively uses the ERK pathway, whereas PACAP selectively uses the cAMP pathway for

8   induction of the LP genes. Considering that LP genes are the common decoders for neurite length

9   in PC12 cells regardless of growth factors (Watanabe et al., 2012), the identified system in this

10   study (except for FosB) reveals the selective NGF- and PACAP-signaling decoding mechanisms

11   for neurite length information.

12   Recently, fluorescence resonance energy transfer probes, optogenetics, and microfluidic

13   devices have been developed to achieve observation and time control of ERK phosphorylation

14   temporal patterns. These methods allow us to focus on quantitative relationships between various

15   ERK phosphorylation temporal patterns and phenotypes such as cell differentiation (Albeck et al.,

16   2013; Aoki et al., 2013; Doupé and Perrimon, 2014; Ryu et al., 2015; Sumit et al., 2017;

17   Toettcher et al., 2013; Zhang et al., 2014). Although the relationship between signal transduction

18   and phenotype has been extensively studied, it remains unclear how the signaling molecules

19   quantitatively regulate the downstream gene expressions over a longer time scale, leading to cell

20   fate decisions. In this study, we revealed the quantitative regulatory mechanism between

21   signaling activation at a short time scale (tens of minutes) and gene expression at a longer time

1   scale (day) by using a system identification method integrating a signal recovery technique and

2   the NARX model based on compressed sensing.

3   A linear or spline interpolation is often used to convert unequally spaced time course data into

4   equally spaced time course data in biological data analysis. However, such interpolation methods

5   are not likely to be reliable because the interpolation methods ignore biochemical property of

6   molecular network. By contrast, the interpolation used in this study is based on the NARX model,

7   which reflects biochemical property. Thus, the proposal method in this study is biologically more

8   plausible than a linear or spline interpolation.

9   There are obscure points for an application of this method to biological data analysis. The

10   relationship between a number of observed time points and accuracy of signal recovery is

11   theoretically unknown. In addition, how to select time points is also unknown. Intuitively, dense

12   time points may be required for transient response, while sparse time points may be sufficient for

13   sustained response. Further study is necessary to address this issue.

14   In molecular and cellular biology, molecular networks—the *I-O* relationship in this

15   study—are generally examined by gene disruption or pharmacological perturbation experiments,

16   meaning that the *I-O* relationship is examined using static and qualitative data. In this study, we

17   used *Inputs–Outputs* time series data for system identification, allowing us to determine the *I-O*

18   relationship using dynamic and quantitative data. Our system identification method does not

19   require detailed knowledge of pathway information, which means that it can be used as a

20   pathway finder directly from time series data. Moreover, additional information such as

21   sensitivity with graded or switch-like response, time delay, and gain can be obtained. One of the

22   advantages of using time series data rather than static perturbation experiments is simultaneously

1   obtaining the *I-O* relationship, sensitivity, and time constant, which characterize the system

2   behavior. This is based on the idea that input–output time series data implicitly include

3   information on the *I-O* relationship. However, we must note that the *I-O* relationship obtained by

4   using this method may be an apparent relationship inferred from time series data and is not

5   necessarily the direct *I-O* relationship. Therefore, the obtained *I-O* relationship should be

6   validated by gene disruption or pharmacological perturbation experiments, as shown in Figure 6.

7   In conclusions, we have devised a system identification method using unequally spaced

8   sparse time series data by signal recovery. Because of technical and budget limitations in

9   biological experiments, it is generally difficult to obtain sufficient numbers of equally spaced

10   dense time series data of molecules with different time scales. Thus far, system identification

11   based on time series data has been limited to phenomena with similar time scales. However, our

12   system identification method can solve this time-scale problem and can be applied to any

13   biological system with different time scales, such as the cell cycle, development, regeneration,

14   and metabolism involving ion flux, metabolites, phosphorylation, and gene expression.

1 **STAR METHODS**

2 Detailed methods of this study are provided and include the following:

3 ● KEY RESOURCES TABLE

4 ● CONTACT FOR REAGENT AND RESOURCE SHARING

5 ● EXPERIMENTAL MODEL AND SUBJECT DETAILS

6 ○ Cell Culture and Treatments

7 ● METHOD DETAILS

8 ○ Quantitative Image Cytometry (QIC)

9 ○ qRT-PCR Analysis

10 ○ NARX Model and Data Representation

11 ○ Extension of ARX System Identification from Unequally Spaced Time Series Data to the

12 NARX System

13 ○ Procedure for System Identification by Integrating Signal Recovery and NARX Model

14 ○ Calculation of Gain and Time Constant from Linear ARX Model

15 ○ Simulation of the Integrated NARX Model

16

17 **Author Contributions:**

18 T.T., K. Konishi, and S.K. conceived the project and designed the experiments. T.T., N.M., and K.

19 Kunida performed the experiments. T.T., M.F., K. Kunida, S.U., and K. Konishi derived the

20 system identification algorithm by integrating signal recovery and the NARX model. T.T., M.F.,

21 and K. Konishi analyzed the data. T.T., M.F., N.M., H.K., K. Konishi, and S.K. contributed

23

1   reagents/materials/analysis tools. T.T., M.F., K. Konishi, and S.K. wrote the manuscript.

2

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. IEEE Trans. Automat. Control *19*, 716–723.

Akimoto, Y., Yugi, K., Uda, S., Kudo, T., Komori, Y., Kubota, H., and Kuroda, S. (2013). The extraction of simple relationships in growth factor-specific multiple-input and multiple-output systems in cell-fate decisions by backward elimination PLS regression. PLoS One *8*, e72780.

Albeck, J.G., Mills, G.B., and Brugge, J.S. (2013). Frequency-modulated pulses of ERK activity transmit quantitative proliferation signals. Mol. Cell *49*, 249–261.

Aoki, K., Kumagai, Y., Sakurai, A., Komatsu, N., Fujita, Y., Shionyu, C., and Matsuda, M. (2013). Stochastic ERK activation induced by noise and cell-to-cell propagation regulates cell density-dependent proliferation. Mol Cell *52*, 529–540.

Behar, M., and Hoffmann, A. (2010). Understanding the temporal codes of intra-cellular signals. Curr Opin Genet Dev *20*, 684–693.

Brightman, F.A., and Fell, D.A. (2000). Differential feedback regulation of the MAPK cascade underlies the quantitative differences in EGF and NGF signalling in PC12 cells. FEBS Lett. *482*, 169–174.

Candès, E.J., and Fernandez-Granda, C. (2014). Towards a mathematical theory of super-resolution. Commun. Pure Appl. Math. *67*, 906–956.

Candès, E.J., and Recht, B. (2009). Exact matrix completion via convex optimization. Found. Comput. Math. *9*, 717–772.

1 Candès, E.J., and Wakin, M.B. (2008). An introduction to compressive sampling. IEEE Signal

2   Proces. Mag. *25*, 21–30.

3 Candès, E.J., Wakin, M.B., and Boyd, S.P. (2008). Enhancing sparsity by reweighted ℓ 1

4   minimization. J. Fourier Anal. Appl. *14*, 877–905.

5 Chung, J., Kubota, H., Ozaki, Y., Uda, S., and Kuroda, S. (2010). Timing-dependent actions of

6   NGF required for cell differentiation. PLoS One *5*, e9011.

7 Donoho, D.L. (2006). Compressed sensing. IEEE Trans. Inform. Theory *52*, 1289–1306.

8 Doupé, D.P., and Perrimon, N. (2014). Visualizing and manipulating temporal signaling

9   dynamics with fluorescence-based tools. Sci. Signal. *7*, re1.

10 Farah, C.A., and Sossin, W.S. (2012). The role of C2 domains in PKC signaling. In Calcium

11   Signaling, M.S. Islam, ed. (Berlin: Springer), pp. 663–683.

12 Fazel, M. (2002). Matrix rank minimization with applications (PhD thesis, Stanford University).

13 Filippi, S., Barnes, C.P., Kirk, P.D., Kudo, T., Kunida, K., McMahon, S.S., Tsuchiya, T., Wada, T.,

14   Kuroda, S., and Stumpf, M.P. (2016). Robustness of MEK-ERK dynamics and origins of

15   cell-to-cell variability in MAPK sgnaling. Cell Rep. *15*, 2524–2535.

16 Gerdin, M.J., and Eiden, L.E. (2007). Regulation of PC12 cell differentiation by cAMP signaling

17   to ERK independent of PKA: Do all the connections add up? Science's STKE. *2007*, pe15.

18 Gilmartin, A.G., Bleam, M.R., Groy, A., Moss, K.G., Minthorn, E.A., Kulkarni, S.G., Rominger,

19   C.M., Erskine, S., Fisher, K.E., Yang, J., et al. (2011). GSK1120212 (JTP-74057) is an

20   inhibitor of MEK activity and activation with favorable pharmacokinetic properties for

21   sustained in vivo pathway inhibition. Clin. Cancer Res. *17*, 989–1000.

1 Gotoh, Y., Nishida, E., Yamashita, T., Hoshi, M., Kawakami, M., and Sakai, H. (1990).

2   Microtubule-associated-protein (MAP) kinase activated by nerve growth factor and

3   epidermal growth factor in PC12 cells. Eur. J. Biochem. *193*, 661–669.

4 Hill, A.V. (1910). The possible effects of the aggregation of the molecules of haemoglobin on its

5   dissociation curves. J. Physiol. (Lond) *40*, 4–7.

6 Janes, K.A., and Lauffenburger, D.A. (2006). A biological approach to computational models of

7   proteomic networks. Curr. Opin. Chem. Biol. *10*, 73–80.

8 Janes, K.A., and Lauffenburger, D.A. (2013). Models of signalling networks: what cell biologists

9   can gain from them and give to them. J. Cell Sci. *126*, 1913–1921.

10 Janes, K.A., and Yaffe, M.B. (2006). Data-driven modelling of signal-transduction networks. Nat.

11   Rev. Mol. Cell Biol. *7*, 820–828.

12 Kholodenko, B., Yaffe, M.B., and Kolch, W. (2012). Computational approaches for analyzing

13   information flow in biological networks. Sci. Signal., *5*, re1.

