PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Synthesizing developmental trajectories

Paul Villoutreix[1☯], Joakim Andén[2☯], Bomyi Lim[1,3], Hang Lu[4,5], Ioannis G. Kevrekidis[2,3], Amit Singer[2,6], Stanislav Y. Shvartsman[1,2]**

1 Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, United States of America, **2** Program in Applied and Computational Mathematics, Princeton University, Princeton, New Jersey, United States of America, **3** Department of Chemical and Biological Engineering, Princeton University, Princeton, New Jersey, United States of America, **4** School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, Georgia, United States of America, **5** Interdisciplinary Program in Bioengineering, Georgia Institute of Technology, Atlanta, Georgia, United States of America, **6** Department of Mathematics, Princeton University, Princeton, New Jersey, United States of America

☯ These authors contributed equally to this work.
* stas@princeton.edu

Abstract

Dynamical processes in biology are studied using an ever-increasing number of techniques, each of which brings out unique features of the system. One of the current challenges is to develop systematic approaches for fusing heterogeneous datasets into an integrated view of multivariable dynamics. We demonstrate that heterogeneous data fusion can be successfully implemented within a semi-supervised learning framework that exploits the intrinsic geometry of high-dimensional datasets. We illustrate our approach using a dataset from studies of pattern formation in *Drosophila*. The result is a continuous trajectory that reveals the joint dynamics of gene expression, subcellular protein localization, protein phosphorylation, and tissue morphogenesis. Our approach can be readily adapted to other imaging modalities and forms a starting point for further steps of data analytics and modeling of biological dynamics.

Author summary

A wide range of problems in biology require analysis of multivariable dynamics in space and time. As a rule, the multiscale nature and complexity of real systems precludes simultaneous monitoring of all the relevant variables, and multivariable dynamics must be synthesized from partial views provided by different experimental techniques. We present a formal framework for accomplishing this task in the context of imaging studies of pattern formation in developing tissues.

Introduction

The need to synthesize data from different observations into coherent multivariable trajectories is discussed in multiple contexts, from physics to social sciences, but systematic approaches for accomplishing this task have yet to be established [1–5]. Here we address this

OPEN ACCESS

Citation: Villoutreix P, Andén J, Lim B, Lu H, Kevrekidis IG, Singer A, et al. (2017) Synthesizing developmental trajectories. PLoS Comput Biol 13(9): e1005742. https://doi.org/10.1371/journal.pcbi.1005742

Editor: Reka Albert, Pennsylvania State University, UNITED STATES

Received: July 3, 2017

Accepted: August 25, 2017

Published: September 18, 2017

Data Availability Statement: All our data are held publicly available on the github repository: https://github.com/paulvill/data-fusion-images.

Funding: PV was partially supported by the WIN program between Princeton University and the Weizmann Institute. SYS was partially supported by R01GM107103 award from the NIGMS. SYS and IGK were supported by the NSF (https://www.nsf.gov/) award MCB 1516970. IGK and AS were supported by NSF (https://www.nsf.gov/) award CMMI-1310173. JA and AS are supported by Simons Investigator Award and Simons

Collaboration on Algorithms and Geometry from Simons Foundation (https://www.simonsfoundation.org/), and the Moore Foundation Data-Driven Discovery Investigator Award (https://www.moore.org/home) to AS. HL was supported by the GM088333 award from the NIH (https://www.nih.gov/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

task for imaging studies of developing tissues, where patterns of cell fates are established by complex regulatory networks [6–8]. Advances in live imaging continue to provide new insights into the dynamics of individual components in these networks, but imaging more than three reporters at the same time is still challenging and limited to model genetic organisms [9, 10]. Furthermore, in the absence of reliable live reporters, dynamics of some state variables can only be inferred from fixed tissues. Because of these limitations, extracting the multivariable dynamics from the heterogeneous datasets collected by imaging of live and fixed tissues becomes a non-trivial task [11, 12].

The problem can be illustrated by an imaging dataset from the early *Drosophila* embryo (Fig 1A and 1B), a model system in which a graded profile of the nuclear localization of transcription factor Dorsal (Dl) establishes the dorsoventral (DV) stripes of gene expression that control cell fates and tissue deformations [13–15]. Current mechanisms of the DV patterning system invoke multiple state variables, such as the levels of gene expression and protein phosphorylation [16] (Fig 1C). These mechanisms were elucidated in studies that reveal only a small subset of the full state space, most commonly 2-3 variables per experiment. Can these partial views be fused into a consistent multivariable trajectory? This is a general question that applies to essentially all developmental systems.

We realized that this question can be addressed by casting the task of data fusion as a matrix completion problem (Fig 1D). Specifically, an image of a fixed embryo or a frame from a live imaging movie can be viewed as a column in a matrix where rows correspond to the relevant variables, such as developmental time or the level of gene expression at a given position. Because of limitations in the number of states that can be accessed simultaneously, the matrix is incomplete. For example, live imaging of gastrulation provides information about nuclear positions as a function of time, but is silent about the levels of gene expression. On the other hand, an image of a fixed embryo reveals the distribution of an active enzyme but has no direct temporal information. Thus, multivariable data fusion requires completing this matrix, filling in the missing components by estimates informed by the rest of the data. Below we show how this task can be accomplished by solving a suitably posed semi-supervised learning problem. We first provide a closed-form solution to this problem and then demonstrate its successful performance on synthetic and experimental datasets.

Results

Semi-supervised learning framework for matrix completion

We assume here that all experiments contain a common variable, which is sufficient to determine all other variables that can be measured or to be predicted. For instance, this variable is revealed by a signal that reports positions of nuclei. This means that the first row in the matrix is complete. To complete other rows, we must establish the mappings between the common variable and each of the target variables. These mappings can be found within a semi-supervised learning framework, in which the values of the variables in the incomplete rows are estimated from a training dataset [17, 18].

As an example, consider images from fixed embryos that are stained to reveal the spatial pattern of an active enzyme, visualized using a phosphospecific antibody (Fig 2). They provide labeled data points that contain information about the common variable and a specific target variable. On the other hand, images without this staining, such as the frames from live imaging of morphogenesis, provide unlabeled data points with only the common variable. By finding a mapping between the common and target variables, we can essentially "color" the frames of a live imaging movie by snapshots of molecular patterns from fixed embryos.