14 Konishi, K., Uruma, K., Takahashi, T., and Furukawa, T. (2014). Iterative partial matrix

15   shrinkage algorithm for matrix rank minimization. Signal Process. *100*, 124–131.

16 Liu, Z., Hansson, A., and Vandenberghe, L. (2013). Nuclear norm system identification with

17   missing inputs and outputs. Syst. & Control Lett. *62*, 605–612.

18 Ljung, L. (1998). System identification. Signal Analysis and Prediction, A. Prochazka et al., eds.

19   (New York: Springer), pp. 163–173.

20 Ljung, L. (2010). Perspectives on system identification. Ann. Rev. Control *34*, 1–12.

21 Lustig, M., Donoho, D.L., Santos, J.M., and Pauly, J.M. (2008). Compressed sensing MRI. IEEE

Signal Process. Mag. *25*, 72–82.

Marshall, C. (1995). Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase activation. Cell *80*, 179–185.

Nakakuki, T., Birtwistle, M.R., Saeki, Y., Yumoto, N., Ide, K., Nagashima, T., Brusch, L., Ogunnaike, B.A., Okada-Hatakeyama, M., and Kholodenko, B.N. (2010). Ligand-specific c-Fos expression emerges from the spatiotemporal control of ErbB network dynamics. Cell *141*, 884–896.

Ongie, G., and Jacob, M. (2016). A fast algorithm for structured low-rank matrix recovery with applications to undersampled MRI reconstruction. Paper presented at: IEEE International Symposium on Biomedical Imaging (ISBI).

Ozaki, Y., Uda, S., Saito, T.H., Chung, J., Kubota, H., and Kuroda, S. (2010). A quantitative image cytometry technique for time series or population analyses of signaling networks. PLoS One *5*, e9955.

Price, N.D., and Shmulevich, I. (2007). Biochemical and statistical network models for systems biology. Curr. Opin. Biotechnol. *18*, 365–370.

Purvis, J.E., and Lahav, G. (2013). Encoding and decoding cellular information through signaling dynamics. Cell *152*, 945–956.

Qiu, M.-S., and Green, S.H. (1992). PC12 cell neuronal differentiation is associated with prolonged p21 ras activity and consequent prolonged ERK activity. Neuron *9*, 705–717.

Ravni, A., Bourgault, S., Lebon, A., Chan, P., Galas, L., Fournier, A., Vaudry, H., Gonzalez, B., Eiden, L.E., and Vaudry, D. (2006). The neurotrophic effects of PACAP in PC12 cells:

control by multiple transduction pathways. J. Neurochem. *98*, 321–329.

Ryu, H., Chung, M., Dobrzynski, M., Fey, D., Blum, Y., Lee, S.S., Peter, M., Kholodenko, B.N., Jeon, N.L., and Pertz, O. (2015). Frequency modulation of ERK activation dynamics rewires cell fate. Mol. Syst. Biol. *11*, 838.

Saito, T.H., Uda, S., Tsuchiya, T., Ozaki, Y., and Kuroda, S. (2013). Temporal decoding of MAP kinase and CREB phosphorylation by selective immediate early gene expression. PLoS One *8*, e57037.

Santos, S.D., Verveer, P.J., and Bastiaens, P.I. (2007). Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. Nat. Cell Biol. *9*, 324–330.

Sasagawa, S., Ozaki, Y., Fujita, K., and Kuroda, S. (2005). Prediction and validation of the distinct dynamics of transient and sustained ERK activation. Nat. Cell Biol. *7*, 365–373.

Sumit, M., Takayama, S., and Linderman, J. (2017). New insights into mammalian signaling pathways using microfluidic pulsatile inputs and mathematical modeling. Integrative Biol., *9*, 6–21

Takahashi, T., Konishi, K., and Furukawa, T. (2012). Rank minimization approach to image inpainting using null space based alternating optimization. Paper presented at: 2012 19th IEEE International Conference on Image Processing (IEEE).

Takahashi, T., Konishi, K., Uruma, K., and Furukawa, T. (2016). Adaptive image inpainting algorithm based on generalized principal component analysis. Paper presented at: Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), 2016 IEEE 12th (IEEE).

1   Toettcher, J.E., Weiner, O.D., and Lim, W.A. (2013). Using optogenetics to interrogate the

2   dynamic control of signal transmission by the Ras/Erk module. Cell *155*, 1422–1434.

3   Traverse, S., Gomez, N., Paterson, H., Marshall, C., and Cohen, P. (1992). Sustained activation

4   of the mitogen-activated protein (MAP) kinase cascade may be required for differentiation of

5   PC12 cells: Comparison of the effects of nerve growth factor and epidermal growth factor.

6   Biochem. J. *288*, 351–355.

7   Uda, S., Saito, T.H., Kudo, T., Kokaji, T., Tsuchiya, T., Kubota, H., Komori, Y., Ozaki, Y., and

8   Kuroda, S. (2013). Robustness and compensation of information transmission of signaling

9   pathways. Science *341*, 558–561.

10   Vaudry, D., Stork, P., Lazarovici, P., and Eiden, L. (2002). Signaling pathways for PC12 cell

11   differentiation: making the right connections. Science *296*, 1648–1649.

12   Watanabe, K., Akimoto, Y., Yugi, K., Uda, S., Chung, J., Nakamuta, S., Kaibuchi, K., and

13   Kuroda, S. (2012). Latent process genes for cell differentiation are common decoders of

14   neurite extension length. J. Cell Sci. *125*, 2198–2211.

15   Watanabe, M., Sowa, Y., Yogosawa, M., and Sakai, T. (2013). Novel MEK inhibitor trametinib

16   and other retinoblastoma gene (RB)-reactivating agents enhance efficacy of 5-fluorouracil on

17   human colon cancer cells. Cancer Sci. *104*, 687–693.

18   Yamaguchi, T., Yoshida, T., Kurachi, R., Kakegawa, J., Hori, Y., Nanayama, T., Hayakawa, K.,

19   Abe, H., Takagi, K., Matsuzaki, Y., et al. (2007). Identification of JTP-70902, a

20   p15(INK4b)-inductive compound, as a novel MEK1/2 inhibitor. Cancer Sci. *98*, 1809–1816.

21   Yang, J., Wright, J., Huang, T.S., and Ma, Y. (2010). Image super-resolution via sparse

30

1   representation. IEEE Trans. Image Process. *19*, 2861–2873.

2   Zechner, C., Seelig, G., Rullan, M., and Khammash, M. (2016). Molecular circuits for dynamic

3   noise filtering. Proc. Natl. Acad. Sci. USA *113*, 4729–4734.

4   Zhang, K., Duan, L., Ong, Q., Lin, Z., Varman, P.M., Sung, K., and Cui, B. (2014).

5   Light-mediated kinetic control reveals the temporal effect of the Raf/MEK/ERK pathway in

6   PC12 cell neurite outgrowth. PLoS One *9*, e92917.

7

Figure legends

Figure 1. Signal recovery based on compressed sensing technology from unequally spaced data. (Top) An example of recovered equally spaced image data from unequally spaced image data by the signal recovery technique using rank minimization of the Hankel-like matrix Y composed of signals in pixels. We assume that the value of each pixel is represented by a linear combination of those of its neighboring pixels, which is mathematically represented by an autoregressive (AR) model. (Bottom) An example of recovered equally spaced time series data from unequally spaced time series data by the signal recovery technique using rank minimization of the Hankel-like matrices Y and U, composed of time series data of input molecules (*Inputs*) and output molecules (*Outputs*), respectively. We assume that the value at a certain time is represented not only by the linear combination of values of the output molecule at past points but also by the linear combination of the values of the input molecule at past points, which is mathematically represented by an autoregressive exogenous (ARX) model. The recovered time series input–output data have the equally spaced time series data with the same time points even if the missing time points of input and output are different.

Figure 2. System identification by integrating signal recovery and the NARX model. (A) The nonlinear ARX (NARX) model consists of static nonlinear conversion of input signal by the Hill equation, followed by time delay by ARX model (Saito et al., 2013). The former gives the sensitivity with a graded or switch-like response and the latter gives the time constant. (B)

32

Algorithm flowchart for system identification by integrating signal recovery and the NARX model. See details in "Procedure for System Identification by Integrating Signal Recovery and the NARX Model" section in the STAR METHODS.

Figure 3. Transformation of *Inputs* by the Hill equation and signal recovery followed by the ARX model. Signal transformation in the nonlinear ARX model of c-Fos is shown. The signals of pERK and pCREB were transformed by the Hill equations and recovered. Then the transformed signals were temporally transformed by the linear ARX model. The sum of the transformed signals by the linear ARX model was c-Fos, the final output of the nonlinear ARX model of c-Fos.

Figure 4. Experimental data of growth factor–dependent changes of signaling molecules and gene expression in PC12 cells. PC12 cells were stimulated with NGF (50 ng/ml, red), PACAP (100 nM, blue), or PMA (100 nM, green). Phosphorylation of signaling molecules, such as ERK and CREB, the product of IEGs such as c-Jun, c-Egr1, c-Jun, FosB, and JunB, and mRNA expression of LP genes such as *Metrnl*, *Dclk1*, and *Serpinb1a* were measured with different time points. These data were used for system identification by the NARX model in Figure 5–C.

Figure 5. System identification of *I-O* relationships between the signaling, IEGs, and LP genes. (A) The sets of combinations of *Inputs* and *Outputs* for system identification. (B) The

33

identified *I-O* relationships. Arrows indicate the estimated *I-O* relationships for each set of inputs and outputs. The colors of the arrows indicate the same sets of combinations of inputs and outputs as in (A). Dots, experimental data; pluses, the recovered signal data; solid lines, simulation data of the NARX model; red, NGF stimulation; blue, PACAP stimulation; green, PMA stimulation. (C) The dose-response curves obtained by the Hill equation. For each panel, conversion of *Inputs* by the identified Hill equation is shown. The colors of the arrows and plotted lines indicate the same *Inputs*, respectively.

Figure 6. Prediction and validation of the identified system by pharmacological perturbation. (A) The predictive simulation result and experimental result by PACAP stimulation in the presence (black) or absence (blue) of trametinib. Lines, simulation; dots, experimental and recovered data. Experimental and recovered data of pERK and pCREB, and the simulated data of c-Jun, c-Fos, Egr1, FosB, and JunB are given as *Inputs*, and simulation was performed using the NARX model in Figure 5 (see Method). In the experiment, PC12 cells were treated in the absence (blue dots) or in the presence (black dots) of trametinib (10 μM) added at 30 min before stimulation with PACAP (100 nM). Note that the PACAP stimulation data are used, as in Figure 4. (B) Simulation using experimental and recovered data as *Inputs*. For each set of the *Inputs* (left panel for each) and *Outputs* (right panel for each), the unequally spaced time series data were recovered (pluses) (right panel for each), and the responses of *Outputs* were simulated by the NARX model identified in Figure 5A–C (solid lines) (right panel for each).

Table

Table 1. The identified *I-O* relationships, parameters of the Hill equation, and gain and time constant calculated from the linear ARX model in Figure 4.

I-O relationship		Hill equation		Linear ARX model	
Input	Output	*K* (EC50)	*n* (Hill coefficient)	Gain	Time constant
pERK	c-Jun	0.9999	1	4.4315	56.4
pERK	c-Fos	0.2274	1.276	1.137	20.8
pCREB	c-Fos	0.7189	6.118	0.8218	22
pERK	Egr1	0.998	1.225	3.2058	21.3
c-Fos	FosB	0.792	81.79	4.682	413
c-Jun	FosB	0.0813	78.58	0.1215	6.5
pCREB	JunB	0.9733	1.58	2.9857	108.9
c-Fos	Metrnl	0.9948	60.7	0.8347	16.8
FosB	Metrnl	0.5998	3.01	1.1256	10.7
JunB	Metrnl	0.8735	89.61	0.2128	21.7
JunB	Serpinb1a	0.3161	7.594	0.8797	275.4
JunB	Dclk1	0.2584	2.234	0.6889	4

STAR METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Mouse anti-phospho-ERK1/2 (Thr 202/Tyr 204)	Cell Signaling Technology	Cat: 9106; RRID: AB_331768
Rabbit anti-phospho-CREB (Ser 133)	Cell Signaling Technology	Cat: 9198; RRID: AB_2561044
Rabbit anti-c-Jun	Cell Signaling Technology	Cat: 9165; RRID: AB_2130165
Rabbit anti-c-Fos	Cell Signaling Technology	Cat: 2250; RRID: AB_2247211
Rabbit anti-Egr1	Cell Signaling Technology	Cat: 4154; RRID: AB_2097035
Rabbit anti-FosB	Cell Signaling Technology	Cat: 2251; RRID: AB_2106903
Rabbit anti-JunB	Cell Signaling Technology	Cat: 3753; RRID: AB_2130002
Alexa 488 goat anti-mouse IgG (H+L)	Invitrogen	Cat: A11029; RRID: AB_138404

Alexa 546 goat anti-rabbit IgG (H+L)	Invitrogen	Cat: A11035; RRID:AB_143051
Chemicals, Peptides, and Recombinant Proteins		
DMEM	Sigma	Cat: D6046
Fetal bovine serum	Sigma	Cat: 172012
Horse serum	Gibco	Cat: 16050-122
poly-L-lysine	Sigma	Cat: P4707
NGF	R&D Systems	Cat: 1156-NG
PACAP	Sigma	Cat: A1439
PMA	Sigma	Cat: P1585
Trametinib	Selleckchem	Cat: S2673
Can Get Signal immunostain solution A	Toyobo	Cat: NKB-501
Hoechst 33342	Invitrogen	Cat: H-3570
Power SYBR Green PCR Master Mix	Applied Biosystems	Cat: 4367659
Critical Commercial Assays		
Agencourt RNAdvance Tissue Kit	Beckman Coulter	Cat: 32646
High Capacity RNA-to-cDNA Kit	Applied Biosystems	Cat: 4387406
Experimental Models: Cell Lines		
PC12 rat adrenal pheochromocytoma cells	Masato Nakafuku (Ohio) Sasagawa et al., 2005	N/A

Oligonucleotides		
Primers for qRT-PCR, see Table S2	This paper	N/A
Software and algorithms		
Iterative partial matrix shrinkage algorithm	Konishi et al., 2014	http://dx.doi.org/10.1016/j.sigpro.2014.01.014
MATLAB	MathWorks	https://www.mathworks.com/
7300 System SDS software version 1.3.1.21	Applied Biosystems	N/A
Other		
Biomek NX Span-8 liquid handling system	Beckman Coulter	N/A
Thermoshake heater-shaker	Variomag	Model number: 7100146-B
Robotic incubator STX-40	Liconic	N/A
CellInsite NTX	Thermo Fisher Scientific	N/A
7300 Real Time PCR System	Applied Biosystem	N/A

CONTACTS FOR REAGENT AND RESOURCE SHARING

Further information and requests for reagents and resources should be directed to and will be fulfilled by the Lead Contact, Shinya Kuroda (skuroda@bs.s.u-tokyo.ac.jp).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell Culture and Treatments

PC12 cells (kindly provided by Masato Nakafuku, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA) (Sasagawa et al., 2005) were cultured at 37°C under 5% CO_2 in complete medium, Dulbecco's modified Eagle's medium (DMEM) (Sigma, Zwijndrecht, The Netherlands) supplemented with 10% fetal bovine serum (Sigma) and 5% horse serum (Gibco, Bethesda, MD, USA). For stimulation, PC12 cells were plated on poly-L-lysine-coated 96-well microplates (0.5×10^4 cells/well) in the complete medium for 24 h and then treated with the complete medium in the presence or absence of the indicated doses of NGF (R&D Systems, Minneapolis, MN, USA), PACAP (Sigma), and PMA (Sigma) (Saito et al., 2013; Uda et al., 2013). Stimulations for cells seeded in 96-well microplates were performed by using a liquid handling system (Biomek NX Span-8, Beckman Coulter, Fullerton, CA, USA) with an integrated heater-shaker (Variomag, Daytona Beach, FL, USA) and robotic incubator (STX-40, Liconic, Mauren, Liechtenstein). For the inhibitor experiment, we stimulated cells with PACAP in the presence of 10 μM trametinib (Selleckchem, Houston, TX, USA). The inhibitor was added 30 min before stimulation with PACAP.

METHOD DETAILS

Quantitative Image Cytometry

QIC was performed as previously described (Ozaki et al., 2010). Briefly, after stimulation by the growth factors, the cells were fixed, washed with phosphate-buffered saline, and permeabilized

39

with blocking buffer (0.1% Triton X-100, 10% fetal bovine serum in phosphate-buffered saline). The cells were washed and then incubated for 2 h with primary antibodies diluted in Can Get Signal immunostain Solution A (Toyobo, Osaka, Japan). The cells were washed three times and then incubated for 1 h with second antibodies. After immunostaining, the cells were stained for the nucleus by incubating with Hoechst 33342 (Invitrogen, Carlsbad, CA, CA). The images of the stained cells were acquired by using a CellInsite NTX (Thermo Fisher Scientific) automated microscope with a 20× objective lens. For QIC analyses, we acquired different field images of the cells in each well, until the number of obtained cells exceeded 1000. Liquid handling for the 96-well microplates was performed using a Biomek NX Span-8 liquid handling system. Intensities of the signaling activity and the IEGs between experiments were normalized by an internal control of each 96-well plate in QIC. Note that for the QIC assays, all the cells within a plate were fixed simultaneously to prevent the exposure of cells to formaldehyde vapor during the treatment.

qRT-PCR analysis

Reverse transcription–polymerase chain reaction (RT-PCR) was performed as previously described (Watanabe et al., 2012). Briefly, total RNA was prepared from PC12 cells using an Agencourt RNAdvance Tissue Kit according to the manufacturer's instructions (Beckman Coulter, La Brea, CA, USA). RNA samples were reverse transcribed by using a High Capacity RNA-to-cDNA Kit (Applied Biosystems, Carlsbad, CA, USA) and the resulting cDNAs were used as templates for qRT-PCR. qRT-PCR was performed with Power SYBR Green PCR Master

Mix (Applied Biosystems); the primers are shown in supplementary information Table S2. As an internal control for normalization, the β-actin transcript was similarly amplified using the primers. qRT-PCR was conducted using a 7300 Real Time PCR System (Applied Biosystems), and the data were acquired and analyzed by the 7300 System SDS software version 1.3.1.21 (Applied Biosystems). The sequences of the primers for the LP genes are shown in Table S2 (Watanabe et al., 2012).

NARX model and data representation

Assuming that the input molecules (*Input*) and output molecules (*Output*) signals satisfy the following NARX model, Eqs. (1) and (2), the system identification is performed by estimating unknown parameters in the NARX model,

$$y_k^{p,s} = \sum_{i=1}^{m_y} a_i^p \, y_{k-i}^{p,s} + \sum_{q \in \mathcal{M}_p} \sum_{j=1}^{m_u} b_j^p f(u_{k-j}^{q,s}, n_p, K_{p'}) \tag{1}$$

$$f(u, n, K) = \frac{u^n}{u^n + K^n}, \tag{2}$$

where $u_k^{q,s}$ and $y_k^{p,s}$ are experimental values of *Input* and *Output* at time step k, p and q respectively denote indices of *Output* and *Input* defined in the following sets,

$$p \in P = \{\text{c-Jun, Egr1, c-Fos, FosB, JunB}, Metrnl, Serpinb1a, Dclk1A\}, \tag{3}$$

$$q \in \mathcal{M}_p \subseteq \mathcal{M} = \{\text{pERK, pCREB, c-Jun, Egr1, c-Fos, FosB, JunB}\}, \tag{4}$$

and s is an index of stimulation conditions defined as follows,

$$s \in \{\text{NGF, PACAP, PMA}\} \tag{5}$$

41

$\mathcal{M}_p$ is the index set of *Input* defined for each *Output* $p \in P$ (Fig. 5A). The nonlinear function $f(x)$ in Eq. (2) is the Hill equation that is one of the steady state solutions of biochemical reaction and widely used in the field of biology (Hill, 1910). The coefficients a_i^p and b_j^p, the orders m_y and m_u in Eq. (1), n_p and K_p in Eq. (2), and set $\mathcal{M}_p$ are unknown parameters. For each molecule under stimulation condition s (NGF, PACAP, and PMA), the unequally spaced time series data are obtained in this study. We consider them as equally spaced time data $u_k^{q,s}$ and $y_k^{p,s}$ with missing time points and identify the unknown NARX parameters after recovering missing time points based on the low rankness of the Hankel-like matrix, which is described in the next section.