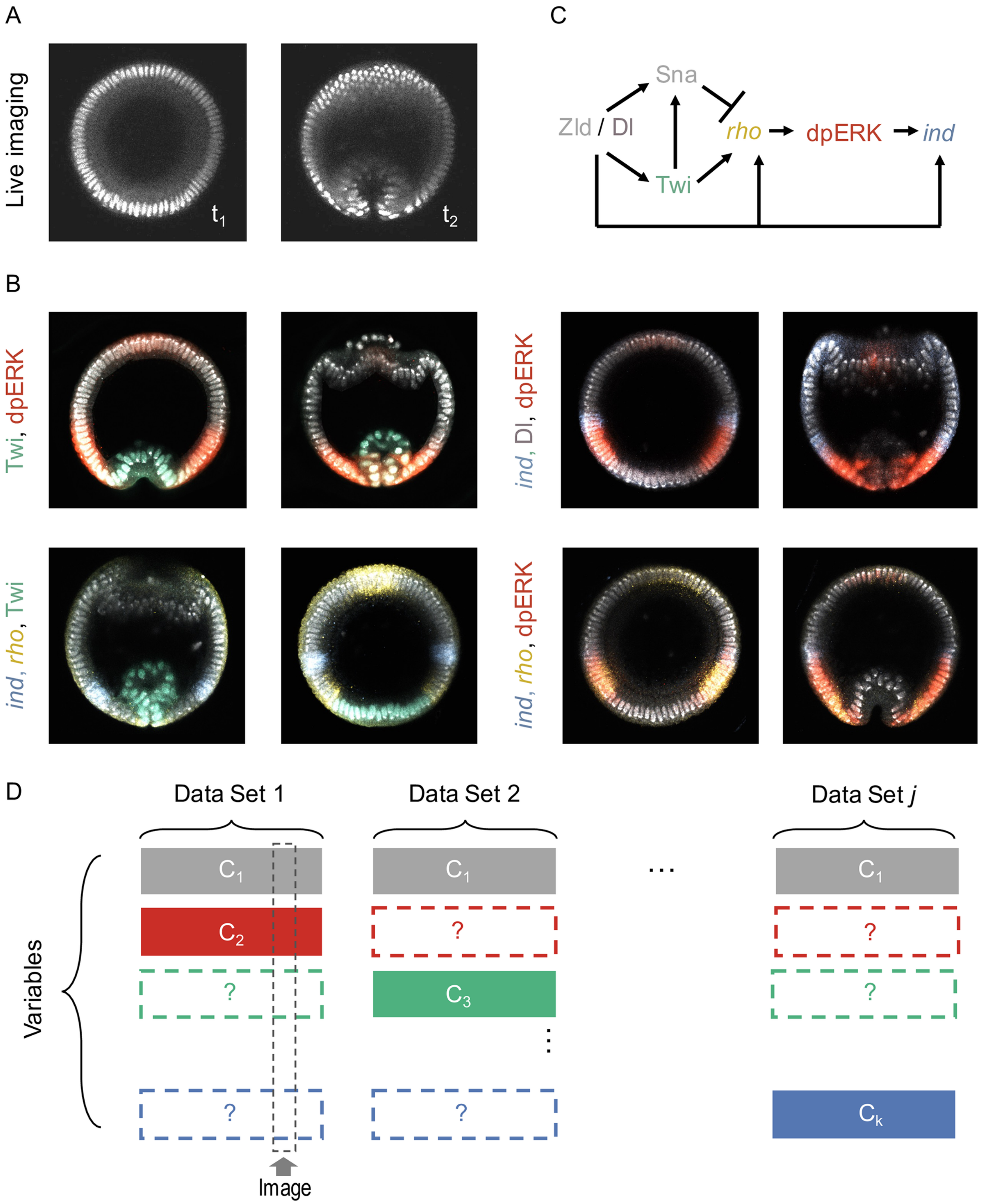

Fig 1. Stating the problem of data fusion. (A-B) Example datasets of molecular signals and morphology during the DV patterning of *Drosophila* embryo; all images are collected from optical cross-section along the DV axis, ∼ 15% from the posterior pole of the embryo. (A) Frames from a live imaging movie, showing positions of nuclei during the early stages of gastrulation. (B) Images of fixed embryos, stained with probes and antibodies revealing the spatial patterns of nuclear Dl (pink), Twi (green), dually phosphorylated ERK (red), and transcripts of *rho* (yellow) and *ind* (blue). (C) A fragment of a DV patterning network in the early *Drosophila* embryo. (D) Data fusion as a

matrix completion problem: Each row corresponds to a variable, e.g. nuclear positions, gene expression levels, time stamp, revealed by visualizing different molecular or cellular components, nuclei, transcripts, or protein phosphorylation. Each column of the matrix corresponds to an image giving access to some of the states through various channels. The remaining states, labeled with a question mark, must be estimated from other datasets.

https://doi.org/10.1371/journal.pcbi.1005742.g001

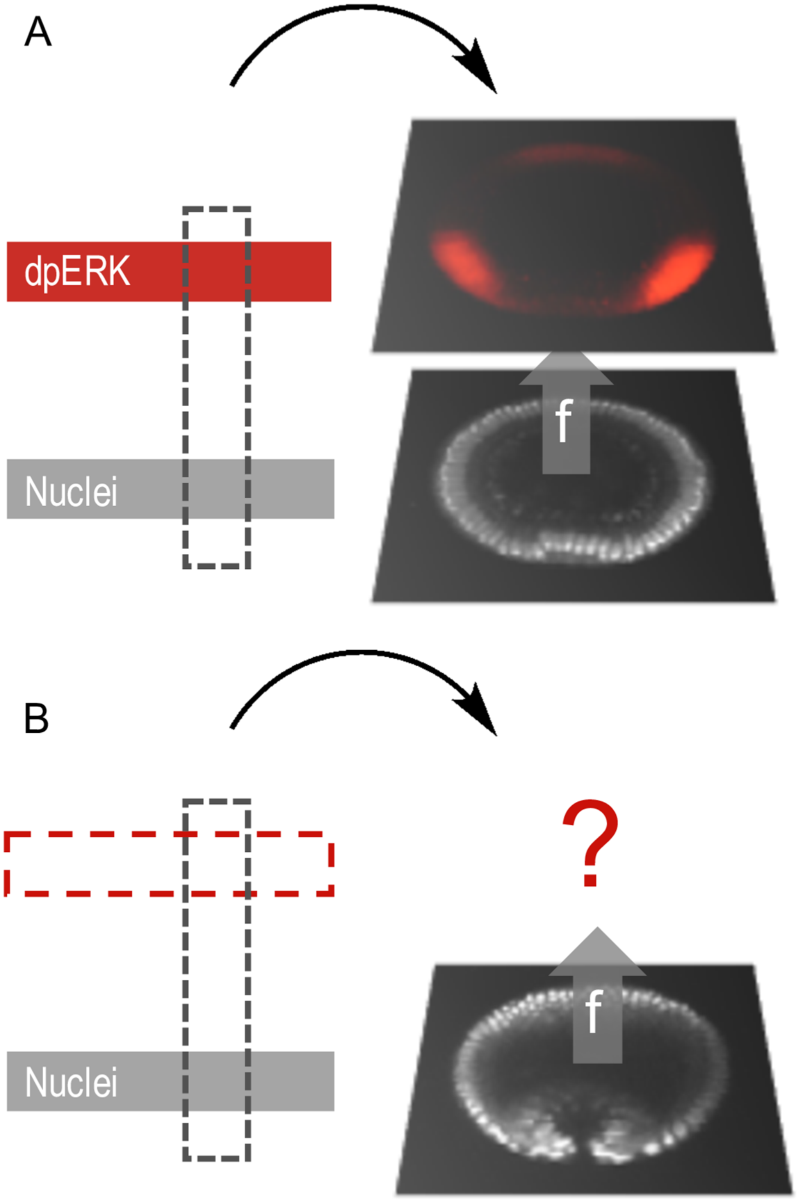

Fig 2. Learning a mapping from a common channel. An experimental image can be decomposed into various channels. E.g., red: dpERK, visualized with a phosphospecific antibody, gray: nuclei, visualized through either DAPI (in fixed images) or Histone-RFP (in live imaging). The training ensemble of labeled images (A) is used to predict the labels on a set of unlabeled images (B) using common information, the morphology obtained through the nuclei signal in this case. Morphological proximity yields similar labels.

https://doi.org/10.1371/journal.pcbi.1005742.g002

A critical assumption in finding the mappings is that the multivariable dynamics of the patterning process are both low-dimensional and smooth with respect to the underlying parameters. This assumption is supported by studies with mathematical models of specific biological systems and by computational analysis of datasets from imaging studies of development [19, 20]. More formally, we consider a set of data points $(x_1, \ldots, x_l, x_{l+1}, \ldots, x_{l+u})$ belonging to a space $\mathcal{X}$. These points correspond to the values of the common variable in the complete row. On the other hand, a row corresponding to any one of the target variables is incomplete. The values in the filled columns of this row are called labels. These are denoted $(y_1, \ldots, y_l)$ and belong to a target space $\mathcal{Y}$. The semi-supervised learning techniques transfer the information contained in the labeled data points $((x_1, y_1), \ldots, (x_l, y_l))$ to the unlabeled points $(x_{l+1}, \ldots, x_{l+u})$, while preserving the intrinsic structure of the dataset [18]. Stated otherwise, these techniques *learn* the mapping $y = f(x)$ assuming that the considered process is smooth, which means that similar values of x give rise to similar values of $f(x)$.

The missing values, corresponding to the unlabeled data points in each of the incomplete rows, are found by solving the following optimization problem:

$$\vec{f} = \underset{\substack{\vec{f} \in \mathcal{Y}^{l+u} \\ \forall i \leq l, f_i = y_i}}{\operatorname{argmin}} \sum_{i,j=1}^{l+u} w_{i,j} \| f_i - f_j \|^2 \qquad (1)$$

where $\vec{f} = (f_1, \ldots, f_{l+u})$ are the values of the target variable on the data points $(x_1, \ldots, x_{l+u})$, the considered norm in $\mathcal{Y}$ is the Euclidean distance and the weights $w_{i,j}$ represent the similarity between two data points x_i and x_j. The norm in the space $\mathcal{X}$ can for example be the Euclidean distance in the space where each dimension corresponds to an image pixel or some coordinate in an arbitrary feature transform of that image. Other distances are possible. For example, the one-norm (or L^1 distance) can be used, which increases robustness to outliers in the data. That being said, as long as these distances preserve the low-dimensional manifold structure, they will yield similar results as the number of points goes to infinity.

This quadratic optimization problem, known as harmonic extension, has a unique solution that relates the unlabeled data points $f_{l+1}, \ldots, f_{l+u}$ to the labels $y_1, \ldots, y_l$ where $\mathcal{Y} = \mathbb{R}$ [17, 21]. The explicit solution reads:

$$\vec{f}^u = (D_u - W_{uu})^{-1} W_{ul} Y \qquad (2)$$

where $Y = (y_1, \ldots, y_l)$ and $\vec{f}^u = (f_{l+1}, \ldots, f_{l+u})$, and $d_i = \sum_{j=1}^{l+u} w_{i,j}$, $D_u = \operatorname{diag}(d_{l+1}, \ldots, d_{l+u})$, $W_{uu} = (w_{i,j})_{l+1 \leq i,j \leq l+u}$, and $W_{ul} = (w_{i,j})_{\substack{l+1 \leq i \leq l+u \\ 1 \leq j \leq l}}$ (S1 Text).

Illustrative example

To illustrate our method, we considered a one-dimensional nonlinear trajectory in a three-dimensional space. The trajectory is given by the set of equations

$$\begin{cases} x^{(1)}(t) & = at\left(\cos(bt) + \epsilon^{(1)}\right) \\ x^{(2)}(t) & = at\left(\sin(bt) + \epsilon^{(2)}\right) \\ y(t) & = ct\exp(-d(t-e)^2) \end{cases} \qquad (3)$$

where a, b, c, d, e are constants, $\epsilon^{(1)}$ and $\epsilon^{(2)}$ are Gaussian noise sources and t is a real-valued parameter. The set of points $(x^{(1)}(t), x^{(2)}(t))$ forms a one-dimensional non-linear manifold embedded in the two dimensional plane and it is parameterized by t. These points are analogs of the embryo morphology. In the absence of noise, this mapping from t to the 2D plane can be inverted as $t = \frac{1}{|a|}\sqrt{(x^{(1)})^2(t) + (x^{(2)})^2(t)}$. The signal $y(t)$ is a smooth function of t and is thus a smooth function of $(x^{(1)}, x^{(2)})$ by composition. In this example, y corresponds to the target modality that we would like to estimate.

To mimic the setting of data fusion with three modalities, $((x^{(1)}, x^{(2)}), t, y)$, we consider the following situation: suppose that one acquires a set of labeled points, i.e. a set of l triplets, $(((x^{(1)}(t_1), x^{(2)}(t_1)), y(t_1)), \ldots, ((x^{(1)}(t_l), x^{(2)}(t_l)), y(t_l)))$ and a set of u unlabeled, but time-stamped, points, $(((x^{(1)}(t_{l+1}), x^{(2)}(t_{l+1})), t_{l+1}), \ldots, ((x^{(1)}(t_{l+u}), x^{(2)}(t_{l+u})), t_{l+u}))$, as shown in Fig 3A and 3B. The pairwise similarity measures $w_{i,j}$ are computed using Euclidean norm between pairs of data points $(x^{(1)}(t_i), x^{(2)}(t_i))$ and $(x^{(1)}(t_j), x^{(2)}(t_j))$. In this case, there are no outliers, so the standard Euclidean distance is well suited and there is no need to consider other distance measures.

Then, using Eq (2) it is possible to estimate $y = f(x)$ on the set of unlabeled data points using the harmonic extension algorithm. The results are shown in Fig 3C. We then directly obtain y as a function of t by composition using the known time stamps $(t_{l+1}, \ldots, t_{l+u})$.