Extension of ARX System Identification from Unequally Spaced Time Series Data to the Nonlinear ARX System

To deal with the nonlinear ARX system, we extend ARX system identification from unequally spaced time series data to the nonlinear ARX system. First consider the simple case of the linear ARX model and then extend it to the NARX model. To perform system identification from unequally spaced time series data, equally spaced time series data are generated by signal recovery of unknown values of missing time points. Because the *Input* and *Output* data of a linear system are missing and the order of the system is unknown as in this study, the system identification using the recovered *Input* and *Output* data based on the low rankness of the Hankel-like matrix has been proposed (Liu et al., 2013). This is a method to recover missing data by solving the matrix rank minimization problem and to generate equally spaced sampled data.

In this study, we apply this matrix rank minimization approach to simultaneously identify the NARX model and recover missing data.

First, for simplicity, let us consider the case of the linear ARX model with single *Input* and single *Output* described by

$$y_k = \sum_{i=1}^{m_y} a_i y_{k-i} + \sum_{j=1}^{m_u} b_j u_{k-j} + v_k, \tag{6}$$

where y_k and u_k are the *Output* and *Input* at time step k, and v_k is the noise. When only $\{u_k\}_{k \in \Omega_u}$ and $\{y_k\}_{k \in \Omega_y}$ are obtained, that is, the part of the *Input* and *Output* data $\{u_k\}_{k=1}^{N}$ and $\{y_k\}_{k=1}^{N}$, we consider the problem of recovering unknown *Input* and *Output* data. Here, Ω_u and Ω_y are index sets and are a subset of the set $\{1, 2, \dots, N\}$. We define Hankel-like matrices Y and U by Eqs. (7) and (8), where we assume that N is sufficiently larger than r.

$$Y = \begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_r \\ y_2 & y_3 & y_4 & \cdots & y_{r+1} \\ y_3 & y_4 & y_5 & \cdots & y_{r+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{N-r+1} & y_{N-r+2} & y_{N-r+3} & \cdots & y_N \end{bmatrix} \tag{7}$$

$$U = \begin{bmatrix} u_1 & u_2 & u_3 & \cdots & u_r \\ u_2 & u_3 & u_4 & \cdots & u_{r+1} \\ u_3 & u_4 & u_5 & \cdots & u_{r+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u_{N-r+1} & u_{N-r+2} & u_{N-r+3} & \cdots & u_N \end{bmatrix}. \tag{8}$$

Hankel-like matrices Y and U are matrices called Hankel matrices if they are square matrices, and they are matrices in which the same components are entered from the lower left to the upper right in the matrix. Considering $v_k = 0$ in Eq. (6), that is, considering an ideal case without noise, Eq. (9) holds for the matrix $[Y\ U]$ in which the matrices Y and U are arranged horizontally (Fig. S1B).

$$\mathrm{rank}[Y\ U] = \mathrm{m}_y + r < 2r \tag{9}$$

43

Thus, the matrix $[Y\ U]$ is a low-rank matrix whose rank is determined by the order of the system. If m_y is known in Eq. (9), the missing data can be recovered by restoring the unknown components of the matrix so that the rank of the matrix $[Y\ U]$ becomes $m_y + r$. Because the order m_y is unknown in this study, we recover the unknown components so as to minimize the rank of the matrix $[Y\ U]$ based on the idea that it is better to describe the system with as few parameters as possible. That is, the missing data are recovered by solving the matrix rank minimization problem as follows,

$$\text{Minimize rank}[Y\ U]$$

$$\text{subject to } y_k = \bar{y}_k \text{ for all } k \in \Omega_y \qquad (10)$$

$$u_k = \bar{u}_k \text{ for all } k \in \Omega_u,$$

where $\bar{y}_k$ and $\bar{u}_k$ are observed values. Eq. (10) is a nonconvex optimization problem, which is generally a Non-deterministic Polynomial time (NP)-hard problem in the field of the computational complexity theory. Therefore, we relax this problem in Eq. (11) in which the objective function is replaced by the nucleus norm, the sum of the singular values of the matrix, and obtain a low-rank matrix by solving this optimization problem with the iterative partial matrix shrinkage (IPMS) algorithm (Konishi et al., 2014).

$$\text{Minimize } \|[Y\ U]\|_{*,r}$$

$$\text{subject to } y_k = \bar{y}_k \text{ for all } k \in \Omega_y \qquad (11)$$

$$u_k = \bar{u}_k \text{ for all } k \in \Omega_u,$$

where $\|\cdot\|_{*,r}$ represents the sum of singular values that are smaller than the rth greater singular value. The IPMS algorithm is a technique to provide a low-rank solution of Eq. (10) by solving

44

Eq. (11) repeatedly for increasing r by 1, starting at $r = 0$, and provides recovered data with small energy loss after recovery and less distortion by preferentially estimating from a singular value of a large value (Konishi et al., 2014).

In the case of a multi-*Input* system, for each *Input*, a Hankel-like matrix U_l corresponding to the matrix U is generated, and by solving the matrix rank minimization problem of matrices arrayed side by side such as $[Y\ U_1\ ...\ U_L]$, *Inputs* and *Output* data can be similarly recovered. Also, when data under multiple stimulation conditions are obtained, *Input* and *Output* data can also be recovered by arranging the matrices vertically for each stimulation condition. For example, when there is a data set of NGF stimulation and PACAP stimulation and simulation condition s is $s \in \{\mathrm{NGF}, \mathrm{PACAP}\}$, a matrix composed of y_k^s and u_k^s is vertically arranged for each stimulation condition s to construct Y and U, and *Input* and *Output* data can be recovered by solving the matrix rank minimization problem for $[Y\ U]$.

$$[Y] = \begin{bmatrix} y_1^{NGF} & y_2^{NGF} & y_3^{NGF} & \cdots & y_r^{NGF} \\ y_2^{NGF} & y_3^{NGF} & y_4^{NGF} & \cdots & y_{r+1}^{NGF} \\ y_3^{NGF} & y_4^{NGF} & y_5^{NGF} & \cdots & y_{r+2}^{NGF} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{N-r+1}^{NGF} & y_{N-r+2}^{NGF} & y_{N-r+3}^{NGF} & \cdots & y_N^{NGF} \\ y_1^{PACAP} & y_2^{PACAP} & y_3^{PACAP} & \cdots & y_r^{PACAP} \\ y_2^{PACAP} & y_3^{PACAP} & y_4^{PACAP} & \cdots & y_{r+1}^{PACAP} \\ y_3^{PACAP} & y_4^{PACAP} & y_5^{PACAP} & \cdots & y_{r+2}^{PACAP} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{N-r+1}^{PACAP} & y_{N-r+2}^{PACAP} & y_{N-r+3}^{PACAP} & \cdots & y_N^{PACAP} \end{bmatrix} \tag{12}$$

$$[U] = \begin{bmatrix} u_1^{NGF} & u_2^{NGF} & u_3^{NGF} & \cdots & u_r^{NGF} \\ u_2^{NGF} & u_3^{NGF} & u_4^{NGF} & \cdots & u_{r+1}^{NGF} \\ u_3^{NGF} & u_4^{NGF} & u_5^{NGF} & \cdots & u_{r+2}^{NGF} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u_{N-r+1}^{NGF} & u_{N-r+2}^{NGF} & u_{N-r+3}^{NGF} & \cdots & u_N^{NGF} \\ \\ u_1^{PACAP} & u_2^{PACAP} & u_3^{PACAP} & \cdots & u_r^{PACAP} \\ u_2^{PACAP} & u_3^{PACAP} & u_4^{PACAP} & \cdots & u_{r+1}^{PACAP} \\ u_3^{PACAP} & u_4^{PACAP} & u_5^{PACAP} & \cdots & u_{r+2}^{PACAP} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u_{N-r+1}^{PACAP} & u_{N-r+2}^{PACAP} & u_{N-r+3}^{PACAP} & \cdots & u_N^{PACAP} \end{bmatrix} \tag{13}$$

In the NARX model used in this study, because the observed *Input* data is nonlinearly transformed using the nonlinear function *f* in Eq. (2) and the *Input* data after transformation and the *Output* data follow the ARX system, signal recovery and system identification can be performed on the nonlinearly transformed *Input* data and *Output* data by the above method. Based on this idea, we performed nonlinear ARX system identification.

Procedure for System Identification by Integrating Signal Recovery and the NARX model

To estimate an *I-O* relationship, we prepare data sets of all combinations of input molecules (*Inputs*) for each output molecule (*Output*). For each data set, leave-one-out cross-validation is performed by preparing all combinations with only one test data set and the rest as the training data set. We have three stimulation conditions, NGF, PACAP, and PMA, and use two of them as the training data set and the other one as the test data set. Therefore, there are three combinations to divide the test and training data sets.

In nonlinear systems such as the NARX model in this study, even if all the *Input* and *Output* data are known, obtaining n_p and K_p is a nonconvex optimization problem, for which it is

46

difficult to obtain an exact solution. Therefore, n_p and K_p are estimated by 500 trials with multiple random initial values. By repeating the following procedures from step i to step v, n_p and K_p are estimated so as to minimize the AIC for the training data set, while *Inputs* and *Output* of the NARX model are recovered. Subsequently, signal recovery of the test data set is performed in step vi, and the residual sum of square (RSS) is calculated for a test data set in step vii. Step vii is performed with all three combinations of training and test data sets, and take the sum of RSS for test data sets. Step viii is performed with all combinations of *Input*, and then in step ix a combination of *Input* with the minimum sum of RSS for test data sets is selected. This combination of *Inputs* is used for the *I-O* relationship. Using the combination of the *Input* molecules in step x and the data set of all stimulation conditions as the training data set, we estimate the parameters of the NARX model, which is used as the finally obtained NARX model (Figures 5 and 6).