The accuracy of the estimated multivariable dynamics can be assessed using a K-fold validation strategy on the labeled samples (S1 Fig and Materials and methods). For the chosen set of parameters and the size of the dataset, the error is $\sim 1\%$. As expected for the semi-supervised learning framework, the error decreases with the addition of new unlabeled data points. This example demonstrates how the proposed approach successfully recovers multivariable dynamics from heterogeneous datasets that combine continuous views for part of the state variables and snapshots that report several states without direct temporal information.

Fusion of imaging datasets

As a representative dataset from imaging studies of multivariable dynamics in living systems, we use a collection of $\sim$ 1000 images each of which reveals the spatial position of the nuclei and either a timestamp or the distribution of one or several components of the DV patterning network (Fig 1D). To apply the semi-supervised learning approach to data fusion to this dataset we need to compute pairwise similarities between the images using the common channel. Prior to this, we took several preprocessing steps that aim to minimize image variability associated to sample handling, microscope calibration and imaging. First, the images were registered to align their ventral-most points. The images were then resized and cropped such that the embryos occupy 80% of the image. All images were resized to 100 by 100 pixels. To overcome local variations of image intensity, we computed a local average using a Gaussian kernel, and then renormalized the image by that value. We also applied a logistic function to the images to handle contrast variability, S2 Fig.

Most importantly, to ensure that pairwise differences between images are insensitive to small translations or deformations, we applied the scattering transform [22] and compared the resulting transform vectors. The scattering transform of an image is a signal representation obtained by alternating wavelet decompositions and pointwise modulus operators. We found that second-order scattering coefficients with an averaging scale of 64 pixels provided sufficient invariance. These are computed using the ScatNet toolbox [23, 24]. The result is a vector of dimension 784 for each image. The point clouds corresponding to each of the 11 datasets were centered separately. It has been shown that the Euclidean distance on the scattering transform is locally invariant to translation and stable to deformation of the original image [22]. For this reason, we compare these 784-dimensional vectors using the Euclidean norm. The corresponding low-dimensional manifold on which the data points lie is shown on S4 Fig.

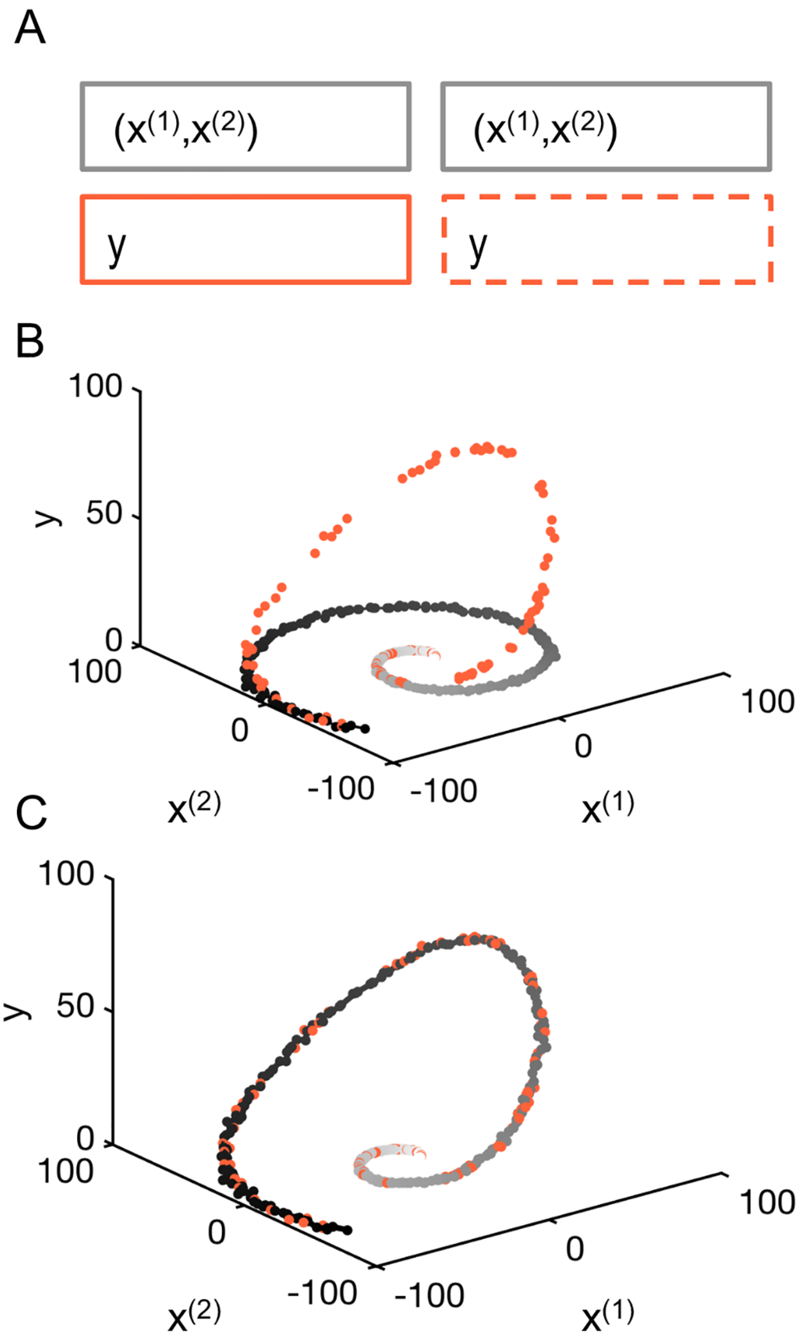

Fig 3. Illustrative example. (A) Matrix formulation of the problem with 120 labeled samples, $(((x^{(1)}(t_1), x^{(2)}(t_1), y(t_1)), \ldots, ((x^{(1)}(t_l), x^{(2)}(t_l), y(t_l)))$, and 300 unlabeled samples $(((x^{(1)}(t_{l+1}), x^{(2)}(t_{l+1})), t_{l+1}), \ldots, ((x^{(1)}(t_{l+u}), x^{(2)}(t_{l+u})), t_{l+u}))$. (B) The points are distributed on a non-linear 1-dimensional manifold in the $(x^{(1)}, x^{(2)})$ - plane. Some points, the snapshots, contain a value for the signal. (C) Result of the interpolation on the nonlinear manifold using the harmonic extension algorithm.

https://doi.org/10.1371/journal.pcbi.1005742.g003

For each of the 512x512 pixels of each live movie frames, there is a common channel reporting the nuclei spatial position and there are 5 channels that we would like to complete. These channels contain the information about the spatial distributions of one enzyme (dpERK), two transcription factors (Twist and Dorsal), and transcripts of two genes (*ind* and *rho*). We thus

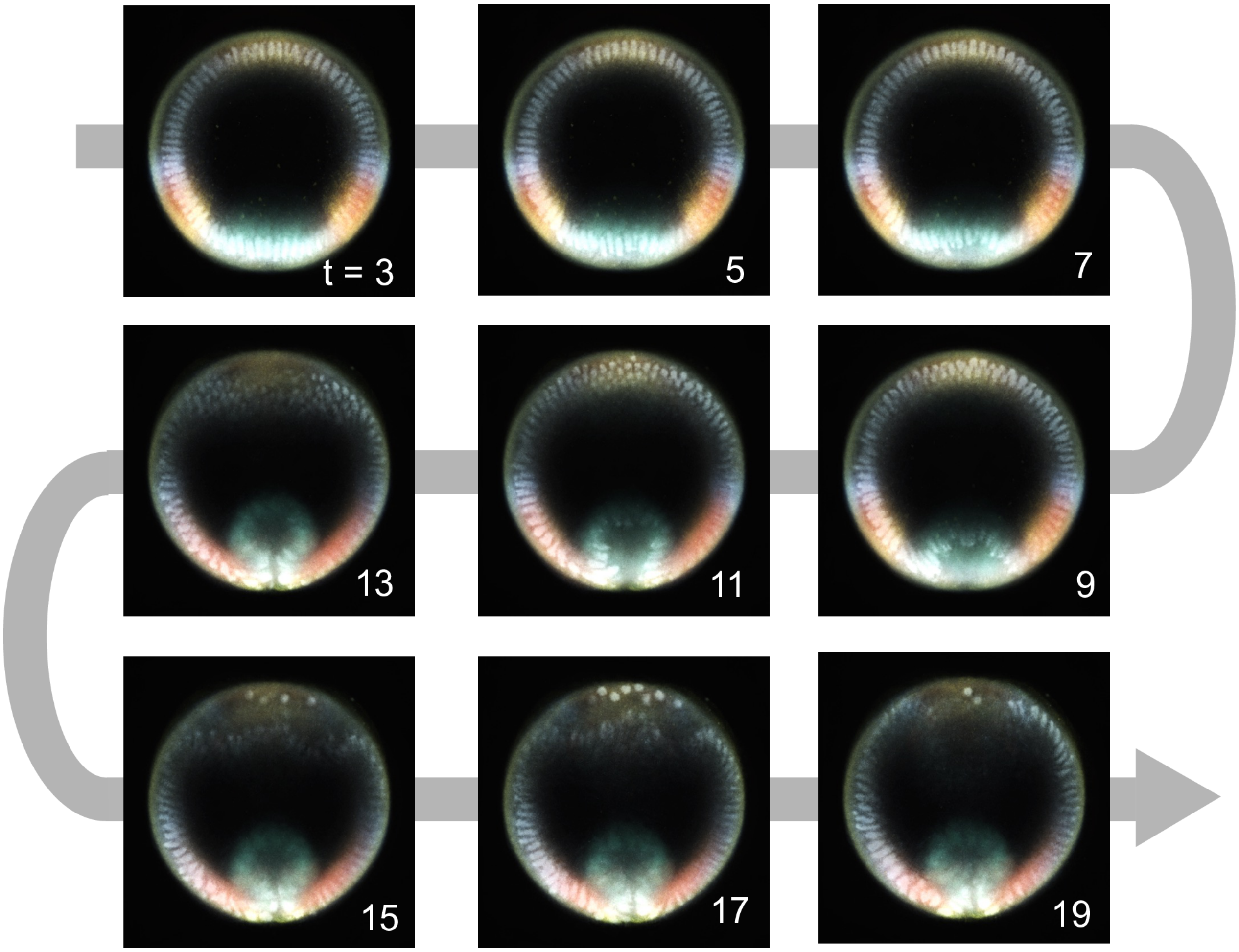

Fig 4. Colored movie frames obtained with our data fusion algorithm. The temporal resolution is 2 min, extracted from a 30 s resolution movie. The time stamps on the images are in min and indicate elapsed time from the start of the live movie. The colors correspond to dpERK (red), Dl (pink), *rho* (yellow), *ind* (blue), Twi (green).

https://doi.org/10.1371/journal.pcbi.1005742.g004

solved the data fusion problem for each pixel and each channel, leading to 5x512x512 semi-supervised learning solutions. The combination of labeled and unlabeled datasets is described on S2 Table. The result is a multivariable trajectory for the joint dynamics of tissue shape and five molecular components within the regulatory network that patterns the DV axis of the embryo (Fig 4). To evaluate the accuracy of the method, we computed the cross-validation error for each pixel and averaged over the entire images. We found that the normalized absolute error is of 0.9–2.5% of the signal range when considering the various modalities of the entire experimental datasets (S3 Table). We show how the algorithm performs on several examples in Fig 5.

Discussion

We presented a formal approach to synthesizing developmental trajectories. By posing the task of data fusion as a semi-supervised learning problem, we obtained a closed-form expression for the estimated values of all variables using harmonic extension. The reconstructed trajectories provide the basis for the more advanced mechanistic studies of multivariable processes