Step i: Nonlinear transformation of *Input* data by the Hill equation. $u_k^{q,s}$, which is *Input* q at time step k under the stimulation condition s, is transformed into $x_k^{q,s} = f(u_k^{q,s})$ by Eq. (2), the Hill equation. The initial values of n_p and $K_p^{n_p}$ are given by $n_p = 1$ and a uniform random number between 0 to 1, respectively, for each *Input q*. Using the observed *Output* $y_k^{p,s}$ and the nonlinearly transformed *Input* $x_k^{q,s}$, the following Hankel-like matrix is constructed for *Output* p while assigning the previous closest observation value to the initial value of missing points. Note that this is a notation in the case of a single *Input*. Hereafter, two training data sets and one test data set are referred as $training\ 1$ and $training\ 2$ and $test$, respectively.

$$[Y] = \begin{bmatrix} y_1^{training\ 1} & y_2^{training\ 1} & y_3^{training\ 1} & \cdots & y_r^{training\ 1} \\ y_2^{training\ 1} & y_3^{training\ 1} & y_4^{training\ 1} & \cdots & y_{r+1}^{training\ 1} \\ y_3^{training\ 1} & y_4^{training\ 1} & y_5^{training\ 1} & \cdots & y_{r+2}^{training\ 1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{N-r+1}^{training\ 1} & y_{N-r+2}^{training\ 1} & y_{N-r+3}^{training\ 1} & \cdots & y_N^{training\ 1} \\ \\ y_1^{training\ 2} & y_2^{training\ 2} & y_3^{training\ 2} & \cdots & y_r^{training\ 2} \\ y_2^{training\ 2} & y_3^{training\ 2} & y_4^{training\ 2} & \cdots & y_{r+1}^{training\ 2} \\ y_3^{training\ 2} & y_4^{training\ 2} & y_5^{training\ 2} & \cdots & y_{r+2}^{training\ 2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{N-r+1}^{training\ 2} & y_{N-r+2}^{training\ 2} & y_{N-r+3}^{training\ 2} & \cdots & y_N^{training\ 2} \end{bmatrix} \tag{14}$$

$$[U] = \begin{bmatrix} x_1^{training\ 1} & x_2^{training\ 1} & x_3^{training\ 1} & \cdots & x_r^{training\ 1} \\ x_2^{training\ 1} & x_3^{training\ 1} & x_4^{training\ 1} & \cdots & x_{r+1}^{training\ 1} \\ x_3^{training\ 1} & x_4^{training\ 1} & x_5^{training\ 1} & \cdots & x_{r+2}^{training\ 1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N-r+1}^{training\ 1} & x_{N-r+2}^{training\ 1} & x_{N-r+3}^{training\ 1} & \cdots & x_N^{training\ 1} \\ \\ x_1^{training\ 2} & x_2^{training\ 2} & x_3^{training\ 2} & \cdots & x_r^{training\ 2} \\ x_2^{training\ 2} & x_3^{training\ 2} & x_4^{training\ 2} & \cdots & x_{r+1}^{training\ 2} \\ x_3^{training\ 2} & x_4^{training\ 2} & x_5^{training\ 2} & \cdots & x_{r+2}^{training\ 2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N-r+1}^{training\ 2} & x_{N-r+2}^{training\ 2} & x_{N-r+3}^{training\ 2} & \cdots & x_N^{training\ 2} \end{bmatrix} \tag{15}$$

Step ii: Signal recovery of training data. Solve the matrix $[Y\ U]$ rank minimization problem of

Eq. (11) by the IPMS algorithm and recover converted *Input* data $x_k^{q,s}$ and *Output* data $y_k^{p,s}$.

Note that, in the case of multi-*Input*, for each *Input*, a matrix U_l corresponding to the matrix U

is generated, and by solving the matrix rank minimization problem of matrices arrayed side by

side such as $[Y\ U_1 \ldots U_L]$, *Inputs* and *Output* data can be similarly recovered.

Step iii: Calculate ARX parameters, a and b. Based on the relationship between the

48

Hankel-like matrix and ARX parameters (Fig. S1B), obtain the ARX parameters a_i^p and b_j^p in Eq. (1) for *Output* p and each *Input* q using the recovered transformed *Input* data $x_k^{q,s}$ and *Output* data $y_k^{p,s}$. The order of the system, the lag order of the ARX model, is determined based on the matrix rank obtained in step ii.

Step iv: Estimate n_p and $K_p^{n_p}$ using the recovered data and ARX parameters. Using the inverse function *f* of Eq. (2), recover the missing time point data of *Input* before transformation by using Eq. (2). For the recovered $x_k^{q,s}$ and $y_k^{p,s}$, n_p in Eq. (2) is given again by uniform random numbers >1 and ≤100 and $K_p^{n_p}$ ≥0.001 and ≤1, and 200 combinations of n_p and $K_p^{n_p}$ are generated. For each combination, perform simulation of the ARX model and calculate AIC for the training data set, $AIC_{training}$. Select the combination of n_p and $K_p^{n_p}$ with the minimum $AIC_{training}$. Using this n_p and $K_p^{n_p}$, *Input* and *Output* data in the matrix $[Y\ U]$ composed of Y and U in Eqs. (14) and (15) is recovered again by the IPMS algorithm.

Step v: Select NARX parameters with the minimum $AIC_{training}$. Repeat steps i to iv 500 times. Select n_p and $K_p^{n_p}$ and ARX parameters that minimize $AIC_{training}$.

Step vi: Signal recovery of test data. Using the n_p, $K_p^{n_p}$ and ARX parameters selected in step v, add test data to the recovered matrix $[Y\ U]$ in Eqs. (14) and (15) like in Eqs. (16) and (17). Test data are also recovered by solving the test data added matrix $[Y\ U]$ rank minimization problem with the IPMS algorithm. Note that training data sets have already been recovered until

49

step v, and signal recovery of only test data is performed in this step.

$$[Y] = \begin{bmatrix} y_1^{training\ 1} & y_2^{training\ 1} & y_3^{training\ 1} & \cdots & y_r^{training\ 1} \\ y_2^{training\ 1} & y_3^{training\ 1} & y_4^{training\ 1} & \cdots & y_{r+1}^{training\ 1} \\ y_3^{training\ 1} & y_4^{training\ 1} & y_5^{training\ 1} & \cdots & y_{r+2}^{training\ 1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{N-r+1}^{training\ 1} & y_{N-r+2}^{training\ 1} & y_{N-r+3}^{training\ 1} & \cdots & y_N^{training\ 1} \\ \\ y_1^{training\ 2} & y_2^{training\ 2} & y_3^{training\ 2} & \cdots & y_r^{training\ 2} \\ y_2^{training\ 2} & y_3^{training\ 2} & y_4^{training\ 2} & \cdots & y_{r+1}^{training\ 2} \\ y_3^{training\ 2} & y_4^{training\ 2} & y_5^{training\ 2} & \cdots & y_{r+2}^{training\ 2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{N-r+1}^{training\ 2} & y_{N-r+2}^{training\ 2} & y_{N-r+3}^{training\ 2} & \cdots & y_N^{training\ 2} \\ \\ y_1^{test} & y_2^{test} & y_3^{test} & \cdots & y_r^{test} \\ y_2^{test} & y_3^{test} & y_4^{test} & \cdots & y_{r+1}^{test} \\ y_3^{test} & y_4^{test} & y_5^{test} & \cdots & y_{r+2}^{test} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{N-r+1}^{test} & y_{N-r+2}^{test} & y_{N-r+3}^{test} & \cdots & y_N^{test} \end{bmatrix} \qquad (16)$$

$$[U] = \begin{bmatrix} x_1^{training\ 1} & x_2^{training\ 1} & x_3^{training\ 1} & \cdots & x_r^{training\ 1} \\ x_2^{training\ 1} & x_3^{training\ 1} & x_4^{training\ 1} & \cdots & x_{r+1}^{training\ 1} \\ x_3^{training\ 1} & x_4^{training\ 1} & x_5^{training\ 1} & \cdots & x_{r+2}^{training\ 1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N-r+1}^{training\ 1} & x_{N-r+2}^{training\ 1} & x_{N-r+3}^{training\ 1} & \cdots & x_N^{training\ 1} \\ \\ x_1^{training\ 2} & x_2^{training\ 2} & x_3^{training\ 2} & \cdots & x_r^{training\ 2} \\ x_2^{training\ 2} & x_3^{training\ 2} & x_4^{training\ 2} & \cdots & x_{r+1}^{training\ 2} \\ x_3^{training\ 2} & x_4^{training\ 2} & x_5^{training\ 2} & \cdots & x_{r+2}^{training\ 2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N-r+1}^{training\ 2} & x_{N-r+2}^{training\ 2} & x_{N-r+3}^{training\ 2} & \cdots & x_N^{training\ 2} \\ \\ x_1^{test} & x_2^{test} & x_3^{test} & \cdots & x_r^{test} \\ x_2^{test} & x_3^{test} & x_4^{test} & \cdots & x_{r+1}^{test} \\ x_3^{test} & x_4^{test} & x_5^{test} & \cdots & x_{r+2}^{test} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N-r+1}^{test} & x_{N-r+2}^{test} & x_{N-r+3}^{test} & \cdots & x_N^{test} \end{bmatrix} \qquad (17)$$

Step vii: NARX model simulation and calculate RSS_{LOO}^s for test data set s. Simulate the

test data using equally spaced time series data recovered in step vi and parameters of the Hill

equation and ARX parameters. Calculate the RSS_{LOO}^s, the residual sum of square for the

stimulation condition of the test data set.

Step viii: Calculate RSS_{LOO} by taking the sum of RSS_{LOO}^s for each stimulation s. Perform

steps i to vii for all three combinations of training and test data sets. Let RSS_{LOO} be the sum of

RSS_{LOO}^s for each stimulation s of test data set.

Step ix: Obtain the *Input* combination with minimum RSS_{LOO}. Perform steps i to viii for all

combinations of *Inputs*. Select the combination of *Inputs* with the minimum RSS_{LOO} for the *I-O*

relationship.