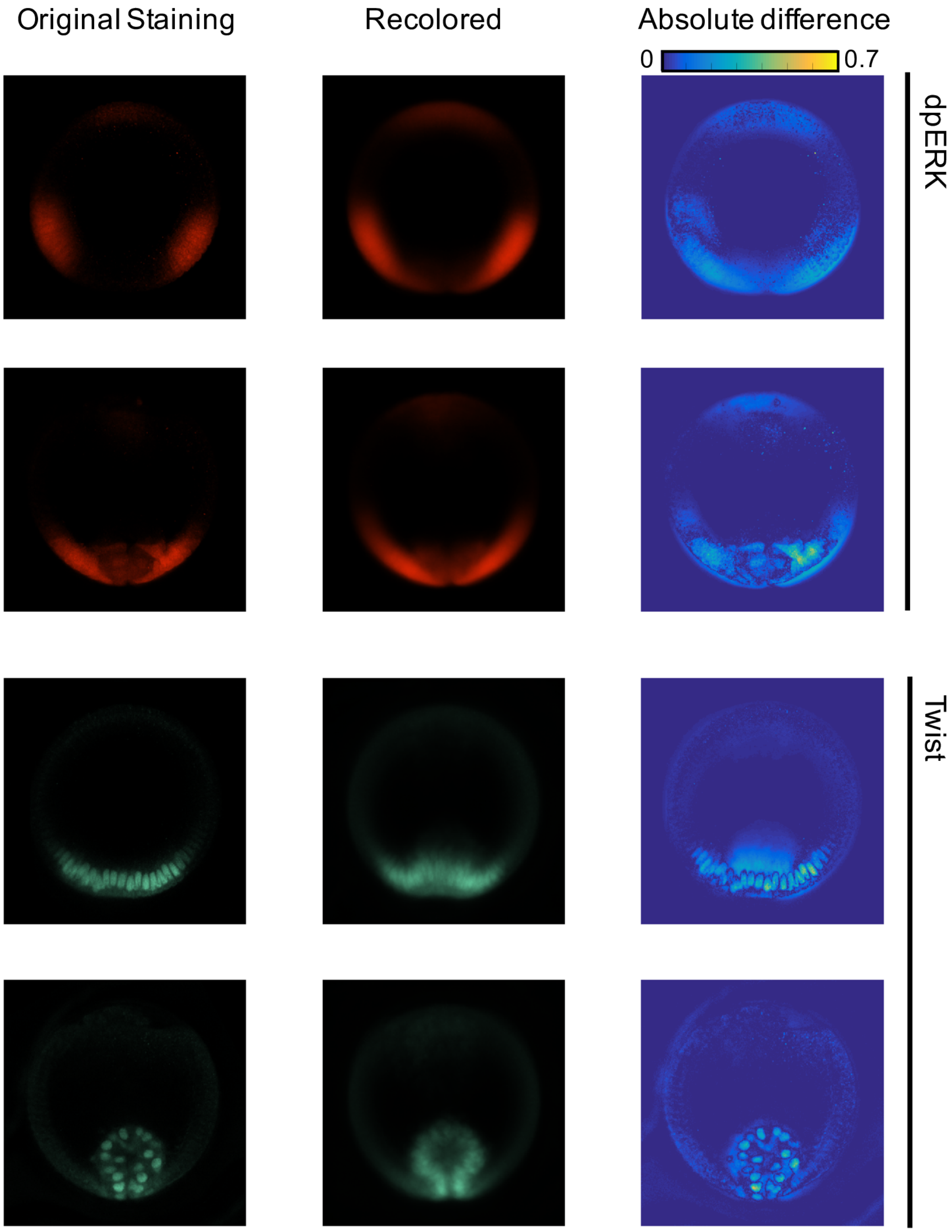

Fig 5. Examples showing how the algorithm performs on four fixed samples. The first column shows the original measurement, the second column shows the result of recoloring the snapshot through K-fold cross validation, and the third column shows the absolute difference between the original and recolored images, normalized by the signal range.

https://doi.org/10.1371/journal.pcbi.1005742.g005

responsible for the highly reproducible dynamics of developmental pattern formation. Our approach can also be extended using other semi-supervised learning methods [25], if the dimensionality of the intrinsic geometry is greater than one or if there is no unique common channel among all experiments.

Most of the previous attempts to accomplishing this task explored specific features of developmental systems, such as the expression level of a particular gene, and used a discrete number of temporal classes, usually defined in ad hoc way [16, 26]. Our approach reconstructs continuous time dynamics and relies on the intrinsic geometry of multidimensional datasets. Some limitations might appear when considering fluorescent reporters for intrinsically variable processes, and thus not smooth, such as MS2 reporters for nascent transcripts [27]. However, our method is readily applicable to datasets stored in established public databases of gene expression patterns such as the BDGP Resources [28] or the FlyEx database [29] and could serve to animate other pathways such as the segmentation cascade in the early fly development.

We conclude by pointing out two directions for the future extensions and applications of the presented approach. First, while there are no conceptual limitations in using the presented matrix completion framework to studies of pattern formation and morphogenesis problems in three dimensions [30], it is important to increase the computational efficiency of our approach, which can be done at multiple levels, starting with dimensionality reduction at the preprocessing step. At the same time, for a large class of patterning processes that happen on the surfaces of epithelial sheets, one can use the recently developed "tissue cartography" approach to first flatten the three-dimensional images [5], which should make our approach directly applicable. Second, following the step of data fusion, one can attempt to model the observed multivariable dynamics. Here one can employ several modeling methodologies, from mechanistic modeling of specific molecular and tissue-level processes [31–35], to equation-free approaches, which aim to deduce the underlying mechanisms directly from data [36, 37].

Materials and methods

Extended Materials and Methods are presented in S1 Text.

Image datasets

All images are cross-sections of *Drosophila* embryos taken at $\sim 90\mu$m from the posterior pole. Time-lapse movies were obtained using a Nikon A1-RS confocal microscope with a 60x Plan-Apo oil objective. The nuclei were stained with Histone-RFP. A total of 7 movies was acquired with a time resolution of 30 seconds per frame. All movies start about 2.5 hr after fertilization and end after about 20 min after gastrulation starts (about 3.3 hr after fertilization). Four datasets of fixed images were acquired to visualize nuclei, protein expression of dpERK, Twist, and Dorsal, and mRNA expression of ind and rho. Immunostaining and fluorescent in situ hybridization protocols were used as described before [16]. DAPI (1:10,000; Vector laboratories) was used to visualize nuclei. Rabbit anti-dpERK (1:100; Cell Signaling), mouse anti-Dorsal (1:100; DSHB), rat anti-Twist (1:1000; gift from Eric Wieschaus, Princeton University), sheep anti-digoxigenin (1:125; Roche), and mouse anti-biotin (1:125; Jackson Immunoresearch) were used as primary antibodies. Alexa Fluor conjugates (1:500; Invitrogen) were used as secondary antibodies. Stained embryos were imaged using Nikon A1-RS confocal microscope with a 60x Plan-Apo oil objective. Embryos were mounted in a microfluidic device for end-on imaging, as described previously [16, 38]. The first dataset contains 108 images stained with rabbit anti-dpERK and rat anti-Twist antibodies. The second dataset contains 59 images stained with mouse anti-Dorsal antibody, rabbit anti-dpERK antibody, and ind-DIG probe. The third dataset contains 58 images stained with ind-biotin probe, rho-DIG probe, and rabbit-dpERK antibody. The fourth dataset contains 30 images stained with rat anti-Twist antibody, ind-biotin probe, and rho-DIG probe. The distribution of the datasets as labeled and unlabeled data depending on the considered variable is summarized on S2 Table. Raw images can be found in

Supplementary Files on the public github repository https://github.com/paulvill/data-fusion-images, see S2 Text.

The affinity matrix

The affinity matrix $W = (w_{i,j})$ is computed using a Gaussian kernel $w_{i,j} = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma_i \sigma_j}\right)$ with scaling parameters σ_i and σ_j computed locally as the average of the distance with respect to the 10 closest neighbors as described in S1 Text. We used the Euclidean norm in the space of scattering transformed data. The resulting affinity matrix is shown on S3 Fig. The corresponding underlying one-dimensional manifold is shown on S4 Fig.

Computing the cross validation error

The K-fold cross validation error was computed by extracting subsamples of the labeled data points and the semi-supervised learning framework was used to predict the value of the labels on them. For the image datasets, we computed the absolute error between the actual value of pixel intensity to the predicted one. The absolute error was then normalized by the range of the signal computed from the entire set of images for a given channel. The number of bins K was chosen so that the number artificially unlabeled data points was about 20. The results for each dataset are shown in S3 Table and described in S1 Text.

Movie coloring

The result of data fusion led to multimodal time lapses of developing embryo showing nuclei and the spatio-temporal dynamics of dpERK, Dl, *rho*, *ind*, and Twi. The images were colored using the color code shown in S4 Table, i.e. dpERK (red), Dl (pink), *rho* (yellow), *ind* (blue), Twi (green). A resulting colored movie is provided in Supplementary Files 2.

Code implementation

The semi-supervised framework used to accomplish the task of data fusion is completely implemented in the open-source MATLAB library and fully runs in GNU Octave. It is available as Supplementary Software on the public github repository https://github.com/paulvill/data-fusion. See S2 Text for a description of the main components of the library.

Supporting information

S1 Fig. K-fold cross validation on the illustrative example. A) Setting with $K = 5$, there are 120 labeled points and the number of unlabeled points varies from 0 to 300. B) The normalized absolute error as a function of the number of unlabeled points. There are 100 repetitions for each number of unlabeled data points.
(TIF)

S2 Fig. Illustration of the image preprocessing steps applied on the nuclei channel. The first line shows images resulting from rotation and centering steps. The second line shows images resulting from intensity renormalization. The third line shows images resulting from contrast increase. The first two columns show early and later stages from movie frames stained with Histone-RFP. The last two columns represent early and later stages from fixed samples stained with DAPI.
(TIF)

S3 Fig. Affinity matrix $W = (w_{i,j})$ obtained by comparing images as described by equation (6) in S1 Text is shown as a heatmap. The white squares identify each of the 11 datasets. The first 7 correspond to live movies, the last 4 correspond to the datasets of fixed images.
(TIF)

S4 Fig. Low-dimensional embedding of the 11 datasets obtained by diffusion maps. Each dot is a point and each color is a different dataset. The top left panel shows the points obtained by embedding the points in the first three diffusion map coordinates. The top right panel shows the data points in the plane formed by the first two diffusion map coordinates, while the two bottom panels show the embedding in the planes obtained with the first and third (left) or second and third (right) diffusion map coordinates. Some outliers were filtered out for visualization purposes if their closest neighbor distance was at least twice the median closest neighbor distance, leading to a very well-defined 1-dimensional manifold.
(TIF)

S1 Table. Values of the parameters for intensity renormalization and contrast increase for each of the experimental datasets (S1 Text).
(PDF)

S2 Table. Distribution of the datasets into labeled and unlabeled sets depending on the modality. We refer to $\Omega(m)$ as the set of labeled datapoints, while $\overline{\Omega(m)}$ is the set of unlabeled data points for the mth modality.
(PDF)

S3 Table. Normalized Absolute Error obtained by K-fold cross-validation for each modality of each dataset. In each case, we performed 10 repetitions, where the labeled samples are distributed randomly among the K bins, and the 309 unlabeled data points are chosen randomly. The error is then averaged over 10 repetitions. More details about the Normalized Absolute Error can be found in S1 Text.
(PDF)

S4 Table. Color scheme used to color the final movie.
(PDF)

S1 Text. Detailed description of the semi-supervised learning framework and its applications to the illustrative example and the experimental datasets.
(PDF)

S2 Text. Detailed description of the supplementary software and the supplementary files.
(PDF)

Acknowledgments

We thank Carmeline Dsilva, Mahim Misra, Roy Lederman and David Sroczynski for helpful discussions.

Author Contributions

Conceptualization: Paul Villoutreix, Joakim Andén, Ioannis G. Kevrekidis, Amit Singer, Stanislav Y. Shvartsman.

Data curation: Bomyi Lim.

Formal analysis: Paul Villoutreix, Joakim Andén.

Funding acquisition: Ioannis G. Kevrekidis, Amit Singer, Stanislav Y. Shvartsman.

Investigation: Paul Villoutreix, Joakim Andén, Ioannis G. Kevrekidis, Amit Singer, Stanislav Y. Shvartsman.

Methodology: Paul Villoutreix, Joakim Andén, Ioannis G. Kevrekidis, Amit Singer, Stanislav Y. Shvartsman.

Project administration: Stanislav Y. Shvartsman.

Resources: Hang Lu, Ioannis G. Kevrekidis, Amit Singer, Stanislav Y. Shvartsman.

Software: Paul Villoutreix, Joakim Andén.

Supervision: Ioannis G. Kevrekidis, Amit Singer, Stanislav Y. Shvartsman.

Validation: Paul Villoutreix, Joakim Andén.

Visualization: Paul Villoutreix, Joakim Andén.

Writing – original draft: Paul Villoutreix, Joakim Andén, Stanislav Y. Shvartsman.

Writing – review & editing: Paul Villoutreix, Joakim Andén, Ioannis G. Kevrekidis, Amit Singer, Stanislav Y. Shvartsman.

References

1. Meijering E, Carpenter AE, Peng H, Hamprecht FA, Olivo-Marin JC. Imagining the future of bioimage analysis. Nature Biotechnology. 2016; 34(12):1250–1255. https://doi.org/10.1038/nbt.3722 PMID: 27926723
2. Crosetto N, Bienko M, Van Oudenaarden A. Spatially resolved transcriptomics and beyond. Nature Reviews Genetics. 2015; 16(1):57–66. https://doi.org/10.1038/nrg3832 PMID: 25446315
3. de Bakker BS, de Jong KH, Hagoort J, de Bree K, Besselink CT, de Kanter FE, et al. An interactive three-dimensional digital atlas and quantitative database of human development. Science. 2016; 354(6315):aag0053. https://doi.org/10.1126/science.aag0053 PMID: 27884980
4. Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, et al. A multi-modal parcellation of human cerebral cortex. Nature. 2016; 536(7615):171–178. https://doi.org/10.1038/nature18933 PMID: 27437579
5. Heemskerk I, Streichan SJ. Tissue cartography: compressing bio-image data by dimensional reduction. Nature methods. 2015; 12(12):1139. https://doi.org/10.1038/nmeth.3648 PMID: 26524242
6. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature biotechnology. 2014; 32(4):381–386. https://doi.org/10.1038/nbt.2859 PMID: 24658644
7. Haghverdi L, Buettner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nature Methods. 2016; 13(10):845–848. https://doi.org/10.1038/nmeth.3971 PMID: 27571553
8. Rizvi AH, Camara PG, Kandror EK, Roberts TJ, Schieren I, Maniatis T, et al. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. Nature Biotechnology. 2017; 35(6):551–560. https://doi.org/10.1038/nbt.3854 PMID: 28459448
9. Cutrale F, Trivedi V, Trinh LA, Chiu CL, Choi JM, Artiga MS, et al. Hyperspectral phasor analysis enables multiplexed 5D in vivo imaging. Nature. 2017; 201:7.
10. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nature biotechnology. 2015; 33(5):495–502. https://doi.org/10.1038/nbt.3192 PMID: 25867923
11. Thompson CL, Ng L, Menon V, Martinez S, Lee CK, Glattfelder K, et al. A high-resolution spatiotemporal atlas of gene expression of the developing mouse brain. Neuron. 2014; 83(2):309–323. https://doi.org/10.1016/j.neuron.2014.05.033 PMID: 24952961
12. Castro-González C, Luengo-Oroz MA, Duloquin L, Savy T, Rizzi B, Desnoulez S, et al. A digital framework to build, visualize and analyze a gene expression atlas with cellular resolution in zebrafish early embryogenesis. PLoS Comput Biol. 2014; 10(6):e1003670. https://doi.org/10.1371/journal.pcbi.1003670 PMID: 24945246