Step x: Estimate the NARX model with signal recovery using all data sets. Using the

combination of *Input* determined in step ix, estimate the NARX parameter with signal recovery

by the procedure from steps i to v using all stimulation conditions as training data sets.

Note that, when simulating with the ARX model, set the value of *Output* to 0 before time 0,

otherwise the value of the *Output* obtained by the simulation is used to obtain the next time value.

For stimulation in Figure 6, signal recovery was performed by step vi using experimental data

with trametinib as the test data set.

51

Calculation of Gain and Time Constant from the Linear ARX Model

Gain and time constant τ were calculated from the frequency response function obtained from the linear ARX model. For simplicity, we consider here the case of a single *Input* single *Output* ARX model like Eq. (6), which can be re-described as follows,

$$y_k - a_1 y_{k-1} - a_2 y_{k-2} - \cdots - a_{m_y} y_{k-m_y} = b_1 u_{k-1} + b_2 u_{k-2} + \cdots + b_{k-m_u},$$

(18)

and its *Z*-transform are given by

$$\left(y_k - a_1 z^{-1} - a_2 z^{-2} - \cdots - a_{m_y} z^{-m_y}\right) y(z)$$
$$= \left(b_1 z^{-1} + b_2 z^{-2} + \cdots + b_{m_u} z^{-m_u}\right) u(z).$$

(19)

Then a discrete-time transfer function, a function to convert *Input* to *Output* through the system, $G(z)$ can be described using these ARX parameters,

$$G(z) = \frac{y(z)}{u(z)} = \frac{b_1 z^{-1} + b_2 z^{-2} + \cdots + b_{m_u} z^{-m_u}}{1 - a_1 z^{-1} - a_2 z^{-2} - \cdots - a_{m_y} z^{-m_y}} .$$

(20)

To consider the frequency response function and calculation of gain and phase, z is substituted by $i\omega$,

$$G(i\omega) = \frac{ib_1 \omega^{-1} - b_2 \omega^{-2} + \cdots + b_{m_u}(i\omega)^{-m_u}}{1 - ia_1 \omega^{-1} + a_2 \omega^{-2} - \cdots - ia_{m_y}(i\omega)^{-m_y}}$$

(21)

$$gain = |G(i\omega)|, phase = \angle G(i\omega)$$

(22)

where i is an imaginary unit and ω is frequency. Therefore, gain and phase can be calculated from ARX parameters. The frequency response curve and phase diagram at each *Input* and *Output* of the identified linear ARX model are shown in Figure S4. Note that gain indicated in Table 1 is steady-state gain. From the frequency response function, cutoff frequency f_{cutoff}, an

inverse of time constant τ, is obtained by calculating the frequency at which the gain

corresponds to $\frac{1}{\sqrt{2}}$ of the steady-state gain. Because Eq. (23) is established between f_{cutoff} and

the time constant τ, we can obtain τ from the ARX parameters through the above procedure.

$$\tau = \frac{1}{2\pi f_{cutoff}} \tag{23}$$

Simulation of the Integrated NARX Model

The simulation of the integrated NARX model was performed as follows. Experimental and

recovered data of pERK and pCREB, and the simulated data of c-Jun, c-Fos, Egr1, FosB, and

JunB were given as *Input* data and simulation was performed using the NARX model in Figure

5.

AR model
$$y_k = \sum_{j=1}^{r} a_j y_{k-j}$$

Low rank matrix
$$\text{rank}[\bar{Y}] < r$$

Original signals
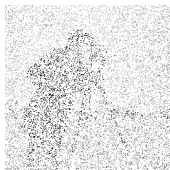
Recovered signals

$$\text{rank}[\bar{Y}] \to \min$$

y_i^* : measured signals
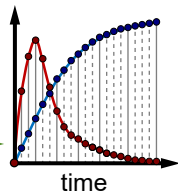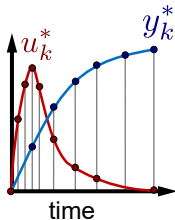
$\bar{y}_i$: missing signals

$$\bar{Y} = \begin{bmatrix} y_1^* & \bar{y}_2 & y_3^* & \cdots & \bar{y}_r \\ \bar{y}_2 & y_3^* & \bar{y}_4 & \cdots & y_{r+1}^* \\ y_3^* & \bar{y}_4 & \bar{y}_5 & \cdots & \bar{y}_{r+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_N^* & \bar{y}_{N+1} & \bar{y}_{N+2} & \cdots & y_{N+r-1}^* \end{bmatrix}$$

u_i^* : measured signals

$\bar{u}_i$: missing signals

$$\bar{U} = \begin{bmatrix} \bar{u}_1 & u_2^* & u_3^* & \cdots & u_r^* \\ u_2^* & u_3^* & \bar{u}_4 & \cdots & \bar{u}_{r+1} \\ u_3^* & \bar{u}_4 & \bar{u}_5 & \cdots & \bar{u}_{r+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u_N^* & \bar{u}_{N+1} & u_{N+2}^* & \cdots & \bar{u}_{N+r-1} \end{bmatrix}$$

u_k^* y_k^*
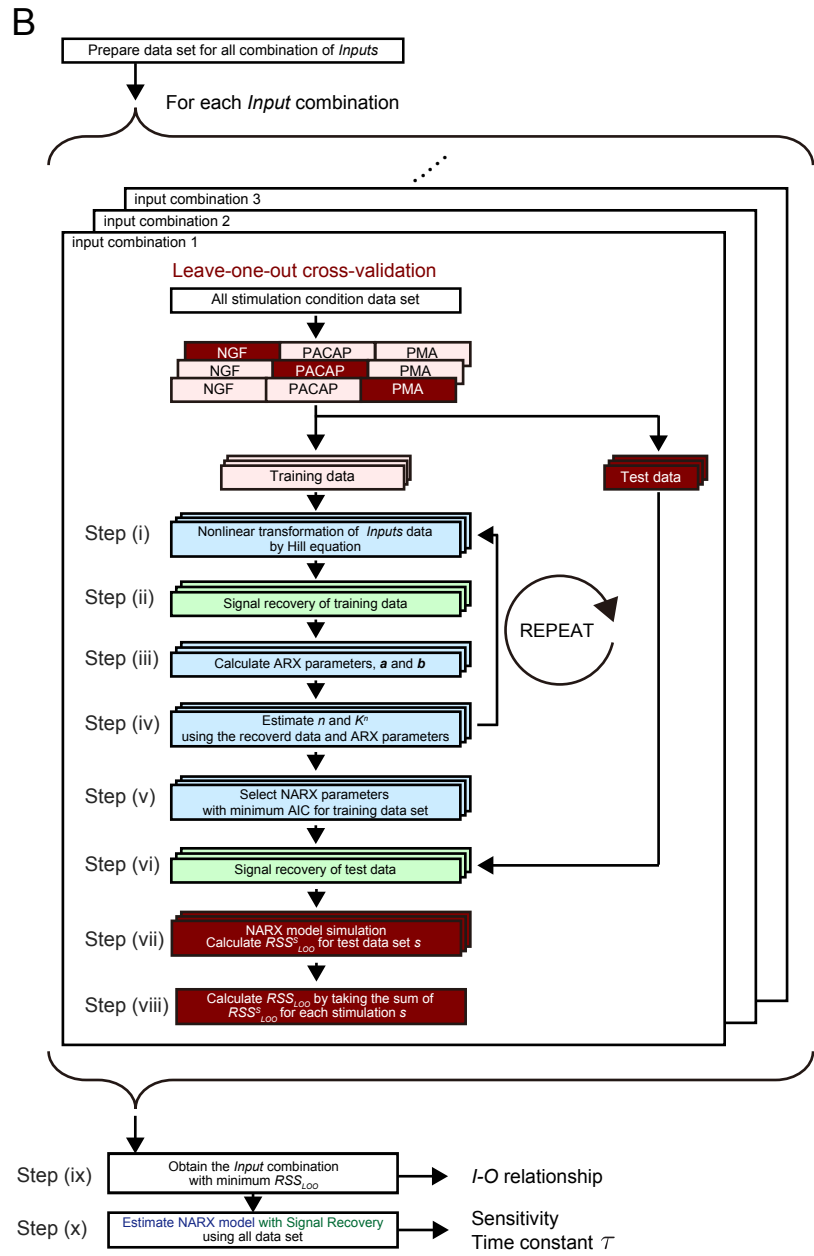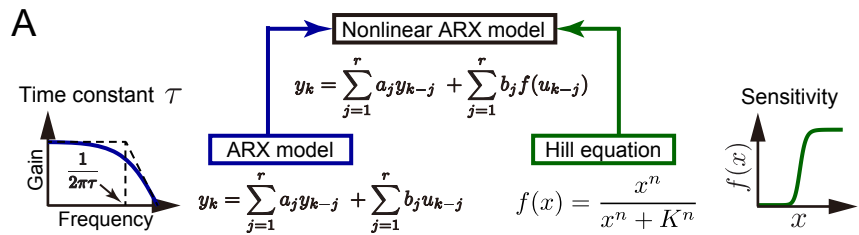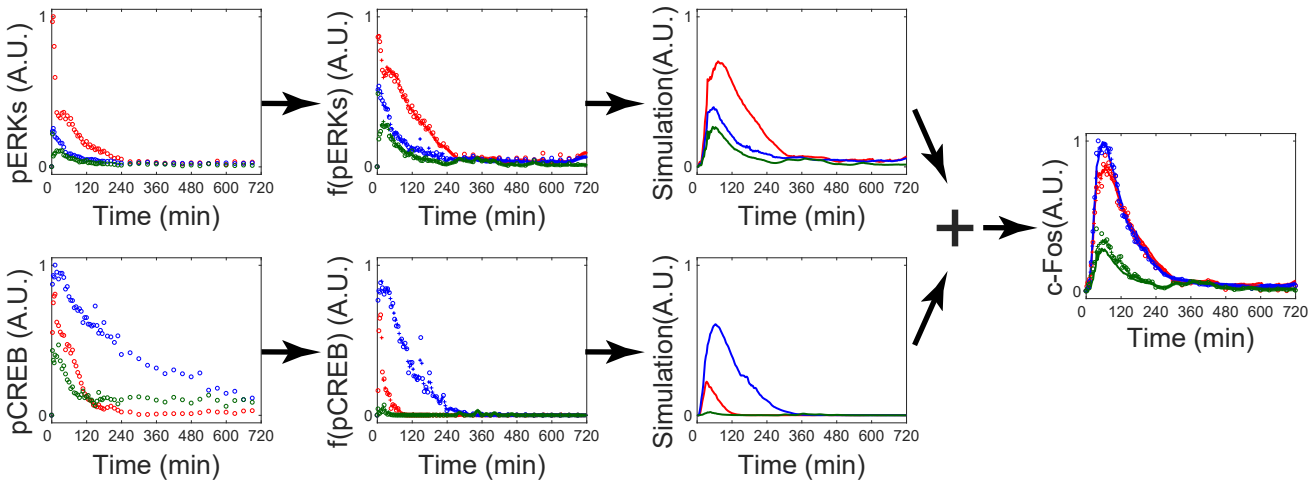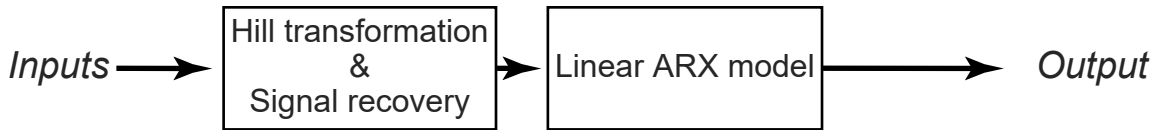
time