13. Lim B, Levine M, Yamakazi Y. Transcriptional pre-patterning of Drosophila gastrulation. Current Biology. 2017; 27(2):286–290. https://doi.org/10.1016/j.cub.2016.11.047 PMID: 28089518

14. Martin AC, Kaschube M, Wieschaus EF. Pulsed contractions of an actin–myosin network drive apical constriction. Nature. 2009; 457(7228):495–499. https://doi.org/10.1038/nature07522 PMID: 19029882

15. Gilmour D, Rembold M, Leptin M. From morphogen to morphogenesis and back. Nature. 2017; 541(7637):311–320. https://doi.org/10.1038/nature21348 PMID: 28102269

16. Lim B, Dsilva CJ, Levario TJ, Lu H, Schüpbach T, Kevrekidis IG, et al. Dynamics of inductive ERK signaling in the Drosophila embryo. Current Biology. 2015; 25(13):1784–1790. https://doi.org/10.1016/j.cub.2015.05.039 PMID: 26096970

17. Zhu X, Ghahramani Z, Lafferty JD. Semi-supervised learning using gaussian fields and harmonic functions. In: Proceedings of the 20th International conference on Machine learning (ICML-03); 2003. p. 912–919.

18. Belkin M, Niyogi P. Semi-supervised learning on Riemannian manifolds. Machine learning. 2004; 56(1–3):209–239. https://doi.org/10.1023/B:MACH.0000033120.25363.1e

19. Surkova S, Spirov AV, Gursky VV, Janssens H, Kim AR, Radulescu O, et al. Canalization of gene expression in the Drosophila blastoderm by gap gene cross regulation. PLoS Biol. 2009; 7(3): e1000049. https://doi.org/10.1371/journal.pbio.1000049 PMID: 19750121

20. Dsilva CJ, Lim B, Lu H, Singer A, Kevrekidis IG, Shvartsman SY. Temporal ordering and registration of images in studies of developmental dynamics. Development. 2015; 142(9):1717–1724. https://doi.org/10.1242/dev.119396 PMID: 25834019

21. Zhu X. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences. 2005;.

22. Mallat S. Group Invariant Scattering. Comm Pure Appl Math. 2012; 65(10):1331–1398. https://doi.org/10.1002/cpa.21413

23. Sifre L, Mallat S. Rotation, scaling and deformation invariant scattering for texture discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2013. p. 1233–1240.

24. Andén J, Mallat S. Deep scattering spectrum. IEEE Transactions on Signal Processing. 2014; 62(16): 4114–4128. https://doi.org/10.1109/TSP.2014.2326991

25. Moscovich A, Jaffe A, Boaz N. Minimax-optimal semi-supervised regression on unknown manifolds. In: Singh A, Zhu J, editors. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. vol. 54 of Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR; 2017. p. 933–942. Available from: http://proceedings.mlr.press/v54/moscovich17a.html.

26. Fowlkes CC, Hendriks CLL, Keränen SV, Weber GH, Rübel O, Huang MY, et al. A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm. Cell. 2008; 133(2):364–374. https://doi.org/10.1016/j.cell.2008.01.053 PMID: 18423206

27. Fukaya T, Lim B, Levine M. Enhancer control of transcriptional bursting. Cell. 2016; 166(2):358–368. https://doi.org/10.1016/j.cell.2016.05.025 PMID: 27293191

28. Hammonds AS, Bristow CA, Fisher WW, Weiszmann R, Wu S, Hartenstein V, et al. Spatial expression of transcription factors in Drosophila embryonic organ development. Genome biology. 2013; 14(12): R140. https://doi.org/10.1186/gb-2013-14-12-r140 PMID: 24359758

29. Poustelnikova E, Pisarev A, Blagov M, Samsonova M, Reinitz J. A database for management of gene expression data in situ. Bioinformatics. 2004; 20(14):2212–2221. https://doi.org/10.1093/bioinformatics/bth222 PMID: 15059825

30. Royer LA, Lemon WC, Chhetri RK, Wan Y, Coleman M, Myers EW, et al. Adaptive light-sheet microscopy for long-term, high-resolution imaging in living organisms. Nature biotechnology. 2016; 34(12): 1267–1278. https://doi.org/10.1038/nbt.3708 PMID: 27798562

31. Goyal Y, Jindal GA, Pelliccia JL, Yamaya K, Yeung E, Futran AS, et al. Divergent effects of intrinsically active MEK variants on developmental Ras signaling. Nature Genetics. 2017; 49(3):465–469. https://doi.org/10.1038/ng.3780 PMID: 28166211

32. Félix MA, Barkoulas M. Pervasive robustness in biological systems. Nature Reviews Genetics. 2015; 16(8):483–496. https://doi.org/10.1038/nrg3949 PMID: 26184598

33. Lei J, Levin SA, Nie Q. Mathematical model of adult stem cell regeneration with cross-talk between genetic and epigenetic regulation. Proceedings of the National Academy of Sciences. 2014; 111(10): E880–E887. https://doi.org/10.1073/pnas.1324267111

34. Kicheva A, Bollenbach T, Ribeiro A, Valle HP, Lovell-Badge R, Episkopou V, et al. Coordination of progenitor specification and growth in mouse and chick spinal cord. Science. 2014; 345(6204):1254927. https://doi.org/10.1126/science.1254927 PMID: 25258086

35. Wunderlich Z, DePace AH. Modeling transcriptional networks in Drosophila development at multiple scales. Current opinion in genetics & development. 2011; 21(6):711–718. https://doi.org/10.1016/j.gde.2011.07.005

36. Yair O, Talmon R, Coifman RR, Kevrekidis IG. Reconstruction of normal forms by learning informed observation geometries from data. Proceedings of the National Academy of Sciences. 2017; p. 201620045. https://doi.org/10.1073/pnas.1620045114

37. Brunton SL, Proctor JL, Kutz JN. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proceedings of the National Academy of Sciences. 2016; 113(15): 3932–3937. https://doi.org/10.1073/pnas.1517384113

38. Levario TJ, Zhan M, Lim B, Shvartsman SY, Lu H. Microfluidic trap array for massively parallel imaging of Drosophila embryos. Nature protocols. 2013; 8(4):721–736. https://doi.org/10.1038/nprot.2013.034 PMID: 23493069