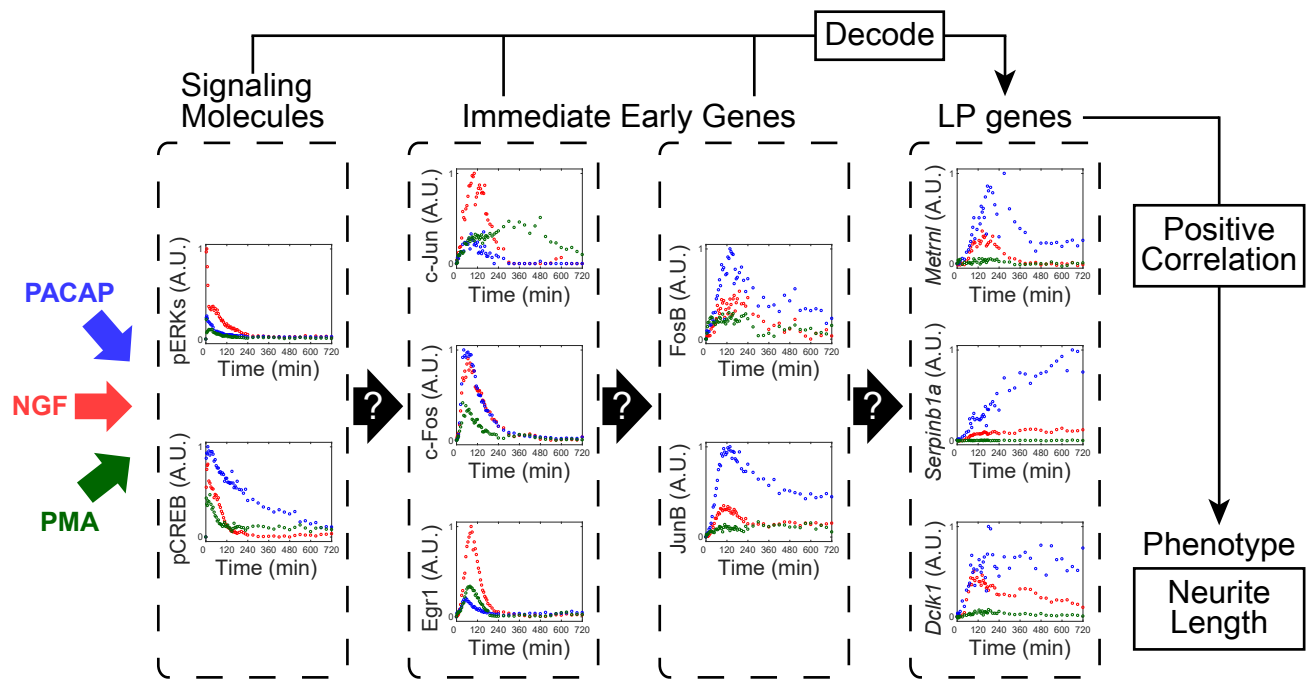$$\text{rank}[\bar{Y} \ \bar{U}] \to \min$$

time

ARX model
$$y_k = \sum_{j=1}^{r} a_j y_{k-j} + \sum_{j=1}^{r} b_j u_{k-j}$$

Low rank matrix
$$\text{rank}[\bar{Y} \ \bar{U}] < 2r$$

A

Time constant τ

Gain $\frac{1}{2\pi\tau}$

Frequency

Nonlinear ARX model

$$y_k = \sum_{j=1}^{r} a_j y_{k-j} + \sum_{j=1}^{r} b_j f(u_{k-j})$$

ARX model

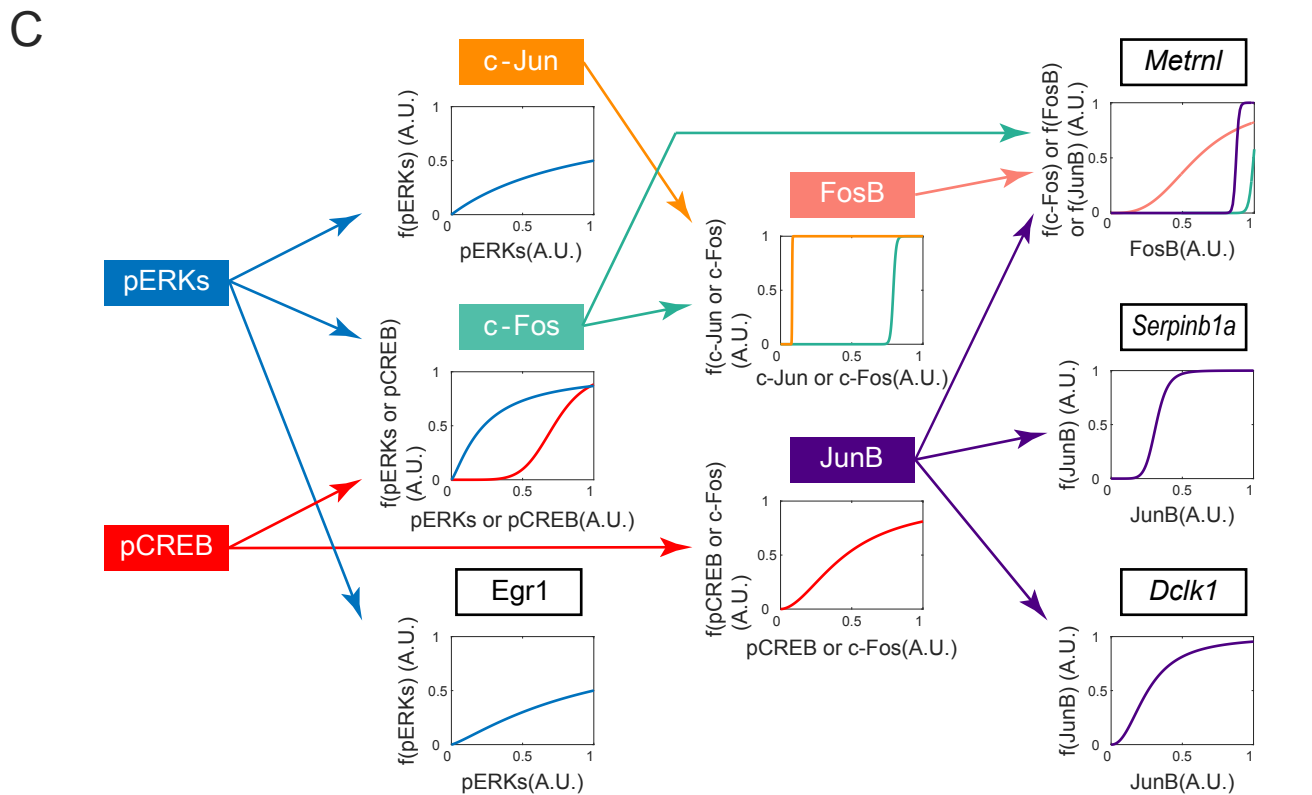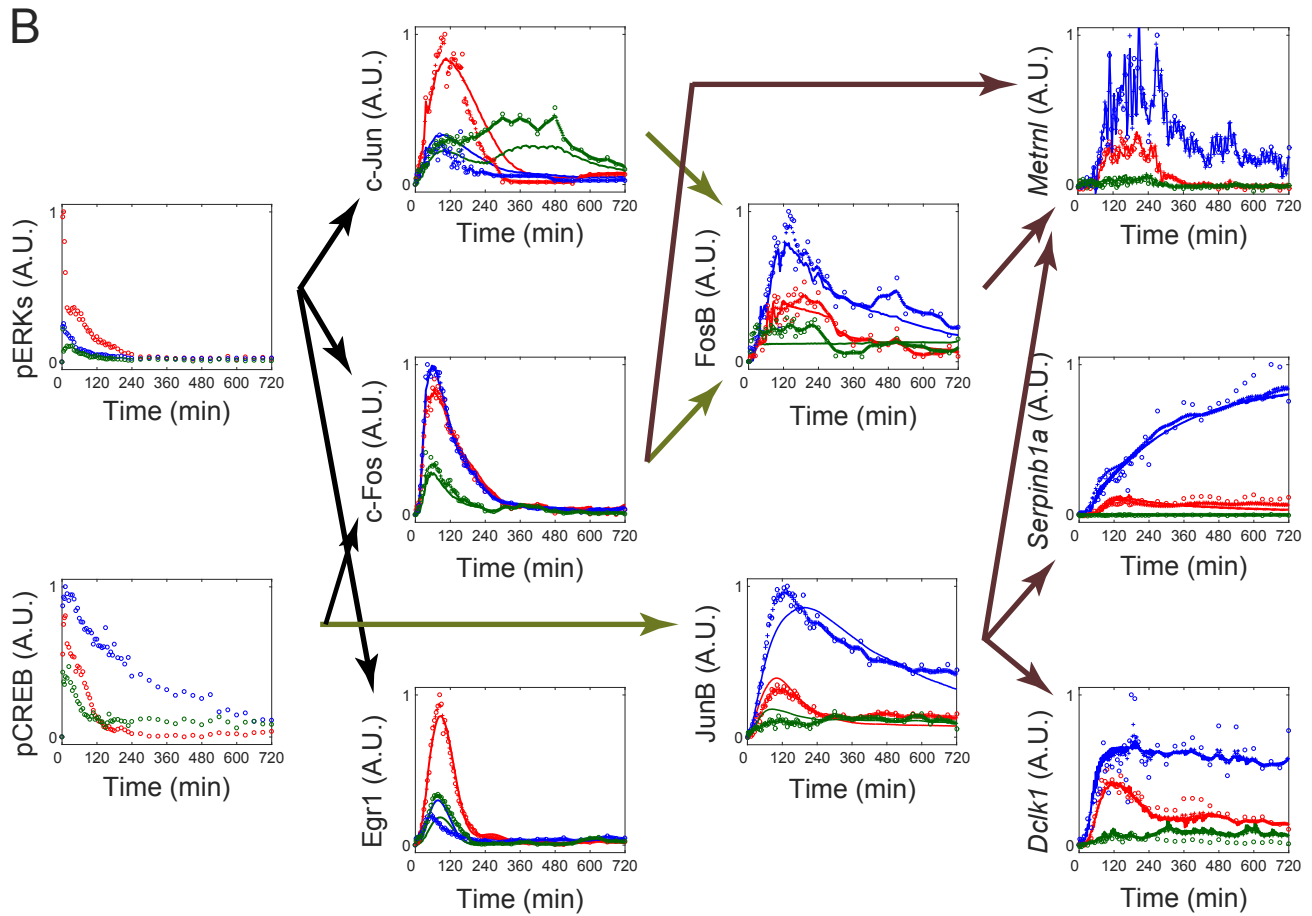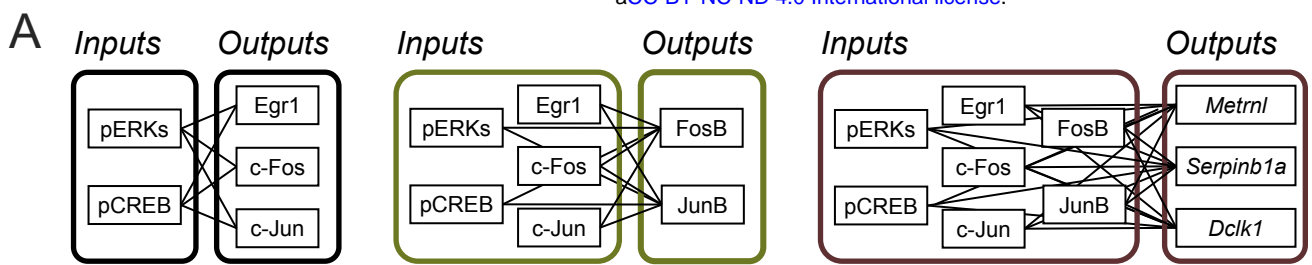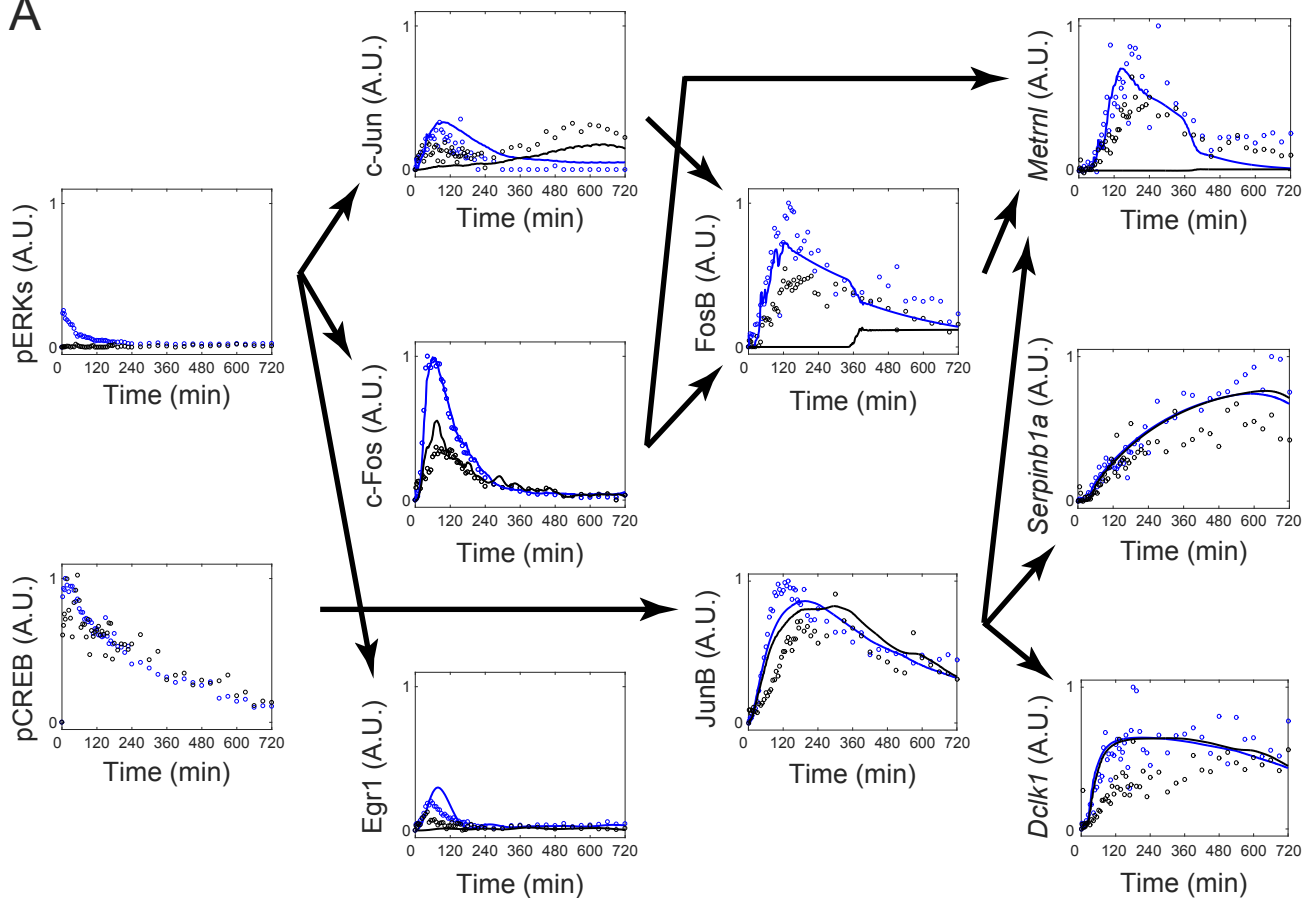$$y_k = \sum_{j=1}^{r} a_j y_{k-j} + \sum_{j=1}^{r} b_j u_{k-j}$$

Hill equation

$$f(x) = \frac{x^n}{x^n + K^n}$$

Sensitivity

$f(x)$

x

B

Prepare data set for all combination of *Inputs*

For each *Input* combination

input combination 3

input combination 2

input combination 1

Leave-one-out cross-validation

All stimulation condition data set

NGF	PACAP	PMA
NGF	PACAP	PMA
NGF	PACAP	PMA

Training data

Test data

Step (i) — Nonlinear transformation of *Inputs* data by Hill equation

Step (ii) — Signal recovery of training data

Step (iii) — Calculate ARX parameters, **a** and **b**

Step (iv) — Estimate n and K^n using the recoverd data and ARX parameters

REPEAT

Step (v) — Select NARX parameters with minimum AIC for training data set

Step (vi) — Signal recovery of test data

Step (vii) — NARX model simulation Calculate RSS^s_{LOO} for test data set s

Step (viii) — Calculate RSS_{LOO} by taking the sum of RSS^s_{LOO} for each stimulation s

Step (ix) — Obtain the *Input* combination with minimum RSS_{LOO} → *I-O* relationship

Step (x) — Estimate NARX model with Signal Recovery using all data set → Sensitivity Time constant τ

Inputs → Hill transformation & Signal recovery → Linear ARX model → Output
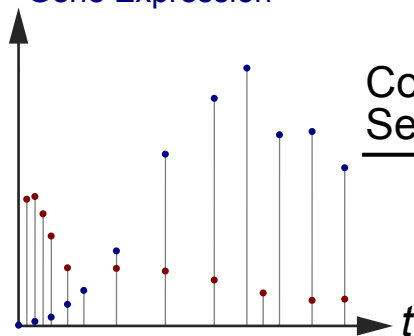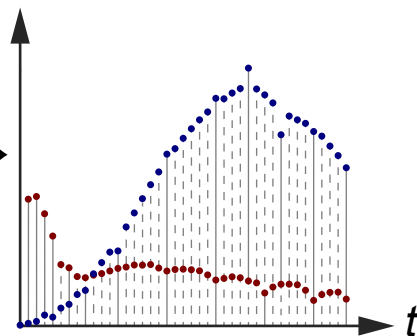
Experimental data

Signaling Molecules
Gene Expression

Recovered data

Compressed
Sensing

t

t

System
Identification

1. *I-O* relationship

pERK → c-Jun
pERK → c-Fos
c-Jun → FosB
c-Fos → FosB
FosB → *Metrnl*
c-Jun → *Metrnl*
pCREB → c-Fos
pCREB → JunB
pCREB → Egr1
JunB → *Serpinb1a*
JunB → *Dclk1*
JunB → *Metrnl*

2. Sensitivity

f(Input)

Input

3. Time constant τ

Gain

$\dfrac{1}{2\pi\tau}$

Frequency